BIG DATA – НОВЕ ЯВИЩЕ В ЗБЕРІГАННІ ТА АНАЛІЗІ ДАНИХ: ПЕРСПЕКТИВИ ТА МОЖЛИВОСТІ

О.Л.Єршова, канд. екон. наук, доцент, в. о. завідувача кафедри інформаційних систем і технологій, Національна академія статистики, обліку та аудиту Державної служби статистики України

П.С.Єршов, студент магістратури НТУУ КПІ, Python Django Developer KIT XXI

Дається короткий огляд труднощів, пов'язаних з Big Data, технологій та підходів для подолання труднощів отримання значимої інформації з Big Data. Дійсно, створення та підтримка сховищ об'ємом в терабайт, петабайт і більше стало можливим завдяки технологіям розподілених файлових систем. У розподілених системах, замість зберігання даних в одній файловій системі, дані зберігаються і індексуються на декількох (і навіть тисячах) жорстких дисках і серверах. Створюється також «карта» (map), де міститься інформація про те, де саме знаходяться ті чи інші дані. З точки зору реалізації, аналітична платформа для роботи з Big Data повинна вміти використовувати нові технології map-reduce.

"Big data" - сучасний термін, що фігурує майже на всіх професійних конференціях, присвячених аналізу даних, прогностичній аналітиці, інтелектуальному аналізу даних (data mining), CRM. Термін використовується у сферах, де актуальна робота з великими обсягами даних, де постійно відбувається збільшення швидкості потоку даних в організаційних процесах: економіці, банківської діяльності, виробництві, маркетингу, телекомунікаціях, веб-аналітиці, медицині та ін.

Разом зі стрімким накопиченням інформації швидкими темпами розвиваються і технології аналізу даних. У той же час, існують ситуації, коли захоплення новими технологіями може призвести і до розчарування. Наприклад, іноді розріджені дані (Sparse data), що дають важливе розуміння дійсності, є набагато ціннішими, ніж Великі дані (Big Data), що описують потоки, найчастіше, не суттєвої інформації.

У сучасних обговореннях поняття Big Data описують як дані обсягу в порядках терабайт.

На практиці (якщо мова йде про гігабайти або терабайти), такі дані легко зберігати і керувати ними за допомогою «традиційних» баз даних і стандартного устаткування (сервера баз даних).

Як правило, обговорення Big Data зосереджено навколо сховищ даних (і проведенні аналізу, заснованого на таких сховищах), обсягом набагато більше, ніж просто кілька терабайт. Існують галузі, де дані збираються і накопичуються дуже інтенсивно.

Крім того, за останні кілька років, впроваджуються так звані "smart grid" технології, що дозволяють комунальним службам вимірювати споживання електроенергії окремми сім'ями кожну хвилину або щосекунди.

Для такого роду додатків, в яких дані повинні зберігатися роками, накопичені дані класифікуються як Extremely Big Data [1]. Зростає і кількість додатків Big Data серед комерційних і державних секторів, де обсяг даних у сховищах, може становити сотні терабайт або петабайт.

Сучасні технології дозволяють «відстежувати» людей і їх поведінку різними способами. Наприклад, коли ми користуємося Інтернетом, робимо покупки в Інтернет-магазинах або великих мережах магазинів, таких як Walmart (згідно Вікіпедії, сховище даних Walmart оцінюється більш ніж в 2 петабайт), або пересуваємося з включеними мобільними телефонами - ми залишаємо слід наших дій, що призводить до накопичення нової інформації.

Аналогічним чином, сучасні медичні технології генерують великі обсяги даних, що відносяться до надання медичної допомоги (зображення, відео, моніторинг у реальному часі).

Існують три типи завдань пов'язаних з Big Data:

1. Зберігання і управління - обсяг даних в сотні терабайт або петабайт не дозволяє легко зберігати і управляти ними за допомогою традиційних реляційних баз даних.

Big Data зазвичай зберігаються і організовуються в розподілених файлових системах. У загальних рисах, інформація зберігається на декількох (іноді тисячах) жорстких дисках, на стандартних комп'ютерах. Так звана «карта» (map) відстежує, де (на якому комп'ютері та / або диску) зберігається конкретна частина інформації.

Для забезпечення відмовостійкості та надійності, кожну частину інформації звичайно зберігають кілька разів, наприклад - тричі.

2. Неструктурована інформація - більшість всіх даних Big Data є неструктурованими і не однотипними.

Велика частина зібраної інформації в розподіленій файловій системі складається з неструктурованих даних, таких як текст, зображення, фотографії або відео. Це має свої переваги і недоліки.

Перевага полягає в тому, що можливість зберігання великих даних дозволяє зберігати "всі дані", не турбуючись про те, яка частина даних актуальна для подальшого аналізу та прийняття рішення (N = «всі») [2].

Недоліком є те, що в таких випадках для здобуття корисної інформації потрібна подальша обробка цих величезних масивів даних.

3. Аналіз Big Data - як аналізувати неструктуровану інформацію? Як на основі Big Data складати прості звіти, будувати і впроваджувати поглиблені прогностичні моделі?

При аналізі сотні терабайт або петабайт даних, не представляється можливим витягнути дані в яке-небудь інше місце для аналізу (наприклад, в STATISTICA Enterprise Analysis Server).

Процес перенесення даних по каналах на окремий сервер або сервера (для паралельної обробки) займе дуже багато часу і вимагає занадто великого трафіку. Замість цього, аналітичні обчислення повинні бути виконані фізично близько до місця, де зберігаються дані.

Алгоритм Map-reduce являє собою модель для розподілених обчислень. Принцип його роботи полягає в наступному: відбувається розподіл вхідних даних на робочі вузли (individual nodes) розподіленої файлової системи для попередньої обробки (map-крок) і, потім, згортка (об'єднання) вже попередньо оброблених даних (reduce-крок).

Таким чином, скажімо, для обчислення підсумкової суми, алгоритм буде паралельно обчислювати проміжні суми в кожному з вузлів розподіленої файлової системи, і потім підсумовувати ці проміжні значення.

Важливо, що, незважаючи на те, що набори даних можуть бути дуже великими, інформація, що міститься в них, має значно меншу розмірність.

Наприклад, у той час як дані накопичуються щосекунди або щохвилини, багато параметрів (температура газів і печей, потоки, положення заслінок і т.д.) залишаються стабільними на великих інтервалах часу. Інакше кажучи, дані, що записуються кожну секунду, є в основному повтореннями однієї і тієї ж інформації. Наприклад, StatSoft брав участь у проектах, пов'язаних з аналізом текстів (text mining) з твітів, що відображають, наскільки пасажири задоволені авіакомпаніями і їх послугами.

Критика Big Data: зберігання Big Data не завжди приводить до отримання вигоди, швидкість оновлення даних і «актуальний» часовий інтервал не завжди розумно порівнянні.

З точки зору реалізації, аналітична платформа для роботи з Big Data повинна вміти використовувати нові технології map-reduce. Платформа STATISTICA Enterprise і Decisioning надає всі можливості для ефективної роботи з Big Data, а також дозволяє управляти тисячами моделей, застосовуваних у відношенні таких даних.

Використані джерела

1. Революция Big Data: Как извлечь необходимую информацию из «Больших Дан*ных»? http://statsoft.ru/products/Enterprise/big-data.php#critics*

2. В. Майер-Шенбергер, К. Кукьер. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим. http://www.mann-ivanov-ferber.ru/books/paperbook/big-data/