МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Національний авіаційний університет

Факультет економіки та бізнес-адміністрування

Кафедра бізнес-аналітики та цифрової економіки

КОНСПЕКТ ЛЕКЦІЙ

з дисципліни:

«Інструменти BigData для бізнесу»

Освітньо-професійна програма: «Митна та біржова справа»

Галузь знань: 07 Управління та адміністрування

Спеціальність: 076 Підприємництво, торгівля та біржова справа

Курс – 4 Семестр – 7

Київ 2023

Представлені орієнтовні матеріали до лекційних занять з дисципліни **«Інструменти BigData для бізнесу»**, питання для самоперевірки знань здобувачами вищої освіти, література, що рекомендується для додаткового опрацювання.

Призначена для студентів спеціальності 076 Підприємництво, торгівля та біржова справа, ОПП «Митна та біржова справа» .

Укладач: старший викладач
кафедри бізнес-аналітики та
цифрової економіки _____ Валентина АБЛАМСЬКА

ЗМІСТ

ВСТУП

Навчальна дисципліна призначена для підготовки нового покоління професіоналів в області ведення бізнесу, які володіють методами обробки великих даних, сучасними аналітичними інструментами, методами та моделями прийняття управлінських рішень, які інформаційно та інноваційно зорієнтовані на створення нових цінностей для клієнтів.

За допомогою найновіших цифрових технологій (хмарні обчислення, мобільні технології, соціальні технології, машинне навчання, інтернет речей) менеджери/бізнес-аналітики здатні ідентифікувати, збирати, аналізувати великі масиви даних, інтерпретувати та трансформувати їх для глибшого розуміння бізнесу в реальному часі з метою більш швидкого прийняття кращих рішень, підвищення ефективності роботи компанії, оптимізації її конкурентоспроможності, пом'якшення ризиків.

Предметом вивчення навчальної дисципліни є інструменти та аналітичні методи для використання даних для збору та впорядкування даних у масштабі та отримають розуміння того, як аналіз даних може допомогти інформувати зміни в організаціях.

Мета навчальної дисципліни **«Інструменти BigData для бізнесу» -** підготувати фахівців зі знаннями у галузі великих даних; надання фахівцям навичок у галузі діяльності з удосконалення організації праці, виробництва та управління даними; вивчити принципи, методи та форми організації управління великими даними.

Завдання вивчення дисципліни:

Застосовувати знання фундаментальних і природничих наук, системного аналізу та технологій моделювання, стандартних алгоритмів та дискретного аналізу при розв'язанні задач проектування і використання ІСТ.

Демонструвати знання сучасного рівня технологій інформаційних систем, практичні навички програмування та використання прикладних і спеціалізованих комп'ютерних систем та середовищ з метою їх запровадження у професійній діяльності.

Видобувати знання шляхом інтеграції та аналізу великих даних, отриманих з різноманітних та різнорідних джерела інформації. Вміти обґрунтовувати вибір абстрактних типів даних та структур даних при проектуванні програмного забезпечення ІСТ.

Лекція 1

ІТ-ФАКТОР ВПЛИВУ В СУЧАСНИХ УПРАВЛІНСЬКИХ ЗАВДАННЯХ

План:

1.1Теоретичні аспекти ІТ-факторів в економіці та управлінні.
1.2 ІТ -підтримка завдань управління

1.3 Методичні принципи удосконалення управління підприємством на основі інформаційних технологій

1.1. Теоретичні аспекти ІТ-факторів в економіці та управлінні.

У загальному розумінні інформаційна технологія (ІТ) − це цілеспрямована організована сукупність інформаційних процесів з використанням засобів обчислювальної техніки, що забезпечують високу швидкість обробки даних, швидкий пошук інформації, розосередження даних, доступ до джерел інформації незалежно від місця їх розташування.

Відповідно до визначення, прийнятого ЮНЕСКО, ІТ − це комплекс взаємозалежних, наукових, технологічних, інженерних дисциплін, що вивчають методи ефективної організації праці людей, зайнятих обробкою й зберіганням інформації; обчислювальна техніка й методи організації й взаємодії з людьми й виробничим устаткуванням, їхні практичні додатки, а також зв'язані з усім цим соціальні, економічні й культурні проблеми. Основними рисами сучасних ІТ є комп'ютерна обробка інформації, зберігання великих обсягів інформації на машинних носіях та передача інформації на будь-які відстані в найкоротші терміни.

Сучасні інформаційні технології дозволяють знайти рішення, які покращують бізнес-операції, і проявляються кількісно та якісно: кількість послуг, вартість ІТ-послуг і рівень надійності, швидкість виведення нових послуг на ринок або створення нових умов для вже існуючих і т.д.

Основним продуктом функції ІТ є ІТ-послуга. Створення нового бізнес-процесу, на основі ІТ-можливостей, вимагає чіткого розуміння обов’язків підрозділів компанії в інформаційних системах. ITIL визначає ІТ-послуги в якості основного елемента, що діє в рамках бізнес-процесу, і / або ІТ-проекту.

В управління ІТ-послугами можна виділити 3 напрямки:

– Співставлення ІТ-послуг з поточними і майбутніми потребами бізнесу
– Підвищення якості ІТ-послуг;
– Оптимізація довгострокових витрат на надання ІТ-послуг.

Метою ІТ управління є задоволення інформаційних потреб усіх без винятку суб'єктів регіональної економіки, особливо суб'єктів, що здійснюють управління регіональною економікою та приймають рішення щодо її розвитку та соціально-економічного зростання регіону загалом.

ІТ-процес – це сукупність дій, спрямованих на досягнення конкретного результату. Кожен процес, як правило, складається з декількох суб-процесів, а також входів і виходів. Ефективність процесу може бути визначена за рівнем зрілості.

Зрілість ІТ-процесів вказує на зрілість ІТ-функції. Рівень зрілості є еволюційним кроком на шляху до досягнення зрілості ІТ-функції. Таким чином, характеристика зрілості є постійною, в той час як рівень зрілості є дискретним і може приймати значення від 0 до 5.

Впровадження кращих практив управління ІТ, фактичне перетворення внутрішньої ІТ-функції в сервісну організацію можливо тільки, якщо це питання добре розуміють представники вищого керівництва і, звичайно ж, керівник функції ІТ.

Практично у кожному куточку світу конкуренція є на тій стадії, коли доходи і EBITDA є менш залежними від продажів, а також грошей, витрачених на маркетинг, але є все більш залежними від інноваційних підходів і бізнес-трансформацій. Одна з таких функцій є ІТ. Більшість компаній, які мають високий рівень зрілості бізнес-процесів розуміють важливість ІТ, але не дуже багато людей усвідомлюють, як бізнес-цілі пов'язані з конкретними ІТ-процесами, і навпаки.

1.2 ІТ -підтримка завдань управління

Згідно з дослідженнями Університету управління Антверпена COBIT 5, існує очевидна взаємозалежність між бізнес-цілями – цілями ІТ та ІТ-процесами. Сьогодні на будь-якому підприємстві, яке використовує інформаційні засоби і технології, бізнес-цілі не можуть бути досягнуті без досягнення цілей ІТ та ІТ-цілі не можуть бути досягнуті без зрілих ІТ-

процесів. Це означає, що IT бере безпосередню участь в формуванні EBITDA компанії чи то у позитивному, чи то негативному ключі.

Існує 17 бізнес-цілей, які пов'язані з 17 IT-цілей і 37 технологічних процесів.

Згідно збалансованої системи показників (Balanced Scorecard), ці 17 бізнес-цілей реалізуються у 4 сферах:

Фінансова перспектива:

– цінність бізнес-інвестицій для stakeholders;
– портфель конкурентоспроможної продукції та послуг;
– керованість бізнес-ризиками;
– дотримання законів і нормативних актів;
– фінансова прозорість.

Перспектива замовника:

– клієнто-орієнтована культура обслуговування;
– безперервність та доступність бізнес послуг;
– гучкі відповіді у мінливому бізнес-середовищі;
– інформація на основі стратегічних рішень;
– оптимізація витрат на послуги.

Внутрішня перспектива:

– оптимізація функціональних можливостей бізнес-процесів;
– оптимізація витрат бізнес-процесів;
– програми змін у бізнесі;
– операційна продуктивність та продуктивність персоналу.

Навчання і зростання:

– дотримання внутрішніх політик;
– кваліфіковані та мотивовані люди;
– культура бізнес-інновацій та нових продуктів.

Як визначити основні бізнес-цілі підприємства?

Є декілька способів, зокрема:

1) Якщо компанія має чітку бізнес-стратегію, то очевидно, що основні цілі бізнесу випливають із стратегії.

2) Якщо компанія не має задокументовану бізнес-стратегію, визначення бізнес-цілей підприємства будуть визначені завдяки прямому опитуванню, а потім інтерв'ю з керівниками компанії (CXO) і її акціонерів.

Згідно нашої практики існує ряд цікавих закономірностей і моделей:

– У 40% компаній, акціонери і менеджери переслідують різні цілі, тобто очікування від IT є абсолютно різними. У цих компаніях IT знаходиться в складній ситуації.
– У 40% компаній, менеджери та акціонери думають у різних напрямках, і таким чином очікування від IT нейтральні. У таких компаніях основний принцип "аби не було гірше".
– У 20% компаній акціонери і менеджери мають спільну думку та переслідують спільні бізнес-цілі, що забезпечує адекватні очікування від IT. В зазначеній ситуації, IT є найбільш ефективним.

У свою чергу бізнес цілі відповідають 17 IT-цілям і IT-цілі відповідають 37 IT-процесам, які розподілені на 5 областей:

1. Governance: EDM
2. План: APO
3. Побудова: BAI
4. Запуск: DSS
5. Монітор: MEA

Треба сказати, що станом на сьогодні домен EDM є функцією управління IT, які повинні бути представлені на рівні CIO+.

Приклад перетворення бізнес-цілей в IT-процеси:

Узгодження бізнес-цілей і IT-цілей

Залежно від стратегії компанії, або результатів інтерв'ю, ми можемо зрозуміти основні бізнес-цілі компанії. Якщо бізнес-цілі не прописані в явному вигляді, або не відображені в стратегії компанії, їх ідентифікація може бути виконана за допомогою опитувальників та інтерв'ю (як було зазначено раніше). Після того, як визначені бізнес-цілі підприємства, вони мають бути розділені на три категорії: первинні, вторинні і незначні. Відповідно до процедури, ми можемо перевести бізнес-цілі в IT цілі за допомогою каскадної методології, яка допоможе нам визначити основні IT-цілі компанії.

Вирівнювання IT-цілей і IT-процесів

ІТ-процеси є інструментом досягнення цілей ІТ. Після того, як ми визначили 5-7 ІТ-цілей, ми можемо визначити основні ІТ-процеси підприємства. У той же час, ми можемо визначити ІТ-процеси, які не беруть участі в досягненні бізнес-цілей, оскільки інвестиції в ці процеси не дають очікуваного результату і прибутку.

Таким чином, ми можемо визначити ключові ІТ-процеси і процеси, які не беруть участі в досягненні бізнес-цілей.

Наступним кроком є визначення рівня зрілості ключових ІТ-процесів і довести їх до необхідного рівня зрілості.

Взагалі існує 6 рівнів зрілості процесу.

Опис та характеристики процесу:

(0) Відсутній: Процес не існує.
(1) Початковий: Діяльність іноді здійснюється хаотично. Управління процесом не організоване.
(2) Повторюваний: Ті ж самі завдання вирішуються різними людьми, але з подібним підходом. Там немає прописаних процедур і обов'язків, але існує висока залежність від деяких люде.
(3) Визначений: Процедури стандартизовані і задокументовані, проте, є відхилення від процедур, які не завжди відслідковуються. Процедури описують існуючу практику.
(4) Керований: Процес контролюється і вимірюється. Відповідні заходи вживаються у тому випадку, якщо процес є неефективним. Можуть бути застосовані засоби автоматизації процесу.
(5) Оптимізований: Розвиток рівня передової практики відбувається в результаті постійних поліпшень і порівняльного аналізу з іншими компаніями.

Слід зазначити, що цей інструмент не може застосовуватися у всіх компаній, і керівництво не повинно використовувати цей інструмент в якості механічного перетворення, тому що кожне підприємство має свою специфіку з точки зору власності (державної, приватної), обмеження (наприклад, в залежності від діяльності), промисловості та інших компонентів. Це вимагає більш детального занурення та вивчення ситуації на кожному конкретному підприємстві.

1.3 Методичні принципи удосконалення управління підприємством на основі інформаційних технологій

Основні напрями реорганізації структур управління в умовах інформаційної економіки зводяться до їх децентралізації для досягнення гнучкості, адаптації (пристосування) до мінливих умов зовнішнього середовища, вирівнювання. Основними критеріями оптимізації організаційних структур на основі ІТ є: швидкість прийняття рішень, гнучкість, складність, надійність, здатність до швидкої інтеграції, рішучість. Процес удосконалення системи управління підприємством на основі ІТ може бути зведений до процесів глобальної інтеграції як всередині фірмової мережі постачальників, так і в зв'язку елементів мережі "постачальник - споживач", тобто ІТ повинні забезпечити трансформацію корпоративних структур в мережеві структури. Мережеві структури повинні легко вбудовуватися в віртуальну ланцюжок "постачальник - споживач", входити в ділові альянси і виходити з них.

На основі дослідження даної проблеми в економічній літературі слід виділяти дві стратегії впровадження ІТ в систему управління підприємства.

1. Інформаційні технології пристосовуються до організаційної структури і здійснюють локальну модернізацію усталених процесів управління (реінжиніринг), комунікація не розвивається, виконується автоматизація робочих місць менеджерів, відбувається злиття процесів збору інформації (фізичний потік інформації) з функцією прийняття рішень (інформаційний потік рішення). Наприклад, технології ERP і CRM.

2. Організаційна структура трансформується для оволодіння моделями електронного бізнесу B2B та B2C, основою стратегії є розробка і розвиток комунікацій, а також нових організаційних взаємодій. У цій ситуації ІТ забезпечують, крім реалізації стандартних функцій на основі систем ERP і CRM, обмін інформацією (електронними даними) на основі системи EDI, проведення електронних торгів, формування єдиної ланцюжка "постачальник - споживач", систему електронних платежів Internet-banking та ін.

Таким чином, ІТ є потужними інструментами організаційних змін, що дозволяють підприємству покращувати свою структуру, комунікації, продукти, послуги та ін У залежності від ступеня входження в глобальний інформаційний простір можна виділити наступні види ІТ:

о глобальні мережі - міжнародний поділ праці. Дистанція фірм розширена до глобальної. Зниження витрат глобальної координації. Зниження операційних витрат;

о мережі підприємства - групова, бригадна робота. Координація роботи поза межами структурних підрозділів. Зниження витрат на управління. Зміна ділових процесів;

о розподілене обчислення - робочі групи володіють необхідними знаннями. Ділові процеси раціоналізовані. Вартість управління знижена. Централізація і децентралізація збалансовані;

о переносне обчислення - віртуальні організації. Робота не прив'язана до географічному місцезнаходженню. Робота стає пересувний. Знання та інформація можуть бути доставлені туди, де вони необхідні і в будь-який час. Зниження організаційних витрат через зниження потреби в нерухомості підприємства, що використовується працівниками;

о графічні інтерфейси користувача - полегшується доступ до корпоративних знань, які можуть бути доповнені усіма службовцями. Зниження організаційних витрат, так як трудові процеси рухаються від паперів до цифрових зображень, документів і голосу.

В основі побудови і взаємодії "нових" компаній, що функціонують в умовах інформаційної економіки, лежить не вузька функціональна спеціалізація, а інтеграційні процеси в управлінській діяльності, що забезпечують взаємодію не тільки по вертикалі, але і по горизонталі - між співробітниками різних підрозділів одного рівня ієрархії. Ці процеси породжують нові структури, що характеризують підприємства "без внутрішніх перегородок", підприємства "без кордонів".

Нові корпоративні моделі управління базуються на розширенні зв'язків між споживачами, постачальниками і конкурентами, застосовують сучасні інформаційно-комунікаційні технології, автоматизовані системи управління і виробництва, сучасну обчислювальну техніку. Такий підхід до побудови систем управління перетворює підприємства з закритих систем, що використовують такі традиційні структури управління, як бюрократичні, ієрархічні та механістичні, відкриті, засновані на мережеві методи управління.

В залежності від етапу організаційної зрілості компанії різна ступінь використання інформації та інформаційних технологій у бізнес-процесах. Розвиток неможливо без організації цільового управління і ефективного використання всіх ресурсів організації.

Якщо простежити життєвий цикл будь-якої організації, то можна помітити, що в своєму розвитку вона проходить декілька фаз: від

слаборозвиненою і слабоорганизованной структури до ефективної системи, яка характеризується правильним підходом до управління ресурсами організації і процесами, що протікають у ній.

Використовуючи підходи, розроблені інститутом Карнегі - Меллона, можна скласти класифікацію фаз розвитку та існування компанії в залежності від того, як вона обробляє та використовує інформацію у процесі своєї діяльності.

В основу цієї класифікації покладено вимоги до організації бізнес-процесів, що визначаються ступенем цільового управління. Рівні управління розрізняються наявністю цільової функції і ступенем використання інформації, що накопичується в компанії. Виділені наступні рівні розвитку системи управління підприємством: початковий, повторюваний, фіксований, керований, оптимізується.

Аналіз цих рівнів наведено щодо двох особливостей: характеристики бізнес-процесів та інформаційних потоків, що взаємодіють між собою. Розвиток інформаційних потоків на основі впровадження РІТ обумовлює вдосконалення функції планування (перехід до стратегічного планування не на показниках минулих років, а на прогнозах майбутнього розвитку), прийняття рішень ґрунтується на моніторингу думок покупців і загальні тенденції розвитку.

У наукових публікаціях існує достатня кількість моделей взаємодії розвитку систем управління і використання інформаційних технологій. Так, наприклад, у літературі наводяться моделі Нолана, Ерла, Бхабута, Хиршхайма. Спільними для цих моделей є виділення трьох етапів у розвитку інформаційних технологій: спочатку підприємство планує ІТ для отримання поточної інформації про стан бізнесу, потім розвиток і становлення ІТ пов'язано з підтримкою процесів прийняття рішень, і в кінці свого розвитку ІТ орієнтовані на стратегічне планування конкурентної переваги, адаптацію до мінливих умов зовнішнього і внутрішнього середовища, моніторингу попиту та ін.

Виходячи з викладеного вище молено виділити наступні напрямки удосконалення систем управління на основі ІТ:

о трансформація організаційної структури підприємства;

о впровадження стратегічного планування на основі прогнозів майбутнього стану національних, міжнародних, глобальних ринків;

о децентралізація управління;

о мотивація персоналу зростанням особистої компетентності.

Основними методичними принципами модифікації компаній і структур управління ними на основі IT є наступні.

1. Інформаційна інтеграція, освоєння інтегрованих моделей управління (Integrated Management/Information Technology - IM/IT).

2. Трансформація організаційних структур підприємств із пірамідальних у плоскі, із мінімальною кількістю рівнів між вищим керівництвом і безпосередніми виконавцями, так як управління по горизонталі більш дієво, ніж по вертикалі.

3. Скорочення кількості ієрархічних рівнів, більш переважними є не великі централізовані компанії, а низка дрібних із гнучкими спеціалізованими формами праці мережі компаній.

4. Мережеві форми зв'язку між самою компанією та іншими підприємствами, наприклад, шляхом створення внутрішніх ринків.

5. Інноваційна діяльність, створення в рамках великих компаній інноваційних венчурних фірм, зорієнтованих на виробництво і самостійне просування на ринках нових виробів та технологій (бренд-компаній).

6. Стандартизація бізнес-процесів, продуктів, послуг, обліку, звітності та ін., відхід від вузької функціональної спеціалізації у змісті й характері самої управлінської діяльності, стилі управління.

7. Децентралізація функцій управління, насамперед виробничих і збутових. З цією метою в рамках компаній створюються напівавтономні або автономні відділення, стратегічні бізнес-одиниці, що повністю відповідають за прибутки і збитки.

8. Бенчмаркінг (освоєння стратегії "від кращого до кращого і великого").

9. Підвищення компетентності персоналу.

Реалізація наведених вище принципів потребує організації єдиного інформаційного простору, що сприяло б інформаційного взаємодії суб'єктів, що беруть участь у виробництві однотипних продуктів.

Розглянемо наступні організаційні принципи побудови системи IT.

1. Розвиток IT визначається потребами основної діяльності компанії, а не технологічними нововведеннями.

Призначення керівників бізнес-підрозділів відповідальними за ІС означає, що IT-відділ підтримує нові розробки і відповідає за організацію економічної інфраструктури. Керівництво, зі свого боку, має володіти достатніми знаннями, щоб підтримувати конструктивний діалог зі своїм IT-відділу. Це означає, що співробітники IT-відділу повинні використовувати бізнес-термінологію, а не технічний жаргон. Завдяки цьому, керівники IT-відділів та бізнес-підрозділів зможуть оцінювати ефективність пропонованих рішень і спільно проводити необхідні коригування у разі невдач.

2. Фінансування рішень в області IT приймається виходячи з їх фінансової вигоди.

"Мудрі" компанії уникають великих одноразових капіталовкладень, вважаючи за краще постійно оновлювати свої системи і щорічно інвестувати кошти в їх вдосконалення на регулярній основі.

3. Інформаційна система має просту і гнучку структуру.

"Мудрі" компанії забезпечують простоту і гнучкість своєї

технологічного середовища за рахунок жорсткого визначення стандартів архітектури і глибокого аналізу реальних плюсів і мінусів у кожному конкретному випадку відхилення від цих стандартів. Їм вдається зберегти простоту системи через скорочення числа використовуваних технологій і платформ, а також завдяки побудови гнучких і простих в реалізації архітектур. При створенні ІС враховуються і комерційні аспекти, а саме: які стандарти прийняті в галузі і наскільки гарантована підтримка даних технологій в майбутньому, так як підтримання морально застарілої системи обходиться надзвичайно дорого.

4. Розробки починають приносити користь практично з моменту впровадження.

"Мудрі" компанії використовують скрізь, де тільки можливо, стандартне програмне забезпечення і вносять мінімальні зміни в програми, воліючи замість цього раціоналізувати свої процеси. "Золоте" правило: програмне забезпечення варто модифікувати тільки в тому випадку, якщо в перший же рік інвестиції в розробку окупляться у чотирикратному розмірі. Тільки при такому співвідношенні будуть покриті майбутні витрати, пов'язані з підтриманням нестандартних програм.

5. Проводяться планомірні поліпшення продуктивності системи.

Більшість "мудрих" компаній оцінює продуктивність інформаційних центрів і глобальних мереж з еталонним тестів.

6. Відділ інформаційних технологій добре розбирається в бізнесі, а бізнес-підрозділи в IT.

Бізнес-підрозділу і IT-відділ повинні спільно працювати над прийняттям рішень у сфері інформатизації, щоб забезпечити їх обґрунтованість. Для цього співробітники компанії повинні мати базові знання в області IT, а фахівці IT-відділу - знання про основний діяльності компанії. У "мудрих" організаціях структура IT-відділів проста. Невелике число співробітників займається підтримкою, а основний упор зроблений на продуктивність. В таких організаціях розуміють, що вони не можуть тримати фахівців за всіма напрямами, які їм можуть знадобитися, тому обмежуються лише тими, потреба в яких особливо значна або важлива, а за іншими послугами звертаються до зовнішнім організаціям.

Таким чином, основні напрями реорганізації структур управління в умовах інформаційної економіки зводяться до їх децентралізації для досягнення гнучкості, адаптації (пристосування) до мінливих умов зовнішнього середовища, вирівнювання.

Питання для самоконтролю:

1. Дайте визначення IT.
2. Які напрями можно виділити в управління IT-послугами?
3. У яких 4х сферах реалізуються 17 бізнес-цілей згідно збалансованої системи показників (Balanced Scorecard)?
4. В чому полягає вирівнювання IT-цілей і IT-процесів
5. Які стратегії впровадження IT в систему управління підприємства ви знаєте?
6. Назвіть напрямки удосконалення систем управління на основі IT.

Лекція 2

МОДЕЛЮВАННЯ ТА АНАЛІЗ ДАНИХ

План:

2.1 Введення в аналіз даних

2.2 Аналітичний і інформаційний підходи до моделювання

2.3 Етапи аналізу даних

2.1 Введення в аналіз даних

Аналіз даних - широке поняття. Сьогодні існують десятки його визначень. У найзагальнішому сенсі аналіз даних - це дослідження, пов'язані з обрахуванням багатовимірної системи даних, що має безліч параметрів. У процесі аналізу даних дослідник виробляє сукупність дій з метою формування визначених уявлень про характер явища, що описується цими даними. Як правило, для аналізу даних використовуються різні математичні методи.

Аналіз даних не можна розглядати тільки як обробку інформації після її збору. Аналіз даних - це перш за все засіб перевірки гіпотез і рішення задач дослідника.

Людство у своїй діяльності (науковій, освітній, технологічній, художній) постійно створює й використовує моделі навколишнього світу. Строгі правила побудови моделей сформулювати неможливо, однак людство заощадило багатий досвід моделювання різних об'єктів і процесів.

Відоме протиріччя між обмеженими пізнавальними здібностями людини і нескінченністю Всесвіту змушує нас використовувати моделі і моделювання, тим самим спрощуючи вивчення об'єктів, що цікавлять, явищ і систем.

Слово «модель» (лат. Modelium) означає «міра», «спосіб», «схожість з какойто річчю». Побудова моделей - універсальний спосіб вивчення навколишнього світу, дозволяє виявляти залежності, прогнозувати, розбивати на групи і вирішувати безліч інших завдань. Основна мета моделювання в тому, що модель повинна досить добре відображати функціонування модельованої системи.

Модель - об'єкт або опис об'єкта, системи для заміщення (при певних умовах, припущеннях, гіпотезах) однієї системи (тобто оригіналу) іншою

системою для кращого вивчення оригіналу або відтворення будь-яких його властивостей.

Моделювання - універсальний метод отримання, опису та використання знань. Застосовується в будь-якій професійній діяльності.

По виду моделювання моделі ділять на:

емпіричні - отримані на основі емпіричних фактів, залежностей;

теоретичні - отримані на основі математичних описів, законів;

змішані, напівемпіричні - отримані на основі емпіричних залежностей і математичних описів.

Нерідко теоретичні моделі з'являються з емпіричних, наприклад, багато закони фізики спочатку були отримані з емпіричних даних.

Таким чином, аналіз даних тісно пов'язаний з моделюванням. Відзначимо важливі властивості будь-якої моделі.

Спрощеність. Модель відображає тільки істотні сторони об'єкта і, крім того, повинна бути проста для дослідження або відтворення.

Кінцівка. Модель відображає оригінал лише в кінцевому числі його відносин, і, крім того, ресурси моделювання кінцеві.

Наближеність. Дійсність відображається моделлю грубо або наближено.

Адекватність. Модель повинна успішно описувати моделируемую систему.

Цілісність. Модель реалізує деяку систему (тобто ціле).

Замкнутість. Модель враховує і відображає замкнену систему необхідних основних гіпотез, зв'язків і відносин.

Керованість. Модель повинна мати хоча б один параметр, змінами якого можна імітувати поведінку модельованої системи в різних умовах.

2.2 Аналітичний і інформаційний підходи до моделювання

Модель в традиційному розумінні являє собою результат відображення однієї структури (вивченої) на іншу (маловивчену). Так, відображаючи фізичну систему (об'єкт) на математичну (наприклад, математичний апарат рівнянь), отримаємо фізико-математичну модель системи, або математичну модель фізичної системи.

Будь-яка модель будується і досліджується при певних припущеннях, гіпотезах. Робиться це звичайно з допомогою математичних методів. Використовувати таку модель легко: маючи дані про продажі за попередні місяці, за формулою ми отримаємо прогноз на майбутній місяць.

Такий підхід до моделювання в літературі називають аналітичним.

Аналітичний підхід до моделювання базується на тому, що дослідник при вивченні системи відштовхується від моделі (рис. 2.1). В цьому випадку він з тих чи інших міркувань вибирає потрібну модель. Як правило, це теоретична модель, закон, відома залежність, представлена найчастіше всього в функціональному вигляді (наприклад, рівняння, що зв'язує вихідний параметр y з вхідними впливами x1, x2 ...). Варіювання вхідних параметрів на виході дасть результат, який моделює поведінку системи в різних умовах.

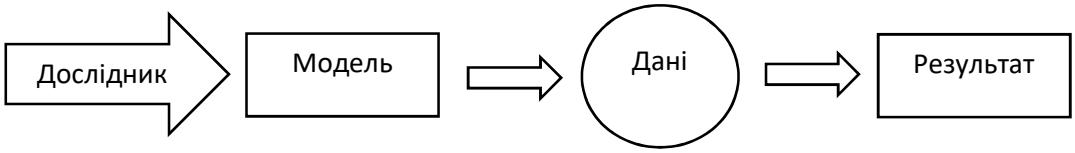


Рис. 2.1 Рух від моделі до результату

Результат моделювання може відповідати дійсності, а може і не відповідати. В останньому випадку дослідникові нічого не залишається, окрім як вибрати іншу модель або інший метод її дослідження. Нова модель, можливо, буде більш адекватно описувати розглянуту систему.

При аналітичному підході не модель «підлаштовується» під дійсність, а ми намагаємося підібрати існуючу аналітичну модель таким чином, щоб вона адекватно відображала реальність.

Модель завжди досліджується будь-яким методом (чисельним, якісним і т.п.). Тому вибір методу моделювання часто означає вибір моделі.

При використанні в бізнесі традиційного аналітичного підходу неминуче виникнуть проблеми через невідповідність між методами аналізу і реальністю, яку вони покликані відображати. Існують труднощі, пов'язані з

формалізацією бізнес-процесів. Тут фактори, що визначають явища, настільки різноманітні і численні, їх взаємозв'язку так «переплетені», що майже ніколи не вдається створити модель, яка задовольняє таким же умов. просте накладення відомих аналітичних методів, законів, залежностей на досліджувану картину реальності не принесе успіху.

У складності і слабкої формалізації бізнес-процесів головним чином «винен» людський фактор, тому буває важко судити про характер закономірностей апріорі (а іноді і апостеріорі, після реалізації якого-небудь математичного методу). З однаковим успіхом описувати ці закономірності можуть різні моделі. Використання різних методів для вирішення однієї і тієї ж завдання нерідко призводить дослідника до протилежних висновків. який метод вибрати? Отримати відповідь на подібне питання можна, лише глибоко проаналізувавши як сенс розв'язуваної задачі, так і властивість використовуваного математичного апарату.

Тому в останні роки набув поширення інформаційний підхід до моделювання, орієнтований на використання даних. Його мета - звільнення аналітика від рутинних операцій і можливих складнощів в розумінні і застосуванні сучасних математичних методів. При інформаційному підході реальний об'єкт розглядається як «чорний ящик », який мав низку входів і виходів, між якими моделюються деякі зв'язку. Іншими словами, відома тільки структура моделі (наприклад, нейронна мережа, лінійна регресія), а самі параметри моделі «підлаштовуються» під дані, які описують поведінку об'єкта.

