УДК 004.422.61

Антон Вітушко,
наук. співроб. НЮБ НБУВ

РЕАЛІЗАЦІЯ І ФОРМУВАННЯ ІНФОРМАЦІЙНО-МОНІТОРИНГОВИХ СИСТЕМ ЯК ІТ-ПРОДУКТУ НА БАЗІ АНАЛІТИЧНИХ СЛУЖБ

У роботі проведено дослідження використання інформаційно-аналітичних систем, їх архітектур і завдань функціонування. Розглянуто перспективи використання інформаційно-моніторингової системи на базі аналітичних служб, як окремої системи, так і в структурі інформаційно-аналітичної системи; її реалізація в ІТ-продукт служби інформаційно-аналітичної діяльності.

Ключові слова: інформаційно-моніторингові системи, інформаційно-аналітичні системи, Text Mining, Data Mining, Knowledge Management, OLAP, Oltp-системи, Etl-інструменти, Data Warehouse.

В останні роки на інформаційно-програмному ринку з'явилася велика кількість інформаційно-аналітичних систем, що дасть можливість проводити інтелектуальний аналіз великих інформаційних масивів, у першу чергу – накопичених у мережі Інтернет у вигляді професійних БД, веб-ресурсів і т. ін. Такі системи здатні автоматизувати процеси збору й аналізу інформації, забезпечити оптимізацію процесу створення інформаційно-аналітичного продукту.

У цілому суть роботи аналітичних служб полягає в збиранні вихідних даних і первинної інформації, її узагальненні, установленні причинно-наслідкових зв'язків впливу одних факторів на інші. На підставі отриманих результатів аналізу й наявного досвіду – агрегування даних, підготовка аналітичних матеріалів, звітів – в остаточному підсумку – в прогнозуванні розвитку ситуації.

Сукупність масивів необхідної інформації, знань і досвіду аналітика, ефективного аналітичного інструментарію становлять так звані корпоративні знання, які необхідно формувати, зберігати й керувати ними для підтримки на високому рівні основних аналітичних функцій інформаційної служби й забезпечення ефективної оперативної реакції на зміни в інформаційному середовищі.

Одним з найважливіших компонентів корпоративних знань є інструментарій аналітика або інформаційно-аналітична система.

Треба зазначити, що основне завдання інформаційно-аналітичних служб полягає не у володінні інформацією, а в наявності і вдосконаленні навичок її використання, правильному розумінні кола завдань і, відповідно, виборі інструменту аналітика.

Увесь процес збору, обробки, аналізу інформації й синтезу знань являє собою ряд послідовних заходів.

До його складу, як правило, входять такі основні етапи:

– усвідомлення – виявлення зовнішніх ознак змін, що відбуваються;
– планування;
– збір даних і потенційно значущої інформації;
– структурування зібраної інформації;
– обробка інформації – її аналіз за допомогою відповідних методів і інструментів;
– забезпечення доступу до інформації;
– аналіз і синтез інформації;
– використання інформації в процесі прийняття й виконання рішень;
– поширення отриманих знань.

Зібраний матеріал на першому етапі являє собою первинні дані, після обробки він перетворюється в інформацію, і тільки після аналізу інформації й синтезу на основі її висновків експертів вона стає інформаційним продуктом.

Найважливішою умовою успішної роботи експерта-аналітика є наявність інформаційного поля досліджуваної предметної галузі, яке повинне являти собою ряд структурованих і неструктурованих інформаційних масивів, потрібних для витягу з них необхідних відомостей.

Інформаційне поле містить у собі як дані, одержувані із зовнішніх джерел, так і дані, одержувані з внутрішніх джерел.

Основна технологія аналітика – це встановлення причинно-наслідкових зв'язків між різного роду даними та їх дослідження. Побудова причинно-наслідкових ланцюжків дає змогу оброблені дані перетворити на інформацію й зробити висновки в предметній галузі.

Визначивши основні принципи організації роботи аналітичної групи, ми можемо сформувати вимоги до функціональності програмних засобів, які повинні забезпечити якісну роботу аналітиків.

Проведення досліджень у напрямах формувань моделі аналітика, створення штучного інтелекту дали розроблювачам засоби й методики для створення програмного продукту в галузі нейронних технологій,

інтелектуального пошуку в неструктурованій текстовій інформації (Text Mining), системи витягу даних і систем розпізнання образів (Data Mining).

