УДК 004.94 https://doi.org/10.15407/pp2019.02.081

О. Захарова

ВИКОРИСТАННЯ МЕТАДАНИХ ДЛЯ ВИРІШЕННЯ ЗАДАЧ ВЕЛИКИХ ДАНИХ

Насьогодні обсяги даних, якими оперують прикладні системи, експоненціально безперервно зростають та уже давно досягли таких розмірів, що не можуть оброблятися традиційними системами. Так виник термін «Великі дані». Головні проблеми таких наборів даних пов'язані, перш за все, не лише з їх об'ємом, але й з різноманітністю, різнорідністю та складністю інформації, яку вони містять. Таким чином, разом із зростанням обсягів даних і кількості ініціатив великих даних, на перший план виходять метадані, як найважливіший пріоритет успіху проектів великих даних. Підприємства усвідомлюють, що повне використання ділового та операційного потенціалу машинного навчання, глибокого навчання та штучного інтелекту вимагає, щоб необроблені дані були доповнені метаданими. Метою даної роботи є аналіз впливу метаданих на вирішення комплексу проблем великих даних, визначення основних категорій даних, що підлягають анотуванню метаданими, та основних типів метаданих, що для цього використовуються. Насьогодні метадані є засобом класифікації, впорядкування та характеристики даних або їх вмісту. Залежно від ролі, яку вони відіграють у вирішенні задач великих даних, NISO поділяє їх на чотири типи, а саме: адміністративні, описові, структурні та мови розмітки. Різні типи метаданих можуть бути використані певним чином для ефективного вирішення задач управління, пошуку, інтеграції даних тощо. Окремим питанням є способи їх створення/автоматичної генерації, тому що ручне створення метаданих є процесом досить трудомістким, а їх обсяг часто у кілька разів перевищує обсяг самих даних.

Ключові слова: аналітика великих даних, управління великими даними, метадані, анотування, машинне навчання, Hadoop, класифікація метаданих, структурні метадані, описові метадані, адміністративні метадані, інтеграція даних, онтології, зв'язані дані, семантики даних.

Вступ

Головні проблеми при використанні великих даних пов'язані з різнорідністю, різноманітністю та складністю інформації: величезні інформаційні активи; розрізнені інформаційні сховища, які можуть розповсюджуватися на декілька бізнес-одиниць; напівструктурований та неструктурований контент даних, що досить часто розподіляється за багатьма внутрішніми операційними системами, бізнес-застосунками, мережами, серверами та інтелектуальним обладнанням.

Організація ефективного управління постійно зростаючими обсягами структурованих даних та контентом неструктурованих даних підвищує конкурентоздатність підприємства. За думкою багатьох експертів, нереалізація такого управління згодом суттєво ускладнить обслуговування вимог клієнтів, що постійно змінюються, або навіть можливо буде коштувати таким підприємствам частки ринку. Треба навчитися мати користь з великих даних та використовувати їх міць для прийняття важливих майбутніх бізнес-рішень. Здатність ефективно організовувати та класифікувати інформацію забезпечить більшу інтелектуальність у бізнесі, надаючи можливості оперативного прийняття рішень.

Таким чином, разом із зростанням обсягів даних та кількості ініціатив великих даних, на перший план виходить цінність метаданих [1]. Метадані стають найважливішим пріоритетом успіху великих даних, та їх вплив не можна недооцінювати. Важливість метаданих [2] зростає ще й в наслідок того, що підприємства усвідомлюють, що повне використання ділового та операційного потенціалу машинного навчання, глибокого навчання та штучного інтелекту вимагає, щоб необроблені дані були доповнені метаданими. Зростання обсягів фактичних даних обумовлює існування ще більшої кількості даних, або метаданих, про використання та джерела цих фактичних даних.

Роль метаданих у вирішенні проблем великих даних

При впровадженні кожного нового проекта великих даних повинна бути можливість ідентифікувати ці великі дані. Важ-

ливою можливістю для розробки та розвитку сервісів обробки великих даних є створення комплексної програми управління метаданими підприємства.

Згідно звіту, що був опублікований International Data Corporation (IDC), метадані є одним з найшвидше зростаючих підсегментів управління корпоративними даними. Однак, хоча метадані швидко зростають, вони все одно не встигають за швидким зростанням проектів великих даних. Цю проблему IDC називає як «прогалина великих даних» (“Big Data Gap”) [3]. Метадані можуть значно спростити та вдосконалити процеси збору, інтеграції та аналізу джерел великих даних. За відсутністю метаданих підприємства можуть втратити глибоке розуміння того, що саме можуть дати великі дані. Метадані можуть керувати всім життєвим циклом даних, процесами, процедурами, а також клієнтами або користувачами, які впливають на певну бізнес-інформацію. Метадані великих даних є основою для збору величезних обсягів даних з нових розрізнених джерел та інформаційних сховищ, перш ніж вони стануть некерованими.