Для коригування параметрів моделі використовується зворотний зв'язок - відхилення результату моделювання від дійсності, а процес налаштування моделі часто носить ітеративний (тобто циклічний) характер (рис. 2.2).

Таким чином, *при інформаційному підході відправною точкою є дані, що характеризують досліджуваний об'єкт, і модель «підлаштовується» під дійсність.*

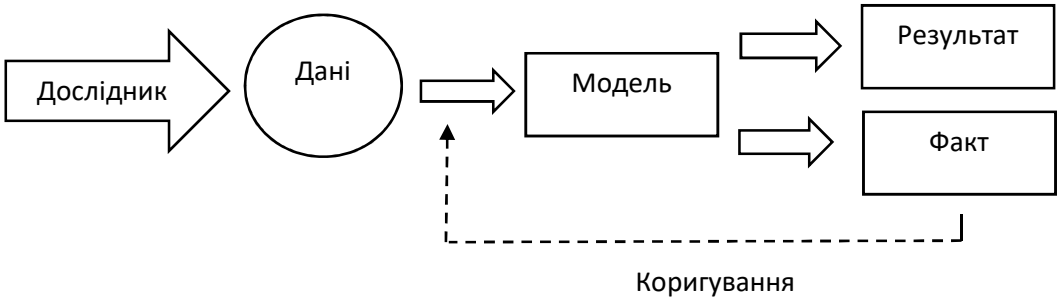


Рис. 2.1 Рух від моделі до результату

Якщо при аналітичному підході ми можемо вибрати модель, навіть не маючи ніяких експериментальних даних, що характеризують властивості системи, і почати її використовувати, то при інформаційному підході без даних неможливо побудувати модель, так як її параметри повністю визначаються ними.

Моделі, отримані за допомогою інформаційного підходу, враховують специфіку об'єкта, що моделюється, явища, на відміну від аналітичного підходу. Для бізнес-процесів остання якість дуже важливо, тому інформаційний підхід ліг в основу більшості сучасних промислових технологій і методів аналізу даних: Knowledge Discovery in Databases, Data Mining, машинного навчання.

Однак концепція «моделей від даних» вимагає ретельного підходу до якості вихідних даних, оскільки помилкові, аномальні і зашумлені дані можуть привести до моделей і висновків, які не мають ніякого відношення до дійсності. Тому в інформаційному моделюванні важливу роль грають консолідація даних, їх очищення і збагачення.

Модель, побудована на деякій множині даних, що описують реальний об'єкт або систему, може виявитися не працює на практиці, тому в інформаційному моделюванні використовуються спеціальні прийоми: поділ даних на навчальне і тестове безлічі, оцінка навчальної та узагальнюючої здібностей моделі, перевірка самий корінь сили моделі.

2.3 Етапи аналізу даних

Аналіз даних можна описати як процес, що складається з декількох кроків, в яких сирі дані перетворюються і обробляються з метою створити візуалізації і зробити передбачення на основі математичної моделі.

Аналіз даних - це всього лише послідовність кроків, кожен з яких відіграє ключову роль для наступних.

Цей процес схожий на ланцюг послідовних, пов'язаних між собою етапів:

Визначення проблеми;

Витяг даних;

Підготовка даних - очищення даних;

Підготовка даних - перетворення даних;

Дослідження і візуалізація даних;

Передбачувальна модель;

Перевірка моделі, тестування;

Розгортання - візуалізація та інтерпретація результатів;

Розгортання - розгортання рішення.

Визначення проблеми.

Процес аналізу даних починається задовго до збору сирих даних. Він починається з проблеми, яку необхідно спершу визначити, а потім і вирішити. Визначити її можна тільки зосередившись на досліджуваної системі: механізм, додатку або процесі в цілому. Дослідження може бути призначене для кращого розуміння функціонування системи, але його краще спроектувати так, щоб зрозуміти принципи поведінки і згодом будувати припущення чи вибір (усвідомлений).

Процеси визначення та документації результатів наукової проблеми або бізнесу потрібні для того, щоб зосередити аналіз на отриманні результатів.

Насправді, всеосяжне і вичерпне дослідження системи - це складний процес, і майже завжди немає достатньої кількості інформації, з якої можна почати. Тому визначення проблеми і особливо планування призводять до появи керівних принципів, яких необхідно дотримуватися протягом всього проекту. Коли проблема визначена і задокументована, можна рухатися до етапу планування проекту аналізу даних. Планування необхідно для розуміння того, які професіонали і ресурси знадобляться для виконання вимог проекту максимально ефективно. Таким чином завдання - розглянути ті питання в області, які стосуються вирішення цієї проблеми Необхідно знайти фахівців з різними інтересами і встановити ПО, потрібне для аналізу даних.

Побудова хорошої команди - один з ключових чинників успішного аналізу даних.

Також під час фази планування вибирається ефективна команда. Такі команди повинні бути міждисциплінарними, щоб у них була можливість вирішувати проблеми, розглядаючи дані з різних точок зору.

Витяг даних.

Коли проблема визначена, перший крок для проведення аналізу - отримання даних. Вони повинні бути обрані з однієї базової метою - побудова предсказательной моделі. Тому вибір даних - також важливий момент для успішного аналізу.

Дані повинні максимально відображати реальний світ - то, як система реагує на нього. Наприклад, використання великих наборів сирих даних, які були зібрані неграмотно, це привести або до невдачі, або до невизначеності.

Тому недостатня увага, приділена вибору даних або вибір таких, які не уявляють систему, призведе до того, що моделі не будуть відповідати досліджуваним системам.

Пошук і вилучення даних часто вимагає інтуїції, кордони якої лежать за межами технічних досліджень і вилучення даних. Цей процес також вимагає розуміння природи і форми даних, надати яке може тільки досвід і знання практичної області проблеми. Незалежно від кількості і якості необхідних даних важливе питання - використання кращих джерел даних. Якщо середовищем вивчення виступає лабораторія (технічна або наукова), а згенеровані дані експериментальні, то джерело даних легко визначити. У цьому випадку мова йде виключно про самих експериментах.

Але при аналізі даних неможливо відтворювати системи, в яких дані збираються виключно експериментальним шляхом, у всіх областях застосування. Багато області вимагають пошуку даних в навколишньому світі, часто покладаючись на зовнішні експериментальні дані або навіть на збір їх за допомогою інтерв'ю та опитувань. У таких випадках пошук хорошого джерела даних, здатного надати всі необхідні дані, - завдання не з легких.

Часто необхідно отримувати дані з декількох джерел даних для усунення недоліків, виявлення розбіжностей і з метою зробити дані максимально загальними.

Інтернет - гарне місце для початку пошуку даних. Але більшу частину з них не так просто взяти. Не всі дані зберігаються у вигляді файлу або бази даних. Вони можуть міститися у файлі HTML або іншому форматі. Тут на допомогу приходить техніка парсинга. Він дозволяє збирати дані за

допомогою пошуку певних HTML-тегів на сторінках. При появі таких збігів спеціальний софт витягує потрібні дані. Коли пошук завершено, у вас є список даних, які необхідно проаналізувати.

Підготовка данних

З усіх етапів аналізу підготовка даних здається найменш проблемним кроком, але насправді потребує найбільшої кількості ресурсів і часу для завершення.

Дані часто збираються з різних джерел, кожен з яких може пропонувати їх в своєму нинішньому вигляді або форматі. Їх потрібно підготувати для процесу аналізу.

Підготовка даних включає такі процеси:

отримання,

очищення,

нормалізація,

перетворення в оптимізований набір даних.

Зазвичай це таблична форма, яка ідеально підходить для цих методів, що були заплановані на етапі проектування. Багато проблем можуть виникнути при появі недійсних, двозначних або відсутніх значень, повторенні полів або даних, які не відповідають допустимому інтервалу.

Вивчення даних / візуалізація.

Вивчення даних - це їх аналіз в графічній або статистичної репрезентації з метою пошуку моделей або взаємозв'язків. Візуалізація - кращий інструмент для виділення подібних моделей. За останні роки візуалізація даних розвинулася так сильно, що стала незалежною дисципліною. Численні технології використовуються виключно для відображення даних, а багато типів відображення працюють так, щоб отримувати тільки кращу інформацію з набору даних.

Дослідження даних складається з попереднього вивчення, яке необхідно для розуміння типу і значення зібраної інформації. Разом з інформацією, зібраною при визначенні проблеми, така категоризація визначає, який метод аналізу даних найкраще підійде для визначення моделі.

Ця фаза, на додаток до вивчення графіків, складається з наступних кроків:

Узагальнення даних;

Угруповання даних;

Дослідження відносин між різними атрибутами;

Визначення моделей і тенденцій;

Побудова моделей регресійного аналізу;

Побудова моделей класифікації.

Як правило, аналіз даних вимагає узагальнення заяв щодо досліджуваних даних.

Узагальнення - процес, при якому кількість даних для інтерпретації зменшується без втрати важливої інформації.

Кластерний аналіз - метод аналізу даних, який використовується для пошуку груп, об'єднаних спільними атрибутами (також називається угрупованням).

Ще один важливий етап аналізу - ідентифікація відносин, тенденцій та аномалій в даних. Для пошуку такої інформації часто потрібно використовувати інструменти і проводити додаткові етапи аналізу, але вже на візуалізації. Інші методи пошуку даних, такі як дерева рішень і асоціативні правила, автоматично отримують важливі факти або правила з даних. Ці підходи використовуються паралельно з візуалізацією для пошуку взаємовідносин даних.

Передбачувальна (предиктивна) модель.

Передбачувальна аналітика - це процес в аналізі даних, який потрібен для створення або пошуку підходящої статистичної моделі для передбачення ймовірності результату.

Після вивчення даних у вас є вся необхідна інформація для розвитку математичної моделі, яка кодує відносини між даними. Ці моделі корисні для розуміння досліджуваної системи і використовуються в двох напрямках.

Перше - передбачення про значення даних, які створює система. У цьому випадку мова йде про регресійних моделях.

Друге - класифікація нових продуктів. Це вже моделі класифікації або моделі кластерного аналізу.

Насправді, можна розділити моделі відповідно до типу результатів, до яких ті призводять:

Моделі класифікації: якщо отриманий результат - якісна змінна.

Регресивні моделі: якщо отриманий результат числовий.

Кластерні моделі: якщо отриманий результат описовий.

Прості методи генерації цих моделей включають такі техніки: лінійна регресія, логістична регресія, класифікація, дерево рішень, метод k-найближчих сусідів.

Але таких методів багато, і у кожного є свої характеристики, які роблять їх придатними для певних типів даних і аналізу. Кожен з них призводить до появи певної моделі, а їх вибір відповідає природі моделі продукту.

Деякі з методів надаватимуть значення, які стосуються реальної системі і їх структурам. Вони зможуть пояснити деякі характеристики досліджуваної системи простим способом. Інші будуть робити хороші прогнози, але їх структура буде залишатися «чорним ящиком» з обмеженою здатністю пояснити характеристики системи.

Перевірка моделі.

Перевірка (валідація) моделі, тобто фаза тестування, - це важливий етап. Він дозволяє перевірити модель, побудовану на основі початкових даних. Він важливий, тому що дозволяє дізнатися достовірність даних, створених моделлю, порівнявши їх з реальною системою.

Але в цей раз ви берете за основу початкові дані, які використовувалися для аналізу. Як правило, при використанні даних для побудови моделі ви будете сприймати їх як тренувальний набір даних (датасета), а для перевірки - як валідаційні набір даних. Таким чином порівнюючи дані, створені моделлю і створені системою, ви зможете оцінювати помилки.

За допомогою різних наборів даних оцінювати межі достовірності створеної моделі. Правильно передбачені значення можуть бути достовірні тільки в певному діапазоні або мати різні рівні відповідності в залежності від діапазону враховуються значень. Цей процес дозволяє не тільки в числовому вигляді оцінювати ефективність моделі, але також порівнювати її з іншими.

Є кілька подібних технік; найвідоміша - перехресна перевірка (крос-валідація). Вона заснована на поділі навчального набору на різні частини. Кожна з них, у свою чергу, буде використовуватися в якості валідаційні набору. Всі інші - як тренувального. Так ви отримаєте модель, яка поступово вдосконалюється.

Розгортання (Деплой).

Це фінальний крок процесу аналізу, завдання якого - надати результати, тобто висновки аналізу.

У процесі розгортання бізнес-середовища аналіз є вигодою, яку отримає клієнт, який замовив аналіз. У технічній або науковій середовищах результат видає конструкційні рішення або наукові публікації.

Розгортання - це процес використання на практиці результатів аналізу даних.

Є кілька способів розгортання результатів аналізу даних або Майнінг даних. Зазвичай розгортання складається з написання звіту для керівництва або клієнта. Цей документ концептуально описує отримані результати. Він повинен бути спрямований керівництву, яке буде приймати рішення. Потім воно використовує висновки на практиці.

У документації від аналітика повинні бути детально розглянуті такі теми:

Результати аналізу;

Розгортання рішення;

Аналіз ризиків;

Вимірювання впливу на бізнес.

Коли результати проекту включають генерацію Передбачуваної моделі, вони можуть бути використані в якості окремих додатків або вбудовані в ПЗ.

Питання для самоконтролю:

1. Дайте визначення Аналіз даних – це?
2. Що назива.ть процесом Моделювання ?
3. В чому полягає аналітичний підхід до моделювання?
4. Розкрийте сутність інформаційного підходу.
5. Назвіть етапи аналізу даних

Лекція 3
АРХІТЕКТУРА ТА МЕТОДИ ЗБЕРІГАННЯ ДАНИХ
План:

3.1. Архітектура бази даних: поняття, визначення
3.2 Інструменти бізнес-аналітики
3.3. Сучасний бізнес-аналіз.

3.1. Архітектура бази даних: поняття, визначення

Архітектура бази даних - комплекс структурних компонентів БД, а також засобів, що забезпечують їх взаємодію як один з одним, так і з кінцевим користувачем, системним персоналом.

Це визначення відображає одну з найважливіших функцій сховищ інформації - забезпечення можливості абстракції відомостей БД. Вона і формує підхід до архітектури даних, що склався в наші дні.

Звідси виникає нове питання: у чому суть, призначення абстракції даних? Ці системи (абстракції) будуть основним засобом підтримки незалежності ведення сховищ інформації (іншими словами, БД) різними групами кінцевих користувачів. По-іншому це називається незалежністю даних системи.

Види БД

Архітектура систем управління базами даних буде різною залежно від різновиду останніх. На сьогодні виділяється два види БД:

централізований;

розподілений.

Централізовані бази даних

Головна відмінність цих БД: вони зберігаються в пам 'яті однієї обчислювальної системи. Але якщо база, в свою чергу, буде компонентом мереж ЕВМ, то стає можливим розподілений доступ до баз даних. Тобто БД буде відкритою для користувачів електронно-обчислювальних машин,

підключених до цієї мережі. Подібне використання характерне для локальних систем ЕВМ, створюваних на базі організацій, компаній.

Розподілені бази даних

Такі БД складаються з декількох частин, що зберігаються в різних ЕВМ однієї мережі. Можливо, інформація тут буде дублюватися, перетинатися. Що зручно, користувачеві розподіленої бази даних не потрібно знати, яким чином елементи сховища інформації розміщені у вузлах подібної мережі. Найчастіше він сприймає цей комплекс відомостей як єдине ціле.

Як здійснюється робота з подібною БД? За допомогою системи управління розподіленими базами даних (СУРБД). Її системний довідник буде описувати інформацію, що міститься в сховищі даних, основи її розміщення в мережі. У свою чергу, сам довідник може бути декомпозований, розміщений в різних вузлах загальної мережі.

Складові частини розподіленої БД розміщуються на окремих підключених до неї ЕВМ. Ними керують вже власні (локальні) СУБД електронно-обчислювальних пристроїв. Що важливо зазначити, подібні локальні системи управління сховищами інформації необов 'язково повинні бути однаковими в різних вузлах загальної мережі. Однак об 'єднання таких різних локальних баз даних в єдину систему - досить складне науково-технічне завдання. Для її успішного вирішення знадобився цілий комплекс експериментальних заходів, теоретичних розробок.

Типи БД за способом доступу до них

Архітектура бази даних також буде відрізнятися за способом доступу до інформації, що знаходиться в сховищі:

Доступ локальний.

Доступ віддалений (мережевий).

Останній тип доступу передбачає поділ архітектури подібних систем ще на дві варіації:

Тип "файл-сервер".

Тип "клієнт-сервер".

ПД "файл-сервер" "

Подібна архітектура комплексів баз даних передбачає виділення одного з пристроїв мережі ЕВМ в якості центрального. Він буде вважатися сервером файлів. На головній машині зберігається спільно використовувана централізована база даних. Інші ж пристрої мережі виступають робочими станціями, які підтримують користувальницький доступ до основної БД.

У системі "" файл-сервер "" кожен користувач має можливість запускати програму, що знаходиться на головній машині. Притому на його пристрої буде відкриватися тільки копія даної програми.

За користувальницькими запитами файли центральної бази даних (що знаходиться на сервері) передаються на комп 'ютери - робочі станції. Там і відбувається обробка інформації. У користувачів, які працюють із загальною БД, на комп 'ютерах з' являється локальна її копія. Остання періодично оновлюється в міру наповнення основного сховища на сервері свіжою інформацією.

Подібна архітектура систем БД найбільше характерна для мереж, до яких підключено невелике число користувачів. Для її реалізації типово використання персональних СУБД (наприклад, Paradox, DBase). Браком архітектури є критично низька продуктивність системи при одночасному доступі декількох користувачів до одних і тих же даних.

БД "" клієнт-сервер ""

Тут також передбачається наявність машини в мережі, яка буде головною. Однак архітектура бази даних "" клієнт-сервер "" має і власну особливість. Головний комп 'ютер не тільки зберігає централізовану БД, але і забезпечує основну частину обробки необхідних користувачеві даних.

Технологія розділяє систему на дві частини: серверну і клієнтську. Остання буде забезпечувати інтерактивний сервіс, а серверна - поділ інформації, управління даними, безпеку і адміністрування.

Що передбачає архітектура клієнт-серверних баз даних? Клієнтський додаток тут оформляє і надсилає запит віддаленому комп 'ютеру-серверу, де розташоване централізоване сховище інформації. Він (запит) складений спеціальною мовою SQL - стандартом доступу до сервера при використанні реляційних БД.

Після отримання запиту віддалений сервер перенаправляє його SQL-серверу. Так називається програма, відповідальна за керування віддаленою

базою даних. Вона забезпечує виконання запиту, надає клієнту необхідні результати по ньому.

Таким чином, вся обробка запитів тут буде проходити на віддаленому сервері. Щоб реалізувати подібну архітектуру, необхідно задіяти багаторівневі СУБД. Друга їх назва - промислові. Такі СУБД здатні організувати масштабну інфосистему, що складається з великого числа користувачів.

Три рівні архітектури БД

Архітектура баз даних підрозділюється на три основних рівні - три ступені опису елементів БД:

Зовнішній. На даному рівні інформація сприймається користувачами.

Внутрішній. На цьому рівні інформація сприймається операційними системами, СУБД (системами управління базами даних).

Концептуальний. Тут здійснюється відображення зовнішнього рівня архітектури системи баз даних на внутрішній, забезпечення необхідної їх незалежності один від одного.

Зовнішній рівень

Зовнішній рівень архітектури систем баз даних - це надання інформації з позиції людей-користувачів.

Рівень описує власну частину баз даних (що належать до кожного користувача). У свою чергу, вона складатиметься з декількох зовнішніх подань сховищ інформації, БД.

Що зручно, кожен користувач тут має справу з таким чином "" реального світу "", який найбільше адаптований під нього. Зовнішнє уявлення буде містити в собі тільки ті сутності, зв 'язки і атрибути, що цікаві і корисні конкретному "юзеру" ".

Не варто вважати, що непотрібні для користувача атрибути, сутності і зв 'язки не існують в базі даних. Вони є, але "юзер" "найчастіше не підозрює про їхнє існування.

Якщо звернутися до термінології ANSI/SPARC (Американського національного інституту стандартів), то подання кожного окремого користувача тут буде називатися зовнішнім. До нього входитиме вміст БД - таке, яким його бачить конкретний "" юзер "". Кожне таке зовнішнє уявлення

визначається за допомогою зовнішньої системи. Вона ж складається з визначення запису кожного типу, присутнього в зовнішньому уявленні.

Концептуальний рівень

Він включає узагальнююче уявлення про сховище інформації. Буде описувати, які саме відомості зберігаються в базі даних, а також які зв 'язки, їх об' єднуючі.

З точки зору адміністратора, сховище містить в собі логічну структуру БД. Даний рівень архітектури бази даних - це фактично повне подання вимог інформації з боку компанії, підприємства, яке не залежатиме від будь-яких міркувань щодо способу, методики її (інформації) зберігання.

Елементи концептуального рівня

Перелічимо компоненти, представлені на концептуальному рівні архітектури:

Сукупність сутностей, атрибутів, зв 'язків між ними.

Обмеження, що можуть бути накладені на дані.

Семантична інформація про відомості в БД (пов 'язана з їх сенсом і значенням).

Інформація щодо заходів забезпечення безпеки зберігання даних, загальної підтримки їх цілісності.

Концептуальний рівень покликаний підтримувати кожне із зовнішніх уявлень. Будь-яка доступна користувачеві інформація з БД повинна міститися (або може бути обчислена) саме на даному рівні. Однак слід пам 'ятати, що інформація про методи зберігання даних у системі тут не зберігається.

Внутрішній рівень

Рівень призначений для опису фізичної реалізації бази даних. Крім того, з його допомогою досягається оптимальна продуктивність, забезпечується економне використання дискового простору комп 'ютерної системи.

Містить опис структур даних, організації певних файлів, які використовуються для реалізації зберігання інформації на дискових просторах, що запам 'ятовують пристрої. Тут, на внутрішньому рівні, СУБД взаємодіє з методами, способами доступу операційних систем, допоміжним функціоналом зберігання і вилучення записів відомостей. Мета всього

перерахованого: розміщувати інформацію на запам 'ятовуючих пристроях, витягувати дані, створювати індекси тощо.

Нижче даного буде знаходитися фізичний рівень. Його контролює вже операційна система, проте все ж під контролем СУБД.

3.2 Інструменти бізнес-аналітики

Інструменти бізнес -аналітики допомагають компаніям залишатися конкурентоспроможними та максимізувати потоки доходу. Організації будь - якого розміру та етапу використовують програмне забезпечення BI для аналізу, управління та візуалізації бізнес -даних. Ось десять переваг програмного забезпечення для бізнес -аналітики, які можна додати до будь - якого бізнесу:

Швидка і точна звітність

Цінна інформація про бізнес

Конкурентний аналіз

Краща якість даних

Підвищене задоволення клієнтів

Визначення тенденцій ринку

Підвищена оперативна ефективність

Покращені, точні рішення

Збільшення доходу

Нижчі поля

Кожен бізнес має свої примхи, і відповідне програмне забезпечення для бізнес -аналітики (BI), як правило, враховує ці особливості та пропонує повне, індивідуальне рішення. Однією з основних функцій програмного забезпечення для бізнес -аналітики є аналіз даних, і тип аналізу, який потрібен компанії, залежить від її цілей.

Аналіз даних може допомогти компаніям описати свій бізнес, подивитися на причини, чому сталися позитивні чи негативні події, створити інформацію, якої вони можуть не мати, та порадити щодо можливих планів дій. Інструменти бізнес -аналітики можуть допомогти компаніям виконувати ці типи аналізу, контролювати ключові показники ефективності (KPI) та

формувати точні звіти. Коли аналітики використовують ці інструменти, щоб донести свої висновки до зацікавлених сторін, вони можуть перетворити своє уявлення на дії.

Інструменти програмного забезпечення BI, такі як Tableau , також можуть допомогти надати компаніям перевагу перед конкурентами, аналізуючи ринкові тенденції, висвітлюючи нові можливості та розробляючи нові стратегії. Ці інструменти також можуть допомогти організаціям зрозуміти потреби своїх клієнтів та клієнтів та оптимізувати їх послуги-від бізнесу до бізнесу (B2B) до бізнесу до споживача (B2C). Компанії також можуть використовувати ці інструменти внутрішньо для моніторингу продуктивності працівників у режимі реального часу.

Системи Business Intelligence-це набори програмних засобів для інтеграції та аналізу великих наборів даних, з яких збираються аналітичні програми для різних завдань (рис 3.1).

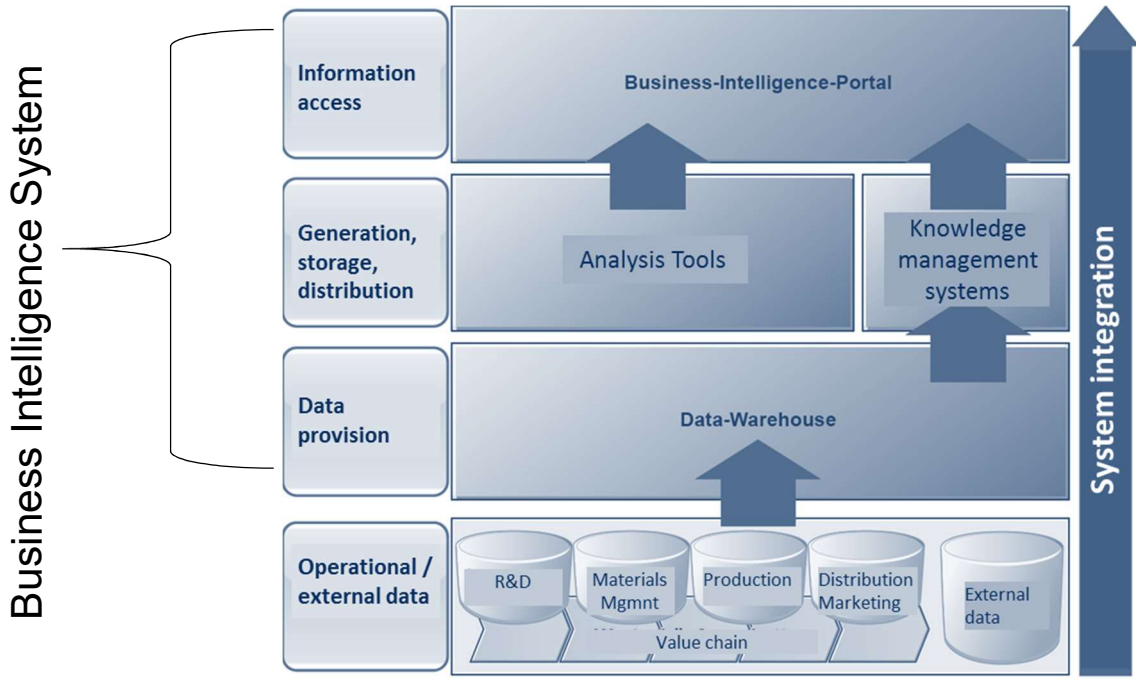


Рис. 3.1 Системи Business Intelligence

Як працює Business Intelligence?

Бізнес-аналіз є основою будь-якої короткострокової та довгострокової бізнес-стратегії. BI використовується як фраза "перехоплення всіх", оскільки вона не застосовується до будь -якого конкретного типу аналізу. Він відноситься до інструментів та процесів бізнес -аналітики, які використовуються для вилучення інформації з необроблених даних для допомоги у прийнятті бізнес -рішень. Організації використовують ці дані, щоб

випередити конкурентів та оптимізувати загальну продуктивність. Ці інструменти необхідні більшості аналітиків бізнес -аналітиків, але також існує цілий ряд інструментів, які можуть допомогти співробітникам з різних відділів.

Деяке програмне забезпечення BI може інтегруватися з інструментами для конкретних галузей бізнесу, таких як роздрібна торгівля, подорожі та медіа -послуги. Звіти про BI та аналітика BI можуть допомогти цим користувачам знайти рішення для інформування свого повсякденного бізнесу, використовуючи інформаційні панелі, складну аналітичну обробку та потужні візуалізації. Звітність про BI також є важливою складовою бізнес-аналітики, оскільки вона допомагає керівникам приймати своєчасні рішення на основі даних.

Але як насправді працює бізнес -аналітика? У більшості компаній дані зберігаються в різних місцях, але вони не можуть відстежувати цю інформацію або уніфікувати ці різні джерела даних. Інструменти BI можуть надавати швидку і точну інформацію особам, які приймають рішення, використовуючи різноманітні джерела даних без допомоги IT -відділу для складання складних звітів. Ці джерела даних можуть бути отримані з маркетингової або збутової аналітики, ефективності операцій, із системи програмного забезпечення для управління відносинами з клієнтами (наприклад, Salesforce) або даних ланцюжка поставок. Як правило, програмне забезпечення BI може об’єднувати всі ці джерела разом, щоб надати історичні, поточні та прогнозні уявлення, які допоможуть у плануванні бізнесу.

Компаніям важливо враховувати, як вони працюють, включаючи поточні процеси інтеграції даних, звітності та обміну інформацією. Після цього вони можуть визначити свої цілі, а потім розробити стратегію розгортання нової системи. Для індивідуальних рішень програмне забезпечення BI можна розгортати поетапно. Підприємства можуть спочатку розпочати, наприклад, з даних клієнтів, а потім побудувати наступний розділ, наприклад, фінансові дані. Цей процес є більш керованим для більшості компаній, тому вони зосереджуються лише на певних уявленнях у певний час і усувають перевантаженість занадто великою кількістю даних одночасно.

Business Intelligence VS Business Analytics

Люди, як правило, не погоджуються щодо відмінностей між бізнес -аналітикою та бізнес -аналітикою. Причина, чому ніхто не може погодитися? Бізнес -аналітика та бізнес -аналітика, або BA, також дуже схожі

і пов’язані між собою. Деякі люди вважають, що основна відмінність залежить від часу: бізнес-аналітика допомагає у повсякденних операціях і в тому, як справи виглядають у сьогоденні, тоді як бізнес-аналітика допомагає планувати бізнес на майбутнє, наприклад, за допомогою прогнозної аналітики, щоб з'ясувати, чому відбуваються і як все буде виглядати в майбутньому.

Особи, які займаються бізнес -аналітикою, зазвичай мають досвід роботи в галузі математики чи статистики або онлайн -ступінь магістра з бізнес -аналітики . Експерти з бізнес -аналітики також повинні добре володіти цифрами та мати аналітичний розум.

Пат Рош, віце -президент з інженерії компанії Magnitude Software, вважає: «BI потрібна для ведення бізнесу, тоді як Business Analytics потрібна для зміни бізнесу». Бізнес -аналітика та бізнес -аналітика, як правило, добре працюють у парі. Оскільки бізнес -аналітик допомагає компаніям керувати та оптимізувати свої повсякденні операції, бізнес -аналітика може продовжити, де зупинився BI, та визначити шляхи покращення майбутніх показників компанії.

Слідуючи логіці Роша, щоб розбити її ще далі, бізнес -аналітик більше зосереджується на перших двох типах бізнес -аналітики - описовій та діагностичній, а бізнес -аналітика зосереджується на аналітиці прогнозування та припису.