Розробки в цих галузях привели до створення технології керування знаннями (Knowledge Management, KM). Це фактично підвело розроблювачів програмного забезпечення до автоматизації галузей людської діяльності, що важко піддаються формалізації, до яких можна віднести процеси пошуку й аналізу інформації.

Більшість програмних засобів інформаційного моніторингу реалізують, як правило, типові функції:

– пошук і збір даних із джерел різних форматів (БД, неструктуровані джерела і т. д.);

– нагромадження й зберігання даних;

– рубрикацію архівів;

– пошук даних, у тому числі й нечіткий пошук;

– побудова звітів у різних зрізах вибірки, у тому числі й багатомірний аналіз даних;

– побудова причинно-наслідкових ланцюжків даних, що дають змогу визначати тенденції й напрями подальшого аналізу.

Таким чином, розв'язання в галузі програмного забезпечення для автоматизації процесів пошуку й аналізу інформації можна класифікувати за рядом ознак (див. рис. 1).

Завданням будь-якої інформаційно-аналітичної системи є ефективне зберігання, обробка й аналіз даних. На сьогодні вже накопичений значний досвід у цій сфері.

Ефективне зберігання інформації досягається наявністю в складі інформаційно-аналітичної системи цілого ряду джерел даних. Обробка й об'єднання інформації досягається застосуванням інструментів витягу, перетворення й завантаження даних. Аналіз даних здійснюється за допомогою сучасних інструментів аналізу даних [1].

Різноманітність джерел даних і необхідність їх використання в кожному конкретному випадку пояснюється потребою по-різному зберігати інформацію, залежно від поставлених перед службою завдань. Якщо спробувати класифікувати джерела даних за їх типами та за призначенням, то кожний з них можна умовно віднести до однієї з трьох груп: транзакційні джерела даних, сховища даних, вітрини даних.

Дані в систему можуть заноситися як вручну, так і автоматично. На етапі первісної фіксації дані надходять через системи збору й обробки інформації в так звані транзакційні бази даних, яких в організації може бути декілька.

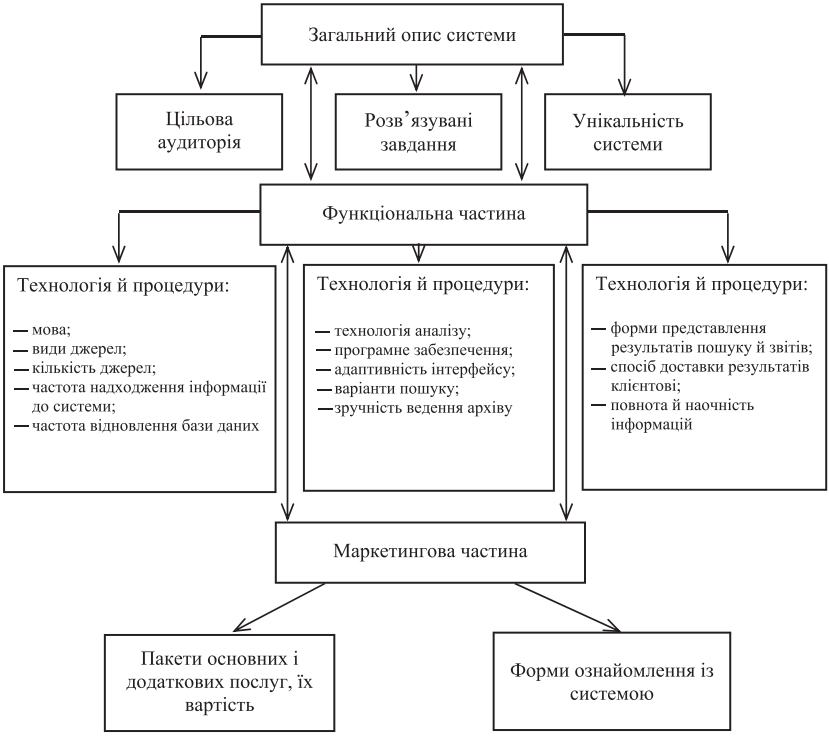

Рис.1. Класифікація інформаційно-аналітичних систем

Оскільки транзакційні джерела даних, як правило, не погоджені один з одним, то для аналізу таких даних потрібно їх об'єднання й перетворення. Тому на наступному етапі вирішується завдання консолідації даних, їх перетворення й очищення, у результаті чого дані надходять у так звані аналітичні бази даних.