Метадані, «маленькі» та великі дані

Метадані – це інформація, яка описує інші дані – «дані про дані» [3]. Це можуть бути описові, адміністративні та структурні дані. Це просте визначення використовувалося спеціалістами з обробки даних на протязі десятиріч. Тим не менш, метадані визначають атрибути, властивості та теги, які будуть описувати та класифікувати інформацію. Доречніше визначити метадані як «інформацію про дані». Метадані можуть бути представлені у вигляді будь-якої кількості характеристик, що пов’язані з цінними інформаційними даними, такими як тип даних, автор, дата створення, стан робочого процесу та використання в межах підприємства. Після того, як метадані визначені, вони забезпечують цінність та призначення вмісту даних та, таким чином, стають ефективним інструментом для швидкого пошуку інформації – обов’язковою умовою для аналізу великих даних та формування звітності. Але, метадані також можуть ідентифікувати «маленькі дані», які у кінцевому підсумку забезпечують структуру того, що стає великими даними. У роботі [4] (Harvard Business Review) були визначені три головні відмінності великих та невеликих даних:

- «Великі дані» спрямовані на просунення організаційних цілей, а «Маленькі дані» допомагають людям досягти особистих цілей.
- Окремі індивіди не можуть бачити великі дані.
- Великі дані контролюються організаціями, а маленькі дані контролюються приватними особами. Компанії надають фізичним особам дозвіл на доступ до великих даних, водночас як фізичні особи, навпаки, надають організаціям доступ до маленьких даних.

Щоб усвідомити істинне значення, яке метадані вносять у великі дані, треба поглянути на визначення структури, яке допомагає знаходити дані в процесі вирішення задачі виявлення даних, а також інтерпретувати та використовувати великі дані точно і правильно.

Для структурованих даних модель метаданих є «рідною» за структурою. Структуровані джерела даних можуть забезпечити логічну структуру через легко отримувані метадані. Але у великих даних немає такої доступності «власних» метаданих, тому для розкриття їх значення використовуються метадані з зовнішніх джерел даних. Великі дані потребують обробки за допомогою певної аналітики, для побудови нових визначень метаданих. Наприклад, при використанні Hadoop для збору даних не потрібно визначати метадані під час збору даних, необхідно просто визначити унікальний ключ, щоб мати можливість звертатися до даних у випадку необхідності. Однак, у кінцевому підсумку все одно треба буде визначити метадані, й Hadoop використовує для цього HCatalog. Після того, як метадані визначені, вони можуть бути співвіднесені з метаданими, що визначені в інших традиційних (структурованих) джерелах даних, забезпечуючи загальну модель метаданих для всієї організації.

Метадані можуть зв'язувати важливу інформацію організації, зв'язуючи відповідні критерії. Це дозволяє об'єднувати зв'язками аналогічні набори даних та, з іншого боку, роз'єднувати різнорідні набори даних різних джерел великих даних. Включення для великих даних атрибутів метаданих, які мають важливе значення, до напівструктурованих даних та неструктурованого контенту робить ці набори даних більш значущими, в результаті чого непотрібна інформація може бути відхилена в процесі пошуку.

Метадані забезпечують точнішу картину даних у межах всього підприємства в цілому та належний рівень погодженості даних для аналітики великих даних та бізнес-застосунків, а також керують багатьма аспектами діяльності компаній.

Так, наприклад, дослідження Стенфордського університету показали, що метадані телефонних дзвінків розкривають значні обсяги особистої інформації без доступу до реальних голосових записів. Аналіз графіків метаданих телефонних дзвінків може виявити частоту, актуальність, силу та характер взаємовідносин між людьми.

Amazon збирає метадані з продаж та в подальшому використовує їх для надання рекомендацій клієнтам та поліпшення їх взаємодії з постачальниками. Amazon також створює метадані про продукти, що доступні іншим сайтам, які надбудовують на ньому власні сервіси, збільшуючи, тим самим, обсяги продажу через Amazon.

В основі соціальних мереж також лежать метадані. Користувачі Facebook створюють метадані, коли керують списками друзів, постять статуси чи додають описи до медіа-файлів та «лайки» до статусів друзів, розділяють раніше викладений контент і надають оригінальне відео. Відстежуючи ці дії Facebook аналізує популярні теми та просуває рекламу, яка приносить користь. Згенеровані метадані використовуються для побудови пошукового індексу та рекомендацій для контенту, що є цікавим для користувачів.

Користувачі Instagram визначають підписи до зображень, які вони викладають та до яких розділяють доступ з обліковими записами інших користувачів. Instagram використовує ці дані для вдосконалення реклами. Користувачі Twitter організують людей, яким вони слідують, у списки, постять текст та відео, використовують хештеги для визначення коментарів у твіті та пов'язують їх один з одним, ретвітять чужий контент з або без коментарів, виділяють "улюблені" твіти, керують властивостями, такими як «список популярних тем». Глибина даних про суспільство, що представлена контентом у Твітері, та його метадані призвели у 2010 році до заключення угоди про те, що Бібліотека Конгресу США буде архівувати цей цінний матеріал для досліджень [5].

Щоб краще уявити, що таке є метадані та чому це важливо, розглянемо невеличкий приклад [6] метаданих (рисунок), що пов'язані з твітом лише з 140 символів. 140 символів не представляють великих обсягів даних, однак обсяги даних вибухають, якщо зв'язати твіт з усіма метаданими, що необхідні для розуміння цих 140 символів у контексті розмови.