Як бізнес -аналітика та бізнес -аналітика підтримують прийняття рішень?

Незалежно від того, чи є подібності та відмінності між бізнес -аналітикою та бізнес -аналітикою, компаніям, можливо, доведеться з'ясувати, які інструменти використовувати та в яке програмне забезпечення інвестувати, щоб отримати цінні дані, які їм потрібні, та приймати своєчасні та точні рішення. Деякі можуть стверджувати, що найважливішою перевагою BI та BA є покращення прийняття рішень. BI та BA можуть допомогти особам, які приймають рішення на кожному рівні, зрозуміти свій бізнес, збільшити прибуток та приймати рішення, які підтримуються не лише кишковими інстинктами, а фактичними статистичними міркуваннями та даними.

Підприємства можуть використовувати ці дані для прийняття рішень, таких як вихід на нові ринки та заходи, які необхідно вжити для зменшення ризиків. Вони також можуть приймати рішення щодо операційних процесів або реструктуризації відділів.

Чому бізнес -аналітика настільки важлива?

У сучасному світі, керованому даними, компанії стикаються з перевантаженням інформацією, а компанії, які зацікавлені працювати розумніше, інвестують у способи контролю та розуміння цієї інформації. Настала ера великих даних. Насправді, ми виробляємо стільки даних, що 90% даних зібрано за останні кілька років . Хоча впровадження нової технології може здатися складним завданням, програмне забезпечення BI, як правило, добре окупається, навіть якщо переваги не видно відразу.

Business Intelligence допомагає компаніям відстежувати тенденції, адаптуватися до мінливих ринкових умов та покращує прийняття рішень на всіх рівнях організації. Інструменти BI, якими користується компанія, залежать від їх цілей. Деякі компанії зацікавлені в отриманні уявлень про покупки споживачів, інші компанії зацікавлені в підвищенні продуктивності працівників або подивіться, хто найкраще працює. Існує нескінченна кількість способів розгортання рішення для бізнес -аналітики. Ось лише десять способів, як бізнес -аналітика може покращити бізнес.

Переваги програмного забезпечення Business Intelligence

Швидка та точна звітність: працівники можуть використовувати шаблони або персоналізовані звіти для моніторингу показників ефективності за допомогою різних джерел даних, включаючи фінансові дані, дані про операції та дані про продажі. Ці звіти формуються в режимі реального часу і використовують найрелевантніші дані, щоб підприємства могли швидко діяти. Більшість звітів включають легкочитані візуалізації, такі як графіки, таблиці та діаграми. Деякі звіти про програмне забезпечення BI є інтерактивними, тому користувачі можуть грати з різними змінними або отримати доступ до інформації ще швидше.

Цінна інформація про бізнес: компанії можуть оцінювати продуктивність праці, дохід, загальний успіх, а також показники діяльності відділу. Він може виявити сильні та слабкі сторони, оскільки інструменти BI допомагають організаціям зрозуміти, що працює, а що ні. Налаштування сповіщень просте і може допомогти відстежити ці показники та допомогти зайнятим керівникам залишатися на першому місці за показниками ефективності, які мають найбільше значення для їхнього бізнесу.

Конкурентний аналіз: Здатність керувати та маніпулювати великою кількістю даних сама по собі є конкурентною перевагою. Крім того, складання бюджету, планування та прогнозування є неймовірно потужним способом

випередити конкурентів, виходить далеко за рамки стандартного аналізу, а також легко виконується за допомогою програмного забезпечення BI. Підприємства також можуть відстежувати продажі та маркетингові результати конкурентів та навчитися диференціювати товари та послуги.

Краща якість даних: Дані рідко є скрипучими, і існує багато способів, за допомогою яких можна виявити розбіжності та неточності - особливо за допомогою зламаної «бази даних». Підприємства, які піклуються про збір, оновлення та створення якісних даних, як правило, більш успішні. За допомогою програмного забезпечення BI компанії можуть об’єднувати різні джерела даних для повної картини того, що відбувається з їх бізнесом.

Підвищення задоволеності клієнтів: програмне забезпечення BI може допомогти компаніям зрозуміти поведінку та закономірності клієнтів. Більшість компаній отримують відгуки клієнтів у режимі реального часу, і ця інформація може допомогти підприємствам утримати клієнтів та охопити нових. Ці інструменти також можуть допомогти компаніям визначити моделі покупки, що допомагає співробітникам споживачів передбачити потреби та забезпечити кращий сервіс.

Визначення ринкових тенденцій: Визначення нових можливостей та розроблення стратегії з підтримкою даних може надати бізнесу конкурентну перевагу, безпосередньо вплинути на довгострокову прибутковість та надасть повну інформацію про те, що відбувається. Співробітники можуть використовувати зовнішні ринкові дані з внутрішніми даними, щоб виявляти нові тенденції продажів, аналізуючи дані про клієнтів та ринкові умови, а також виявляючи проблеми бізнесу.

Підвищена операційна ефективність: інструменти BI об'єднують безліч джерел даних, які допомагають загальній організації бізнесу, так що керівники та співробітники витрачають менше часу на відстеження інформації та можуть зосередитися на складанні точних та своєчасних звітів. Отримуючи актуальну та точну інформацію, співробітники можуть зосередитися на своїх короткострокових та довгострокових цілях та проаналізувати вплив своїх рішень.

Покращені, точні рішення: конкуренти швидко рухаються, і компаніям важливо приймати рішення якомога швидше. Якщо не вирішити проблеми з точністю та швидкістю, це може призвести до втрати клієнтів та доходу. Організації можуть використовувати наявні дані для надання

інформації потрібним зацікавленим сторонам у потрібний час, оптимізуючи час прийняття рішення.

Збільшення доходу: збільшення доходу - важлива мета будь -якого бізнесу. Дані з інструментів бізнес -аналітики можуть допомогти компаніям задавати кращі запитання про те, чому це сталося, шляхом порівняння різних аспектів та виявлення слабких місць у продажах. Коли організації прислухаються до своїх клієнтів, спостерігають за своїми конкурентами та покращують свою діяльність, швидше за все збільшиться дохід.

Зниження рентабельності: Норми прибутку - це ще одна проблема для більшості підприємств. На щастя, інструменти BI можуть аналізувати неефективність та допомагати збільшити маржу. Сукупні дані про продажі допомагають компаніям зрозуміти своїх клієнтів і надають командам продажів можливість розробляти кращі стратегії щодо того, де потрібно витрачати бюджети.

3.3. Сучасний бізнес-аналіз.

Останнім часом, у четвертій галузевій переоцінці, існує дуже велика кількість створених та сформованих даних за допомогою комп'ютерних машин, таких як GPS, датчики, веб -сайти чи системи додатків, або людьми через соціальні мережі (Twitter, Facebook, Instagram або LinkedIn). Щомиті сервери даних зберігають величезну кількість даних, які виробляються організаціями. Це величезна кількість даних, що надходить із веб -сайтів, соціальних мереж, відстеження, додатків Інтернету речей, датчиків та новин в Інтернеті. Крім того, прогрес у галузі обчислювальних та комунікаційних технологій спростив збір великого обсягу неоднорідних даних з різних джерел. Ці дані складаються зі структурованої та неструктурованої, складної та простої інформації.

В даний час бізнес отримує дохід від аналізу таких даних у неструктурованій формі до 80%. Таким чином, організація може покращити бізнес -виробничий процес завдяки цьому аналізу неструктурованих даних, що містять цінну інформацію. Крім того, він важливий для освіти, безпеки, охорони здоров'я та виробництва. Цього можна досягти за допомогою аналізу великих даних, штучного інтелекту та управління даними для досягнення бізнес -аналізу.

BI - це технології, інструменти, системи та програми для складання, аналізу, комбінації та виставки бізнес -звіту з активним способом виконання бізнес -рішень. Цей спосіб надасть необмежену допомогу в отриманні, вивченні та контролі їхніх даних для подальшого прийняття рішень щодо розвитку бізнес -процесів та процедур. Крім того, BI можна охарактеризувати як здатність фірми створювати значущі дані, за допомогою яких кожен день збираються бізнес -процеси та операції.

Бізнес -аналітика (BI) відіграє важливу роль, допомагаючи маркеру прийняття рішень отримати уявлення про покращення продуктивного або кращого та швидкого прийняття рішень. Крім того, BI може покращити та сприяти ефективності операційних правил та їх впливові на прийняття рішень на корпоративному рівні, систему нагляду, адміністрування, складання бюджету та фінансовий облік, що дає кращі стратегічні альтернативи в динамічному бізнес-середовищі. Крім того, BI може покращити організаційні показники, виявляючи нові можливості, розкриваючи нові бізнес -ідеї, виділяючи потенційні загрози та покращуючи процеси прийняття рішень серед багатьох інших переваг.

Перше питання в бізнесі - це управління великими даними з різними форматами даних, що є серйозною проблемою управління через те, що сучасні інструменти не є достатніми для управління такими великими обсягами великих даних. Нові виклики щодо складності інтеграції даних, місткості сховища, відсутності управління та аналітичних інструментів надають значення для вирішення великої проблеми управління великими даними, пов'язаної з попередньою обробкою, обробкою, безпекою та зберіганням. Керування великими даними у великій кількості даних, створених гетерогенними джерелами для використання у BI та прийнятті рішень, є складним процесом. Тому деякою формою великих даних можуть керувати 75% організацій. Метою управління великими даними є забезпечення ефективності безпеки, зберігання та аналітичних застосувань великих даних.

На жаль, практичні наслідки використання аналітики великих даних для покращення бізнес-аналітики залишаються порівняно незрілими та недостатньо дослідженими, оскільки існуючі моделі дослідження зосереджені переважно на перевагах та проблемах бізнес-аналітики та великих даних. Отже, найважливішими питаннями є вивчення впливу аналізу великих даних на бізнес -аналітику для збирання даних з різних джерел та вивчення

майбутніх напрямків пошуку подальших подій у використанні аналітики великих даних для аналізу бізнесу.

Друге питання у BI - це визначення найбільш підходящої техніки інтелектуального аналізу даних, що є одним з найважливіших обов'язків. Виходячи з ділового характеру та труднощів, що зазнали, або виду об'єктів у бізнесі виникає необхідність у визначенні оптимальної техніки видобутку даних. У процесі видобутку даних більшість основних методів визначають характер варіанта рекультивації даних та його процес видобутку. На основі отриманих результатів техніка видобутку даних буде високопродуктивною. Існує багато методів інтелектуального аналізу даних, оскільки правило асоціації, кластеризація, класифікація, дерево рішень та нейромережі є надзвичайно успішними та практичними.

Data Mining має на меті інтерпретувати величезні обсяги даних та витягти знання з різних об'єктів. Для деяких підприємств цілі видобутку даних визнаються для виявлення різних особливостей, розвитку маркетингових здібностей та прогнозування перспектив на основі попередніх спостережень та сучасних схильностей. Існує вимога щодо перевірки даних для підтримки розпродажів та додаткових цілей підприємця. Крім того, можна продовжувати практику видобутку даних для розпізнавання незвичайної продуктивності та виявити дивну поведінку представників, які практикують деякі технології.

Третє питання в BI - штучний інтелект (AI). ШІ - це головний крок у еволюції технологій, до якої активно вдаються з тих пір, як британський математик і порушник коду Алан Тьюрінг передбачив чіткий шлях у своєму новаторському документі 1950 року «Обчислювальна техніка та інтелект». У той час комп'ютерні технології не могли йти в ногу з ідеями Тьюринга. Але завдяки прогресу в обчислювальній техніці був створений ШІ. В Оксфордському університеті Інститут майбутнього людства представив звіт за 2018 рік для опитування групи дослідників ШІ щодо термінів для Сильного ШІ. У цьому звіті було виявлено, що через 45 років 50% ймовірність ШІ перевершить людей у всіх завданнях, а через 120 років вона автоматизує всі робочі місця людини. Крім того, ШІ відкриє багато можливостей для створення нових робочих місць. Також, як стверджують багато експертів, усунення необхідності виконувати нудні та повторювані завдання - одна з цінностей ШІ. Натомість користувачі можуть зосередитися на своїх основних навиках та цінностях. Щоб зменшити людські помилки, скоротити витрати на оплату праці, а згодом збільшити прибуток, було спрямовано застосування технологій у багатьох галузях та бізнесі. Це справедливо для прогресів,

досягнутих під час четвертої промислової революції (FIR), аж до народження комп'ютера, і все ще вірно для епохи ШІ.

У цьому розділі буде представлено та обговорено важливість аналітики великих даних, інтелектуального аналізу даних, штучного інтелекту для побудови сучасної BI та вдосконалення. Крім того, виклики та можливості для створення цінності даних шляхом створення сучасних процесів BI.

Business Intelligence (BI) можна охарактеризувати як автоматизований процес отримання моделей та уявлень із необроблених даних, які збираються з неоднорідних джерел даних та систематизовано систематизовано для вдосконалення ділових операцій та процесів. У корпоративних архітектурах BI найкращою практикою є розподіл процесів збору даних та організації даних, які пов'язані з внутрішньою архітектурою, від аналізу даних та їх відображення користувачеві через зовнішній інтерфейс. У BI оброблені транзакції генерують дані, які зберігаються в джерелах оперативних даних, які називаються серверами обробки онлайн -транзакцій (OLTP). За допомогою OLTP дані зберігаються у структурованому сховищі даних, що називається сховищем даних, після процесів вилучення та перетворення. Зі сховищем даних, для прискорення аналізу даних та виконання аналітичного запиту можна застосувати різні методи оптимізації запитів. Для досягнення цього прискорення сховище даних створює підмножини сховища даних, які називаються базами даних. Також механізми звітності для доступу до даних транзакцій, що зберігаються у сховищі даних, використовуються в традиційних системах BI. Тому аналіз даних транзакцій може допомогти нам виявити закономірності та передбачити тенденції бізнесу.

Останнім часом джерелами даних BI є не лише традиційні джерела даних як дані транзакцій, але вони включають сучасні джерела даних як мобільні пристрої та дані датчиків, а також веб -повідомлення, які надсилалися інтрамережами компанії та профілями працівників та клієнтів. Більшість сучасних джерел даних є неструктурованими, наприклад, розміщені повідомлення в онлайн -соціальних мережах (OSN) та дані з різних датчиків. Тому основним викликом є те, як підтримувати ці сучасні джерела даних як традиційні реляційні бази даних та досягати ефективності запитів. З точки зору аналізу даних, додаткові дані означають додаткові можливості для отримання додаткової інформації. Проте великі проблеми з аналітичної точки зору залишаються великою проблемою.

Завдяки збільшенню даних, в рамках BI є розширені можливості, які є не лише механізмом аналізу історичних тенденцій даних, але й можуть

поєднувати дані датчиків та іншу особисту інформацію в режимі реального часу для отримання висновків, які не є загальнодоступними це називається ситуаційним БІ. Для комерційних операцій BI називається операційним BI, який надає уявлення про ці операції в режимі реального часу як отримання миттєвого зворотного зв'язку для роботи колл -центру як переваги від їх роботи. Крім того, правила аналітики можуть бути складені залежно від метаінформації викритих даних для його/її, які можна розглядати як BI для самообслуговування. Тому цими новими підходами до BI слід керувати обережно, щоб не порушувати моделі відповідності та управління підприємством.

Трирівнева архітектура традиційної системи BI показана на рис. 3.2 . Ця архітектура складається з трьох шарів: 1) рівня презентації, 2) рівня додатків і 3) рівня бази даних. Основним викликом цієї трирівневої архітектури є те, як виконати такі цілі рівня обслуговування, як мінімальна швидкість проходження та максимальний час відгуку. Це пояснюється тим, що управління зберіганням даних на низькорівневих шарах приховано від рівня додатка, що ускладнює передбачення часу виконання.

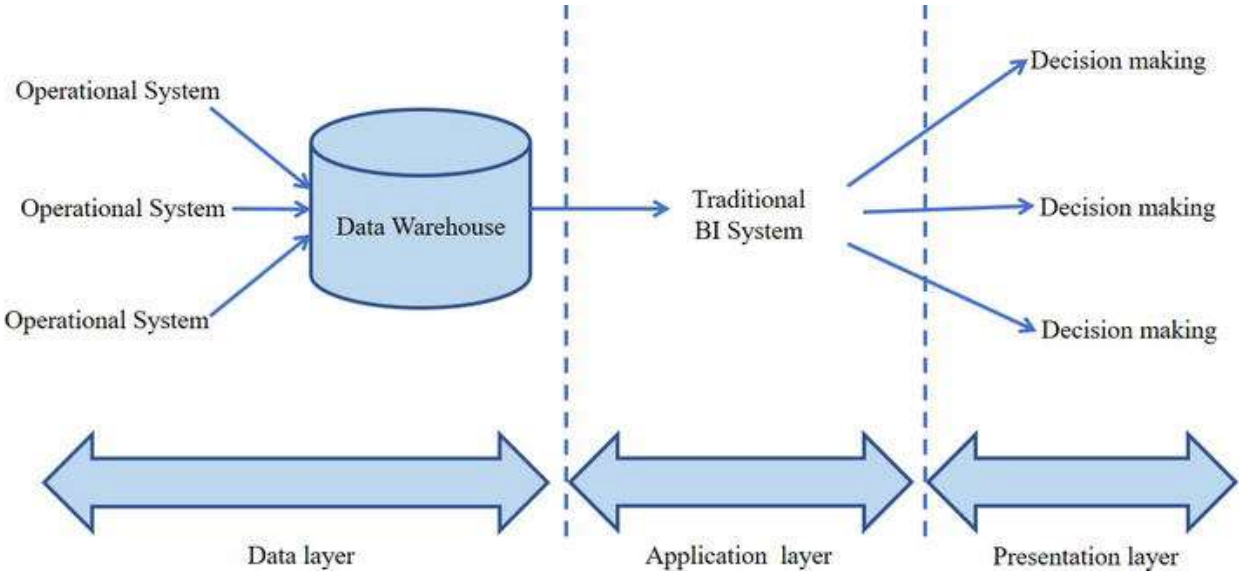


Рис. 3.2 Традиційна система BI.

Однак традиційні системи BI ефективні у вилученні та аналізі даних, але вони жорсткі, повільні, потребують багато часу та потребують знань фахівців для обслуговування. Тому було зроблено багато дослідницьких робіт щодо додавання сучасних функцій для вдосконалення трирівневої архітектури, яка створить BI наступного покоління.

Питання для самоконтролю:

1. Дайте визначення Архітектура бази даних .
2. Види БД.
3. Назвіть три рівні архітектури БД.
4. Які інструменти бізнес-аналітики ви знаєте?
5. Як працює Business Intelligence?
6. Назвіть переваги програмного забезпечення Business Intelligence
7. В чому полягає сучасний бізнес-аналіз?

Лекція 4

ВИДОБУВАННЯ, ПЕРЕТВОРЕННЯ ТА НАВАНТАЖЕННЯ (ETL)

План:

4.1 Процес ETL у сховищі даних

4.2 Дизайн процесу ETL та підтримка інструментів

4.1 Процес ETL у сховищі даних

Сховище даних - це сукупність величезних обсягів даних, що надають інформацію діловим користувачам за допомогою інструментів бізнес-аналітики.

ETL - це процес видобування, перетворення та завантаження даних для зберігання в сховищі даних, у той час як сховище даних є центральним місцем, яке використовується для зберігання консолідованих даних з декількох джерел даних.

Витяг

Видобуток є першим кроком. Він включає вилучення даних з різних джерел даних, таких як бази даних. Одним з основних фактів, які слід відзначити під час виконання вилучення є те, що він не повинен впливати на продуктивність або час відгуку вихідного джерела даних. Тому існують різні стратегії вилучення даних.

Повний видобуток - Це передбачає вилучення всіх даних з усіх джерел даних. Основною метою цієї стратегії є завантаження сховища даних на початковому етапі або завантаження, коли важко ідентифікувати змінені дані.

Часткове вилучення (з повідомленням про оновлення) - Ця стратегія простіше і швидше, ніж повне вилучення. Це включає вилучення лише змінених даних.

Часткове вилучення (без сповіщення про оновлення) - Вона передбачає вилучення даних на основі певних ключових особливостей. Наприклад, якщо вже вчора видобуті дані, можна витягти дані сьогодні і визначити зміни в них.

Трансформація

Витягнуті дані є необробленими даними, тому це не дуже корисно. Тому перетворення даних відбувається на наступному етапі. Вона включає

очищення, відображення та перетворення даних. Основні завдання трансформації такі:

Вибір - Вибір необхідних даних

Мап - Пошук даних з різних файлів пошуку та відповідності даних, які потребують трансформації

Очищення даних - Очищення даних для їх стандартизації

Узагальнення - Агрегація та консолідація даних

Основними завданнями перетворення даних є наступні.

Стандартизація - Оскільки дані надходять з різних джерел, це вимагає стандартизації

Перетворення набору символів і обробка кодування - Перетворення даних у визначене кодування

Обчислення значень - Розрахунок і виведення нових стовпців з існуючих стовпців.

Пролиті та об'єднані поля - Розбиття поля на кілька полів або об'єднання декількох полів в одне поле на основі вимог.

Перетворення одиниць вимірювань - Залучення переходів часу даних і т.д.

Узагальнення - Агрегація та консолідація даних.

Видалення дублювання - Видалення дубльованих даних, отриманих з декількох джерел.

Завантаження

Це процес вибірки отриманих даних і зберігання їх у сховищі даних. Існують різні методи навантаження.

Початкове навантаження - Завантаження сховища даних вперше.

Додаткове навантаження - Періодичне застосування постійних змін.

Повне оновлення - Повністю стирання вмісту однієї або декількох таблиць і перезавантаження з новими даними.

Сховище даних - це система, яка підтримує процес бізнес-аналітики. Він перетворює дані в значущу інформацію для аналізу бізнесу. Тому це цінний ресурс для керівництва організації при прийнятті рішень.

Крім того, організація має різні бази даних, такі як MySQL і MSSQL. Всі ці дані витягуються, трансформуються і завантажуються в сховище даних. Потім дані інтегруються і обробляються. Нарешті, аналітики даних, вчені даних та менеджери використовують ці дані для отримання ділової інформації.

Крім того, дані в сховищі даних поділяються на вітрини даних. Кожен з них містить дані для конкретних користувачів. Вони покращують безпеку та цілісність даних. Зазвичай сховище даних розташоване в окремому місці від звичайних оперативних баз даних.

У розрізі, основною відмінністю між ETL і сховищем даних є те, що ETL є процесом вилучення, перетворення і завантаження даних для зберігання в сховищі даних, в той час як сховище даних є центральним місцем, яке використовується для зберігання консолідованих даних з декількох джерел даних.

4.2 Дизайн процесу ETL та підтримка інструментів

Використовуючи встановлену ETL-схему, можна збільшити шанси досягнення кращого підключення та масштабованості. Хороший інструмент ETL повинен мати можливість спілкуватися з багатьма різнимиреляційними базами даних і прочитати різні формати файлів, які використовуються в організації. Інструменти ETL почали мігрувати в Інтеграцію корпоративних програм[en] або навіть в Інтеграційну шину даних, які зараз охоплюють значно більше, ніж просто виймання, перетворення та завантаження даних. Багато постачальників ETL тепер мають профілі даних[en], якість даних[en] та можливості метаданих. Звичайне використання для інструментів ETL включає перетворення файлів CSV у формат, який можна зчитувати реляційними базами даних. Типовий переклад мільйонів записів полегшує ETL інструменти, які дозволяють користувачам вводити канали / файли даних csv і імпортувати їх у базу даних з якнайменш можливою кількістю коду.

Інструменти ETL, як правило, використовуються широким колом професіоналів — від студентів, що вивчають інформатику, які бажають швидко імпортувати великі обсяги даних до архітекторів баз даних,

відповідальних за управління обліковими записами компанії, ETL інструменти стали зручним інструментом, на який можна покластися, щоб отримати максимальну ефективність. Інструменти ETL в більшості випадків містять графічний інтерфейс, який допомагає користувачам зручно перетворювати дані, використовуючи маппер візуальних даних, на відміну від написання великих програм для аналізу файлів та модифікації типів даних.

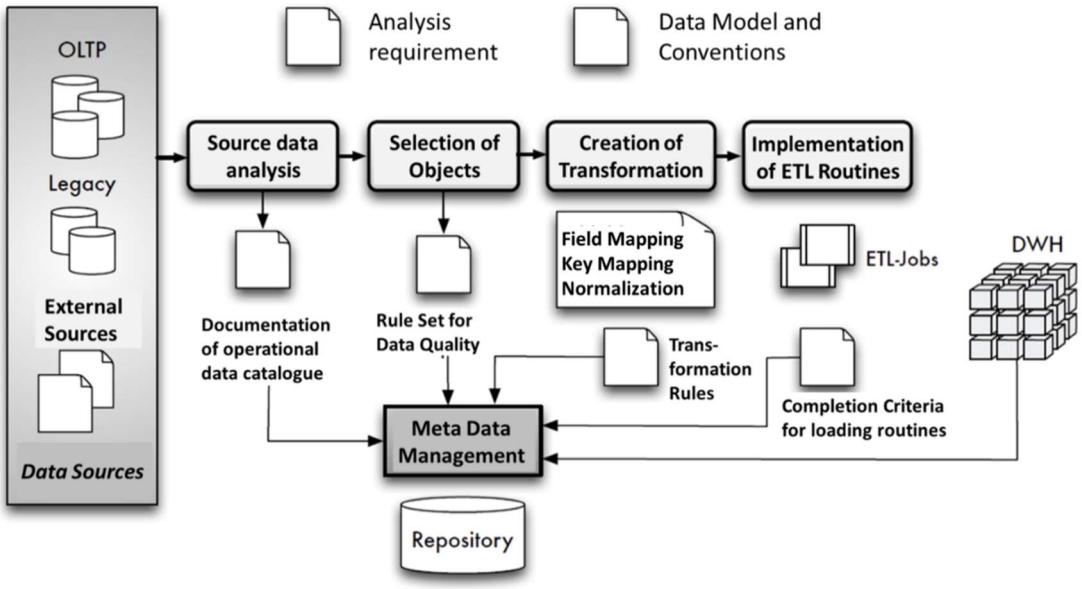


Рис. 4.2 ETL - Дизайн

Хоча інструменти ETL традиційно були для розробників та І. Т. персоналу, нова тенденція полягає в тому, щоб забезпечити ці можливості діловим користувачам, щоб вони могли самостійно створювати зв'язки та інтеграцію даних, коли це потрібно, а не йти до І. Т. персоналу. Гартнер називає цих нетехнічних користувачів громадянськими інтеграторами.

У сучасну технологічну еру слово 'дані' є дуже важливим, оскільки більша частина бізнесу ведеться навколо цих даних, потоку даних, формату даних тощо. Сучасні програми та робоча методологія вимагають даних у реальному часі для цілей обробки та для того, щоб для досягнення цієї мети на ринку доступні різні інструменти ETL.

Використання таких баз даних та інструментів ETL значно полегшує завдання управління даними та одночасно покращує зберігання даних.

Нижче наведено список найкращих програм з відкритим кодом та комерційних програм ETL із деталями порівняння.

1) Hevo - Рекомендований інструмент ETL

Hevo, платформа передачі даних без коду може допомогти вам переміщати дані з будь-якого джерела (бази даних, хмарні програми, SDK та потокове передавання) до будь-якого пункту призначення в режимі реального часу.

Основні характеристики:

Простота реалізації: Hevo можна налаштувати і запустити всього за кілька хвилин.

Автоматичне виявлення та відображення схеми: Потужні алгоритми Hevo можуть виявляти схеми вхідних даних і копіювати їх у сховищі даних без будь-якого втручання вручну.

Архітектура в режимі реального часу: Hevo побудований на потоковій архітектурі в режимі реального часу, яка забезпечує завантаження даних на ваш склад у режимі реального часу.

ETL та ELT: Hevo має потужні функції, які дозволяють очищати, трансформувати та збагачувати дані як до, так і після переміщення їх на склад. Це гарантує, що у вас завжди є дані, готові до аналізу.

Захист корпоративного рівня: Hevo відповідає вимогам GDPR, SOC II та HIPAA.

Попередження та моніторинг: Hevo надає детальні попередження та детальний моніторинг, щоб ви завжди були в курсі своїх даних.

2) Xplenty - це хмарне рішення ETL, що забезпечує прості візуалізовані конвеєри даних для автоматизованих потоків даних у широкому діапазоні джерел та напрямків.

Потужні інструменти трансформації компанії на платформі дозволяють своїм клієнтам очищати, нормалізувати та трансформувати свої дані, дотримуючись також найкращих практик відповідності.

Основні характеристики:

Централізуйте та підготуйте дані для BI.

Передача та перетворення даних між внутрішніми базами даних або сховищами даних.

Надсилайте додаткові сторонні дані до Heroku Postgres (а потім до Salesforce через Heroku Connect) або безпосередньо до Salesforce.

Xplenty - єдиний інструмент ETL Salesforce to Salesforce.

Нарешті, Xplenty підтримує з'єднувач Rest API для отримання даних із будь-якого API відпочинку.

3) Скайвія

Скайвія - це хмарна платформа даних для інтеграції, резервного копіювання, управління та доступу до даних без кодування, розроблена Devart. Компанія Devart - це відомий і надійний постачальник рішень для доступу до даних, інструментів баз даних, засобів розробки та інших програмних продуктів, у яких понад 40 000 вдячних клієнтів у двох відділах досліджень та розробок.

Skyvia включає рішення ETL для різних сценаріїв інтеграції даних з підтримкою файлів CSV, баз даних (SQL Server, Oracle, PostgreSQL, MySQL), хмарних сховищ даних (Amazon Redshift, Google BigQuery) та хмарних додатків (Salesforce, HubSpot, Dynamics CRM, та багато інших).

Він також включає хмарний інструмент резервного копіювання даних, онлайн-клієнт SQL та рішення OData сервер як послуга.

Основні характеристики:

Skyvia - це комерційні, безкоштовні тарифні плани на основі передплати.

Конфігурація інтеграції без кодування на основі майстра не вимагає великих технічних знань.

Розширені налаштування відображення з константами, пошуками та потужними виразами для перетворення даних.

Автоматизація інтеграції за графіком.

Можливість збереження відносин вихідних даних у цільовій.

Імпорт без дублікатів.

Двонаправлена синхронізація.

Наперед визначені шаблони для поширених випадків інтеграції.

3)IRI ненажерливість

Ненажерливість - це локальна платформа ETL та управління даними з підтримкою хмарних технологій, найвідоміша завдяки значенню `` доступної швидкості в обсязі ", що лежить в основі механізму CoSort, а також завдяки

багатим можливостям виявлення, інтеграції, міграції, управління та аналітики -in, та на Eclipse.

Voracity підтримує сотні джерел даних, а також подає цілі BI та візуалізації безпосередньо як „аналітичну платформу виробництва“.