При цьому інформаційно-аналітична система повинна забезпечувати користувачам доступ до аналітичної інформації, захищеної від несанкціонованого використання та відкритої як через внутрішню мережу організації, так і користувачам мережі інтранет та Інтернет. Таким чином, архітектура сучасної інформаційно-аналітичної системи має такий вигляд (див. рис. 2).

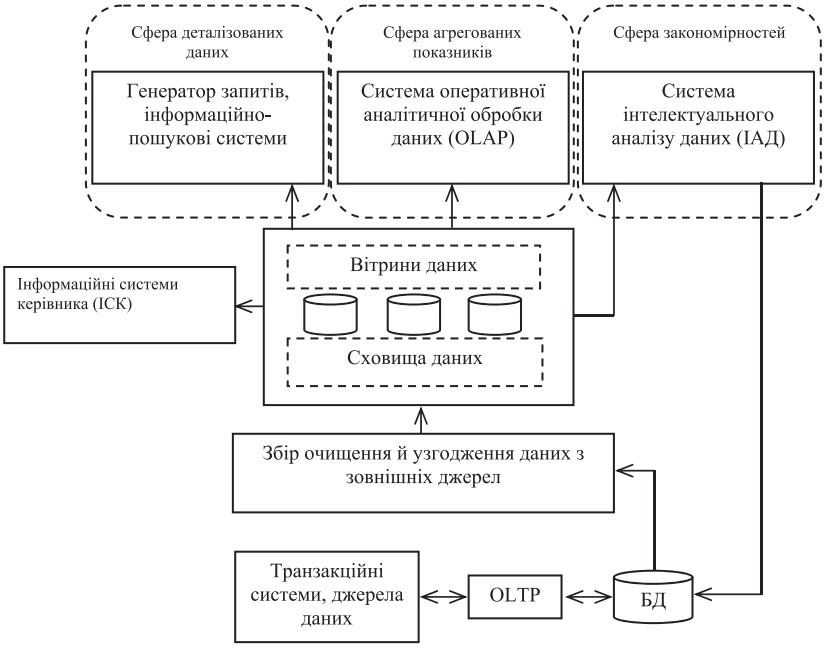

Рис. 2. Архітектура сучасної інформаційно-аналітичної системи

До першого рівня архітектури інформаційно-аналітичної системи належать, як правило, транзакційні або операційні джерела даних, що є частиною так званих OLTP-систем (online transactional processing). Транзакційні бази даних містять у собі джерела даних, орієнтовані на фіксацію результатів повсякденної діяльності організації. Вимоги, що висуваються до транзакційних баз даних, обумовили їхні такі відмінні риси: здатність швидко обробляти дані й підтримувати високу частоту їх змін, орієнтованість на обслуговування одного процесу, а не всієї діяльності організації в цілому.

Інформація в таких базах даних орієнтована на конкретний додаток і керується транзакціями, вона сильно деталізована й постійно коректується. Транзакційні бази даних якісно виконують завдання рутинної обробки щоденної інформації, але нечасто можуть служити джерелами для проведення комплексного аналізу. Отже, сукупність транзакційних джерел даних утворює нижню ланку архітектури інформаційно-аналітичної системи.

Процес витягування, перетворення й завантаження даних підтримується так званими ETL-інструментами (extraction, transformation, loading), призначеними для витягування даних з різних транзакційних джерел нижнього рівня, їх перетворення й консолідації, а також завантаження в цільові аналітичні бази даних – сховища даних і вітрини даних. На етапі перетворення усувається надмірність даних, проводяться необхідні обчислення й агрегування. Процес витягування, перетворення й завантаження повинен здійснюватися на основі встановленого регламенту.

До третього рівня архітектури ІАС належать джерела даних, які називають сховищами даних. Сховища даних містять у собі джерела даних, орієнтовані на зберігання й аналіз інформації. Такі джерела можуть поєднувати інформацію з декількох транзакційних систем і дають змогу аналізувати її в комплексі із застосуванням сучасних програмних інструментів аналізу даних.