Наведений приклад демонструє елементи метаданих.

- Ім'я та ідентифікатор користувача, що відповів автору твіта.
- Дата та час створення твіту.
- Ім'я автора.
- Ім'я користувача.
- Біографія автора.
- URL автора.
- Місцезнаходження автора.
- Надання інформації для автора.
- Дата створення облікового запису.
- Кількість обраних, що має користувач.
- Кількість користувачів, на яких підписаний даний користувач.
- Часовий пояс та зміщення часу для даного користувача.
- Мова, обрана користувачем.
- Чи є користувач захищеним.
- Кількість підписників користувача.
- Ідентифікатор місця.
- Друкована назва цього місця.
- Тип місця.
- Країна.
- Застосунок, який відіслав твіт.

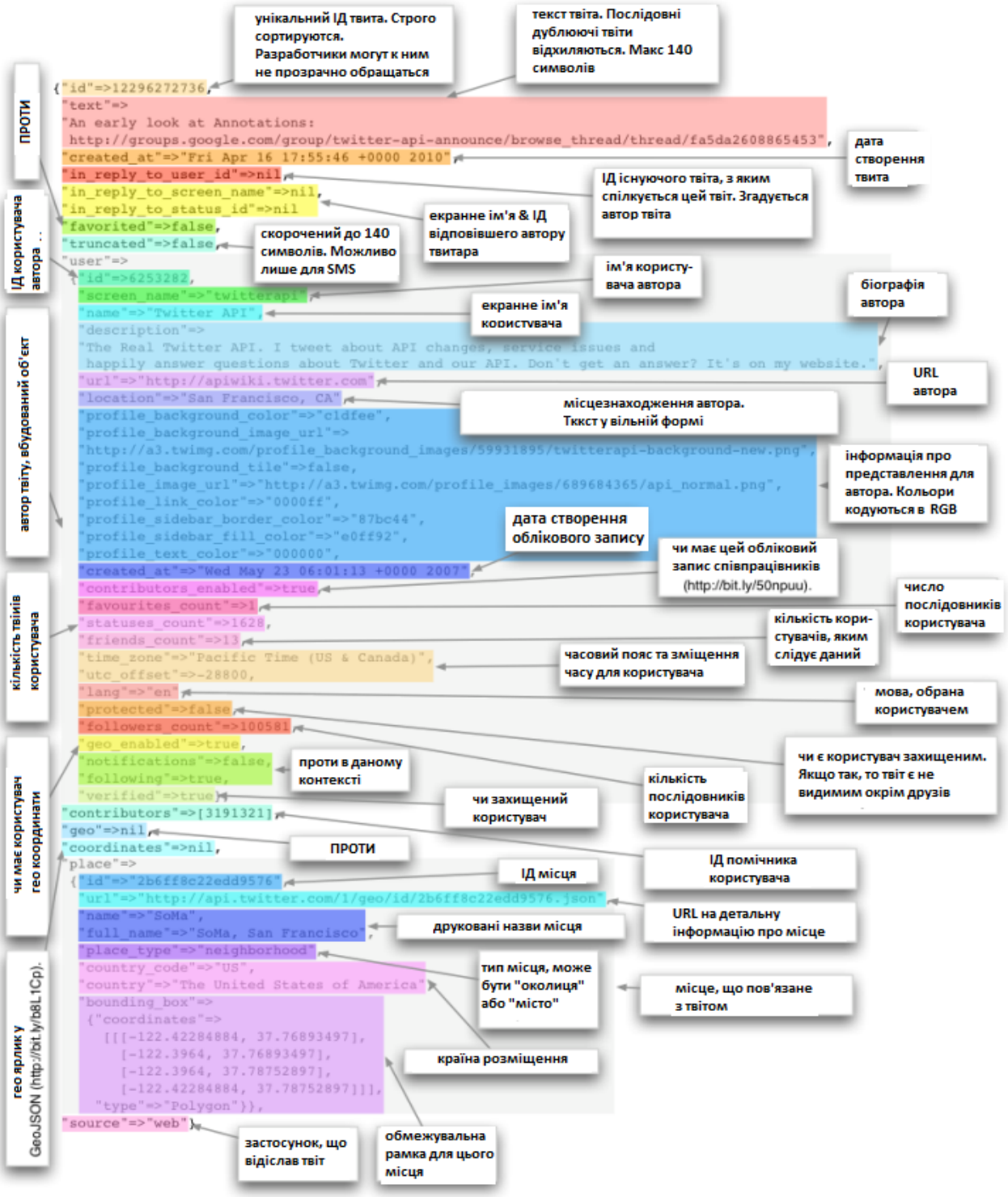

Рисунок. Метадані, пов’язані з твітом

Наведені приклади демонструють нечітку межу між метаданими та інформацією, яку вони описують. У багатьох ситуаціях ця відмінність не має значення, так як метадані часто створюються, зберігаються та обробляються так, якби це дані. Відмінністю є лише семантика.