Користувачі ненажерливості можуть розробляти операції в режимі реального часу або пакетні операції, що поєднують вже оптимізовані операції E, T та L, або використовувати платформу, щоб 'пришвидшити або залишити' існуючий інструмент ETL, такий як Informatica, з міркувань продуктивності чи ціноутворення. Швидкість ненажерливості близька до Аб Ітіо, але її вартість близька до Пентахо.

Основні характеристики:

Різноманітні з'єднувачі для структурованих, напів- та неструктурованих даних, статичних та потокових, застарілих та сучасних, локальних або хмарних.

Маніпуляції з консолідованими даними та IO, включаючи багаторазові перетворення, якість даних та функції маскування, визначені разом.

Трансформації, засновані на багатопотоковому механізмі IRI CoSort, що оптимізує ресурси, або взаємозамінно в MR2, Spark, Spark Stream, Storm або Tez.

Одночасні визначення цілей, включаючи попередньо відсортовані масові навантаження, тестові таблиці, спеціально відформатовані файли, конвеєри та URL-адреси, колекції NoSQL тощо.

Зіставлення даних та міграції можуть переформатувати структури ендіана, поля, запису, файлу та таблиці, додати сурогатні ключі тощо.

Вбудовані майстри для ETL, підмножини, реплікації, збору даних, повільної зміни розмірів, тестування даних тощо.

Функціонал і правила очищення даних для пошуку, фільтрації, уніфікації, заміни, перевірки, регулювання, стандартизації та синтезу значень.

Звіт про те ж проходження, суперечки (для Cognos, Qlik, R, Tableau, Spotfire тощо) або інтеграція зі Splunk та KNIME для аналітики.

Надійне проектування, планування та розгортання завдань, а також управління метаданими з підтримкою Git та IAM.

Сумісність метаданих з Erwin Mapping Manager (для перетворення застарілих завдань ETL) та Мостом інтеграції моделі метаданих.

Прожерливість не є відкритим кодом, але коштує нижче, ніж Talend, коли потрібні кілька двигунів. Ціни на його підписку включають підтримку, документацію та необмежену кількість клієнтів та джерел даних, а також доступні можливості безперервного ліцензування та виконання.

4) Посипати

Посипати є наскрізним управлінням даними та платформою Analytics, що дозволяє користувачам автоматизувати повний шлях до даних, починаючи від збору даних з декількох джерел даних, переміщуючи дані до бажаного сховища даних для створення звітів на ходу. Sprinkle пропонує як SaaS, так і варіант розгортання On-Premise.

Рішення трубопроводу даних у режимі реального часу від Sprinkle дозволяє підприємствам швидше приймати бізнес-рішення і тим самим сприяти загальному зростанню бізнесу. Покращена безпека даних Sprinkle гарантує, що жодні дані не залишать приміщення клієнта, тим самим забезпечуючи 100% безпеку даних.

Платформа без коду Sprinkle робить дані доступними для всіх співробітників в організації незалежно від їх технічних можливостей. Це забезпечує швидше прийняття ділових рішень, оскільки бізнес-командам більше не потрібно покладатися на команду Data Science для надання інформації.

Sprinkle також має додатковий вбудований модуль Advanced Reporting & BI, який можна використовувати для створення інтерактивних інформаційних панелей із звітами перетягування та падіння з деталізацією.

Особливості посипання:

Проковтування нульового коду: Автоматичне виявлення схеми та відображення типів даних до типів складу. Також підтримує дані JSON.

Немає власного коду трансформації: Розбризкування робить ELT (пропонує набагато більше гнучкості та масштабування, ніж застарілий ETL).

5)DBConvert Studio By SLOTIX s.r.o.

DBConvert Studio - це рішення ETL для локальних та хмарних баз даних. Він витягує, трансформує та завантажує дані між різними форматами баз

даних, такими як Oracle, MS SQL, MySQL, PostgreSQL, MS FoxPro, SQLite, Firebird, MS Access, DB2 та Amazon RDS, Amazon Aurora, MS Azure SQL, хмарні дані Google Cloud.

По-перше, студія DBConvert створює одночасне підключення до баз даних. Потім створюється окреме завдання для відстеження процесу міграції / реплікації. Дані можуть бути перенесені або синхронізовані одним або двонаправленим способом.

Копіювання структури бази даних та об'єктів можливо з даними або без них. Кожен об'єкт можна переглянути та налаштувати, щоб запобігти потенційним помилкам.

Основні характеристики:

DBConvert Studio - це інструмент з комерційною ліцензією.

Для тестування доступна безкоштовна пробна версія.

Автоматична міграція схеми та відображення типу даних.

Потрібна маніпуляція без кодування на основі майстра.

Автоматизуйте сеанси / завдання, що виконуються за допомогою планувальника або командного рядка.

Односпрямована синхронізація

Двонаправлена синхронізація

Перегляд подань та запитів.

Він створює журнали міграції та синхронізації для моніторингу процесу.

Він містить функцію масового переміщення великих баз даних.

Можна ввімкнути / вимкнути перетворення кожного елемента як таблиці, поля, індексу, запиту / подання.

IT - PowerCenter

Informatica є лідером в галузі управління хмарними даними Enterprise з більш ніж 500 глобальними партнерами та понад 1 трлн транзакцій на місяць. Це компанія з розробки програмного забезпечення, яка була заснована в 1993 році зі штаб-квартирою в Каліфорнії, США. Дохід компанії - 1,05 млрд. Доларів, а загальна чисельність працівників - близько 4000.

PowerCenter - це продукт, розроблений Informatica для інтеграції даних. Він підтримує життєвий цикл інтеграції даних і забезпечує важливі дані та цінності для бізнесу. PowerCenter підтримує величезний обсяг даних, будь-який тип даних та будь-яке джерело для інтеграції даних.

Основні характеристики:

PowerCenter - це інструмент з комерційною ліцензією.

Це легко доступний інструмент і має прості навчальні модулі.

Він підтримує аналіз даних, міграцію додатків та зберігання даних.

PowerCenter підключає різні хмарні програми та розміщується на веб-службах Amazon і Microsoft Azure.

PowerCenter підтримує гнучкі процеси.

Його можна інтегрувати з іншими інструментами.

Автоматизована перевірка результатів або даних у середовищі розробки, тестування та виробництва.

Людина, яка не є технічною особою, може керувати роботою та контролювати її, що, в свою чергу, зменшує вартість.

7) IBM - Інформаційний сервер Infosphere

IBM - багатонаціональна програмна компанія, заснована в 1911 році зі штаб-квартирою в Нью-Йорку, США, і має офіси у понад 170 країнах. Станом на 2016 рік він має дохід у 79,91 млрд доларів, а загальна кількість працюючих зараз становить 380 000.

Інформаційний сервер Infosphere - це продукт IBM, який був розроблений у 2008 році. Він є лідером на платформі інтеграції даних, яка допомагає зрозуміти та надати критичні значення для бізнесу. Він в основному призначений для компаній з великими даними та великих підприємств.

Основні характеристики :

Це комерційно ліцензований інструмент.

Інформаційний сервер Infosphere - це наскрізна платформа інтеграції даних.

Її можна інтегрувати з Oracle, IBM DB2 та Hadoop System.

Він підтримує SAP через різні плагіни.

Це допомагає вдосконалити стратегію управління даними.

Це також допомагає автоматизувати бізнес-процеси з метою більш економії.

Інтеграція даних у режимі реального часу в декілька систем для всіх типів даних.

Існуючий ліцензований інструмент IBM можна легко інтегрувати з ним.

8) Oracle Data Integrator

Oracle - американська багатонаціональна компанія зі штаб-квартирою в Каліфорнії, яка була заснована в 1977 році. Вона має дохід 37,72 мільярда доларів станом на 2017 рік і загальний штат працівників 138 000.

Oracle Data Integrator (ODI) - це графічне середовище для побудови та управління інтеграцією даних. Цей продукт підходить для великих організацій, які часто потребують міграції. Це всеосяжна платформа для інтеграції даних, яка підтримує великі обсяги даних, послуги передачі даних із підтримкою SOA.

Основні характеристики :

Oracle Data Integrator - це комерційний ліцензований інструмент RTL.

Покращує взаємодію з користувачем за допомогою реконструкції інтерфейсу на основі потоку.

Він підтримує декларативний підхід до трансформації та інтеграції даних.

Швидша та простіша розробка та обслуговування.

Він автоматично виявляє несправні дані та переробляє їх перед переміщенням у цільове додаток.

Oracle Data Integrator підтримує бази даних, такі як IBM DB2, Teradata, Sybase, Netezza, Exadata тощо.

Унікальна архітектура E-LT виключає потребу в сервері ETL, що призводить до економії коштів.

Він інтегрується з іншими продуктами Oracle для обробки та перетворення даних за допомогою існуючих можливостей СУБД.

9) Microsoft - інтегровані служби SQL Server (SSIS)

Microsoft Corporation - американська багатонаціональна компанія, заснована в 1975 році за межами Вашингтона. При загальній чисельності працівників 124 000 він має дохід 89,95 мільярда доларів.

SSIS - це продукт корпорації Microsoft і розроблений для міграції даних. Інтеграція даних відбувається набагато швидше, оскільки процес інтеграції та перетворення даних обробляються в пам'яті. Оскільки продукт Microsoft, SSIS підтримує лише Microsoft SQL Server.

Основні характеристики :

SSIS - це інструмент з комерційною ліцензією.

Майстер імпорту / експорту SSIS допомагає перемістити дані від джерела до пункту призначення.

Він автоматизує обслуговування бази даних SQL Server.

Перетягніть користувальницький інтерфейс для редагування пакетів SSIS.

Перетворення даних включає текстові файли та інші екземпляри сервера SQL.

SSIS має вбудоване середовище сценаріїв, доступне для написання програмного коду.

Його можна інтегрувати з salesforce.com та CRM за допомогою плагінів.

Можливості налагодження та легке управління помилками.

SSIS також може бути інтегрований із програмним забезпеченням для управління змінами, таким як TFS, GitHub тощо.

10) ab initio

Ab Initio - американська приватна компанія, що займається розробкою програмного забезпечення, заснована в 1995 році в штаті Массачусетс, США. Він має офіси у всьому світі у Великобританії, Японії, Франції, Польщі, Німеччині, Сінгапурі та Австралії. Ab Initio спеціалізується на інтеграції додатків та обробці великих обсягів даних.

Він містить шість продуктів для обробки даних, таких як Co> Операційна система, Бібліотека компонентів, Графічне середовище розробки, Підприємство Meta> Середовище, Профайлер даних та Conduct> It. «Ab Initio Co> Операційна система» - це інструмент ETL на основі графічного інтерфейсу з функцією перетягування та перетягування.

Основні характеристики:

Ab Initio - це інструмент з комерційною ліцензією та найдорожчий інструмент на ринку.

Основні особливості Ab Initio легко засвоїти.

Ab Initio Co> Операційна система забезпечує загальний механізм обробки даних та зв'язку між іншими інструментами.

Продукти Ab Initio пропонуються на зручній платформі для додатків паралельної обробки даних.

Паралельна обробка дає можливість обробляти великий обсяг даних.

Він підтримує платформи Windows, Unix, Linux та Mainframe.

Він виконує такі функції, як пакетна обробка, аналіз даних, маніпулювання даними тощо.

Користувачі, які використовують продукти Ab Initio, повинні зберігати конфіденційність, підписуючи NDA.

Поки що ми глибоко розглянули різні інструменти ETL, які доступні на ринку. На сучасному ринку інструменти ETL мають значну цінність, і вони дуже важливі для визначення спрощеного способу видобутку, переробки та завантаження.

Різні інструменти, доступні на ринку, допоможуть вам виконати роботу, але це залежить від вимог.

Кілька компаній використовують концепцію сховища даних, і поєднання технологій та аналітики призведе до постійного зростання сховища даних, що, в свою чергу, збільшить використання інструментів ETL.

Питання для самоконтролю:

1. В чому полягає ETL?
2. Розкрийте сутність дизайна процесу ETL
3. Інструменти ETL

Лекція 5

НАДАННЯ ІНФОРМАЦІЇ (ЗВІТУВАННЯ, ІНФОРМАЦІЙНІ ПАНЕЛІ)

План:

5.1 Проектування та впровадження інформаційних панелей

5.2 Типи звітів, які повільно змінюються

5.1 Проектування та впровадження інформаційних панелей

Інформаційна панель (інфопанель, дашборд) — тип графічного інтерфейсу користувача, що забезпечує наочну презентацію основних показників продуктивності (ОПП), значимих для конкретної цілі чи підприємчого процесу. Інфопанелі є динамічними звітами в режимі реального часу, за допомогою яких керівники та менеджери слідкують за визначеними показниками.

Інформаційні панелі можуть містити графіки, таблиці, картки показників та примітки щодо ефективності підприємчого процесу.

На рівнях аналізу та презентації має бути надано інформаційне забезпечення, орієнтоване на завдання та користувача (рис. 5.1).

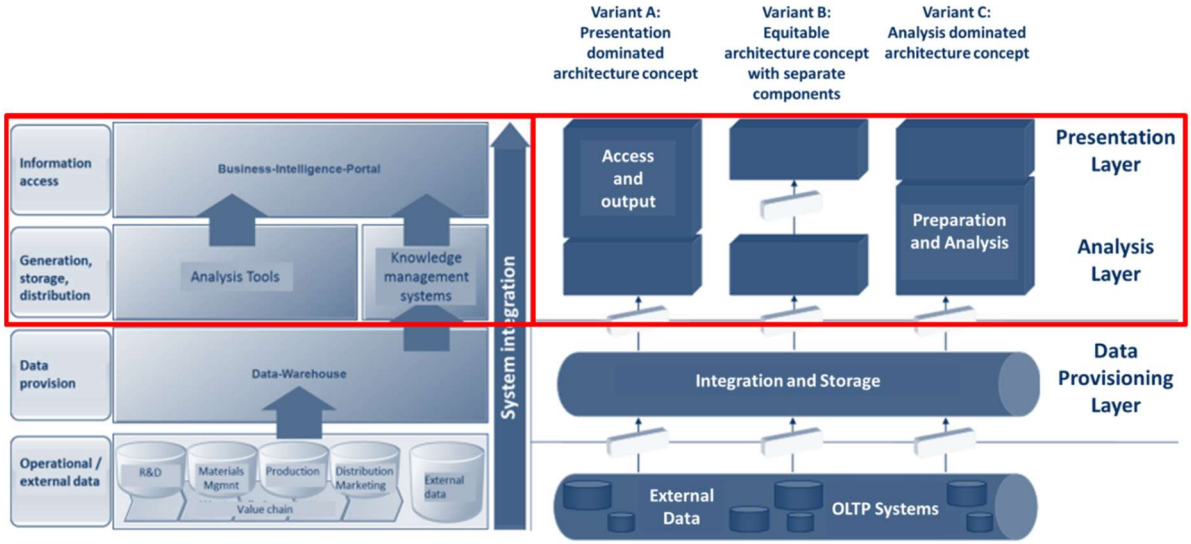


Рис. 5.1 Проектування та впровадження інформаційних панелей

Спектр користувачів на шарі презентації дуже неоднорідний (навички, уподобання), багато різних форм представлення застосовні з різним ступенем взаємодії.

Тип користувача Інформаційний споживач ◊ Звіти, інформаційні панелі

Тип користувача, який використовує переважно інструменти, які обробляють та відображають матеріал даних відповідно до фіксованих шаблонів.

Аналітик типу користувача ◊ OLAP хоче використовувати навігаційно - орієнтований аналіз і хоче рухатися якомога вільніше в наборі даних. Використовуються прості методи та інструменти для відображення / виведення.

Тип користувача Спеціаліст ◊ Видобуток даних покладається насамперед безпосередньо на методи, орієнтовані на здійснення складного аналізу даних. Приймає функціональну складність та менш зручні інтерфейси, не довіряє легкому доступу та засобам редагування (рис. 5.2).

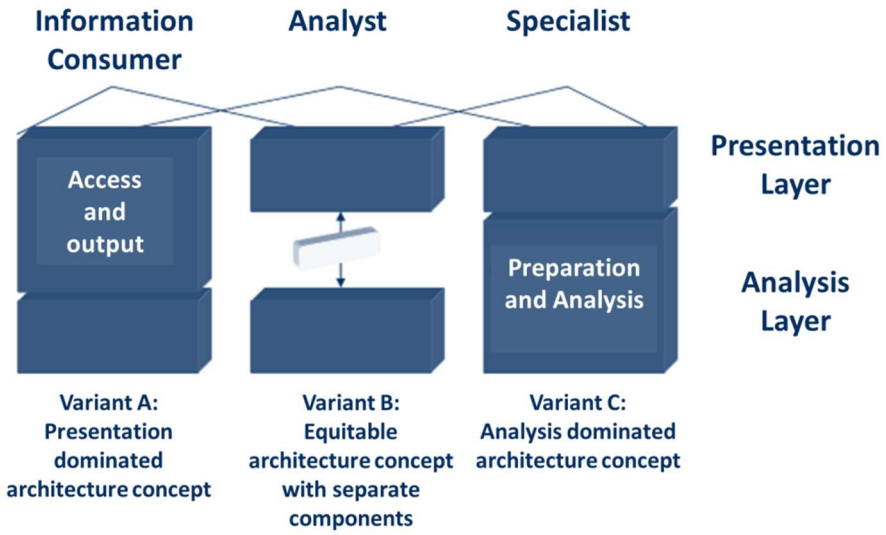


Рис. 5.2 Спектр користувачів на шарі презентації.

Інформаційні панелі містять картки показників, графіки, таблиці або примітки щодо даних про ефективність у вигляді сітки, яку можна налаштувати. Ці картки можна розмістити в будь-якому місці сітки.

Картки показників: відображають ефективність ключових показників.

Графіки й таблиці: дають змогу вставляти звіти з візуальними даними, створені в Редакторі звітів.

Примітки: допомагають користувачам, з якими ви співпрацюєте, краще зрозуміти вашу інформаційну панель.

Картки показників, графіки, таблиці або примітки можна переміщати та міняти їх розмір як потрібно.
Для кожної окремої картки показників, таблиці або графіка можна змінити діапазон дат, щоб переглянути ефективність за певний період часу. Також можна змінити загальну дату всієї інформаційної панелі, щоб побачити її дані за певний день.

Інформаційні панелі дають вам змогу співпрацювати з будь-яким користувачем, який має доступ до вашого облікового запису Google Ads.

Користувачам, які мають доступ лише до електронної пошти, можна надсилати дані інформаційної панелі електронною поштою.

Щоб поділитися даними з користувачами, які не мають доступу до вашого облікового запису, завантажте інформаційну панель у форматі .pdf

Впровадження інформаційних панелей: 60% компаній, що використовують BI, також використовують інформаційні панелі, згідно з дослідженням VBA (Schulz 2015); багато інструментів BI в значній мірі покладаються на інтерактивні та візуальні елементи, щоб доповнити стандартну функціональність BI (рис. 5.3).

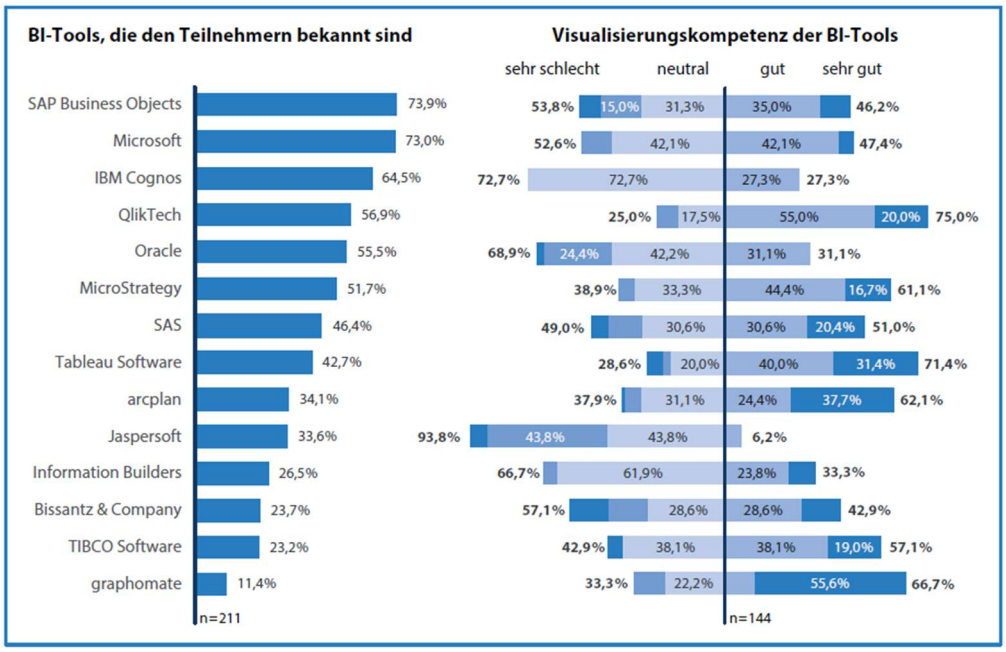


Рис. 5.3 Впровадження інформаційних панелей

5.2 Типи звітів, які повільно змінюються

Кожен маркетолог відчуває тиск на досягнення результатів, будь то успіх, що вимірюється коефіцієнтами конверсії, показниками залучення чи старомодним доходом від продажів. Інформаційні панелі аналітики дають маркетинговим командам інформацію, необхідну для демонстрації рентабельності інвестицій своїм ключовим особам, що приймають рішення.

Хочете показати керівникам компаній, що ваші маркетингові зусилля окупаються? Візуальні діаграми та схеми, включені до цих платформ, є чудовим способом продемонструвати свої маркетингові виграші.

Але припустимо, прогноз не настільки хороший, і ваші маркетингові зусилля не досягають ваших квартальних цілей. Вбудовані інструменти аналітики дозволяють краще зрозуміти, чому ваші стратегії не виконуються на високому рівні і де вони не можуть зв’язатися з цільовою аудиторією та клієнтською базою.

Неможливо виправити корабель, якщо ви не знаєте, чому ви взагалі пішли з курсу. Маркетингові інформаційні панелі надають інформацію та аналіз, щоб повернути вас у правильному напрямку.

Важливе значення має аналітика в режимі реального часу, яка регулярно надає маркетинговим командам оновлення продуктивності. Ви можете динамічно реагувати на зміну тенденцій аудиторії, ринку чи галузі, щоб постійно підтримувати свої маркетингові стратегії.

Загалом, ці інформаційні панелі – це чудовий спосіб згустити вашу маркетингову аналітику та дані у візуально привабливий формат, який легко засвоюється менеджерами з маркетингу та іншими важливими зацікавленими сторонами та особами, що приймають рішення.

Таблиця 5.2. Типи інформаційних панелей

Стратегічні інформаційні панелі	Аналітичні інформаційні панелі	Оперативні інформаційні панелі
використовуються для прийняття переважно складних рішень, які впливають	використовуються для забезпечення взаємодії з даними характеризуються	використовуються для моніторингу або контролю за подіями та обставинами має

на майбутній розвиток компанії чи підрозділу представляють дуже скорочену інформацію (ключові показники ефективності) зазвичай оновлюються щотижня або щомісяця, а не щодня Групи користувачів у компаніях - це керівництво вищого та середнього рівня	здатністю свердлити вниз Деталі повинні стати видимими Зазвичай достатньо щоденного забезпечення	забезпечити швидке втручання у разі виникнення проблем дані повинні бути доступні в режимі реального часу, якщо це можливо, з іншого боку подана інформація має бути легко зрозумілою

Незалежно від типу кампанії існує маркетингова інформаційна панель, яка дасть вам більше розуміння цієї маркетингової діяльності.

Ось декілька найважливіших варіантів, які слід врахувати:

1. Інформаційна панель SEO-маркетингу

Кожен маркетолог хоче бачити рейтинг своїх веб-сайтів у верхній частині сторінок результатів пошуку (SERP) за своїми найбільш бажаними ключовими словами. І оскільки на 3 найвищі позиції припадає 75% усіх кліків пошуку Google як ніколи важливо скористатися інструментами SEO, щоб просунутися вгору по цих SERP і захопити це дорогоцінне нерухоме майно.

SEO – це складна суміш на сторінках, поза сторінками та технічних факторів, які безпосередньо впливають на рейтинг вашого веб-сайту в пошуковій мережі – і, як наслідок, на вашу здатність залучати більше органічного трафіку. Це багато підстав для охоплення, і часто ключові показники ефективності, які вам потрібно відстежувати, можуть варіюватися від базового відвідуваності веб-сайту до ефективності пошуку за ключовими словами конкурента.

Кожній маркетинговій команді потрібна спеціальна інформаційна панель для моніторингу ефективності своєї SEO-кампанії, щоб зацікавлені сторони могли швидко коригувати неефективні стратегії та постійно вдосконалювати та вдосконалювати свої маркетингові зусилля.

Яку платформу ви повинні використовувати для потреб своєї інформаційної панелі ефективності маркетингу SEO?

Google Analytics – очевидний вибір:

Це безкоштовно.

Це можна налаштувати.

Його легко інтегрувати з іншими платформами.

Це зручно для користувачів.

Це підтримується самою компанією Google.

Це галузевий стандарт.

Дійсно, було б дивним, якби команда з цифрового маркетингу не використовувала Google Analytics в якійсь якості для вимірювання та кращого розуміння ефективності веб-сайтів. Звичайно, навіть якщо ви вже використовуєте Google Analytics, завжди є способи отримати більше можливостей від платформи.

Існує безліч маркетингових команд, які використовують Google Analytics для фіксації SEO та ефективності веб-сайтів з високого рівня, але ніколи насправді не заглиблюються в більш детальні показники та тенденції. Скористайтеся спеціальними можливостями звітування платформи для відстеження цінних комерційних цільових сторінок, відстеження конкретних рекламних або електронних кампаній, розбиття конверсій веб-сайтів за типом аудиторії та багато іншого.

Використовуючи колекцію вбудованих віджетів і шаблонів інформаційної панелі, ви можете пристосувати свою інформаційну панель аналітики до нуля щодо маркетингових ключових показників, які є для вас найбільш важливими.

Цей приклад, люб’язно наданий Moz, висвітлює спеціальну інформаційну панель, побудовану для відстеження конверсій з місяця в місяць. Керівникам маркетингових служб легко виявити такі важливі тенденції, як стрибки коефіцієнтів конверсії, без необхідності перебирати купу статистичних даних.

2. Інформаційна панель маркетингу електронною поштою

Електронна пошта – це один з найважливіших маркетингових каналів, що забезпечує середню рентабельність інвестицій 42 долари за кожен витрачений 1 долар . Щоб побачити таку рентабельність інвестицій – або, в ідеалі, перевершити середні показники по галузі – підприємствам слід уважно стежити за кожним аспектом своїх маркетингових кампаній електронною поштою та вносити коригування, коли це необхідно.

Навіть найменший помилковий крок – загальний рядок теми, наприклад, може відхилити передплатника розсилки електронних листів і втратити потенційну перевагу продажів, тому брендам потрібно ретельно аналізувати свої маркетингові кампанії електронною поштою, щоб зрозуміти, які стратегії працюють.

Основні показники маркетингу електронної пошти для відстеження включають:

Відкрита ставка.

Рейтинг кліків.

Коефіцієнт конверсії.

Показник відмов.

Тариф відписки.

Коефіцієнт спільного використання або пересилання електронної пошти.

Темпи зростання списку електронних адрес.

Загальна рентабельність інвестицій.

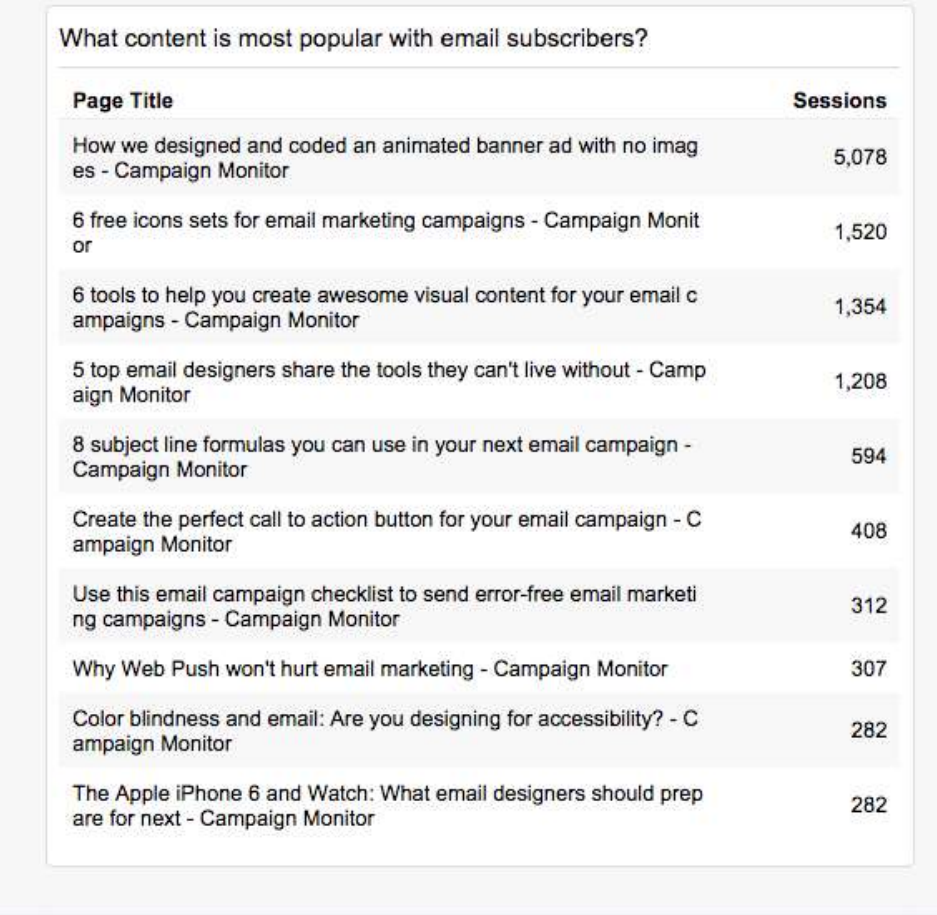
Це лише вершина айсберга.

Також корисно сегментувати мобільні показники з інших форматів, враховуючи взаємодію з користувачами та фактори інтерфейсу, які можуть вплинути на ефективність маркетингу електронною поштою.

Хороша інформаційна панель маркетингу електронної пошти об’єднує всі ці дані та упаковує їх таким чином, щоб їх було легко зрозуміти та засвоїти. Що ще важливіше, це повинно дозволити маркетинговим командам нарізати та нарізати дані, щоб вони могли зосередитися на маркетингових ключових показниках, які є для них найбільш важливими.

Наприклад, якщо ви проводите електронну кампанію в рамках ширшої стратегії маркетингу вмісту, ви можете відстежувати, які типи контенту найбільше резонують у вашої аудиторії.

Google Analytics має безліч віджетів, маркетингових шаблонів інформаційних панелей та інструментів налаштування, які допоможуть вам детально вивчити конкретні показники ефективності. Ось приклад, що показує кількість сеансів, створених різними цільовими сторінками, та частини вмісту, що додаються до розсилок електронною поштою (рис. 5.4.):



What content is most popular with email subscribers?