Характерними рисами сховищ даних є недостатня можливість коригування більшості даних, оновлюваність даних на періодичній основі, єдиний підхід до найменування й зберігання даних незалежно від їх організації у вихідних джерелах.

Сховище даних, що є одним з головних ланок архітектури ІАС, виступає як основне джерело даних для всебічного аналізу всієї наявної в службі інформації. Для того щоб існуючі сховища даних сприяли аналітичному аналізу, аналітик повинен мати розвинуті інструменти доступу до даних сховища та їх обробки.

Автором концепції сховищ даних (Data Warehouse) є Б. Інмон, який визначив їх як «предметно орієнтовані, інтегровані, немінливі набори, що підтримують хронологію, даних, організовані з метою підтримки керування», покликані виступати в ролі «єдиного джерела істини», що забезпечує аналітиків достовірною інформацією, яка необхідна для оперативного аналізу й прийняття рішень щодо вирішення завдань аналізу [3].

В основі концепції сховищ даних лежать дві основні ідеї:

1. Інтеграція раніше роз'єднаних деталізованих даних у єдиному сховищі даних, їх узгодження й можлива агрегація.

2. Розподіл наборів даних, що використовуються для операційної обробки, і наборів даних, що використовуються для вирішення завдань аналізу.

Крім єдиного довідника метаданих, засобів вивантаження, агрегації й узгодження даних, концепція сховищ даних має на увазі інтегрованість, немінливість, підтримку хронології й погодженість даних. І якщо дві перші

властивості (інтегрованість і немінливість) впливають на режими аналізу даних, то останні дві (підтримка хронології й погодженість) істотно звужують список вирішуваних аналітичних завдань. Без підтримки хронології не можна говорити про вирішення завдань прогнозування й аналізу тенденцій. Але найбільш критичними й вразливими є питання, пов'язані з узгодженням даних.

Основною вимогою аналітика є не стільки оперативність, скільки вірогідність відповіді, яка (вірогідність) в остаточному підсумку визначається погодженістю. Поки не проведено роботу із взаємного узгодження значень даних з різних джерел, складно говорити про їх вірогідність.

Практично в будь-якій організації питання про погодженість даних у різних інформаційних системах стоїть надзвичайно гостро. Досить часто аналітики зустрічаються із ситуацією, коли на одне й те саме запитання різні системи дають різні відповіді. Це може бути пов'язано з несинхронністю моментів модифікації даних, відмінностями в трактуванні тих самих подій, понять і даних, зміною семантики даних у процесі розвитку предметної області, елементарними помилками під час введення й обробки, частковою втратою окремих фрагментів архівів тощо. Обчислити й заздалегідь визначити алгоритми дозволу всіх можливих колізій складно. Тим більше що це нереально зробити і в оперативному режимі, безпосередньо в процесі формування відповіді на запит.

До четвертого рівня архітектури ІАС належать джерела даних, які отримали назву вітрин даних (data marts), призначені для проведення цільового ділового аналізу. Вітрини даних будуються, як правило, на основі інформації зі сховища даних, але можуть також формуватися з даних безпосередньо із транзакційних систем, коли сховище даних в організації з будь-яких причин не реалізовано.

За типом зберігання інформації вітрини поділяються на реляційні й багатомірні. Вітрини першого типу організують у вигляді реляційної бази даних зі схемою «зірка», де центральна таблиця, таблиця фактів, призначена в основному для зберігання кількісної інформації, пов'язана з таблицями-довідниками [4].

Багатомірні вітрини організують у вигляді багатомірних баз даних OLAP (Online Analytical Processing), де довідкова інформація представляється у вигляді вимірів, а кількісна – у вигляді показників.

Відмінність вітрин даних від транзакційних баз даних полягає в тому, що перші служать для задоволення потреб кінцевих користувачів, що не є професійними аналітиками. Транзакційні бази даних використову-

ються в основному операторами, відповідальними за введення й обробку первинної інформації, а не за її аналіз, спрямований на підтримку прийняття рішень.