Особливістю наведених вище прикладів є те, що всі метадані у деякій мірі структуровані. Метадані збираються таким чином, щоб вони могли виконувати корисну задачу, та сортуються за відомими категоріями. Саме це поняття структури пере-

творює необроблену інформацію у реальні метадані.

Які дані має сенс анотувати?

Очевидно, що метадані суттєво скорочують обсяги необроблених даних. Саме це відбувається, коли підприємство починає помічати (анотувати) більше своїх транзакцій та взаємодій, щоб отримати додатково уявлення про природу та контекст діалогу та взаємодії. Слід пам'ятати, що метадані при цьому стрімко збільшують обсяги даних, що зберігаються, в цілому, та не всі дані завжди є корисними для аналітики великих даних. Але деякі типи великих даних особливо підходять для аналізу, а їх анотування може суттєво вплинути на ефективне вирішення конкретних, життєво-важливих задач. Розглянемо деякі з них [6].

Відеозаписи спостережень. Загальні метадані (дата, час, місцезнаходження та ін.), як правило, автоматично прикріплюються до відеофайлу. Однак, з розповсюдженням IP-камер з'являється більше можливостей для вбудовування більшого обсягу інформації у камеру, що дозволяє аналізувати та маркувати відзнятий матеріал в режимі реального часу. Такий тип тегів може прискорити кримінальне розслідування, покращити роздрібну аналітику великих даних для моделей споживчого трафіку та покращити військову розвідку, оскільки відео з дронів у різних географічних регіонах порівнюються за кореляцією моделей (шаблонів).

Дані різного роду датчиків. У майбутньому датчики всіх типів (включаючи ті, що можуть бути імплантовані до організму людини) будуть фіксувати життєво важливі та не життєво важливі біометрії, відстежувати ефективність ліків. Це дозволить корелювати фізичну активність зі станом здоров'я, відстежувати потенційні спалахи вірусів у режимі реального часу.

Розваги та соціальні мережі. Тенденції, що засновані на великих групах людей, можуть стати відмінним джерелом великих даних, щоб допомогти вивести на ринок «наступну велику річ», допомогти обрати переможців та переможених на фондовому ринку та навіть передбачити результат виборів – все на основі інформації, яку користувачі вільно публікують через соціальні мережі.

Зображення користувачів. Люди багато розповідають про себе, коли публікують власні фото або фото сімей/друзів. Раніше картинка коштувала тисячі слів, але з появою великих даних з'явився значний множник. Ключовим моментом буде впровадження складних алгоритмів тегування, які зможуть аналізувати зображення або в режимі реального часу, коли знімки зроблені або завантажені, або в масовому порядку після їх агрегування з різних вебсайтів.

Перелічені типи даних є доповненням до звичайних ***транзакційних даних,*** що проходять через корпоративні системи в ході звичайної обробки даних.

Класифікація метаданих

Метадані є засобом класифікації, впорядкування та характеристики даних або їх вмісту. Національна організація з інформаційних стандартів (NISO – The National Information Standards Organization) [7] пропонує класифікацію [8], яка може бути застосована для всіх типів даних або репозиторіїв даних, від бібліотек до веб-сайтів, текстовых та нетекстових даних, даних у цифровій або матеріальній формі.

NISO описує наступні типи метаданих [8].

- *Описові метадані* включають таку інформацію, як, наприклад, контактні дані, заголовок або автор публікації, анотація роботи, ключові слова, географічне розміщення або навіть пояснення методології. Ці дані можуть використовуватися для виявлення, збору або групування ресурсів за загальними для них характеристиками.
- *Структурні метадані* пояснюють склад або організацію ресурсів. Наприклад, цифрову книгу можна публікувати у вигляді зображень окремих сторінок файла PDF або HTML. Ці сторінки або компоненти зазвичай групують у глави. Дані про глави, зміст або відомості про макет сторінок вважаються структурними метаданими. До структурних метаданих відносяться також такі записи, як структу-

рна карта сторінок або інших ресурсів вебсайту, подія вторгнення чи запису відомостей про голосові виклики.

– *Адміністративні метадані* використовуються для управління ресурсом. Дати створення або отримання, права доступу, права або походження, або правила утилізації, такі як зберігання чи видалення, є прикладами прав, які може застосовувати цифровий архівіст, куратор. Такі метадані є корисними для адміністратора бази даних або для адміністраторів, що відповідають за отримання даних з трафіку телекомунікаційних мереж або мереж передачі даних, або журналів систем безпеки, або даних про події.

– *Мови розмітки*. Інтегрують метадані та флаги для інших структурних чи семантичних властивостей в контенті. Змішують разом метадані та контент. Форма метаданих, що найбільш часто використовується. Вбудовані у контент флаги позначають відмічені властивості або особливості. Для текстового ресурсу це може означати маркерування структурних елементів, таких як параграфи; помітка слів семантичною інформацією – наприклад, слово є географічною назвою або певною частиною мови; також це може бути наданням інформації про форматування.