Page Title	Sessions
How we designed and coded an animated banner ad with no images - Campaign Monitor	5,078
6 free icons sets for email marketing campaigns - Campaign Monitor	1,520
6 tools to help you create awesome visual content for your email campaigns - Campaign Monitor	1,354
5 top email designers share the tools they can't live without - Campaign Monitor	1,208
8 subject line formulas you can use in your next email campaign - Campaign Monitor	594
Create the perfect call to action button for your email campaign - Campaign Monitor	408
Use this email campaign checklist to send error-free email marketing campaigns - Campaign Monitor	312
Why Web Push won't hurt email marketing - Campaign Monitor	307
Color blindness and email: Are you designing for accessibility? - Campaign Monitor	282
The Apple iPhone 6 and Watch: What email designers should prepare for next - Campaign Monitor	282

Рис. 5.4. Google Analytics (Джерело: Campaign Monitor)

Завдяки такій універсальності, Google Analytics може функціонувати як інформаційна панель електронного маркетингу та маркетингу контенту, а також багато інших програм.

3. Інформаційна панель маркетингу в соціальних мережах

Підтримка сильної присутності в соціальних мережах є важливою для компаній B2B та B2C. Якщо ви не можете зрозуміти, наскільки ефективним є ваш бренд на різних платформах соціальних мереж, можливо, ви не будуєте

того рівня зацікавленості, який ви думаєте. З огляду на те, як швидко все рухається у світі соціальних медіа, бренди повинні залишатися на висоті своєї аналітики в соціальних мережах цілодобово та без вихідних.

Спеціальна інформаційна панель маркетингу в соціальних мережах може допомогти відстежувати тенденції та події в режимі реального часу та надавати практичну інформацію, яка може бути негайно застосована до вашої стратегії.

Цей інтелект може надати вам змогу бути більш чуйними, змінювати підхід до соціальних мереж на льоту і продовжувати розвиватися разом із цільовою аудиторією. Інформаційні панелі маркетингу в соціальних мережах надають інформацію, необхідну для того, щоб усе це відбулося.

Команди маркетингу можуть знайти всі відповідні метрики маркетингу в соціальних мережах в одному легко зрозумілому інтерфейсі.

Ви можете відстежувати такі ключові показники ефективності, як:

Послідовники.

Лайки.

Охоплення.

Враження.

Середній коефіцієнт залучення.

Темпи зростання аудиторії.

Найпопулярніші посади.

Соціальна частка голосу.

Є безліч хороших варіантів на вибір – ще раз, Google Analytics має інструменти, що допомагають візуалізувати аналітику в соціальних мережах – але одна з найкращих спеціалізованих платформ існує Sprout Соціальна . Ви можете розподілити свої маркетингові показники в соціальних мережах за каналами та сегментувати різні ключові показники ефективності, щоб отримати кращий погляд на ефективність вашого бренду.

Маючи стільки можливостей налаштування для вивчення, маркетологи можуть контролювати свою присутність у соціальних мережах з будь-якої

точки зору, яка їм подобається, будь то огляд високого рівня або поглиблений аналіз певного сегменту вашої аудиторії та послідовників.

Дані повинні керуватися кожним маркетинговим рішенням, яке ви приймаєте, але для тих, хто не має наукового ступеня статистики, пробирання всієї цієї інформації, щоб знайти реальну інформацію, може здатися нездоланним. Інформаційні панелі цифрового маркетингу використовують інтуїтивно зрозумілі наочні засоби для спрощення, а аналітичні дисплеї виводять на поверхню найбільш релевантні та корисні маркетингові дані.

Питання для самоконтролю:

1. Інформаційна панель – це?
2. Типи звітів, які повільно змінюються
3. Які типи інформаційних панелей вам відомі?

Лекція 6

АНАЛІТИЧНИЙ ЖИТТЄВИЙ ЦИКЛ ТА МЕТОДИ: КЛАСТЕРИЗАЦІЯ, КЛАСИФІКАЦІЯ, МАШИННЕ НАВЧАННЯ

План:

6.1. Що таке аналітичний життєвий цикл.
6.2. Сучасна архітектура даних. Машинне навчання.
6.3. Сутність класифікації. Знайомство з методами кластеризації даних

6.1 Що таке аналітичний життєвий цикл.

В даний час системи бізнес -аналітики (BI) широко використовуються в багатьох сферах бізнесу, які ґрунтуються на прийнятті рішень щодо створення вартості. BI-це процес отримання наявних даних для вилучення, аналізу та прогнозування критично важливих для бізнесу результатів. Традиційна BI зосереджена на зборі, вилученні та впорядкуванні даних для забезпечення ефективної та професійної обробки запитів, щоб отримати уявлення з історичних даних. Завдяки наявності великих даних, Інтернету речей (IoT), штучного інтелекту (AI) та хмарних обчислень (CC), BI став більш критичним і важливим процесом і отримав більший інтерес як у галузях промисловості, так і в наукових колах. Основна проблема полягає в тому, як використовувати ці нові технології для створення цінності на основі даних для сучасної BI. У цьому розділі, щоб вирішити цю проблему, важливість аналізу великих даних, видобутку даних, Буде представлено та обговорено ШІ для побудови та вдосконалення сучасного BI. Крім того, виклики та можливості для створення цінності даних шляхом створення сучасних процесів BI.

У сучасному цифровому світі дані мають величезне значення. Протягом свого життя він проходить різні етапи, під час створення, тестування, обробки, споживання та повторного використання. Життєвий цикл Data Analytics намічає ці етапи для професіоналів, які працюють над проектами аналізу даних. Ці фази розташовані у круговій структурі, яка формує життєвий цикл Data Analytics. Кожен крок має своє значення та характеристики (рис. 6.1)

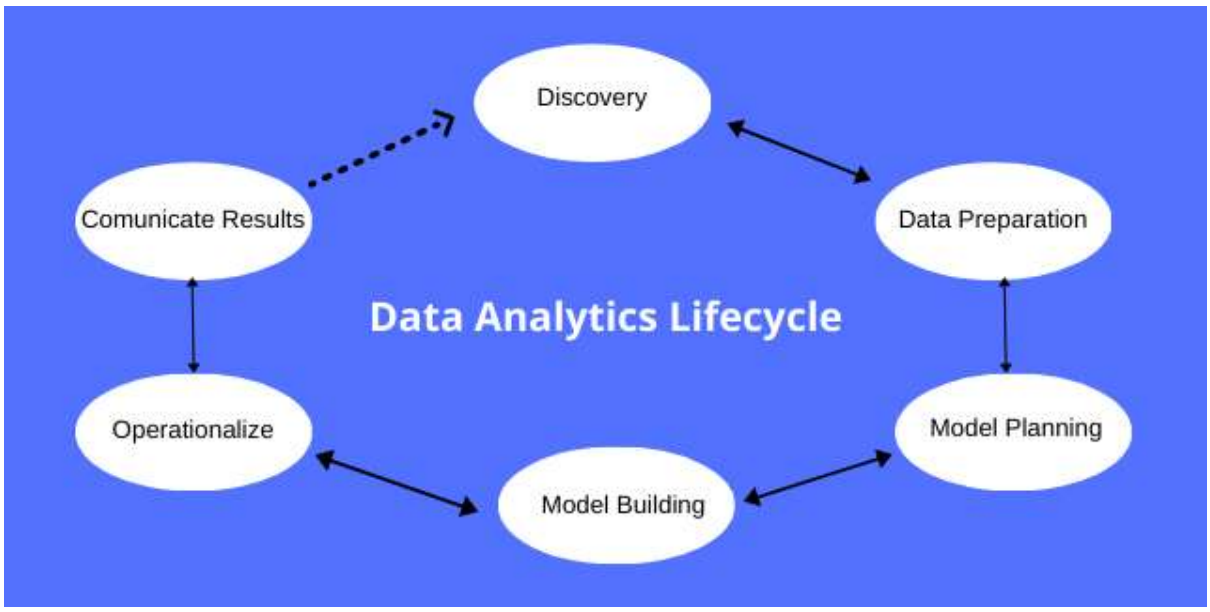


Рис. 6.1 Життєвий цикл Data Analytics

Життєвий цикл аналітики даних визначає дорожню карту того, як дані генеруються, збираються, обробляються, використовуються та аналізуються для досягнення бізнес -цілей. Він пропонує систематичний спосіб управління даними для їх перетворення в інформацію, яка може бути використана для досягнення організаційних та проектних цілей. Процес передбачає напрямки та методи вилучення інформації з даних та продовження у правильному напрямку для досягнення бізнес -цілей. Фахівці в галузі даних використовують кругову форму життєвого циклу для переходу до аналізу даних вперед чи назад. На основі нещодавно отриманої інформації вони можуть вирішити, чи продовжувати своє існуюче дослідження, чи скасувати його, і повторити повний аналіз. Життєвий цикл Data Analytics спрямовує їх у цьому процесі.

Етапи життєвого циклу аналітики даних.

Немає чітко визначеної структури фаз у життєвому циклі аналітики даних, а отже, може не бути однорідності на цих етапах. Деякі фахівці з обробки даних можуть виконувати додаткові кроки, а деякі можуть пропустити деякі етапи взагалі або працювати одночасно на різних етапах. Давайте обговоримо різні фази життєвого циклу аналізу даних.

Розберемо етапи, які є основоположними для кожного процесу аналізу даних. Отже, вони, швидше за все, будуть присутні в життєвому циклі

більшості проектів з аналізу даних. Життєвий цикл Data Analytics в першу чергу складається з 6 фаз.

Фаза 1: Виявлення та формування даних Цей етап стосується визначення мети даних та способів її досягнення до кінця життєвого циклу аналізу даних. Етап полягає у визначенні найважливіших цілей, які бізнес намагається виявити шляхом складання даних. Під час цього процесу команда дізнається про сферу бізнесу та перевіряє, чи підрозділ чи організація працювала над подібними проектами, щоб посилатися на якісь знання. На цьому етапі команда також оцінює технології, людей, дані та час.

Наприклад, під час роботи з невеликим набором даних команда може використовувати Excel. Однак більш складні завдання вимагають більш жорстких інструментів для підготовки та дослідження даних. У таких сценаріях команді потрібно буде використовувати Python, R, Tableau Desktop або Tableau Prep та інші засоби очищення даних. Найважливіші дії цього етапу включають формулювання бізнес -проблеми, формулювання початкових гіпотез для перевірки та початок вивчення даних.

Фаза 2: Підготовка та обробка даних. На цьому етапі фокус уваги експертів переходить від вимог бізнесу до вимог до інформації. Одним з істотних аспектів цього етапу є забезпечення доступності даних для обробки. Етап охоплює збір, обробку та очищення накопичених даних (рис. 6.2).

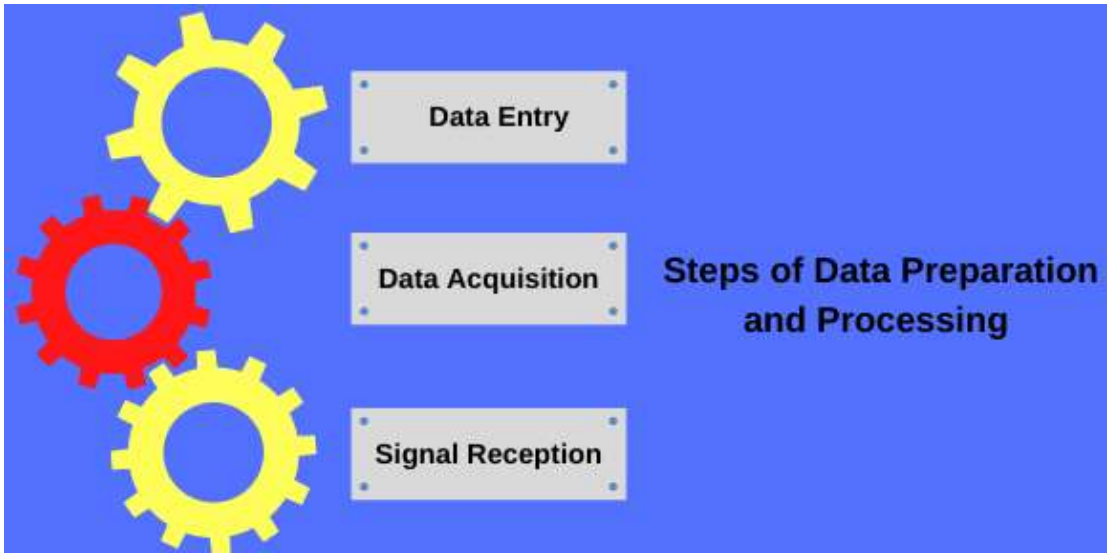


Рис. 6.2 Підготовка та обробка даних.

На початковому етапі цього етапу команда збирає цінну інформацію та продовжує життєвий цикл бізнес -екосистеми. Для цього використовуються різні методи збору даних, наприклад

- Введення даних - Збір останніх даних за допомогою методів введення даних вручну або цифрових систем в організації

- Збір даних - збір даних із зовнішніх джерел

- Прийом сигналу - збір даних з цифрових пристроїв, включаючи Інтернет речей та системи управління.

Фаза 3: Розробка моделі.

Цей етап потребує наявності аналітичної пісочниці для роботи групи з даними та аналізу протягом усього періоду проекту. Команда може завантажувати дані кількома способами.

- Витяг, перетворення, завантаження (ETL) - він перетворює дані на основі набору правил бізнесу, перш ніж завантажувати їх у пісочницю.

- Витяг, завантаження, перетворення (ELT) - завантажує дані у пісочницю, а потім перетворює їх на основі набору правил бізнесу.

- Extract, Transform, Load, Transform (ETLT) - це поєднання ETL та ELT і має два рівні трансформації. Команда визначає змінні для класифікації даних, визначає та виправляє помилки даних. Помилки даних можуть бути будь - якими, включаючи відсутні дані, нелогічні значення, дублікати та орфографічні помилки. Наприклад, команда зараховує середній бал даних для категорій за відсутні значення. Це дозволяє більш ефективно обробляти дані без перекосу даних. Після очищення даних команда визначає методи, методи та робочий процес для побудови моделі на наступному етапі. Команда досліджує дані, визначає відносини між точками даних для вибору ключових змінних і врешті -решт розробляє відповідну модель.

Фаза 4: Побудова моделі.

На цьому етапі команда розробляє набори даних тестування, навчання та виробництва. Крім того, команда будує та виконує моделі ретельно, як це було заплановано на етапі планування моделі. Вони перевіряють дані та намагаються знайти відповіді на поставлені цілі. Вони використовують різні методи статистичного моделювання, такі як методи регресії, дерева рішень, моделювання випадкових лісів та нейромережі, і виконують пробний запуск, щоб визначити, чи відповідає він наборам даних.

Фаза 5: Повідомлення результатів та публікація.

Цей етап має на меті визначити успішність чи невдачу результатів проекту та розпочати співпрацю із значними зацікавленими сторонами. Команда визначає важливі висновки свого аналізу, вимірює відповідну вартість бізнесу та створює узагальнену розповідь, щоб передати результати зацікавлених сторін.

Фаза 6: Вимірювання ефективності.

На цьому останньому етапі команда представляє зацікавленим сторонам детальний звіт із кодуванням, брифінгом, основними висновками та технічними документами та документами. Крім того, дані переміщуються в середовище живого середовища та контролюються для оцінки ефективності аналізу. Якщо результати відповідають поставленій меті, результати та звіти остаточно опрацьовуються. З іншого боку, якщо вони відхиляються від встановленого наміру, команда рухається назад у життєвому циклі до будь - якої попередньої фази, щоб змінити вхідні дані та отримати інший результат.

Круговий процес життєвого циклу Data Analytics складається з 6 основних етапів, які визначають спосіб створення, збирання, обробки, використання та аналізу інформації. Визначення бізнес-цілей та прагнення до їх досягнення проведуть вас на решті етапів.

6.2. Сучасн архітектура даних. Машинне навчання

Традиційна архітектура бізнес -додатків має три рівні: дані, додаток та презентація. У трирівневій архітектурі час виконання дуже важко передбачити через зв’язок між процесами управління низькорівневими даними та операціями високого рівня. Зазвичай рішення для управління робочим навантаженням будуються поверх СУБД загального призначення, яким потрібні затримки часу для виконання паралельних запитів. За допомогою сучасних бізнес -додатків це створить проблеми для функцій оперативної інформації в режимі реального часу. Тому важливі технології, які дозволяють одночасно виконувати господарські операції та аналітичні запити щодо одних і тих же даних. Сьогодні організації використовують ETL для вилучення даних, здійснення передач та завантаження даних, які перетворюються на сховище даних. Ця модель базується на двох типах критичних процесів бізнес -процесів: онлайн -аналітична обробка (OLAP) та онлайн -обробка транзакцій (OLTP). OLTP використовується для управління господарськими операціями,

такими як обробка замовлення. OLAP використовується для підтримки прийняття стратегічних рішень як аналітика продажів.

Виклики:Традиційно навантаження OLAP та OLTP виконуються в одній системі баз даних. Однак навантаження OLAP здебільшого складаються з масового читання даних, які постійно оновлюються OLTP. Тому продуктивність обробки транзакцій може бути несподіваною через конкуренцію за ресурс, коли обидва навантаження виконуються в одній базі даних. Таким чином, необхідно відокремити навантаження від OLAP та OLTP. Рисунок 6.3 описує фонову інформацію на основі ETL, де OLAP і OLTP розділені.

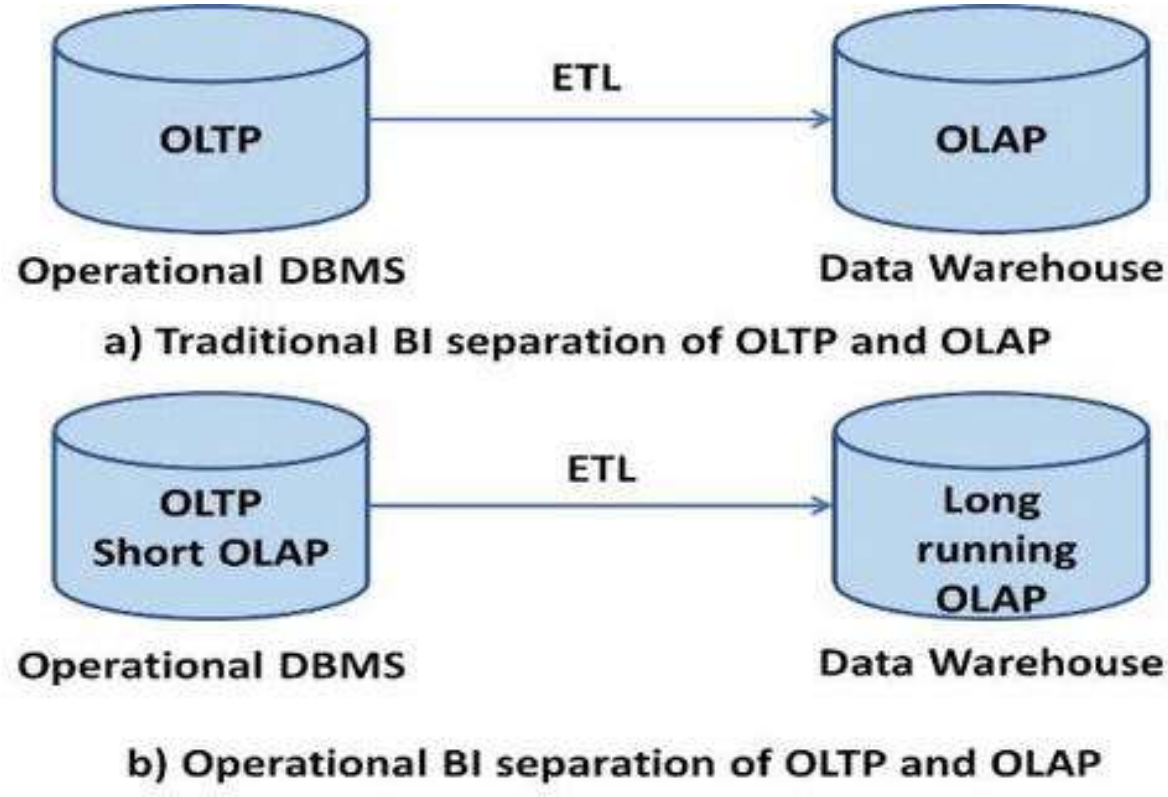


Рис. 6.3. Класифікація систем баз даних.

У цій архітектурі кожне робоче навантаження OLAP має чекати, поки дані в пулі дат будуть повністю оновлені та видимі, що спричинить затримки. Сьогодні, щоб зменшити затримку, операційні системи BI виконують OLTP та короткострокові аналітичні запити разом у СУБД, як показано на рис. 6.3 b . Ці навантаження називаються короткими навантаженнями OLAP. Однак довгострокові навантаження OLAP можуть конфліктувати з багатьма короткими транзакціями OLTP, які вносять зміни до

бази даних. Тому для боротьби з конкуренцією ресурсів потрібна висока синхронізація, що призводить до зниження використання всіх ресурсів.

Також комерційна система управління базами даних (СУБД) використовує спеціальні методи як тіньову копію для обробки змішаних робочих навантажень з меншими накладними витратами. Тобто, у різних логічних версіях даних різні навантаження будуть розділені та виконуватись. Тому додатковий простір може бути збільшено, що збільшить витрати та вимоги до інфраструктури. Тому в поточних СУБД на основі дисків основним викликом є управління цими змішаними робочими навантаженнями (OLAP та OLTP).

Поточні системи BI

Розширені системи традиційного BI: Поточні традиційні технології BI можуть виконувати запити OLAP та транзакції OLTP в одній базі даних, не заважаючи один одному. Поєднання цих змішаних робочих навантажень з однією системою потребує надзвичайного підвищення продуктивності через величезний вибух динамічного розміру даних.

“База даних в пам’яті (IMDB)”: Сьогодні в більшості систем BI змішане робоче навантаження OLAP та OLTP в одній системі можна обробляти за допомогою бази даних In-Memory (IMDB) (також званої Master-Memory). Ця техніка потребує, щоб система зберігала всі дані в основній пам’яті, оскільки вона швидше, ніж оптимізовані бази даних на диску, а внутрішні алгоритми оптимізації використовують менше інструкцій процесора і є простішими. У разі запиту даних ця техніка забезпечує більш передбачувану та швидшу роботу диска за рахунок скорочення часу пошуку. Однак системам IMDB може не вистачати довговічності через втрату збереженої інформації при скиданні пристрою або при відключенні живлення. Багато систем IMDB пропонують різні механізми підтримки довговічності, такі як знімки, енергонезалежний DIMM, енергонезалежна оперативна пам'ять, журналювання транзакцій та висока доступність.

У таблиці 6.1 наведено системи сучасного BI, які використовують різні методи утримання більшості або всіх даних у основній пам’яті для отримання високої продуктивності OLTP. Наприклад, розподілений набір спільних пристроїв використовується для запуску системи H-Store, де дані повністю розташовані в основній пам’яті. H-Store може виконувати обробку транзакцій з високими показниками продуктивності, видаляючи традиційні функції

СУБД, такі як управління буфером, блокування та закриття. Нещодавно прототип H-Store був проданий стартапом під назвою VoltDB.

"Гібриди з дисковою базою даних": основна пам'ять стала достатньо великою для обробки більшості баз даних OLTP, проте це може бути не найкращим вибором. Для навантажень OLTP за допомогою шаблонів доступу, де одні записи є "крутими" (рідко або взагалі не доступні), інші - "гарячими" (часто доступні). Отже, найхолодніші записи зберігаються на швидких вторинних пристроях зберігання даних у сучасних системах для забезпечення гарної продуктивності. Наприклад, Стойка та Айламакі запропонували спосіб міграції даних БД первинної пам’яті на більш дешеве та велике вторинне сховище. Для поліпшення частоти серцевих скорочень пам'яті та зменшення міграції операційної системи вводу -виводу, реляційні структури даних реорганізуються за допомогою статистики доступу для робочих навантажень OLTP. Нещодавно Сибір було впроваджено як холодну систему управління даними в Microsoft Hekaton IMDB. Як і він не вимагає зберігання всієї бази даних у головній пам’яті.

Таблиця 6.1.

Системи сучасного BI, які використовують різні методи для утримання більшості або всіх даних у головній пам’яті.

Система	Тип	Методи	Досягнення
H-магазин	IMDB	Розподілена техніка зберігання рядків	Висока пропускна здатність OLTP
Раду Стоїка	Гібрид	Реорганізація даних	Висока продуктивність, зменшення підкачкового вводу -виводу та покращення швидкості попадання в пам'ять
Сибір	Гібрид	Холодний доступ до даних та механізми міграції	Прийнятні швидкості доступу з втратою пропускної здатності 7–14%

Hekaton зосереджується на тому, як записи переміщуються до холодильної камери та з неї, і як послідовно здійснюють доступ та оновлюють записи в холодному сховищі для транзакцій. Таким чином, лише деякі таблиці можуть бути оголошені та керовані в основній пам’яті Гекатоном. Оцінка досвіду показує, що коли холодне сховище знаходиться на товарній золі, Сибір може призвести до відповідної втрати продуктивності на 7–14%, враховуючи, що швидкість доступу до холодних даних є покращеною базою даних основної пам’яті.

2. Сучасні особливості з Системи BI:Існує три сучасних індикатора обстеження інформації: оперативне біологічне дослідження, ситуаційне тимчасове обстеження та самообстеження самообслуговування. У той час як система H-Store призначена лише для обробки транзакцій OLTP, сучасна система під назвою HyPer може впоратися зі змішаними робочими навантаженнями як OLTP, так і OLAP, що мають надзвичайно високі показники пропускної здатності, використовуючи механізм з низькими накладними витратами для створення диференціальних знімків. У цій системі використовується розблокований підхід, який дозволяє виконувати всі транзакції OLTP послідовно або на спеціальних розділах. Паралельно з обробкою OLTP система HyPer виконує запити OLAP на одному і тому ж знімку та послідовно.

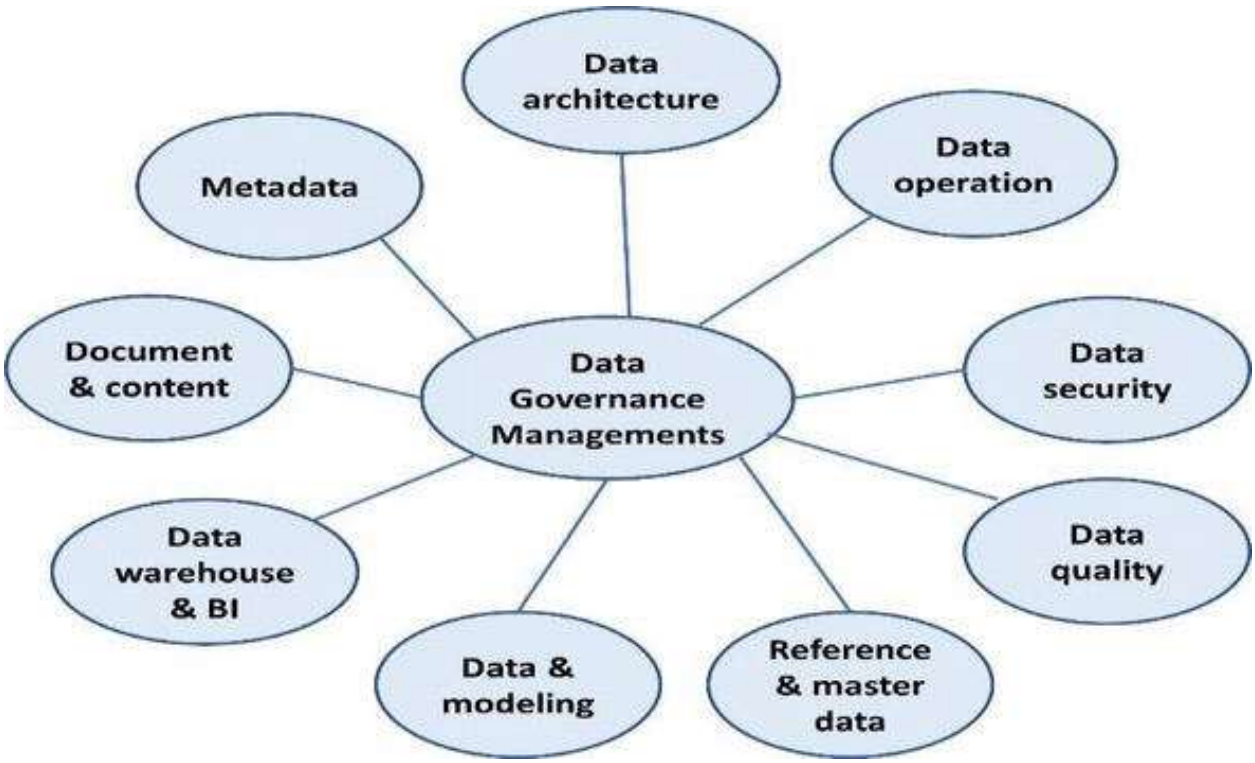
Кастелланос та ін. запропонував нову платформу, покликану повідомляти керівників підприємств про ситуації, які можуть вплинути на їхній бізнес. SIE-OBI інтегрує функції, необхідні для використання відповідної швидкої інформації з Інтернету. Вони запропонували нові схеми для вилучення та пов'язування інформації, отриманої з Інтернету, із збереженими історичними даними у сховищі даних для виявлення шаблонів позицій. Відповідна інформація витягується лише з двох або більше різних неструктурованих джерел даних, зазвичай одного потоку внутрішнього повільного тексту та потоку зовнішнього швидкого тексту. Ця платформа мінімізації часу та зусиль була побудована для побудови повільних та швидких потоків даних, що інтегрують структуровані та неорганізовані потоки, та для їх аналізу майже у реальному часі.

Управління даними

Довідка: у DAMA I управління даними визначається як «здійснення повноважень та контроль за управлінням активами даних, плануванням, наглядом та контролем за управлінням та використанням даних». Управління даними описує відповідальність та роль організації у просуванні бажаної

поведінки у використанні даних. Управління даними відрізняється від управління даними, яке передбачає встановлення стандартів якості даних, прийняття рішень та їх впровадження. Він також відрізняється від «Управління BI», яке має на меті створити спеціальну основу для прийняття рішень шляхом управління усіма видами діяльності в середовищі BI. DAMA I ідентифікує 10 функцій управління даними, як показано в рис. 6.4. Функція управління даними-це нагляд на високому рівні, планування та контроль усіх інших функцій. Існують чотири функції управління даними, пов'язані з наступним поколінням біологічної інформації, які вимагають швидкого доступу до даних, використання зовнішніх даних та загального аналізу даних користувачами. Управління архітектурою даних включає встановлення стандартів даних, підтримку та розвиток структур корпоративних даних та зв'язування прикладних проектів та архітектури. Відділ якості даних зосереджується на плануванні, впровадженні та контролі діяльності, яка застосовує методи управління якістю для вимірювання, оцінки, вдосконалення та забезпечення використання даних. Зберігання даних та управління бізнес - аналітикою зосереджені на наданні даних підтримки прийняття рішень для звітності, запитів та аналізу.