Застосування вітрин даних, багатомірних і реляційних, у комбінації із сучасними інструментами аналізу даних дає змогу перетворити прості дані в корисну інформацію, на основі якої можна одержувати інформаційний продукт високої якості.

Концепція вітрин даних (Data Mart) була запропонована Forrester Research ще в 1991 р. [2]. Вітрини даних – це безліч тематичних БД, які містять інформацію, що належить до окремих аспектів діяльності організації.

Концепція вітрин даних має свої позитивні якості:

1. Аналітики бачать і працюють тільки з тими даними, які їм реально потрібні.

2. Цільова БД вітрини даних максимально наближена до кінцевого користувача.

3. Вітрини даних містять тематичні підмножини заздалегідь агрегованих даних.

4. Для реалізації вітрин даних не потрібна потужна обчислювальна техніка.

Але концепція вітрин даних має й свої недоліки. По суті, тут передбачається реалізація територіально розподіленої інформаційної системи з малоконтрольованою надмірністю, але не пропонується способів, як забезпечити цілісність і несуперечність збережених у ній даних.

Ідея з'єднати дві концепції – сховищ даних і вітрин даних – належить M. Demarest, який у 1994 р. запропонував об'єднати дві концепції і використовувати сховище даних як єдине інтегроване джерело даних для вітрин даних [2].

Сьогодні ж саме таке багаторівневе рішення є найбільш перспективним:

– перший рівень – загальнокорпоративна БД на основі РСУБД із нормалізованою або малонормалізованою схемою (деталізовані дані);

– другий рівень – БД рівня підрозділу (або кінцевого користувача), реалізовані на основі МСУБД (агреговані дані);

– третій рівень – робочі місця кінцевих користувачів, на яких безпосередньо встановлено аналітичний інструментарій.

Поступово стає стандартом, даючи змогу найбільш повно реалізувати й використовувати переваги кожного з підходів:

– компактне зберігання деталізованих даних і підтримка дуже великих БД, забезпечуваних реляційними СУБД;

– простота налаштування й швидкий період відгуку під час роботи з агрегованими даними, забезпечуваними багатомірними СУБД.

Реляційна форма представлення даних, що використовується в центральній загальнокорпоративній БД, забезпечує найбільш компактний спосіб зберігання даних.

Сучасні реляційні СУБД уже вміють працювати навіть із терабайтними базами. І хоча така центральна система, звичайно, не зможе забезпечити оперативного режиму обробки аналітичних запитів, під час використання нових способів індексації й зберігання даних, а так само часткової денормалізації таблиць час обробки заздалегідь регламентованих запитів (як такі можна розглядати й регламентовані процедури вивантаження даних у багатомірні БД) виявляється цілком прийнятним.

У свою чергу, використання багатомірних СУБД у вузлах нижнього рівня забезпечує мінімальні часи обробки й відповіді на нерегламентовані запити користувача. Крім того, у деяких багатомірних СУБД є можливість зберігати дані як на постійній основі (безпосередньо в багатомірній БД), так і динамічно (на час сеансу) завантажити дані з реляційних БД (на основі регламентованих запитів).

Таким чином, є можливість зберігати на постійній основі тільки ті дані, які найбільш часто запитуються в даному вузлі. Для всіх інших зберігаються тільки описи структури й програми їх вивантаження з центральної БД. І хоча при первинному зверненні до таких віртуальних даних час відгуку може виявитися досить тривалим, таке розв'язання забезпечує високу гнучкість і потребує більш дешевих апаратних засобів.

До наступного рівня архітектури ІАС організації належать сучасні програмні системи інтелектуального аналізу даних (ІАД), системи оперативної аналітичної обробки даних (OLAP) та інформаційно-пошукові системи.

Виконуваний аналіз може проводитися в трьох базових сферах:

1. Сфера деталізованих даних. Це сфера дії більшості систем, націлених на пошук інформації. У більшості випадків реляційні СУБД відмінно справляються із завданнями, що виникають. Загальновизнаним стандартом мови маніпулювання реляційними даними є SQL. Інформаційно-пошукові системи, що забезпечують інтерфейс кінцевого користувача в завданнях пошуку деталізованої інформації, можуть використовуватися як надбудови як над окремими базами даних транзакційних систем, так і над загальним сховищем даних.