Ці різні категорії метаданих підтримують різні варіанти використання в інформаційних системах. Структурні метадані найчастіше використовуються при вирішенні задачі пошуку (виявлення) даних, так як вони дозволяють користувачам шукати або переглядати ресурси та інформацію, що їх цікавлять. Багато з властивостей метаданих доцільно відображати користувачам, щоб допомогти їм в ідентифікації або розумінні ресурсу. Вирішення задач інтероперабельності даних, ефективного обміну контентом між системами базується на метаданих, які описують цей контент. Залучені до операції системи можуть ефективно профілювати матеріал, що надходить, та співставляти його зі своїми внутрішніми структурами. Метадані підтримують управління цифровими об'єктами, надаючи інформацію, що є необхідною для належного відтворення цифрового контенту або надання відповідної версії, що задовольняє вимоги користувача.

Задача збереження великих даних може вирішуватися шляхом створення метаданих, які дозволяють перевіряти цілісність контенту після передачі даних та в інших важливих точках, а також сигналізувати у випадку виникнення необхідності вжиття заходів, що пов'язані зі збереженням даних, наприклад таких, як перенесення формату або перевірка цілісності. Нарешті, метадані підтримують навігацію частинами елементів, наприклад, від однієї сторінки або розділу до іншого, та між різними версіями об'єкта, такими як фотографічні зображення різної якості.

Таблиця далі наводить приклади елементів метаданих різних типів та варіантів їх використання в задачах великих даних.

Таблиця

Тип метаданих	**Приклади елементів метаданих**	**Варіанти використання**
Описові	Заголовок Автор Тема Жанр Дата публікації	Виявлення Візуалізація Інтероперабельність
Адміністративні	Технічні: Тип файлу Розмір файлу Дата/час створення Схема стиснення	Інтероперабельність Управління цифровими об'єктами Зберігання
	Метадані зберігання: Контрольна сума Подія збереження	Інтероперабельність Управління цифровими об'єктами Зберігання
	Правові: Статус авторського права Умови ліцензування Правовласник	Інтероперабельність Управління цифровими об'єктами
Структурні	Послідовність Місце в ієрархії	Навігація
Мови розмітки	Параграф Заголовок Список Назва Дата	Навігація Інтероперабельність

Використання метаданих для пошуку даних

Використання метаданих, що прив'язані до алгоритмів пошуку, дозволяють отримувати результати пошуку з високим ступенем достовірності. Це особливо важливо в ініціативах «Великих даних», де автономні результати, що базуються на ключових словах, можуть включати в себе скопичення менш актуальної інформації. Але використання асоціацій метаданих, створює можливість користувачам та аналітикам великих даних швидко знаходити потрібну інформацію, незважаючи на великий контент, який знаходиться у розрізнених репозиторіях.

Даний підхід може бути розповсюджений як на пошук структурованих даних, так й на пошук неструктурованого контенту репозиторіїв підприємств. Метадані можуть зв'язувати весь контент, що пов'язаний з одним або декількома атрибутами метаданих, незалежно від їх місцезнаходження або формату. Як варіант, метадані можуть надавати інформацію про елемент даних (наприклад, продукт), яка однозначно описує цей елемент. Таке поле, як ідентифікатор продукту, також є засобом для зв'язку з іншими джерелами даних з метою інтеграції даних. Окрім цього, за допомогою дескрипторів метаданих можна зв'язувати елементи даних у загальних термінах та використовувати ці метадані для інтеграції та кращого розуміння розрізнених джерел великих даних. Даний підхід надає метадані послідовно на рівні підприємства.

Дуже важливо, щоб метадані дозволяли створювати та підтримувати погодженість даних. Так, наприклад, компанії порізному визначають термін «клієнт». Та якщо компанія має розрізнені сховища даних та розкидані бізнес-одиниці, то цей термін може бути легко неправильно витлумачений по всьому підприємству і, таким чином, оцінений по-різному. Навіть, якщо кожне джерело даних визначене правильно, контекст одного й того самого елемента даних може змінюватися в різних областях застусування. Ця проблема існує в більшості організацій та, якщо її не вирішити, впливає на цілістність звітів та результати пошуку на підприємстві в цілому.

Існує два підходи до вирішення цієї проблеми:

1) перейменувати або помітити терміни застосунків, щоб вони були більш конкретними, або

2) згорнути ці імена застосунків у більш абстрактне ім'я на рівні сектора або навіть підприємства.

Саме тут буде надзвичайно корисним сховище метаданих. Адмініструючи метадані, підприємство може побудувати несуперечливе визначення або бізнес-правило для конкретного атрибуту даних та застосовувати його на рівні даних підприємства, як для структурованих, так й для неструктурованих сховищ даних.

Управління метаданими даних

Управління метаданими повинно бути частиною загальної практики управління даними підприємства. Це важливий компонент будь-якої надійної практики управління даними. Підхід, який це підтримує, полягає у встановленні управління даними для метаданих [3]. Надійні метадані забезпечують погодженість даних для підтримки підприємства та забезпечують прийняття рішень відносно аналізу великих даних. Реалізація практики управління корпоративними даними надає користувачам даних їх значення і контекст для розуміння даних та їх компонентів. Обов'язки управляючого метаданими даних включають документування контексту контенту даних (походження та спадщина даних), а також визначення даних для сутностей й атрибутів сховища даних, ідентифікацію зв'язків між даними та забезпечення перевірки своєчасності, точності та повноти даних. Підтримка належного управління метаданими сприяє успіху ініціативи великих даних та забезпечує повну реалізацію бізнес-значимості даних підприємства.