Розгортання BI наступного покоління в управлінні даними: Управління даними стало життєво важливим для організації, оскільки дані стають невід'ємними. Бізнес отримує свою цінність для бізнесу і приймає рішення на основі інформації, отриманої з даних. Отже, контроль даних необхідний для забезпечення якості даних, що безпосередньо впливає на якість рішень, прийнятих організацією. Більш ефективне управління даними (DG) може призвести до більш високої шкали прийняття рішень. Для досягнення ефективного управління даними моделі зрілості управління даними підприємства допомагають зрозуміти DG та визначити наступний очікуваний план. Було запропоновано багато моделей зрілості управління даними, які спрямовують організацію розуміти, що таке рівень управління даними. У Oracle передбачав, що модель зрілості управління даними допоможе організації знайти її в еволюції системи управління даними, визначити короткострокові кроки, необхідні для виходу на наступний рівень, та покращить можливості управління даними. У моделі Oracle найвищим рівнем зрілості є інтеграція управління даними з BI.



Римс. 6.4. Рамки управління даними, визначені в Dama I

Наступне покоління BI підтримує практично уявлення про реальний час із використанням зовнішньої інформації, яка генерує великий обсяг даних та маніпулювання нею. Отже, це вимагає дуже зрілого Генерального директора для забезпечення якості даних, надійності та цілісності. Три характеристики мають вирішальне значення для отримання точного розуміння за допомогою методів видобутку даних. Наприклад, у "самообслуговуванні" BI (наприклад, Tableau та QlikTech), користувачі можуть отримувати уявлення з багатьох джерел даних без моделювання середовища даних та реалізації складних операцій ETL, що є однією з найбільш трудомістких та складних операцій завдання в BI. Отже, ці нові функції дозволяють користувачам легко отримувати доступ до даних, отримувати швидкі результати та візуальну візуалізацію даних. Щоб забезпечити еволюцію наступного покоління біологічної інформації, управління даними має вирішальне значення для достовірності даних відкритого бачення. Наприклад, у разі самообслуговування BI той факт, що кінцеві користувачі можуть отримати доступ та обробляти свої дані, знижує надійність результатів BI. В управлінні даними можна розглянути корисні функції для забезпечення надійності, такі як відстеження співвідношення даних до джерела та створення записів про те, як дані обробляються або передаються. Однак інтеграція управління даними в наступне покоління біологічної інформації зіткнулася з деякими проблемами

через вимоги гнучких та надійних відповідей, хоча існує величезна кількість зовнішніх даних та залучення громадських користувачів.

Проблеми управління даними: Існує дві основні переваги наступного покоління біологічної інформації, яка впливає на модель управління даними. Прийняття рішень у BI наступного покоління має бути ефективнішим та швидшим між величезною кількістю даних, які надходять із багатьох форматів даних та джерел. Однак дані з багатьох джерел ускладнюють управління даними та ускладнюють належний контроль. Це також може призвести до прийняття неефективних рішень. У разі, коли дані надходять з різних конфліктуючих джерел, особи, які приймають рішення, повинні провести додаткові дослідження та аналіз даних та різних джерел цих даних, щоб визначити, що є правдивим і точним, або його наближення, що буде коштувати дорогих операцій. Тому управління даними через неоднорідні джерела в системі BI наступного покоління є дуже важливим.

Загалом, центральна IT -організація та багато керівників даних беруть участь у ініціативах управління даними та мають сховище метаданих для платформи управління даними та набір засобів управління даними для роботи з різними даними. Заздалегідь вони стандартизують загальні визначення основних даних та довідкових даних, які широко поширені у багатьох корпоративних додатках. Коли вони отримують різні дані, вони відповідають їм, щоб визначити заздалегідь визначені загальні дані, визначити їх якість, визначити, які правила, конвертувати та об’єднати. Однак у наступному поколінні BI користувачі також самі вибирають, обробляють або об’єднують свої імена даних, використовуючи різні інструменти BI для самообслуговування. Вони можуть захотіти завантажити в БД і поділитися своїм баченням з іншими. Участь бізнес -користувачів у процесі обробки даних може призвести до того, що дані потрапили в безлад, де одні й ті ж дані можуть бути перетворені та об’єднані різними способами через менеджерів даних та центральну організацію, що використовує інструменти управління даними, та бізнес -користувачів, які мають інструменти BI для “ самообслуговування". Отже, критерії спільного використання метаданих мають вирішальне значення для передачі спільних даних, загальних назв даних та правил спільної інтеграції.

Модель управління даними для BI наступного покоління: Розробка моделі управління даними спрямована на централізацію та порівняння. Децентралізація та ієрархія проти кооперативу. Центральний дизайн покладає всі повноваження щодо прийняття рішень у центральному

відділі IT, тоді як децентралізований дизайн покладає повноваження на окремі бізнес-підрозділи.

Термін великі дані - це група величезних і складних наборів даних з різних джерел, де управління та традиційні методи обробки додатків стикаються з труднощами їх обробки. Великі дані-це сукупність великої кількості структурованих або неструктурованих даних, які обробляються та аналізуються для прийняття обґрунтованих рішень або оцінки. Ці дані можна взяти з різних джерел, включаючи історію перегляду, географічне розташування, соціальні медіа, медичні документи та записи про покупки. Великі дані складаються зі складних даних, які зменшать обчислювальну здатність традиційних простих систем баз даних. Великими даними є три основні характеристики: (1) Обсяг - це функція, яка використовується для опису величезних обсягів даних, які використовують великі дані. Зазвичай діапазон обсягів даних починається від ГБ до YouTube. Великі дані повинні мати можливість обробляти будь -який обсяг даних, навіть при його очікуваному зростанні. (2) Різноманітність - це функція, яка використовується для опису різних типів джерел даних, які використовуються як частина великої системи аналізу даних. В даний час існує багато форматів зберігання даних, які використовуються комп'ютерами по всьому світу. Один формат - це структуровані дані, такі як бази даних і. CSV, відео, служба коротких повідомлень (SMS) та документи Excel. Неорганізовані дані можуть бути у формі рукописних приміток. Усі дані з цих джерел ідеально використовуватимуться для аналітики великих даних. (3) Швидкість - це функція, яка використовується для опису швидкості, з якою генеруються дані. Він також використовується для опису швидкості обробки сформованих даних. Одним натисканням кнопки інтернет -роздріб може швидко переглянути великі дані про конкретного клієнта. Швидкість також важлива для того, щоб дані оновлювалися та оновлювалися в режимі реального часу, дозволяючи системі працювати якнайкраще. Ця швидкість необхідна, оскільки генерація даних у режимі реального часу допомагає організаціям прискорювати операції. Що може заощадити установам велику суму грошей. Ця швидкість необхідна, оскільки генерація даних у режимі реального часу допомагає організаціям прискорювати операції. Що може заощадити установам велику суму грошей. Ця швидкість необхідна, оскільки генерація даних у режимі реального часу допомагає організаціям прискорювати операції. Що може заощадити установам велику суму грошей.

Сьогодні багато компаній все більше зацікавлені у використанні технологій обміну великими даними для підтримки свого БІ, тому стає дуже

важливим зрозуміти різні практичні питання з попереднього досвіду в системах БІ. Сучасні системи BI досліджують світ і використовують ці точки даних для точного рекомендування найкращих варіантів та прогнозування результатів. Оскільки системи BI продовжують будуватись у режимі реального часу, попит на збір, інтеграцію, обробку та візуалізацію даних зростає майже в режимі реального часу. Системи BI характеризуються високими можливостями чутливості, що спостерігається у датчиках з великою різноманітністю датчиків, починаючи від мобільних телефонів, персональних комп'ютерів та пристроїв для відстеження стану здоров'я до технологій Інтернету речей (IoT), призначених для надання контекстуального та смислового звуку суб'єктам, які раніше не могли внести інтелектуальну допомогу у прийняття ключових рішень. Отже, сьогодні багато компаній аналізують великі дані.

Потрібна велика аналітика даних, яка є технікою машинного навчання через часто розповсюджувані набори даних, а її конфіденційність та розмір є свідченням методів розповсюдження, де дані знаходяться на платформах з різними обчислювальними можливостями та мережами. Переваги різноманітності додатків та аналізу великих даних створюють проблеми. Наприклад, щогодини сервери Walmart обробляють понад мільйон транзакцій для клієнта, і ця інформація зберігається у базах даних, які містять більше 2,5 петабайт даних, що в 167 разів перевищує кількість книг у Бібліотеці Конгресу. У цьому випадку адронний коллайдер CERN виробляє близько 15 петабайт даних щорічно, і цього достатньо для заповнення понад 1,7 мільйона двошарових DVD -дисків щорічно. Аналітика великих даних використовується для освіти, охорони здоров'я, засобів масової інформації, страхування, виробництва та уряду. Аналіз великих даних бізнес-аналітики та систем підтримки прийняття рішень, які дозволяють організаціям охорони здоров'я аналізувати розмір даних, різноманітність та величезну швидкість, були розроблені у широкому спектрі мереж охорони здоров'я для підтримки прийняття рішень та дій на основі фактичних даних. Отже, з дискусії видно, що управління даними та аналіз великих даних важливі в BI з чотирьох причин:

Краще прийняття рішень (BDM): Аналітика великих даних може аналізувати поточні та старі дані для прогнозування майбутнього. Отже, компанії можуть приймати не тільки кращі поточні рішення, а й готуватися до майбутнього.

Зниження витрат (CR): Технології великих даних, такі як хмарні аналітичні дані та Hadoop, пропонують великі економічні переваги при зберіганні великої кількості даних. Крім того, він надав уявлення про вплив різних змінних.

Нові продукти та послуги (NPS): Завдяки можливості вимірювати потреби та задоволеність клієнтів за допомогою аналітики, ми отримуємо силу дати клієнтам те, що вони хочуть. Отже, все більше компаній створюють нові продукти та послуги для задоволення потреб клієнтів.

Розуміння ринкових умов (UMC): Аналізуючи великі дані, ми можемо краще зрозуміти поточні ринкові умови для отримання важливої інформації. Крім того, є кілька особливостей та проблем, які слід враховувати в інструментах та прийомах аналізу великих даних, а також вони включають масштабованість та стійкість до помилок. У наведеній нижче таблиці 6.1 представлено кілька широко використовуваних інструментів з перевагами аналітики великих даних.

Швидкий розвиток бізнес -аналітики та аналізу привернув увагу дослідників. Причина в тому, що організації більше не покладаються на традиційні технології, оскільки дані зростають у геометричній прогресії. Цей величезний обсяг даних вимагає передових аналітичних методів для того, щоб перетворити їх на цінну інформацію, яка допомагає організаційному зростанню. BI&A-це сучасна методологія вилучення цінності з цієї величезної кількості даних, стимулювання прийняття стратегічних рішень, прогнозування та отримання переваг від майбутніх можливостей.

BI&A необхідний у більшості організацій. BI&A зарекомендувала себе як ефективна підтримка у прийнятті рішень. Окрім цього, на дані та IT - інфраструктуру явно впливає належне використання методів BI&A. У наш час бізнес -аналітика та аналіз відіграли важливу роль у більшості установ та секторів через їх цінність та переваги. BI&A допомагає організаціям краще бачити свої особисті дані і тим самим покращує прийняття рішень на основі фактів. Ці методології та аналіз даних також допомагають зберегти конкурентну перевагу на додаток до вирішення технічних проблем та проблем з якістю, що підвищить результативність та продуктивність підприємств.

За даними Abai et al. BI&A допомагає побудувати інтегровану структуру, яка підтримує прискорення діяльності організації. Багато факторів та технологічний розвиток сформували минулі та сучасні тенденції BI&A. Зі стрімким розвитком технологій недостатньо використовувати традиційні

аналітичні методи. Майбутній напрямок бізнес -аналітики та аналізу розшириться, включивши сфери різноманітності. За даними Chen et al.. Можливості успіху, пов'язані з технологіями аналізу даних, викликали майбутній інтерес до бізнес -аналітики та аналітики. Крім того, BI&A містить різні практики та методології, які можуть бути застосовані до різних секторів; Охорона здоров'я, безпека, аналіз ринку, електронне урядування та ін. За словами Мохаммеда та Вестбері, BI&A робить внесок у розвиток систем майбутнього. Перевіривши всі факти, BI&A незабаром стала біотехнологією у містах, що розвиваються, підтримуючи інформацію в режимі реального часу, яка перетворить країни на розумні міста.

Одним з найважливіших обов’язків у процесі видобутку даних є вибір відповідної технології вилучення даних. Характер роботи та тип об’єкта чи труднощі, що виникають у роботі, дають відповідні вказівки для визначення найкращих прийомів. Застосування методів інтелектуального аналізу даних Існують деякі узагальнені підходи, які можуть свідчити про підвищення ефективності та економічної ефективності. Багато основних прийомів, які виконуються в процесі видобутку даних, визначають характер процесу майнінгу та можливість відновлення даних.

Штучний інтелект (ШІ) є кроком в еволюції технологій, до якої активно вдаються з тих пір, як британський математик і ламач коду Алан Тьюрінг був задуманий як чіткий шлях у своєму новаторському дослідженні 1950 року «Обчислювальна техніка та інтелект». У той час комп’ютерні технології не могли йти в ногу з ідеями Тьюрінга. Але з розвитком обчислювальної техніки Amnesty International просунулася вперед. Більшість штучного інтелекту, який ми бачимо сьогодні,-це вузький штучний інтелект (ANI), що означає, що він може виконувати чітко визначене завдання. У звіті Інституту майбутнього людства в Оксфордському університеті за 2018 рік досліджено групу дослідників ШІ у графіках сильного ШІ. Вона виявила «50% ймовірність того, що штучний інтелект перевершить людей у всіх завданнях за 45 років і автоматизує всі функції людини за 120 років». Однак, ШІ також принесе з собою багато можливостей для створення нових можливостей для бізнесу. Як відзначали багато експертів, однією з цінностей штучного інтелекту є його здатність усунути потребу у важких і повторюваних завданнях. Крім того, користувачі можуть зосередитися на своїх основних цінностях та навиках. Технології застосовувалися у багатьох галузях промисловості, головним чином спрямовані на зменшення людських помилок, зменшення витрат на оплату праці, а отже, збільшення прибутку. Це стосувалося прогресу, досягнутого під час промислової революції до народження

комп’ютера, і все ще вірно щодо появи штучного інтелекту. користувачі можуть зосередитися на своїх основних цінностях та навиках. Технології застосовувалися у багатьох галузях промисловості, головним чином спрямовані на зменшення людських помилок, зменшення витрат на оплату праці, а отже, збільшення прибутку. Це стосувалося прогресу, досягнутого під час промислової революції до народження комп’ютера, і все ще вірно щодо появи штучного інтелекту. користувачі можуть зосередитися на своїх основних цінностях та навиках. Технології застосовувалися у багатьох галузях промисловості, головним чином спрямовані на зменшення людських помилок, зменшення витрат на оплату праці, а отже, збільшення прибутку. Це стосувалося прогресу, досягнутого під час промислової революції до народження комп’ютера, і все ще вірно щодо появи штучного інтелекту.

Штучний інтелект за останні кілька років значно просунувся через низку факторів, починаючи з масового збільшення доступних обчислювальних можливостей. Колись навчена модель ШІ тепер займає дні або навіть години з машинним навчанням (про це незабаром). Ще один фактор - ширший доступ до даних. Можливо, ви чули, що це "нова нафта" або щось подібне. Однак дані повинні бути оброблені за допомогою сучасних інструментів, таких як аналізи та алгоритми машинного навчання, щоб виявити корисну інформацію. Ця обробка є місцем, де ШІ в BI стає безцінним інструментом.

Машинне навчання - це двигун систем штучного інтелекту. Він зміцнює моделі штучного інтелекту шляхом аналізу складних наборів даних. Машинне навчання вдосконалює моделі, аналізуючи складні набори даних за допомогою набору правил та знань, що самостійно набуваються, як показано на рис. 6.5. Модель машинного навчання вчиться на основі великих даних та частої взаємодії людей, щоб вона могла надати інформацію та відповіді, що стосуються інтересів або цілей користувача. Великі дані відносяться до дуже великих наборів даних, які можна математично проаналізувати, щоб виявити закономірності, тенденції та кореляції, особливо щодо людської поведінки та взаємодії. У просторі штучного інтелекту глибоке навчання є значним стрибком уперед у технологіях. Як ми нещодавно торкнулися, програмісти пишуть код, який керує пристроєм, як інтерпретувати серію слів, малюнків або команд для прийняття рішення та виконання замовлення. Потім кінцевий користувач вводить запис (дані), тоді як внутрішні інженери можуть визначити більш конкретні правила інтерпретації та аналізу цих даних. Нарешті, система забезпечує результати (аналіз) на основі конкретних вхідних даних та визначених правил.

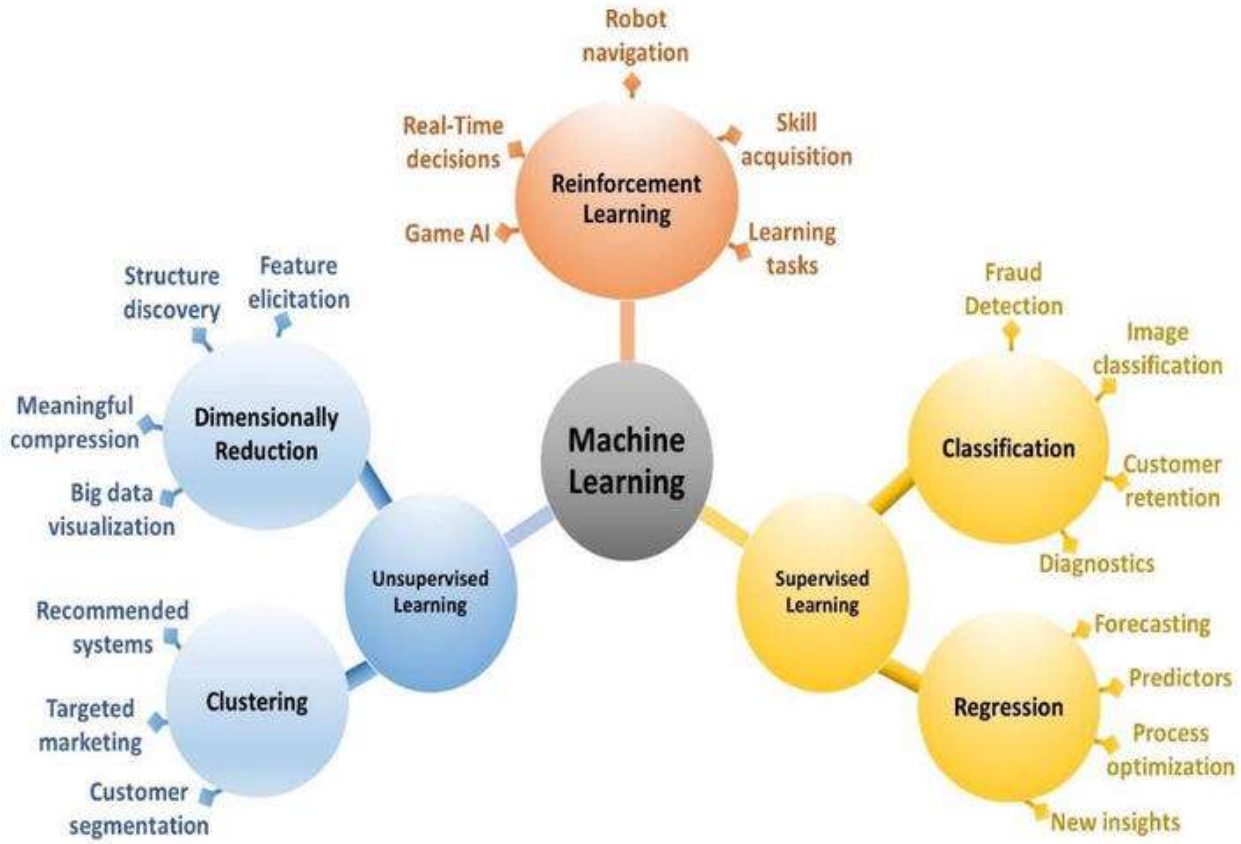


Рис. 6.5 Модель машинного навчання.

Навіщо ВІ потрібен ШІ?

Чи має значення, якщо постійне усвідомлення оригіналу, чи все -таки копія залишиться живою 19? На краще чи на погане, майбутнє настає швидше, ніж ми собі уявляємо. Не буде до або після штучного інтелекту, а повільний перехід протягом десятиліття або більше. Як ми бачили з Google Glass, наразі неможливо здогадатися, як виглядатимуть прийнятні результати. Але наскільки ми можемо довіряти нашим майбутнім помічникам? Вони працюватимуть з нами чи невідомими організаціями? Якщо ми зараз не поставимо правильні запитання, ми отримаємо додаток за умовчанням. Це буде безкоштовно, але що включатиме дрібні відбитки? Доброго ранку, Джон. Ось сьогоднішня програма. Які-небудь питання? Можливо, це все -таки не має значення: використовуючи хороший алгоритм навчання, програма знатиме, що нам потрібно і що нам потрібно робити, краще, ніж ми можемо здогадатися. Сила статистики виграє війну проти богів, а ми втратимо душу. Відомо, що кандидати на роботу можуть рішуче втратити свої шанси, коли вважають, що ніхто не стежить за поганою поведінкою чи відмовою від працівників приймальні та персоналу, що чекає. Як тільки НЛП та інші ШІ стануть широко розповсюджені, незабаром буде введений той самий тест з

літератури. Дивлячись на 2050 рік, майбутнє людства полягає у переході до цивілізації першого роду. Ми - тип 0, вимерлий. Ми ось-ось станемо напівбогами. Швидше за все, ми злиємося з нашою власною технологією обробки, і кожен з нас матиме свій власний віртуальний світ, щоб панувати над ним з абсолютним контролем у кожному його аспекті, і незліченними мільйонами планет «життя», якими ми можемо керувати або зливатися з ними так само. Так само, як програмісти відеоігор мають абсолютний контроль над світами, які вони створюють. Безсмертні, всезнаючі і всюдисущі, всі здатні до наших всесвітів. Звичайно, він також може досліджувати цей Всесвіт, можливо, контактувати безпосередньо зі своєю творчою істотою і знати, що ми - персонажі його гри. Наше останнє питання буде моральність і зрілість. Чи буде у нас лише один Всесвіт? Або сила приводить нас у божевілля і перетворює на "загарбників Всесвіту" і проникає у всесвіти інших, засновані на жадібності, проти бажання більшої сили? Чи буде нам добре? Або зло? Або обидва? Чи зможемо ми досягти мудрості та забезпечити мирне та гармонійне співіснування з усіма іншими напівбогами, або ми підемо на війну? Або ми об’єднаємось в одну надмірну силу? Або ми одного дня втомилися від божественного і знову розпочинаємо фінальну гру, і перетворити себе на всесвіт, який нам доведеться еволюціонувати протягом мільярдів років, щоб нас одного разу відтворили? Можливо, саме це і відбувається.

3.6 Покращення BI за допомогою AI

У цьому розділі ми досліджуємо, як штучний інтелект BI підвищує та покращує спосіб організації, які використовуються для аналізу та інтерпретації життєвого шляху її бізнесу.

Перетворення бізнес -користувачів на експертів даних (TBUDE):Як правило, доступ до даних та їх інтерпретація контролюють бізнес -аналітики (БА) та службовці IT. Хоча ці професії досі є вирішальними. Завдяки засобам штучного інтелекту в сучасних інструментах BI, включаючи LOB, користувачам NLI більше не потрібно розраховувати на аналіз своїх даних від експертів з даних. ШІ дозволяє користувачам легко та безпосередньо отримувати відповіді на дії, щоб допомогти «демократизувати» дані. Іншими словами, це дає користувачам можливість двосторонньої розмови зі своїми даними та відчуває, що вони мають можливість надійно реагувати на відповіді. Ось приклад того, як AI працює на практиці: певна організація впроваджує рішення BI, яке використовує розширений NLI, і замість того, щоб чекати системних адміністраторів або вчених з аналізу даних, менеджер бізнес

-підрозділу безпосередньо звертається до рішення BI. Менеджер надає дані за допомогою дзвінка або завантаження та задає питання простою мовою. Потім користувач отримує уявлення про ці питання разом з інформаційною панеллю та візуальними матеріалами, готовими до презентації, щоб допомогти передати ці відповіді. Попередньо навчена модель ШІ може бути націлена навіть на конкретні завдання БІ, такі як рекомендації щодо візуалізації, сценарії "що якщо", і передбачення, які допоможуть менеджерам приймати важливі рішення для свого бізнесу.

Допомагаючи вам досліджувати свої дані (HYEYD):Дослідити ваші дані за допомогою правильного інструменту штучного інтелекту, який підтримує штучний інтелект, є щось невід'ємне. За лічені хвилини ви можете перейти від завантаження наборів даних до виявлення прихованих фактів у даних та представлення цих результатів у чудовій візуалізації. На початковому моменті дані доступні, штучний інтелект у системі BI важко піднімається шляхом автоматичної сортування, позначення стовпців та об’єднання відповідних даних між групами. Доступ до NLI - це перший крок у дослідженні даних для користувача. Інструмент штучного інтелекту запропонує питання, які можуть бути корисними, якщо ви застрягнете. Ви також можете почати з основ, наприклад, "Як працював відділ роздрібної торгівлі протягом періоду X?" ШІ надасть відповіді та запропонує способи дослідження даних, щоб отримати додаткове уявлення про продуктивність. Дослідження захоплююче, тому що ви можете продовжувати заглиблюватися у бачення, яких може досягти лише ШІ. Те, що втілює уяву користувачів, - це уява. Візуальні зображення є невід'ємною особливістю всіх сучасних рішень для бізнес -аналітики, але завдяки рішенням із штучним інтелектом користувачі отримують запропоновані автоматизовані візуалізації, які найкраще відповідають відповідям на їх запитання.

Навчання у кінцевого користувача (LFEU):Провідні системи ШІ у системах BI налаштовуються та вдосконалюються за допомогою машинного навчання, яке індексує та вивчає традиційні питання та поведінку користувача. Чим більше користувач взаємодіє з інструментом BI, тим краще ШІ буде знати, чого хоче цей користувач у презентації та аналізі. Якщо користувач зазвичай використовує дані прогнозу, система почне готувати та представляти дані у моделі прогнозування через інформаційні панелі.

Автоматичне очищення та підготовка даних (ACPD):Для успішної інтерпретації ваші дані мають бути організовані в єдиному та доступному для пошуку порядку. Як будь -який бізнес добре знає, численні набори даних

викликають численні головні болі. Що робити, якщо імена форматуються як ім'я/прізвище в одній таблиці, а прізвище/ім'я - в іншій? Що робити, якщо є дублікати записів? Що робити, якщо записи є в одному наборі даних, а не в іншому? Що робити, якщо дані в одному наборі щоденні, а в іншому щомісячні?

ШІ в BI зменшує очищення даних і підготовку контакту і забезпечує масивний аспірин для головного болю. Автоматично налаштовуючи дані (один із найбільших штучних інтелектів, що заощаджує час), ви можете перейти від надання даних до роботи з ними за лічені хвилини, а не за години чи дні. Майбутня функція ШІ дозволить користувачам вводити структуровані та неструктуровані дані, не пропускаючи жодного виграшу; Велика зміна, оскільки більшість даних, що створюються сьогодні, таких як фотографії, відео та аудіо, - дезорганізовані. Усунення бар'єрів для ефективного аналізу - один із способів, за допомогою якого розширений інструмент штучного інтелекту у BI допомагає користувачам, які не є дослідниками даних, отримати доступ та інтерпретувати свої дані.

Отримання конкурентної переваги (GCA):Зараз ШІ робить вирішальну різницю між компаніями, які дозволяють їй досягти успіху, і тими, які незабаром залишаться позаду. Gartner прогнозує, що до 2021 року 75% заздалегідь підготовлених звітів, таких як ті, що використовуються для вилучення даних, будуть замінені або зміцнені за допомогою автоматизованого аналізу. Надійний інструмент штучного інтелекту в BI також забезпечує підвищену точність для критичних оперативних звітів про використання. Якщо цього не відбудеться, керівники даних та аналітики повинні планувати негайно запровадити Розширений аналіз (ШІ) у своєму бізнесі по мірі розвитку можливостей платформи. Ріта Саллам, віце -президент Gartner Research, попередила на нещодавній конференції, що «лідери даних та аналітики повинні вивчити потенційний вплив бізнесу» від посилення залежності від прогнозів, використовуючи вдосконалені та автоматизовані аналітичні дані »та відповідно коригуючи бізнес та бізнес -моделі, або ризикуючи втратити конкурентну перевагу тих, хто це робить. «Вже сьогодні штучний інтелект пропонується у рішеннях BI, і ті компанії, які впроваджують технології, мають успіх більш безпечно, ніж ті, що цього не роблять. Розкриваючи тенденції та кореляції у даних та пропонуючи способи інтерпретації результатів природною мовою разом із забезпеченням найкращої координації для представлення цих результатів, ШІ економить час та надає дієві уявлення про збільшення прибутковості та уникнення потенційних проблем до їх виникнення.

Підтримка векторних машин (SVM)

Додавання достатніх розмірів робить все лінійно роздільним.

Тут (x, y) -> (x, y, x^2+y^2) виконує свою роботу.

Для цього доступні ефективні вирішувачі, такі як LibSVM.

6.3 Сутність класифікації. Знайомство з методами кластеризації даних

Метод класифікації – це метод аналізу даних, який дозволяє оцінити ймовірність приналежності екземплярів даних до деякого класу залежно від значень їх атрибутів. Як модель класифікації рекомендується використовувати структуру даних «дерево» (див. Рис. 6.6), в якій кожен вузол являє собою точку прийняття рішення на підставі значень атрибутів даних, що класифікуються.

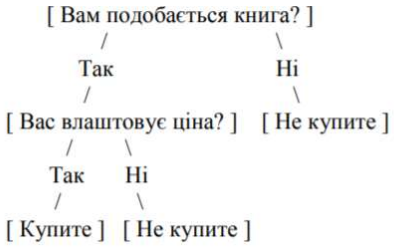


Рис. 6.6 Приклад дерева класифікації

На рис.6.6 наведено дерево класифікації, яке дає відповідь на питання «Ви купите книгу?». У кожному вузлі ставиться уточнююче питання (до атрибуту) з відповідями (значення атрибуту) у гілках, відповідаючи ви переходите до наступного вузла (питання, атрибуту) до тих пір, поки не дійдете до листа зі значенням класу, у прикладі це відповіді «Купите» чи «Не купите» книгу. Перевага класифікаційних дерев полягає у тому, що вони не вимагають надмірної кількості інформації для побудови досить точного та інформативного дерева рішень. Метод класифікації використовує відомі значення атрибутів екземплярів даних та зв'язки між їх значеннями при побудові моделі класифікації. При наявності нових екземплярів даних невідомого класу, до даних застосовується раніше побудована модель класифікації і визначається відповідний клас.