2. Сфера агрегованих показників. Комплексний погляд на зібрану в сховищі даних інформацію, її узагальнення й агрегація, гіперкубічне

представлення й багатомірний аналіз є завданнями систем оперативної аналітичної обробки даних (OLAP). Тут можна або орієнтуватися на спеціальні багатомірні СУБД, або залишатися в рамках реляційних технологій. У другому випадку заздалегідь агреговані дані можуть збиратися в БД зіркоподібного виду, або агрегація інформації може проводитися в процесі сканування деталізованих таблиць реляційної БД.

3. Сфера закономірностей. Інтелектуальна обробка проводиться методами інтелектуального аналізу даних (ІАД, Data Mining), головними завданнями яких є пошук функціональних і логічних закономірностей у накопиченій інформації, побудова моделей і правил, які пояснюють знайдені аномалії і/або прогнозують розвиток деяких процесів [5].

Методи інтелектуального аналізу дають змогу проводити всебічний аналіз інформації, допомагають успішно орієнтуватися в більших обсягах даних, аналізувати інформацію, робити на основі аналізу об'єктивні висновки й ухвалювати обґрунтовані рішення, прогнозувати.

Методи інтелектуального аналізу даних використовуються кінцевими користувачами для доступу до інформації, її візуалізації, багатомірного аналізу й формування як визначених за формою й складу, так і довільних звітів, створюваних аналітиком.

Дуже рідко інформаційно-аналітичні служби починають будувати ІАС з нуля. Як правило, на базі аналітичних служб завжди є діючі інформаційні системи. Бажання використовувати комплексні рішення однієї фірми-виробника натрапляє на прагнення зберегти вже наявні напрацювання, представлені у вигляді окремих систем, виконаних у різний час і в різних середовищах. При цьому відмова від діючих систем найчастіше неможлива, а їх перенесення на платформу обраного виробника призводить до значних витрат.

Крім того, комплексні рішення одного виробника на сьогодні залежні від систем керування базами даних. Це пояснюється тим, що основні виробники програмного забезпечення для ІАС прагнуть максимальної інтеграції пропонованих ними рішень. Тому бажання використовувати один або кілька інструментів змушує службу використовувати інші продукти цього постачальника, що не завжди відповідає очікуваному результату. Також зростає ризик перспектив довгострокового розвитку ІАС від одного виробника.

У процесі реалізації інформаційно-аналітичної системи в аналітичних службах на основі змішаного варіанта вибір систем, що взаємодіють, може бути здійснений за принципом приналежності до рівнів архітектури ІАС. При цьому група інструментів аналізу даних може бути незалежною

від групи інструментів витягу, перетворення, завантаження й зберігання, тобто кожна із цих груп може бути представлена окремим виробником. Інструменти другої групи доцільно вибирати від постачальників СУБД, а інструменти ділового аналізу – від постачальників, що спеціалізуються на спектрі інструментів ділового аналізу даних.

Провівши аналіз проблеми, можна зробити висновок, що аналітичні служби потребують комплексних інформаційних систем для вирішення своїх повсякденних завдань. При виборі програмних засобів для реалізації інформаційно-аналітичних систем у кожному конкретному випадку потрібно шукати якийсь збалансований варіант із залученням консультантів для оцінки техніко-економічних показників і розробки.

Список використаної літератури

1. Белов В. С. Информационно-аналитические системы. Основы проектирования и применения : учеб. пособие, руководство, практикум / В. С. Белов // Моск. госуд. ун-т экономики, статистики и информатики. – М., 2005. – 111 с.

2. Коровкин С. Д. Решение проблемы комплексного оперативного анализа информации хранилищ данных / *С. Д. Коровкин* // СУБД. – 1997. – № 5–6. – С. 47–51.

3. Кречетов Н. Продукты для интеллектуального анализа данных / Н. Кречетов // ComputerWeek-Москва. – 1997. – № 14–15. – С. 32–39.

4. Сахаров А. А. Принципы проектирования и использования многомерных баз данных (на примере Oracle Express Server) / А. А. Сахаров // СУБД. – 1996. – № 3. – С. 44–59.

5. Codd E. F. Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate / E. F. Codd // Associates, 1993, An Introduction to Multidimensional Database Technology. – Kenan Systems Corporation, 1995. – 486 p.