Зі зростанням обсягів використання великих даних, з'являються і будуть з'являтися надалі нові типи метаданих, які відповідають особливим вимогам різних сегментів ринку, які надають великі дані. Реалізація підходів, що засновані на метаданих, а також програми управління для

підтримки як структурованих так й великих даних, має критичне значення у встановленні загальної погодженості даних та забезпеченні кращого розуміння взаємозв'язків даних для підприємства. Реалізація відповідних корпоративних та бізнес-ініціатив дозволить досягти більшої віддачі завдяки більш швидкому доступу до конкретного контенту даних, який знаходиться у великих різноманітних сховищах великих даних, «озерах даних» [9], а також репозиторіях реляційних баз даних, що мають першорядне значення для бізнесу. Для успішного вирішення задач великих даних ключову роль відіграє вирішення проблеми їх інтеграції, в першу чергу, семантичної. Також, слід зазначити, що для управління метаданими в масштабі підприємства, необхідне створення та впровадження репозиторію метаданих.

Інтеграція даних

Розробку метаданих для різного роду прикладної інформації (біологічної, медичної та ін.) на основі стандартів семантичного веб можна розглядати як перспективний підхід для семантичної інтеграції інформації. З іншого боку, онтології, як формальні моделі для представлення інформації з чітко визначеними поняттями й взаємозв'язками між ними, можуть використовуватись для вирішення проблеми неоднорідності у джерелах даних.

Швидкий розвиток та впровадження онтологій у прикладних доменах спонукало дослідницьке товариство використовувати їх для інтеграції даних та інформації. Від моменту виникнення зв'язаних даних вони стали важливою технологією для досліджень у галузі семантики й онтологій. Багато проблем зв'язаних даних подібні проблемам великих даних, тому їх можна розглядати як частину крупнішого ландшафту великих даних. Зв'язуючий компонент зв'язаних даних приділяє особливу увагу інтеграції та об'єднанню даних у декількох джерелах. Таким чином, можна виділити три основні парадигми семантичного веб, які відіграють важливу роль у вирішенні проблем великих даних [10]: семантики, онтології та зв'язані дані.

Використання семантик

Семантичний Веб просуває стандарт анотування та інтеграції даних. Мета полягає у тому, щоб, заохочуючи включення семантичного контенту у дані, що доступні через Інтернет, перетворити існуючу мережу, де переважають неструктуровані та напівструктуровані документи, у мережу даних. Це охоплює публікацію інформації на мовах, які спеціально розроблені для даних, таких як: Resource Description Framework (RDF), Web Ontology Language (OWL), SPARQL [11] (протокол та мова запитів для джерел даних семантичного веб), і Extensible Markup Language (XML) [11]. RDF підтримує модель представлення метаданих та визначає дані трійками суб'єкт- предикат-об'єкт, що відомі як "оператори." Таке представлення гнучко пов'язує дані частинами та за принципом «посилання-за-посиланням», формуючи спрямований розмічений граф. Завдяки цьому дані та метадані тісно пов'язані один з одним, що значно спрощує пошуковий запит. Компоненти кожного RDF оператора можуть бути ідентифіковані за допомогою URI (Uniform Resource Identifiers). Окрім цього, на них можна посилатися через RDF схеми (RDFS) [12], мову веб онтологій (OWL) або інші (що не відносяться до схеми) RDF документи. Зокрема, OWL є родиною мов представлення знань для створення онтологій або баз знань. Мови характеризуються формальними семантиками та серіалізаціями на основі RDF/XML для семантичного веб. Використання інформації RDF ресурсів може здійснюватися за допомогою SPARQL (SPARQL Protocol та RDF Query Language), які здатні виявляти та обробляти дані, які зберігаються у RDF-форматі.

Онтології

Рівень онтології також має величезне значення для підтримки інтеграції даних [13]. Онтології дозволяють відображати відношення між даними, що зберігаються у базі даних. Вони надають формальне представлення набора понять

через описи базових об'єктів, класів, атрибутів та відношень. Завдяки формі представлення у вигляді дерева, онтології дозволяють зв'язувати терміни одного домена, навіть, якщо вони належать різним джерелам, в контексті інтеграції даних та ефективно співставляти різноманітні й віддалені сутності. Це дозволяє не лише покращити інтеграцію даних, але й спростити пошук інформації, а також здійснювати пошук на різних рівнях деталізації. Так, наприклад, прямий запит до терміну «рак» [14] у біомедичному контексті поверне лише дане слово у всіх входженнях, що знайдені в ресурсі. Але використовуючи специфічну онтологію (наприклад, онтологію хвороб людини – DOID [15]), результат виконання запиту буде багатшим, та включатиме такі терміни як саркома та рак, які в іншому випадку просто не будуть знайдені. Інтеграція даних на основі онтологій включає використання онтологій для ефективного об'єднання даних або інформації з декількох різнорідних джерел. Ефективність такої інтеграції багато в чому визначається погодженістю та виразністю онтологій.