Проблеми класифікації логістики

Збалансований розмір навчальних класів

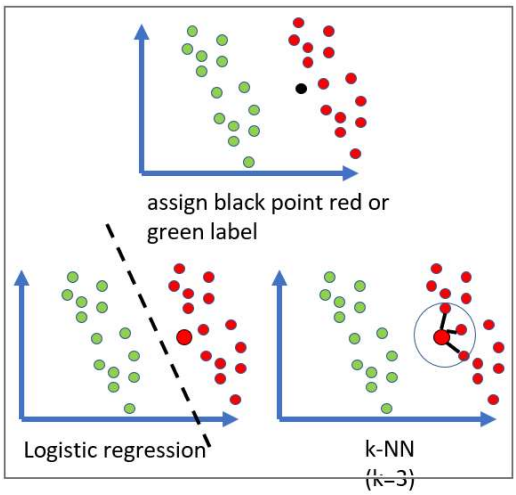
Багатокласова класифікація

Ієрархічна класифікація

Функції розділу

Визначте, який навчальний приклад найбільш подібний до цільового, і візьміть з нього мітку класу. Ключовим питанням тут є створення правильної функції відстані між рядками/точками.

Переваги: простота, інтерпретація та нелінійність. Проблема: відповідна метрика відстані!



Кластеризація - це проблема групування точок за подібністю. Часто елементи надходять з невеликої кількості "джерел" або "пояснень", і кластеризація - це хороший спосіб розкрити це походження. Подібність визначається деякою базовою функцією відстані/метрикою.

Метод кластеризації – це метод аналізу даних, який дозволяє розділити екземпляри даних за значеннями їх атрибутів на класи, кожен з яких має певні ознаки. Кластерний аналіз використовується в тих випадках, коли необхідно автоматично виділити деякі правила, взаємозв'язки або тенденції у сукупності даних. Як модель кластеризації можна використовувати двовимірний простір Евкліда, в якому відносне скупчення екземплярів даних являє собою певний клас.

Чому кластеризація?

Розвиток гіпотези - скільки різних груп населення у ваших даних?

Моделювання для менших груп - побудуйте окремі моделі прогнозування для кожного кластера.

Зменшення даних - замінити/представити кожну групу елементів її центроїдом.

Виявлення викидів - які елементи знаходяться далеко від центрів кластерів або застрягли у крихітних скупченнях?

Просторовий підхід використовує некласифіковані екземпляри для аналізу зв'язків між значеннями їх атрибутів. Побудовану модель кластеризації можна використовувати для оцінювання приналежності нових екземплярів даних до вже відомих кластерів.

Об’єднання схожих об’єктів у групи може бути здійснене різними способами. Виділяють певні групи методів кластерного аналізу (рис. 6.8):

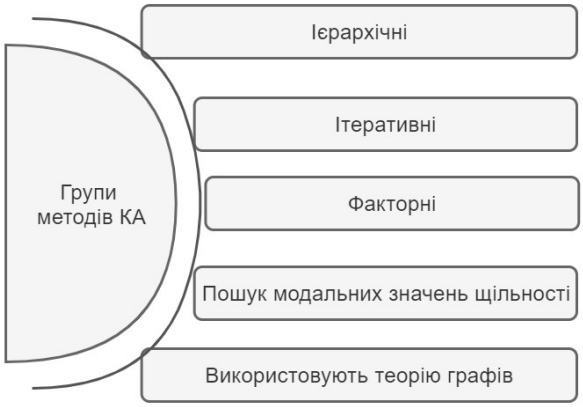


Рис. 6.8. Групи методів кластерного аналізу

Класифікація методів кластерного аналізу

За способом обробки даних:

ієрархічні (агломеративні, дивізимні);

неієрархічні.

За способом аналізу даних:

чіткі;

нечіткі.

За кількістю застосування алгоритмів кластеризації:

з одноетапною кластеризацією;

з багатоетапною кластеризацією.

За можливістю розширення обсягу оброблюваних даних:

масштабовані;

немасштабовані.

За часом виконання кластеризації:

потокові (on-line);

не потокові (off-line).

Розглянемо особливості різних методів кластеризації (рис. 6.9):

Рис. 6.9 Особливості ієрархічних методів кластерного аналізу

Найбільш поширеними є ієрархічні методи, серед яких розрізняють агломеративний і дивізимний методи.

Візуалізація кластерної структури наведена на рис.6.10

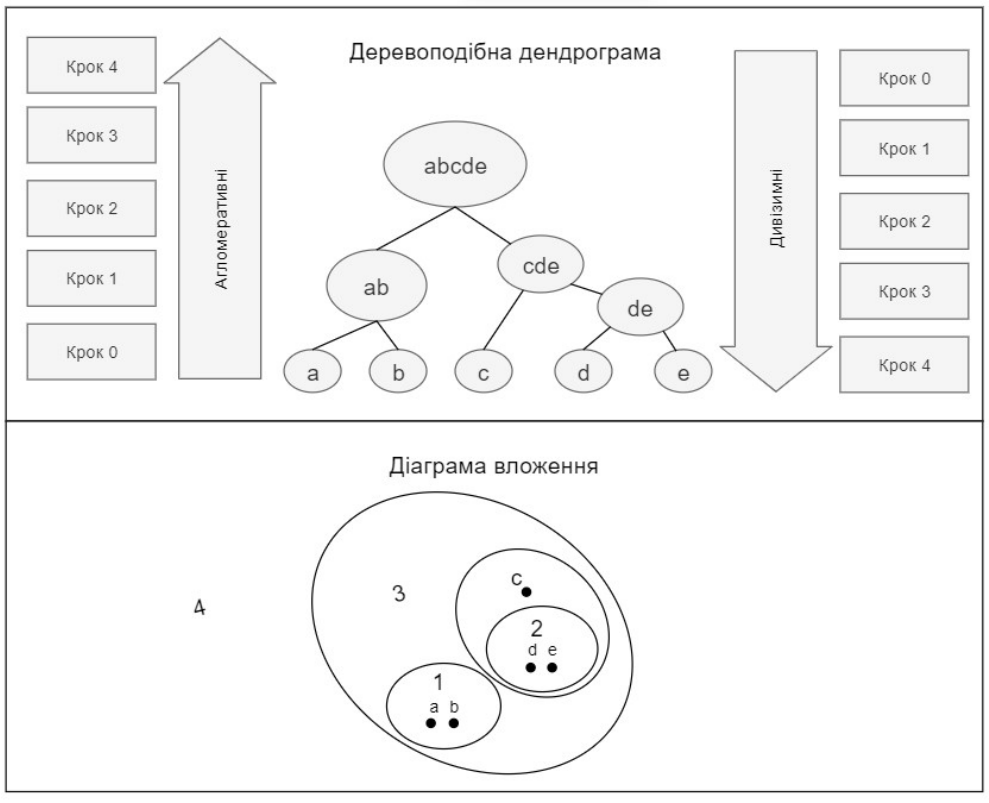


Рис. 6.10 Візуалізація кластерної структури у застосуванні ієрархічних методів

Який алгоритм кластеризації використовувати?

Існує величезна кількість можливих алгоритмів кластеризації, але набагато важливіші рішення: за допомогою функції потрібної відстані належним чином нормалізувати ваші змінні належним чином візуалізувати кінцеві кластери, щоб дізнатися, чи вони хороші.

Агломеративна кластеризація

Імовірнісна групування

Очікування-Максимізація (EM)-Алгоритм

GrabCut - Алгоритм

Алгоритм агломеративної ієрархічної кластеризації:

нормування вихідних даних;

розрахунок матриці відстаней або матриці мір подібності;

знаходиться пара найближчих кластерів. За обраним алгоритмом об'єднуються ці два кластери. Новому кластеру присвоюється менший з номерів об'єднувальних кластерів;

кроки 2, 3 і 4 повторюються, поки всі об'єкти не будуть об'єднані в один кластер або до досягнення заданого «порогу» подібності.

Алгоритм дивізимної ієрархічної кластеризації:

нормування вихідних даних;

розрахунок матриці відстаней або матриці мір подібності;

знаходиться пара найдальніших об'єктів nini, njnj;

оцінюється відстань об'єктів, що залишилися, до виділених об'єктів nini, njnj і визначається до nini або до njnj вони ближче знаходяться;

близькі об'єкти об'єднуються в кластер. Так, початковий єдиний кластер розбивається на два;

кроки 3, 4 і 5 повторюються, поки всі об'єкти не будуть розділені на кластери.

Класифікатори дерева рішень.

Кожен рядок/екземпляр проходить унікальний шлях від кореня до листа до класифікації. Дерево поділяє навчальні приклади на групи відносно однорідного складу, де прийняття рішення стає легким. Зведення кожного прикладу навчання до власного листового вузла означає надто тренування.

Переваги:

Нелінійність

Підтримка категоричних змінних (волосся = руде)

Інтерпретація - люди можуть читати дерево

Міцність - ми можемо створювати ансамблі з різних дерев та голосувати (КОРЗИНА).

Застосування до регресії всередині підмножини листя Найбільший недолік - це відсутність елегантності/математика.

Ансамблі дерев рішень - Випадкові ліси

Ми можемо побудувати сотні дерев рішень, випадковим чином вибравши функцію для поділу.

Голосування серед кількох класифікаторів підвищує надійність і дозволяє оцінити рівень довіри.

Мішка вибирає випадково вибрані підмножини предметів, на яких можна навчити кожне дерево.

Питання для самоконтролю:

1. В яких сферах бізнесу використовуються системи бізнес -аналітики?
2. Життєвий цикл аналітики – це?
3. Назвіть етапи життєвого циклу аналітики даних
4. В чому різниця між сучасною архітектурою даних та традиційною?
5. Назвіть проблеми управління даними
6. В чому полягає модель управління даними для BI наступного покоління?
7. Яким чином відбувається покращення BI за допомогою AI?
8. Назвіть методами кластеризації даних

Лекція 7

МОДЕЛЮВАННЯ ВЕЛИКИХ ДАНИХ

План:

7.1. Big data: які дані вважаються великими

7.2. Big Data в маркетингу

7.3 Важливість моделювання даних у світі великих даних

7.4. Поради створення ефективних моделей великих даних

7.1. Big data: які дані вважаються великими

Уведення терміну “великі дані” належить Кліффорду Лінчу, редакторові журналу Nature, який підготував у вересні 2008 р. спеціальний випуск журналу, де аналізував феномен Великих даних та їх значення для науки. Він зібрав матеріали про явище вибухового зростання обсягу і різноманітності даних, а також технологічних перспектив у парадигмі ймовірного переходу від “кількості до якості”.

Незважаючи на те, що термін уведений в академічному середовищі, первинною була проблема зростання кількості даних і збільшення їх різноманітності у практичних задачах. Станом на 2009 р. термін поширений у діловій пресі, а у 2010 р. з’явилася перша низка продуктів і рішень, що стосуються винятково проблем обробки Великих обсягів даних. До 2011 р. багато найбільших постачальників інформаційних технологій використовують Великі дані для формування бізнес-стратегії, зокрема IBM, Oracle, Microsoft, Hewlett-Packard, EMC.

Проблеми, що виникають під час опрацювання, інтерпретації, збору та організації Великих даних, з'явилися в численних секторах, а також бізнес, промисловість, некомерційні організації. Набори даних, такі як операції замовника у роздрібній торгівлі, моніторинг погоди, бізнес-аналіз, можуть швидко випереджати потужність традиційних методів та інструментів аналізу даних. Тому з’явилися нові методи та інструменти, зокрема бази даних NoSQL, MapReduce, обробка природної мови, машинне навчання, візуалізація тощо.

Завдяки експоненціального зростання можливостей обчислювальної техніки, описаного в законі Мура, обсяг даних не може бути точним критерієм того, чи є вони великими. Наприклад, сьогодні великі дані вимірюються в терабайт, а завтра - в петабайт. Тому головною характеристикою Big Data є ступінь їх структурованості і варіантів представлення.

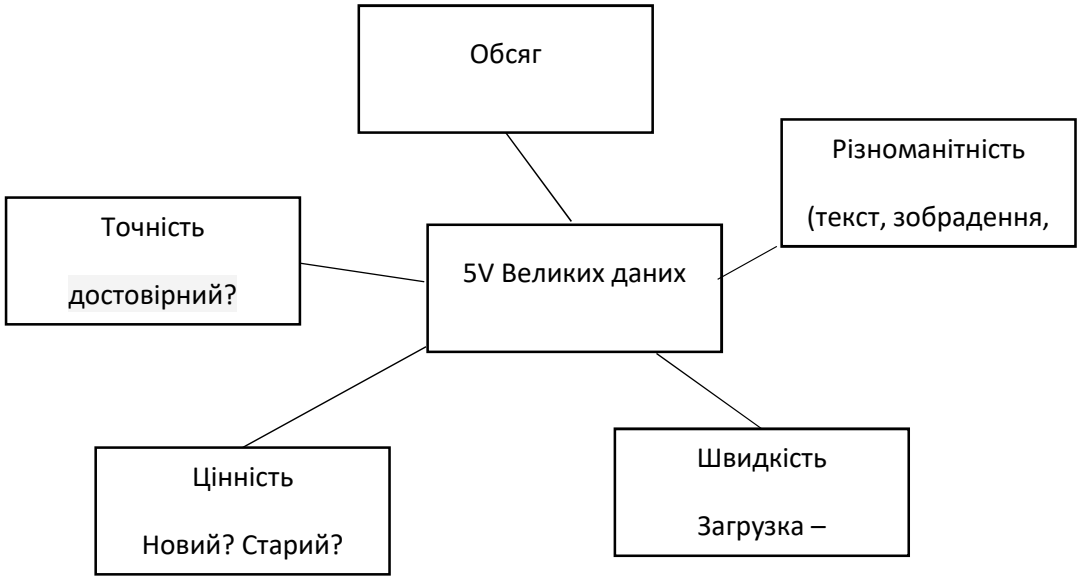


Рис. 7.1. 5V – Основні показники Big Data

Яскрава ілюстрація великих даних - це безперервно надходить інформація з датчиків або пристроїв аудіо- і відеореєстрації, потоки повідомлень з соцмереж, метеорологічні дані, координати геолокації абонентів стільникового зв'язку і т.п.

Таким чином, джерелами великих даних можуть бути:

інтернет - соцмережі, блоги, ЗМІ, форуми, сайти, інтернет речей (Internet of Things, IoT);

корпоративна інформація - транзакції, архіви, бази даних і файлові сховища;

показання приладів - датчиків, сенсорів, реєстраторів та ін.

Аналіз Великих даних починається з їх збору. Інформацію отримують звідусіль: з наших смартфонів, кредитних карт, програмних додатків, автомобілів. Веб-сайти здатні передавати величезні обсяги даних.

Через різних форматів і шляхів виникнення Big Data відрізняються рядом характеристик:

1 Volume. Величезні «обсяги» даних, які організації отримують з бізнес-транзакцій, інтелектуальних (IoT) пристроїв, промислового обладнання, соціальних мереж та інших джерел, потрібно десь зберігати. У минулому це було проблемою, але розвиток систем зберігання інформації полегшило ситуацію і зробило інформацію доступнішою.

2 Velocity. Найчастіше цей пункт відноситься до швидкості приросту, з якої дані надходять в реальному часі. У більш широкому розумінні характеристика пояснює необхідність високошвидкісної обробки через темпів зміни і сплесків активності.

3 Variety. Різноманітність великих даних проявляється в їх форматах: структуровані цифри з клієнтських баз, неструктуровані текстові, відео- і аудіофайли, а також полуструктурированного інформація з кількох джерел. Якщо раніше дані можна було збирати тільки з електронних таблиць, то сьогодні дані надходять в різному вигляді: від електронних листів до голосових повідомлень.

Big Data характеризує великий обсяг структурованих і неструктурованих даних, які щохвилини утворюється в цифровому середовищі. IBM стверджує, що в світі підприємства щодня генерують майже 2,5 квінтильйони байтів даних! А 90% глобальних даних отримано тільки за останні 2 роки. Але важливий не обсяг інформації, а можливості, які дає її аналіз.

Одне з основних переваг Big Data - предіктивне аналіз. Інструменти аналітики Великих даних прогнозують результати стратегічних рішень, що оптимізує операційну ефективність і знижує ризики компанії. Big Data об'єднують релевантну і точну інформацію з кількох джерел, щоб найбільш точно описати ситуацію на ринку. Аналізуючи інформацію з соціальних мереж і пошукових запитів, компанії оптимізують стратегії цифрового маркетингу і досвід споживачів. Наприклад, відомості про рекламні акції всіх конкурентів, дозволяють керівництво фірми запропонувати більш вигідний «персональний» підхід клієнту.

Компанії, урядові установи, постачальники медичних послуг, фінансові та академічні установи - все використовують можливості Великих даних для поліпшення ділових перспектив і якості обслуговування клієнтів. Хоча дослідження показують, що ще майже 43% комерційних організацій до сих пір не володіють необхідними інструментами для фільтрації нерелевантних даних, втрачаючи потенційний прибуток. Тому сьогодні на ринку намітився

курс на модернізацію бізнес-процесів, освоєння нових технологій і впровадження Big Data.

Всіх, хто має справу з великими даними, можна умовно розділити на кілька груп:

Постачальники інфраструктури - вирішують завдання зберігання і предобработки даних. Наприклад: IBM, Microsoft, Oracle, Sap та інші.

Датамайнери - розробники алгоритмів, які допомагають замовникам отримувати цінні відомості. Серед них: Yandex Data Factory, Glowbyte Consulting, CleverData і ін.

Системні інтегратори - компанії, які впроваджують системи аналізу великих даних на стороні клієнта.

Споживачі - компанії, які купують програмно-апаратні комплекси і замовляють алгоритми у консультантів. Це компанії з галузей фінансів, телекомунікацій, рітейлу.

Розробники готових сервісів - пропонують готові рішення на основі доступу до великих даними. Вони відкривають можливості Big Data для широкого кола користувачів.

7.2. Big Data в маркетингу

Навіщо потрібні великі дані в маркетингу? Аналіз масивів інформації про компанії відкриває нові можливості:

Зрозуміти роботу бізнесу в цифрах.

Вивчити конкурентів. Дізнатися своїх клієнтів.

Маркетинг зможе вийти на новий рівень розуміння та аналітики, що дозволить знизити витрати і збільшити продажі.

Вигоди використання технології в маркетингу.

Створення точних портретів цільових споживачів.

Передбачення реакції споживачів на маркетингові повідомлення.

Максимальна персоналізація рекламних повідомлень.

Збільшення крос-продажів, повторних продажів, ремаркетингу.

Пошук і визначення причин популярності затребуваних товарів і продуктів.

Удосконалення продуктів і послуг, підвищення лояльності клієнтів.

Підвищення якості обслуговування.

Попередження шахрайства.

Зниження витрат в роботі з постачальниками і клієнтами. З

авдяки спеціальним сервісів технології великих даних, Big Data знайдеться застосування в будь-якому відділі маркетингу, в тому числі середнього та малого бізнесу. Вам не буде потрібно встановлювати і обслуговувати дороге устаткування і містити фахівця.

Методи та засоби роботи З BIG DATA

До основних методів збору і аналізу великих даних відносять такі:

Data Mining - навчання асоціативним правилами, класифікація, кластерний і регресійний аналіз;

краудсорсінг - категоризація та збагачення даних народними силами, тобто з добровільною допомогою сторонніх осіб; змішання і інтеграція різнорідних даних, таких як, цифрова обробка сигналів і обробка природної мови;

машинне навчання (Machine Learning), включаючи штучні нейронні мережі, мережевий аналіз, методи оптимізації та генетичні алгоритми; розпізнавання образів;

прогнозна аналітика;

імітаційне моделювання;

просторовий і статистичний аналіз;

візуалізація аналітичних даних - малюнки, графіки, діаграми, таблиці.

Програмно-апаратні засоби роботи з Big Data передбачають масштабованість, паралельні обчислення і розподіленість, тому що безперервне збільшення обсягу - це одна з головних характеристик великих даних. До основних технологій відносять нереляційні бази даних (NoSQL), модель обробки інформації MapReduce, компоненти кластерної екосистеми Hadoop, мови програмування R і Python, а також спеціалізовані продукти Apache (Spark, AirFlow, Kafka, HBase і ін.).

Таблиця 7.1. Засоби роботи з Великими даними

Засоби для Великих даних	Опис
Засоби аналізу даних	
Ambari http://ambari.apache.org	Інструмент веб для надання послуг, управління та моніторингу Apache Hadoop кластерів
Avro http://avro.apache.org	Система серіїзації даних
Chukwa http://incubator.apache.org/chukwa	Система колекціонування даних для керування великими розподіленими системами
Hive http://hive.apache.org/	Інфраструктура сховища даних, яка забезпечує агрегацію даних
Pig http://pig.apache.org	Високорівнева мова потоків даних і виконуваний framework для паралельних обчислень
Spark http://spark.incubator.apache.org	Швидкий і генеральний обчислювач для даних Hadoop. Забезпечує просту і виразну модель програмування, яка підтримує широкий спектр додатків, у тому числі ETL, машинне навчання, опрацювання потоків
ZooKeeper	Високо продуктивна служба координації для розподілених

http://zookeeper.apache.org/	додатків
ctian http://www.actian.com/about-us/#overview	Забезепечує зберігання «сирих» даних і готує дані для подальшого аналізу
HPCC http://hpccsystems.com	Забезпечує швидке перетвореня, паралельне опрацювання для застосувань з Великими даними
Засоби Data Mining	
Orange http://orange.biolab.si	Візуалізація та аналіз даних
Mahout http://mahout.apache.org	Бібліотека засобів машинного навчання та видобування даних
KEEL http://keel.es	Еволюційний алгоритм для задач видобування даних
Засоби соціальних мереж	
Apache Kafka	Платформа з високою пропускною здатністю для опрацювання даних в режимі реального часу
Засоби BI	
Talend http://www.talend.com	Інтеграція даних, управління, інтеграція застосувань, засоби і сервіси для Великих даних
Jedox http://www.jedox.com/en	Функції аналізу, звітності, планування
Pentaho http://www.pentaho.com	Інтеграція даних, бізнес-аналіз, візуалізація даних, прогнозування
Rasdaman http://rasdaman.eecs.jacobs-university.de/	Багатовимірні расторові дані (масив) без обмежень на розмір, наявність мови запитів
Засоби пошуку	
Apache Lucene http://lucene.apache.org	Застосування для повнотекстового індексування і пошуку

Apache Solr http://lucene.apache.org/solr	Повнотекстовий пошук, фасетний пошук, динамічна кластеризація, формати документів типу Word, PDF, просторовий пошук
Elasticsearch http://www.elasticsearch.org	Засіб розподіленого повнотекстового пошуку з веб-інтерфейсом і JSON докуметами
MarkLogic http://developer.marklogic.com	NOSQL і XML база даних
mongoDB http://www.mongodb.org	Крос-платформенна документо-орієнтована система управління базами даних з підтримкою JSON та динамічних схем
Cassandra http://cassandra.apache.org	Маштабована база даних без єдиної точки відмови
HBase http://hbase.apache.org	Маштабована розподілена база даних з підтримкою структуровано зберігання даних великого обсягу
InfiniteGraph http://www.objectivity.com	Розподілена графова база даних

7.3 Важливість моделювання даних у світі великих даних

Оскільки все більше організацій охоплюють великі дані та аналітику, щоб отримати уявлення про надзвичайно великі набори даних, інструменти та системи, що використовуються для управління даними, зростають, змінюються та розмножуються. Замість просто реляційних систем баз даних, ми тепер використовуємо бази даних NoSQL та файлові системи Hadoop для зберігання все більших обсягів корпоративних даних.

Ви могли б подумати, що з огляду на високу важливість даних у сучасній сучасній організації, моделювання даних вважатиметься надзвичайно важливим для керівництва та IT-фахівців, тому дещо іронічно, що вік великих даних збігвся з довгостроковим спадом даних адміністрування та

моделювання у багатьох організаціях. Це не та ситуація, яку слід продовжувати терпіти.

Що таке моделювання даних?

Моделювання даних - це процес аналізу «речей», що становлять інтерес для вашої організації, і того, як ці речі пов’язані між собою. Процес моделювання даних призводить до виявлення та документування ресурсів даних вашого бізнесу. Створюючи концептуальні та логічні моделі даних, ви розвиваєте лексикон бізнесу вашої організації.

Модель даних будується з використанням компонентів, які діють як абстракції реальних речей. Найпростіша модель даних складається з сутностей та зв'язків. У міру просування роботи над моделлю даних додаються додаткові деталі та складність, включаючи атрибути, домени, обмеження, ключі, потужність, вимоги, відносини - і, що важливо, визначення всього в моделі даних. Якщо ми хочемо зрозуміти дані, які у нас є, і як їх використовувати, потрібна фундаментальна модель.

Проблеми з великими даними Великі дані та аналітика є важливою частиною сучасної IT. За оцінками аналітиків, кількість даних, якими ми користуємось і якими ми керуємо, щороку подвоюється, і аналіз цих даних може виявити досі невідомі дані, які ведуть до конкурентних переваг. Крім того, великі дані, що використовуються для забезпечення аналітики, адаптуються для використання штучним інтелектом та програмним забезпеченням машинного навчання, що ще більше покращить окупність наших комп’ютерних інвестицій за рахунок автоматизації процесів та завдань, тим самим підвищуючи продуктивність та операційну ефективність.

Але проблеми можуть виникнути при використанні технологій гнучких схем, таких як NoSQL та Hadoop. Така гнучкість часто є вимогою, коли великі обсяги даних відкриваються, приймаються та переміщуються в організацію. Якщо один рядок (або запис) даних може мати іншу схему, ніж наступна, ви не можете застосувати до даних фіксовану модель.

Проте програміст повинен знати, як виглядають дані. Ви не можете просто кинути на когось велику групу даних і сказати: «Ось дані, тепер напишіть мені програму». Ну, ви можете так сказати, але тоді програміст (або хтось інший) повинен проаналізувати та задокументувати структуру даних.

Це дуже нагадує модель даних, чи не так? Ну, це повинно бути, тому що це так. Замість моделювання заздалегідь, перед тим, як написати будь -який

код, як це поширено у реляційному світі, моделювання великих даних іноді виконується на основі запитів додатків у програмному коді або інструментах. Чого ми хочемо уникнути, так це мати всі знання про дані, вбудовані в прикладні програми, як це було звичайно до того, як реляційне стало популярним у 1980 -х роках. І ми повинні намагатися уникати того, щоб розробники змінювали моделі тих самих даних щоразу, коли вони використовуються. Моделювання даних створює систему запису корпоративних даних, доступну всім, а не лише тим, хто розуміє мову програмування du jour.

7.4. Поради створення ефективних моделей великих даних

Великі дані менш передбачувані, ніж традиційні дані, і тому вимагають особливого розгляду при побудові моделей. Ось деякі речі, які слід мати на увазі.

Моделювання даних - це складна наука, яка передбачає організацію корпоративних даних так, щоб вона відповідала потребам бізнес -процесів. Це вимагає розробки логічних зв'язків, щоб дані могли взаємозв'язуватися між собою і підтримувати бізнес. Потім логічні конструкції перетворюються на фізичні моделі, які складаються з пристроїв зберігання даних, баз даних та файлів, що містять дані.

Історично підприємства використовували технологію реляційних баз даних, таку як SQL, для розробки моделей даних, оскільки вона унікально підходить для гнучкого зв’язування ключів набору даних і типів даних разом для забезпечення інформаційних потреб бізнес -процесів.

На жаль, великі дані, які зараз складаються з великого відсотка даних під управлінням, не працюють у реляційних базах даних. Він працює на нереляційних базах даних, таких як NoSQL. Це призводить до переконання, що вам не потрібна модель для великих даних.

Проблема в тому, що вам потрібне моделювання даних для великих даних. Ось шість порад щодо моделювання великих даних:

Не намагайтеся нав’язувати традиційні методи моделювання великих даних Традиційні фіксовані дані стабільні та передбачувані у своєму зростанні. Це робить моделювання відносно простим. Навпаки, експоненціальне зростання великих даних непередбачуване, як і їх незліченна кількість форм та джерел. Коли веб -сайти розглядають можливість

моделювання великих даних, зусилля з моделювання мають зосередитися на створенні відкритих та еластичних інтерфейсів даних, тому що ніколи не знаєш, коли може з’явитися нове джерело даних чи форма даних. Це не є пріоритетом у традиційному світі даних фіксованих записів.

Проектуйте систему, а не схему У традиційній сфері даних схема реляційної бази даних може охоплювати більшість зв'язків і зв'язків між даними, необхідними бізнесу для інформаційного забезпечення. Це не так з великими даними, які можуть не мати бази даних або які можуть використовувати базу даних, наприклад NoSQL, яка не потребує схеми бази даних.

Через це моделі великих даних повинні будуватися на системах, а не на базах даних. Системними компонентами, які повинні містити моделі великих даних, є вимоги до ділової інформації, корпоративне управління та безпека, фізичне сховище, що використовується для даних, інтеграція та відкриті інтерфейси для всіх типів даних, а також здатність обробляти різноманітні типи даних.

3. Шукайте інструменти моделювання великих даних Існують інструменти комерційного моделювання даних, які підтримують Hadoop, а також програмне забезпечення для подання великих даних, наприклад Tableau. Розглядаючи інструменти та методології великих даних, особи, які приймають IT -рішення, повинні включити одну зі своїх вимог у можливість побудови моделей даних для великих даних.

4. Зосередьтеся на даних, які є основою вашого бізнесу Гори великих даних щодня надходять на підприємства, і більшість цих даних є сторонніми. Немає сенсу створювати моделі, що включають усі дані. Кращим підходом є визначення великих даних, необхідних для вашого підприємства, та моделювання цих даних.

5. Подайте якісні дані Покращені моделі даних та взаємозв’язки можуть бути досягнуті для великих даних, якщо організації зосереджуються на розробці обґрунтованих визначень даних та ретельних метаданих, які описують, звідки дані надходили, яке їх призначення тощо. Чим більше ви знаєте про кожну частину даних, тим більше ви можете належним чином розмістити його у моделях даних, які підтримують ваш бізнес.

6. Шукайте ключові місця в даних Одним з найбільш часто використовуваних векторів у великих даних є географічне розташування. Залежно від вашого бізнесу та вашої галузі існують також інші загальні ключі

великих даних, які потрібні користувачам. Чим більше ви зможете визначити ці загальні точки входу у ваші дані, тим краще ви зможете розробити моделі даних, які підтримують ключові шляхи доступу до інформації для вашої компанії.

Питання для самоконтролю:

1. Дайте визначення Big data – це?
2. Назвіть основні показники Big Data.
3. Які вигоди використання IT в маркетингу?
4. Моделювання даних – це?
5. В чому важливість моделювання даних у світі великих даних?

Лекція 8

Архітектура та розгортання

План:

8.1. Стиль архітектури великих даних

8.2. Еволюція моделей розгортання в епоху великих даних

8.3. Моделі розгортання в хмарі великих даних

8.1. Стиль архітектури великих даних

Архітектура великих даних призначена для обробки, обробки та аналізу даних, які є занадто великими або складними для традиційних систем баз даних.