Дані, що анотовані за допомогою онтологій, можуть бути інтегровані з іншими наборами даних, забезпечуючи підтримку семантик для виконання запитів. Публікація таких наборів даних як RDF, разом з їх онтологіями, забезпечує як синтаксичну так і семантичну інтеграцію даних, що була давно обіцяна технологіями семантичного веб.

Зв'язані дані

Ще одним прогресивним підходом до роботи з великими даними є парадигма зв'язаних даних [16]. Даний підхід просуває принципи гіпертексту з мережі документів у мережу «багатих» даних. Зв'язані дані описують метод для публікації структурованих даних таким чином, щоб вони були пов'язані один з одним. Це робить їх взаємозалежності більш ясними. Ця технологія заснована на семантичних веб-технологіях (зокрема, може використовувати HTTP, RDF та URI), але розширює їх для загального користування таким чином, щоб дані могли автоматично читатися IT системами.

Ідея полягає в тому, що після розміщення даних в інтернеті та визначення їх структури у машинночитаємому вигляді, їх необхідно анотувати метаданими з відкритими стандартами від W3C (наприклад, RDF та SPARQL). Це дозволить користувачам зв'язувати дані для надання контекстної інформації. Даний підхід дозволяє створювати явні зв'язки між наборами даних, використовуючи розподілені семантики зі стандартних онтологій та словників, сприяючи збільшенню ступеня інтеграції даних.

Створення метаданих

Описові метадані, як правило, створюються людиною в ручну. Деякі є просто фіксацією властивостей (назва, автор, дата видання), а деякі, такі як, наприклад, фонова інформація про автора, історія створення та ін., вимагають додаткових досліджень і лише після цього оформлення результатів. Інтерпретуюча інформація потребує залучення експертів. Спочатку технології просто дозволяли обмінюватися цими створеними в ручну метаданими. Потім почали з'являтися спеціалізовані системи введення метаданих, а потім для прискорення процесу створення метаданих стали з'являтися такі інструменти, як електронні таблиці. Останнім часом інтерфейси створення метаданих стають більш складними та зручними для користувача. Сьогодні метадані зазвичай створюються за допомогою проміжних кроків, а не безпосередньо за допомогою XML чи RDF [8]. Єдиним виключенням з цього принципу є розмітка контенту метаданими. З'являються надійні методи автоматизованого створення метаданих. Зокрема, більшість форматів файлів включають, хоча б де-яку вбудовану технічну інформацію, яка може допомогти програмному забезпеченню в інтерпретації вмісту. Також використовується системна інформація для додавання додаткової адміністративної інформації до цифрових файлів, такої як дата створення або ідентифікатор користувача, який увійшов до системи під час створення файлу. Мережеві технології та більш широка інте-

грація програмних систем полегшують реалізацію ефективного обміну метаданими, що, у свою чергу, скорочує зусилля, що витрачаються на їх дублювання.

Останнім часом з'явилися процеси для аналізу цифрового контенту та автоматичної генерації метаданих про нього. Автоматична транскрипція мови з аудіо та відео зараз є вже відносно зрілою технологією, особливо для записів, що відзняті у контролюємому середовищі з виділеними звуковими системами. Швидко розвивається технологія розпізнавання осіб для відео та нерухомих зображень. Відносно текстових ресурсів, то схований семантичний аналіз та тематичне моделювання дозволяють напівконтрольовано генерувати теми, які відносяться до текстів, що аналізуються. У дослідницьких середовищах часто використовуються технології розпізнавання частини мови й іменованих об'єктів. Розвивається автоматичне анотування зображень, що використовує алгоритми для ідентифікації об'єктів на фотографіях. Виконуються роботи з обробки сигналів для аудіофайлів, які охоплюють також створення списків відтворення в он-лайн сервісах потокової передачі музики та автоматичну класифікацію жанрів для музичних записів. Спостерігається реальний прогрес та можливості отримання високоякісних даних за допомогою програмних засобів.

Висновки

Існуючі дослідження довели, що метадані мають вирішальне значення як для загальної успішності проекту Big Data, так й для організації архітектури даних будь-якого великого підприємства.

Але, слід зазначити, що метадані суттєво збільшують і так великий об'єм даних, тому необхідно чітко розуміти, які саме дані має сенс анотувати метаданими. Інший важливий момент - метадані мають сенс, якщо вони є зрозумілими прикладним системам та людям, які їх використовують. Тут головну роль відіграють наступні моменти:

1) розробка та використання стандартів метаданих;

2) розробка та використання для створення метаданих розвинених засобів проектування, що забезпечують можливості ефективного співставлення даних;

3) окрім цього, як було зазначено, для управління метаданими в масштабі підприємства, необхідне створення та впровадження репозиторію, або сховища метаданих [6]. Насьогодні існує три підходи до його створення. Найбільш широко використовуваним нині є центральний репозиторій метаданих. Даний підхід забезпечує керовану масштабованість для захоплення нових метаданих та доступ з високою продуктивністю. Другий підхід – розподілене сховище метаданих. Воно розвивалося роками, особливо для підприємств, які мають децентралізовані бізнес-одиниці. Це дозволяє користувачам отримувати метадані з усіх сховищ у режимі реального часу. Та, нарешті, гібридний підхід використовує характеристики двох попередніх. Він підтримує доступ у режимі реального часу з інших репозиторіїв, а також забезпечує центральне джерело для підтримки визначень метаданих всього підприємства в цілому. Однак гібридний підхід розвивається досить повільно. При реалізації будь-якого з перелічених підходів потрібно враховувати семантичну інтеграцію. Незалежно від обраного підходу необхідно зв'язати різноманітний контент великих даних з самою інформацією та точно погодити правила, для яких цей контент інтерпретується. Якщо вдасться створити сховище метаданих підприємства та довести його до певного рівня зрілості, воно зможе забезпечити конкретні переваги, а саме можливість всебічного відстеження, логічні та фізичні визначення і зв'язки, міжпідприємницькі бізнес-терміни, моделі процесів, а також елементи моделі даних. Однак для інтеграції таких конструкцій метаданих необхідні спеціальні навички, та знайти потрібних спеціалістів є проблемою, яку треба вирішувати при первинному впровадженні сховища метаданих підприємства.

Література

1. https://whatis.techtarget.com/definition/metadata

2. https://www.gartner.com/doc/3075917/reasons-big-data-needs-metadata
3. https://www.datasciencecentral.com/profiles/blogs/why-you-need-metadata-for-big-data-success
4. https://hbr.org/2013/05/little-data-makes-big-data-mor
5. hts://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/
6. https://www.datasciencecentral.com/profiles/blogs/importance-of-metadata-in-a-big-data-world
7. http://framework.niso.org/24.html
8. https://groups.niso.org/apps/group_public/download.php/17443/understanding-metadata
9. https://www.i-scoop.eu/big-data-action-value-context/data-lakes/
10. https://groups.niso.org/apps/group_public/download.php/17443/understanding-metadata
11. "OWL Web Ontology Language Overview," W3C Recommendation, 2004, http://www.w3.org/TR/owl-features/.
12. http://www.w3.org/TR/rdf-schema/
13. Blake J. A. and Bult C. J. "Beyond the data deluge: data integration and bio-ontologies". Journal of Biomedical Informatics. 2006. Vol. 39, N 3. P. 314–320, View at Publisher · View at Google Scholar · View at Scopus.
14. Viti F., Merelli I., Calabria A. et al., "Ontology-based resources for bioinformatics analysis," International Journal of Metadata, Semantics and Ontologies. 2011. Vol. 6, N 1. P. 35–45. View at Publisher · View at Google Scholar · View at Scopus.
15. Osborne J. D., Flatow J., Holko M. et al. "Annotating the human genome with disease ontology," BMC Genomics. 2009. Vol. 10, supplement 1, article S6. View at Publisher·View at Google Scholar·View at Scopus.
16. https://www.w3.org/DesignIssues/LinkedData.html

References

1. https://whatis.techtarget.com/definition/metadata
2. https://www.gartner.com/doc/3075917/reasons-big-data-needs-metadata
3. https://www.datasciencecentral.com/profiles/blogs/why-you-need-metadata-for-big-data-success
4. https://hbr.org/2013/05/little-data-makes-big-data-mor
5. hts://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/
6. https://www.datasciencecentral.com/profiles/blogs/importance-of-metadata-in-a-big-data-world
7. http://framework.niso.org/24.html
8. https://groups.niso.org/apps/group_public/download.php/17443/understanding-metadata
9. https://www.i-scoop.eu/big-data-action-value-context/data-lakes/
10. https://groups.niso.org/apps/group_public/download.php/17443/understanding-metadata
11. "OWL Web Ontology Language Overview," W3C Recommendation, 2004, http://www.w3.org/TR/owl-features/.
12. http://www.w3.org/TR/rdf-schema/
13. Blake J. A. and Bult C. J. "Beyond the data deluge: data integration and bio-ontologies," Journal of Biomedical Informatics. 2006. Vol. 39, N 3. P. 314–320, View at Publisher·View at Google Scholar·View at Scopus.
14. Viti F., Merelli I., Calabria A. et al. "Ontology-based resources for bioinformatics analysis," International Journal of Metadata, Semantics and Ontologies 2011. Vol. 6, N 1. P. 35–45. View at Publisher·View at Google Scholar·View at Scopus.
15. Osborne J. D., Flatow J., Holko M. et al. "Annotating the human genome with disease ontology," BMC Genomics. 2009. Vol. 10, supplement 1, article S6, View at Publisher·View at Google Scholar·View at Scopus.
16. https://www.w3.org/DesignIssues/LinkedData.html

Одержано 19.02.2019

Про автора:

Захарова Ольга Вікторівна,
кандидат технічних наук,
старший науковий співробітник.
Кількість наукових публікацій в українських виданнях – 28.
http://orcid.org/0000-0002-9579-2973.

Місце роботи автора:

Інститут програмних систем
НАН України,
проспект Академіка Глушкова, 40.
Тел.: 526 5139.
E-mail: ozakharova68@gmail.com.
Моб. тел.: +38(068)5947560.