Рішення великих даних зазвичай включають один або кілька таких типів навантаження (рис. 8.1.):

Пакетна обробка великих джерел даних у стані спокою.

Обробка великих даних у русі в режимі реального часу.

Інтерактивне дослідження великих даних.

Прогностична аналітика та машинне навчання.

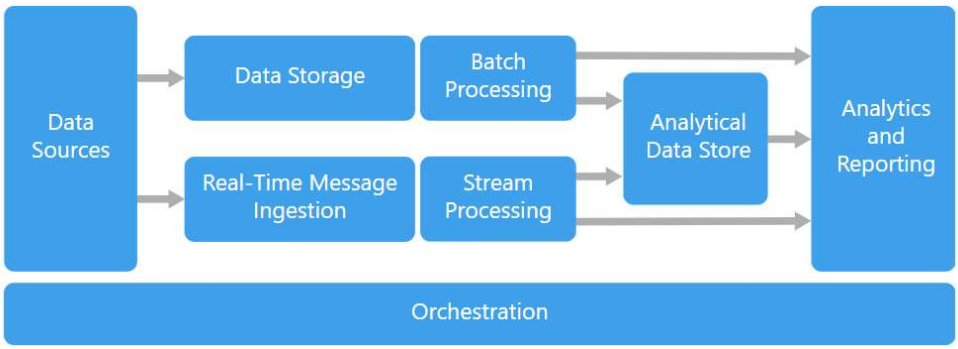


Рис 8.1 Архітектура великих даних

Більшість архітектур великих даних включають деякі або всі наступні компоненти:

Джерела даних: Усі рішення для великих даних починаються з одного або декількох джерел даних.

Приклади включають: Прикладні сховища даних, такі як реляційні бази даних. Статичні файли, створені програмами, наприклад файли журналу веб - сервера.

Джерела даних у режимі реального часу, такі як пристрої IoT. Зберігання даних: Дані для пакетної обробки зазвичай зберігаються у розподіленому сховищі файлів, яке вміщує великі обсяги великих файлів у різних форматах. Такий магазин часто називають озером даних. Варіанти реалізації цього сховища включають Azure Data Lake Store або контейнери BLOB -файлів у сховищі Azure.

Пакетна обробка: оскільки набори даних настільки великі, часто рішення для обробки великих даних має обробляти файли даних за допомогою тривалих пакетних завдань для фільтрації, агрегації та іншої підготовки даних до аналізу. Зазвичай ці завдання включають читання вихідних файлів, їх обробку та запис результатів до нових файлів. Параметри включають виконання завдань U-SQL в Azure Data Lake Analytics, використання Hive, Pig або користувацької карти/Скорочення завдань у кластері HDInsight Hadoop або використання програм Java, Scala або Python у кластері HDInsight Spark.

Прийом повідомлень у режимі реального часу: Якщо рішення містить джерела в режимі реального часу, архітектура повинна містити спосіб збору та зберігання повідомлень у режимі реального часу для потокової обробки. Це може бути просте сховище даних, де вхідні повідомлення потрапляють у папку для обробки. Однак багатьом рішенням потрібне сховище прийому повідомлень, яке буде виконувати роль буфера для повідомлень і підтримувати масштабовану обробку, надійну доставку та іншу семантику черги повідомлень. Параметри включають концентратори подій Azure, концентратори Azure IoT та Kafka.

Обробка потоків: Після збору повідомлень у режимі реального часу рішення має обробляти їх шляхом фільтрації, агрегації та іншої підготовки даних до аналізу. Потім оброблені дані потоку записуються у вихідний потік. Azure Stream Analytics надає послугу обробки керованого потоку на основі постійно виконуваних запитів SQL, які працюють з необмеженими потоками. Ви також можете використовувати потокові технології Apache з відкритим вихідним кодом, такі як Storm та Spark Streaming, у кластері HDInsight.

Зберігання аналітичних даних: Багато рішень для великих даних готують дані для аналізу, а потім подають оброблені дані у структурованому форматі, який можна запитувати за допомогою аналітичних засобів. Сховище аналітичних даних, що використовується для обслуговування цих запитів, може бути сховищем реляційних даних у стилі Кімбола, як це можна побачити в більшості традиційних рішень бізнес-аналітики (BI). Крім того, дані можуть бути представлені за допомогою технології NoSQL з низькою затримкою, такої як HBase, або інтерактивної бази даних Hive, яка забезпечує абстрагування метаданих над файлами даних у розподіленому сховищі даних. Azure Synapse Analytics надає керовану службу для масштабного хмарного зберігання даних. HDInsight підтримує Interactive Hive, HBase та Spark SQL, які також можна використовувати для подання даних для аналізу. Аналіз та звітування: Метою більшості рішень для обробки великих даних є надання уявлення про дані за допомогою аналізу та звітності. Для розширення можливостей користувачів аналізувати дані архітектура може включати рівень моделювання даних, наприклад багатовимірний куб OLAP або табличну модель даних у службах аналізу Azure. Він також може підтримувати BI самообслуговування, використовуючи технології моделювання та візуалізації в Microsoft Power BI або Microsoft Excel.

Аналіз та звітування також можуть мати форму інтерактивного дослідження даних науковцями або аналітиками даних. У цих сценаріях багато служб Azure підтримують аналітичні блокноти, такі як Jupyter, що дозволяє цим користувачам використовувати свої наявні навички з Python або R. Для масштабного дослідження даних можна використовувати Microsoft R Server, як автономний, так і зі Spark.

Організація: Більшість рішень великих даних складаються з повторних операцій з обробки даних, інкапсульованих у робочі процеси, які перетворюють вихідні дані, переміщують дані між кількома джерелами та стоками, завантажують оброблені дані в сховище аналітичних даних або переносять результати прямо у звіт або на інформаційну панель . Щоб автоматизувати ці робочі процеси, можна скористатися такою технологією оркестрування, як Azure Data Factory або Apache Oozie та Sqoop.

Azure містить багато служб, які можна використовувати в архітектурі великих даних.

Вони поділяються приблизно на дві категорії: Керовані служби, включаючи Azure Data Lake Store, Azure Data Lake Analytics, Azure Synapse Analytics, Azure Stream Analytics, Azure Event Hub, Azure IoT Hub та Azure Data

Factory. Технології з відкритим кодом на основі платформи Apache Hadoop, включаючи HDFS, HBase, Hive, Pig, Spark, Storm, Oozie, Sqoop та Kafka. Ці технології доступні в Azure у службі Azure HDInsight.

Ці параметри не виключають один одного, і багато рішень поєднують технології відкритого коду зі службами Azure.

Коли використовувати цю архітектуру. Розгляньте цей стиль архітектури, коли вам потрібно: Зберігати та обробляти дані у занадто великих томах для традиційної бази даних.

Перетворення неструктурованих даних для аналізу та звітності. Захоплюйте, обробляйте та аналізуйте необмежені потоки даних у режимі реального часу або з низькою затримкою. Використовуйте машинне навчання Azure або когнітивні служби Microsoft.

Переваги

Вибір технологій. Ви можете змішувати та поєднувати керовані служби Azure та технології Apache у кластерах HDInsight, щоб скористатися наявними навичками чи технологічними інвестиціями. Виконання через паралелізм. Рішення великих даних використовують переваги паралелізму, дозволяючи високопродуктивні рішення, які масштабуються до великих обсягів даних.

Еластична шкала. Усі компоненти архітектури великих даних підтримують масштабоване надання ресурсів, щоб ви могли налаштувати своє рішення на невеликі або великі робочі навантаження та оплачувати лише ресурси, які ви використовуєте.

Взаємодія з існуючими рішеннями. Компоненти архітектури великих даних також використовуються для обробки IoT та корпоративних рішень BI, що дозволяє вам створювати інтегроване рішення для різних навантажень даних.

Виклики

Складність. Рішення великих даних можуть бути надзвичайно складними, з численними компонентами для обробки даних з різних джерел даних. Створення, тестування та усунення несправностей у процесах великих даних може бути складним. Крім того, може існувати велика кількість налаштувань конфігурації в декількох системах, які необхідно використовувати для оптимізації продуктивності.

Сукупність навичок. Багато технологій великих даних є вузькоспеціалізованими та використовують фреймворки та мови, які не характерні для загальних архітектур додатків. З іншого боку, технології великих даних розвивають нові API, які спираються на більш відомі мови. Наприклад, мова U-SQL в Azure Data Lake Analytics базується на комбінації Transact-SQL і C#. Аналогічно, API на основі SQL доступні для Hive, HBase та Spark.

Технологічна зрілість. Багато технологій, що використовуються у великих даних, розвиваються. У той час як основні технології Hadoop, такі як Hive та Pig, стабілізувалися, нові технології, такі як Spark, вносять значні зміни та вдосконалення з кожним новим випуском.

Керовані служби, такі як Azure Data Lake Analytics та Azure Data Factory, порівняно молоді, порівняно з іншими службами Azure, і, ймовірно, з часом будуть розвиватися. Безпека. Рішення великих даних зазвичай спираються на зберігання всіх статичних даних у централізованому озері даних. Забезпечення доступу до цих даних може бути складним завданням, особливо коли дані мають поглинатися та споживатися різними програмами та платформами.

Кращі практики

Використовуйте паралелізм. Більшість технологій обробки великих даних розподіляють робоче навантаження на декілька одиниць обробки. Для цього потрібно створювати статичні файли даних і зберігати їх у розділеному форматі. Розподілені файлові системи, такі як HDFS, можуть оптимізувати продуктивність читання та запису, а фактична обробка виконується кількома вузлами кластера паралельно, що скорочує загальний час виконання завдань.

Дані розділів. Пакетна обробка зазвичай відбувається за періодичним графіком - наприклад, щотижня або щомісяця. Розділіть файли даних та структури даних, такі як таблиці, на основі часових періодів, які відповідають графіку обробки. Це спрощує введення даних та планування роботи та полегшує усунення несправностей. Крім того, таблиці розділів, які використовуються у запитах Hive, U-SQL або SQL, можуть значно покращити продуктивність запитів.

Застосування семантики "читання схеми при читанні". Використання озера даних дозволяє комбінувати сховище для файлів у різних форматах, будь то структуровані, напівструктуровані чи неструктуровані. Використовуйте семантику при читанні схеми, яка проектує схему на дані під час обробки даних, а не коли дані зберігаються. Це вбудовує рішення у гнучкість і запобігає

виникненню вузьких місць під час введення даних, викликаних валідацією даних та перевіркою типів.

Обробка даних на місці. У традиційних рішеннях BI часто використовується процес вилучення, перетворення та завантаження (ETL) для переміщення даних у сховище даних. Завдяки великим обсягам даних та більшому різноманіттю форматів рішення для великих даних зазвичай використовують варіанти ETL, такі як перетворення, вилучення та завантаження (TEL). За такого підходу дані обробляються в розподіленому сховищі даних, перетворюючи їх у необхідну структуру, перед переміщенням перетворених даних у сховище аналітичних даних.

Використання балансу та витрати часу. Для завдань пакетної обробки важливо враховувати два фактори: вартість одиниці обчислювальних вузлів та похвилинна вартість використання цих вузлів для завершення роботи. Наприклад, пакетне завдання може зайняти вісім годин із чотирма вузлами кластера. Однак може виявитися, що робота використовує всі чотири вузли лише протягом перших двох годин, а після цього потрібні лише два вузли. У цьому випадку виконання всього завдання на двох вузлах збільшить загальний час виконання робіт, але не збільшить його удвічі, тому загальна вартість буде меншою. У деяких бізнес -сценаріях більш тривалий час обробки може бути кращим, ніж більша вартість використання недостатньо використаних ресурсів кластера.

Окремі ресурси кластера. Розгортаючи кластери HDInsight, ви зазвичай досягаєте кращої продуктивності, забезпечуючи окремі ресурси кластера для кожного типу робочого навантаження. Наприклад, хоча кластери Spark включають Hive, якщо вам потрібно виконати обширну обробку з Hive і Spark, вам слід розглянути можливість розгортання окремих виділених кластерів Spark та Hadoop. Аналогічно, якщо ви використовуєте HBase та Storm для обробки потоку з низькою затримкою та Hive для пакетної обробки, розгляньте окремі кластери для Storm, HBase та Hadoop.

Організуйте введення даних. У деяких випадках існуючі бізнес - програми можуть записувати файли даних для пакетної обробки безпосередньо в контейнери BLOB -сховищ зберігання Azure, де вони можуть використовуватися HDInsight або Azure Data Lake Analytics. Однак часто вам доведеться організувати введення даних із локальних або зовнішніх джерел даних в озеро даних. Використовуйте робочий процес або конвеєр оркестрації, наприклад, підтримувані Azure Data Factory або Oozie, щоб досягти цього передбачуваним та централізовано керованим способом.

Стерти чутливі дані завчасно. Робочий процес з введення даних повинен очистити конфіденційні дані на початку процесу, щоб уникнути їх зберігання в озері даних.

Архітектура IoT

Інтернет речей (IoT) - це спеціалізована підмножина рішень для великих даних. Наступна діаграма показує можливу логічну архітектуру IoT. Діаграма підкреслює подієві компоненти архітектури (рис. 8.2).

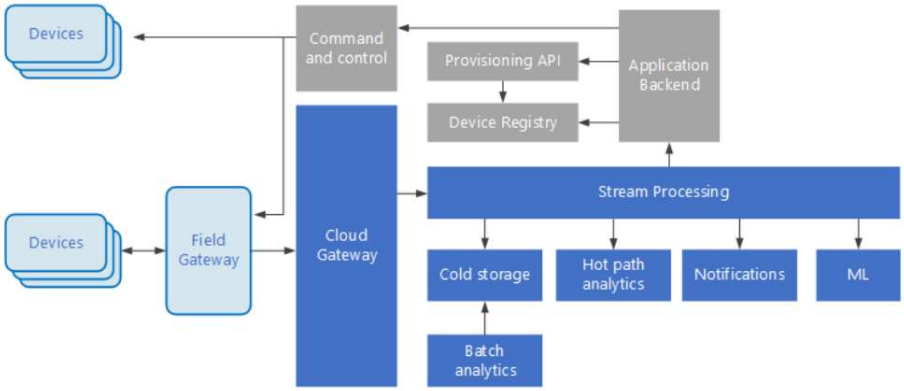


Рис. 8.2. Архітектура IoT

Хмарний шлюз ковтає події пристрою на межі хмари, використовуючи надійну систему обміну повідомленнями з низькою затримкою.

Пристрої можуть надсилати події безпосередньо до хмарного шлюзу або через польовий шлюз. Польовий шлюз - це спеціалізований пристрій або програмне забезпечення, яке зазвичай розміщується разом із пристроями, яке приймає події та пересилає їх до хмарного шлюзу. Шлюз поля також може попередньо обробляти необроблені події пристрою, виконуючи такі функції, як фільтрація, агрегація або трансформація протоколу.

Після введення події проходять через один або кілька потокових процесорів, які можуть направляти дані (наприклад, до сховища) або виконувати аналітику та іншу обробку.

Нижче наведено деякі поширені типи обробки. (Цей список, безумовно, не є вичерпним.) Запис даних про події на холодне сховище для архівування або пакетної аналітики.

Аналіз гарячих шляхів, що аналізує потік подій у (майже) реальному часі, для виявлення аномалій, розпізнавання закономірностей у вікнах часу, що прокручується, або викликати сповіщення, коли в потоці виникає певна умова.

Обробка спеціальних типів нетелеметричних повідомлень від пристроїв, таких як сповіщення та тривоги.

Машинне навчання.

Сірі рамки, затінені сірим кольором, показують компоненти системи IoT, які не мають прямого відношення до потокової передачі подій, але включені сюди для повноти.

Реєстр пристроїв являє собою базу даних наданих пристроїв, включаючи ідентифікатори пристроїв і зазвичай метадані пристрою, такі як місцезнаходження. API забезпечення є загальним зовнішнім інтерфейсом для надання та реєстрації нових пристроїв.

Деякі рішення IoT дозволяють надсилати на пристрої повідомлення команд та керування.

8.2. Еволюція моделей розгортання в епоху великих даних

З появою великих даних моделі розгортання для управління даними змінюються. Традиційне сховище даних зазвичай реалізується на єдиній великій системі в центрі обробки даних. Витрати на цю модель змусили організації оптимізувати ці склади та обмежити обсяг та розмір керованих даних.

Однак, коли організації хочуть використовувати величезний обсяг інформації, що генерується великими джерелами даних, обмеження традиційних моделей більше не працюють. Тому пристрій зберігання даних став практичним методом створення оптимізованого середовища для підтримки переходу до нового управління інформацією.

Модель пристрою великих даних

Коли компаніям потрібно поєднати структуру свого сховища даних з великими даними, модель пристрою може стати однією відповіддю на проблему масштабування. Як правило, пристрій являє собою інтегровану

систему, яка включає апаратне забезпечення (зазвичай у стійці), оптимізоване для зберігання та управління даними.

Оскільки вони є автономними, прилади можуть бути відносно простими та швидкими для впровадження, а також пропонують менші витрати на експлуатацію та обслуговування. Тому система буде попередньо завантажена реляційною базою даних, фреймворком Hadoop, MapReduce та багатьма інструменти, які допомагають збирати та упорядковувати дані з різних джерел.

Він також містить аналітичні механізми та інструменти для спрощення процесу аналізу даних з різних джерел. Тому пристрій є одноцільовою системою, яка зазвичай містить інтерфейси для полегшення підключення до наявного сховища даних.

Модель хмари великих даних

Хмара стає переконливою платформою для управління великими даними і може бути використана в гібридному середовищі з локальним середовищем. Деякі нові інновації у завантаженні та передачі даних уже змінюють потенційну життєздатність хмари як платформи для зберігання великих даних.

Наприклад, компанія Aspera, яка спеціалізується на швидкій передачі даних між мережами, співпрацює з Amazon.com, щоб запропонувати послуги управління хмарними даними. Інші постачальники, такі як FileCatalyst та Data Expedition, також орієнтовані на цей ринок. По суті, ця категорія технологій використовує мережу та оптимізує її для переміщення файлів із зменшеною затримкою.

Оскільки ця проблема затримок у передачі даних продовжує розвиватися, буде нормою зберігати системи великих даних у хмарі, які можуть взаємодіяти зі сховищем даних, яке також є хмарним, або зі складом, який знаходиться у центрі обробки даних.

8.3. Моделі розгортання в хмарі великих даних

Дві ключові моделі хмар важливі для обговорення великих даних - загальнодоступні та приватні хмари. Хмарні обчислення - це метод надання набору спільних обчислювальних ресурсів, що включає програми, обчислення, зберігання, мережу, платформи розробки та розгортання, а також

бізнес -процеси. Хмарні обчислення перетворюють традиційні обчислювальні ресурси на спільні ресурси.

Два типи моделей розгортання для хмарних обчислень - публічні та приватні. Вони пропонуються для обчислювальних потреб загального призначення на відміну від конкретних типів моделей хмарної доставки.

Громадська хмара.

Публічна хмара - це набір апаратного забезпечення, мереж, сховищ, послуг, програм та інтерфейсів, що належать третій стороні та управляються нею для використання іншими компаніями та приватними особами. Ці комерційні постачальники створюють високомасштабований центр обробки даних, який приховує деталі основної інфраструктури від споживача.

Громадські хмари є життєздатними, оскільки вони зазвичай керують відносно повторюваними або простими робочими навантаженнями. Наприклад, електронна пошта - це дуже проста програма. Тому постачальник хмарних послуг може оптимізувати середовище так, щоб він найкраще підходив для підтримки великої кількості клієнтів.

Подібним чином постачальники публічної хмари, що пропонують послуги зберігання чи обчислювальні послуги, оптимізують своє обчислювальне обладнання та програмне забезпечення для підтримки цих конкретних типів навантажень.

На відміну від цього, типовий центр обробки даних підтримує стільки різних програм та робочих навантажень, що його неможливо легко оптимізувати. Публічна хмара може бути дуже ефективною, коли організація виконує складний проект аналізу даних і потребує додаткових обчислювальних циклів для виконання завдання.

Крім того, компанії можуть вибрати зберігання даних у загальнодоступній хмарі, де вартість одного гігабайта порівняно недорога у порівнянні з придбаним сховищем. Найважливішими проблемами загальнодоступних хмар для великих даних є вимоги безпеки та допустима величина затримки.

Усі публічні хмари не однакові. Деякі загальнодоступні хмари - це масштабовані керовані послуги з високим рівнем безпеки та високим рівнем управління послугами. Інші загальнодоступні хмари менш надійні та менш безпечні, але вони набагато дешевші у використанні.

Приватна хмара Приватна хмара - це набір обладнання, мереж, сховищ, послуг, програм та інтерфейсів, якими володіє та управляє організація для використання її співробітниками, партнерами та клієнтами. Приватну хмару може створювати та управляти третя сторона для виключного використання одного підприємства.

Приватна хмара - це висококонтрольоване середовище, не відкрите для загального споживання. Таким чином, приватна хмара знаходиться за брандмауером. Приватна хмара високо автоматизована з акцентом на управління, безпеку та відповідність вимогам. Автоматизація замінює більш ручні процеси управління IT -послугами для підтримки клієнтів.

Таким чином, бізнес -правила та процеси можуть бути реалізовані всередині програмного забезпечення, щоб середовище стало більш передбачуваним та керованим. Якщо організації керують великими проектами даних, які вимагають обробки великої кількості даних, приватна хмара може бути найкращим вибором з точки зору затримки та безпеки.

Питання для самоконтролю:

1. Для чого призначена архітектура великих даних?
2. Розкрийте стиль архітектури великих даних.
3. Еволюція моделей розгортання в епоху великих даних
4. Назвіть моделі розгортання в хмарі великих даних
5. Які ключові моделі хмар важливі для обговорення?

СПИСОК РЕКОМЕНДОВАНИХ ДЖЕРЕЛ

1. Заяць В. М. Роль інформаційних технологій у формуванні стратегічного мислення менеджера / В. М. Заєць // Актуальні проблеми економіки. – 2009. – №6 (96). – С. 280-288.
2. Європейська Бізнес Асоціація: www.eba.com.ua. - Нові підходи в управлінні IT або як бізнес-цілі, пов'язані з IT-процесами.
3. MastersInDataScience.org is owned and operated by 2U, Inc.. Here Are 10 Key Benefits of Business Intelligence Software. [Електроний носій]. - https://www.mastersindatascience.org/learning/benefits-of-business-intelligence/
4. The Inexorable Rise of Self Service Data Integration. [Електроний носій]. - https://blogs.gartner.com/andrew_white/2015/05/22/the-inexorable-rise-of-self-service-data-integration/
5. 15 найкращих інструментів ETL у 2021 році (повний оновлений список). [Електроний носій]. - https://uk.myservername.com/15-best-etl-tools-2021
6. Проектування інформаційних систем: навчальний посібник / В.С. Авраменко, А.С. Авраменко. – Черкаси: Черкаський національний університет ім. Б. Хмельницького, 2017. – 434 с.: іл.
7. Будсвіт Україна » Маркетинг » 3 типи маркетингових інформаційних панелей та способи їх використання. [Електроний носій]. - https://budsvit.net.ua/3-typy-marketyngovyh-informaczijnyh-panelej-ta-sposoby-yih-vykorystannya
8. Data Analytics Lifecycle: An Easy Overview For 2021 [Електроний носій]. - https://www.jigsawacademy.com/blogs/hr-analytics/data-analytics-lifecycle/
9. Ахмед А. А. Гад-Елраб. Сучасний бізнес-аналіз: аналіз великих даних та штучний інтелект для створення цінності, керованої даними. [Електроний носій]. - https://www.intechopen.com/chapters/76332
10. Інтелектуальний аналіз даних: Комп'ютерний практикум [Електронний ресурс] : навч. посіб. для студ. спеціальності 122 «Комп'ютерні науки та інформаційні технології», спеціалізацій «Інформаційні системи та технології проектування», «Системне проектування сервісів» / О. О. Сергеєв-Горчинський, Г. В. Іщенко ; КПІ ім. Ігоря Сікорського. – Електронні текстові данні (1 файл: 1,72 Мбайт). – Київ : КПІ ім. Ігоря Сікорського, 2018. – 73 с.: Іл
11. By Craig S. Mullins. The Importance of Data Modeling in a Big Data World. [Електроний носій]. - https://www.dbta.com/Editorial/Think-

About-It/The-Importance-of-Data-Modeling-in-a-Big-Data-World-145915.aspx
12. Big data architecture style. - [Електроний носій]. -
13. https://docs.microsoft.com/en-us/azure/architecture/guide/architecture-styles/big-data
14. By Judith S. Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman. The Evolution of Deployment Models in the Big Data Era. [Електроний носій]. -https://www.dummies.com/programming/big-data/engineering/big-data-cloud-deployment-models/
15. By Judith S. Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman. Big Data Cloud Deployment Models. [Електроний носій]. - https://www.dummies.com/programming/big-data/engineering/big-data-cloud-deployment-models/
16. Hariri, R.H., Fredericks, E.M. & Bowers, K.M. Uncertainty in big data analytics: survey, opportunities, and challenges. J Big Data 6, 44 (2019). https://doi.org/10.1186/s40537-019-0206-3
17. R. Lovas, A. Farkas, A. C. Marosi et al., "Orchestrated Platform for Cyber-Physical Systems," Complexity, vol. 2018, Article ID 8281079, 16 pages, 2018.
18. R. Y. Zhong, X. Xu, E. Klotz, and S. T. Newman, "Intelligent Manufacturing in the context of industry 4.0: a review," Engineering Journal, vol. 3, no. 5, pp. 616–630, 2017.
19. R. M. Müller, H.-J. Lenz. 2013. Business Intelligence
20. Y. Su, X. Meng, Q. Kang, and X. Han, "Dynamic Virtual Network Reconfiguration Method for Hybrid Multiple Failures Based on Weighted Relative Entropy," Entropy, vol. 20, no. 9, p. 711, 2018.
21. Daniel Keim, Jörn Kohlhammer, Geoffrey Ellis und Florian Mansmann. „Visual Analytics". 2010
22. C. Chen, M. Lin, and X. Guo, "High-level modeling and synthesis of smart sensor networks for Industrial Internet of Things," Computers & Electrical Engineering, vol. 61, pp. 48–66, 2017.
23. Daniel Keim, Jörn Kohlhammer, Geoffrey Ellis und Florian Mansmann. „Visual Analytics". 2010
24. Dimitri P. Bertsekas and John N. Tsitsiklis. Introduction to Probability. Charles Wheelan. Naked Statistics: Stripping the Dread from the Data. W. W. Norton and Company, 2013.
25. F. Liu, Y. Liu, D. Jin, X. Jia, and T. Wang, "Research on Workshop-Based Positioning Technology Based on Internet of Things in Big Data Background," Complexity, vol. 2018, Article ID 875460, 11 pages, 2018.

26. H. Mora, M. Signes-Pont, D. Gil, and M. Johnsson, “Collaborative Working Architecture for IoT-Based Applications,” Sensors, vol. 18, no. 6, p. 1676, 2018.
27. H. Tahaei, R. Salleh, S. Khan, R. Izard, K.-K. R. Choo, and N. B. Anuar, “A multi-objective software defined network traffic measurement,” Measurement, vol. 95, pp. 317–327, 2017.
28. Hariri, R.H., Fredericks, E.M. & Bowers, K.M. Uncertainty in big data analytics: survey, opportunities, and challenges. J Big Data 6, 44 (2019). https://doi.org/10.1186/s40537-019-0206-3
29. INMON, W.H.; LINSTEDT, D.: Data architecture a primer for the data scientist: big data, data warehouse and data vault. 2014.
30. J. Han, M. Kamber. 2011. Data Mining. Concepts and Techniques Visualize This by Nathan Yau
31. J. Pan and J. McElhannon, “Future edge cloud and edge computing for internet of things applications,” IEEE Internet of Things Journal, vol. 5, no. 1, pp. 439–449, 2018.
32. L. J. M. Nieuwenhuis, M. L. Ehrenhard, and L. Prause, “The shift to Cloud Computing: The impact of disruptive technology on the enterprise software business ecosystem,” Technological Forecasting & Social Change, vol. 129, pp. 308–313, 2018.
33. M. Giacobbe, R. Di Pietro, A. Longo Minnolo, and A. Puliafito, “Evaluating Information Quality in Delivering IoT-as-a-Service,” in Proceedings of the 2018 IEEE International Conference on Smart Computing (SMARTCOMP), pp. 405–410, June 2018.
34. M. Osman, “A novel big data analytics framework for smart cities,” Future Generation Computer Systems, vol. 91, pp. 620–633, 2019.
35. Marrone, M. and Hazelton, J. (2019), "The disruptive and transformative potential of new technologies for accounting, accountants and accountability: A review of current literature and call for further research", Meditari Accountancy Research, Vol. 27 No. 5, pp. 677-694. https://doi.org/10.1108/MEDAR-06-2019-0508
36. Osborne, Jason W. “Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data.” 2013
37. R. Lovas, A. Farkas, A. C. Marosi et al., “Orchestrated Platform for Cyber-Physical Systems,” Complexity, vol. 2018, Article ID 8281079, 16 pages, 2018.
38. R. M. Müller, H.-J. Lenz. 2013. Business Intelligence
39. R. Y. Zhong, X. Xu, E. Klotz, and S. T. Newman, “Intelligent Manufacturing in the context of industry 4.0: a review,” Engineering Journal, vol. 3, no. 5, pp. 616–630, 2017.
40. Steven Skiena. “The Data Science Design Manual” http://www.data-manual.com/

41. TURBAN, EFRAIM ; SHARDA, RAMESH ; DELEN, DURSUN ; KING, DAVID: Business intelligence: a managerial approach. Boston, Mass. : Pearson, Prentice Hall, 2011 www.vismaster.eu/wp-content/uploads/2010/11/VisMaster-book-lowres.pdf
42. X. Wang, C. Xu, G. Zhao, K. Xie, and S. Yu, “Efficient Performance Monitoring for Ubiquitous Virtual Networks Based on Matrix Completion,” IEEE Access, vol. 6, pp. 14524–14536, 2018.
43. Y. Guo, Z. Yang, S. Feng, and J. Hu, “Complex Power System Status Monitoring and Evaluation Using Big Data Platform and Machine Learning Algorithms: A Review and a Case Study,” Complexity, vol. 2018, Article ID 8496187, 21 pages, 2018.
44. Y. Su, X. Meng, Q. Kang, and X. Han, “Dynamic Virtual Network Reconfiguration Method for Hybrid Multiple Failures Based on Weighted Relative Entropy,” Entropy, vol. 20, no. 9, p. 711, 2018.

Інформаційні ресурси:

1.Державна служба статистики [Электронный ресурс] Режим доступа: http://ukrstat.gov.ua
2. Національний інститут стратегічних досліджень. Офіційний сайт. [Электронный ресурс] Режим доступа: http://www.niss.gov.ua
3.Національна бібліотека України імені В. І. Вернандського Офіційний сайт. [Электронный ресурс] Режим доступа: http://www.nbuv.gov.ua