Незважаючи на те, що існує декілька альтернатив BigQuery як в домені з відкритим вихідним кодом, так і у вигляді хмарних сервісів, є неможливим відтворити масштаб і продуктивність BigQuery. В першу чергу тому, що Google відмінно об'єднанує інфраструктуру з програмним забезпеченням. Рішення з відкритим вихідним кодом, такі як Apache Drill і Presto, вимагають масштабного проектування інфраструктури та постійних операційних витрат, щоб відповідати продуктивності BigQuery.

Висновок. BigQuery призначений для створення запитів структурованих і напівструктурованих даних з використанням стандартної мови SQL. Сервіс оптимізований до використання і забезпечує надзвичайно високу рентабельність. BigQuery – це повністю кероване хмарне сховище, яке не потребує додаткових витрат на експлуатацію. Підходить для інтерактивних запитів і сценаріїв використання OLAP / BI. Технології хмарної інфраструктури Google є ключовими відмінностями сервісу BigQuery в порівнянні з іншими аналогами.

Список використаної літератури:

1. *Глибоке занурення в архітектуру Google BigQuery – [Електронний ресурс]. Режим доступу: https://panoply.io/data-warehouse-guide/bigquery-architecture/*
2. *Лакшманан В., Тайджані Д. – «Google BigQuery. Все про сховищах даних, аналітику та машинне навчання» - O'Reilly, 2021. - 496 с.*

УДК 004.04

Шевчук В.Ю., *студент 1 курсу спеціальності 122 «Комп'ютерні науки»*
Бабаков Р.М., *к.т.н., доцент, доцент кафедри комп'ютерних наук та інформаційних технологій*

DATA QUALITY ЯК НЕВІД'ЄМНА ЧАСТИНА АНАЛІЗУ ДАНИХ В ПРИКЛАДНИХ ПРОЄКТАХ BIG DATA

Донецький національний університет імені Василя Стуса, м. Вінниця

В основі прикладних проєктів Big Data лежить аналіз і обробка великих об'ємів даних. Проте, перед тим як обробляти певні дані потрібно перевірити чи взагалі дані відповідають певним критеріям, саме цим і займаються спеціалісти з Data Quality.

Data Quality стосується стану якісної чи кількісної інформації. Існує багато визначень Data Quality, але дані, як правило, вважаються високоякісними, якщо вони "придатні для [їх] передбачуваного використання в операціях, прийнятті рішень та плануванні". [1] [2] Більше того, дані вважаються високоякісними, якщо вони правильно відображають реальну конструкцію, на яку вони посилаються. Крім того, крім цих визначень, із збільшенням кількості джерел

даних, питання внутрішньої узгодженості даних стає важливим, незалежно від придатності для використання для будь-якої конкретної зовнішньої мети. Погляди людей на якість даних часто можуть бути розбіжними, навіть коли обговорюється однаковий набір даних, що використовується з тією ж метою. У цьому випадку управління даними використовується для формування узгоджених визначень та стандартів якості даних. У таких випадках для забезпечення якості даних може знадобитися очищення даних, включаючи стандартизацію.

Визначити якість даних важко через багато контекстів, в яких використовуються дані, а також різну точку зору серед кінцевих користувачів, виробників та зберігачів даних. [3]

З точки зору споживача, якість даних:

- "дані, придатні для використання споживачами даних"
- дані "що відповідають або перевищують очікування споживачів"
- дані, які "відповідають вимогам передбачуваного використання"
- З точки зору бізнесу якість даних:
- дані, які "придатні для використання "за передбачуваними оперативними функціями, процесами прийняття рішень та іншими ролями" або які демонструють "відповідність встановленим стандартам", щоб забезпечити придатність до використання"
- дані, які "придатні для використання за призначенням у операціях, прийнятті рішень та плануванні"
- "здатність даних задовольняти заявленим діловим, системним та технічним вимогам підприємства"
- З точки зору стандартів якість даних:
- "ступінь, до якого набір властивих характеристик (якісних розмірів) об'єкта (даних) відповідає вимогам"
- "корисність, точність та правильність даних для їх застосування"

Можливо, у всіх цих випадках "якість даних" - це порівняння фактичного стану певного набору даних із бажаним станом, причому бажаний стан зазвичай називають "придатним для використання", "до специфікації", "" задоволення очікувань споживачів, "" без дефектів "або" відповідність вимогам ".

Виходячи з цього, щоб аналізувати безліч файлів або записів Big Data, ці інформаційні набори повинні володіти не тільки певною структурою, а й відповідати наступним характеристикам[4]:

- актуальність - відповідність даних відображають реальний стан цільового об'єкта в поточний період часу;
- об'єктивність - точність відображення даними реального стану цільового об'єкта, яка залежить від методів і процедур збору інформації, а також від щільності реєстрованих даних;
- цілісність - повнота відображення даними реального стану цільового об'єкта, яка показує, наскільки повні, безпомилкові і несуперечливі дані за

змістом і структурою (формату) зі збереженням їх правильної ідентифікації та взаємної пов'язаності;
- релевантність - відповідність даних про реальний стан цільового об'єкта і важливість справ, що характеризує можливість їх застосування з урахуванням змісту, структури і формату;
- сумісність - процедурний показник, який характеризує можливість обробляти і аналізувати дані в подальшому, не тільки в рамках поточного завдання;
- вимірність - якісні чи кількісні характеристики реального стану цільового об'єкта і кінцевий обсяг набору цифрових даних;
- керованість - можливість цільовим і осмисленим чином обробити, передати і контролювати дані про реальний стан цільового об'єкта, на основі структури і формату датасета;
- прив'язка до джерела даних - пов'язана і достовірна ідентифікація ланцюжка постачання даних, наприклад, вказівка авторства, джерела генерації та інші атрибути походження даних (Data Provenance);
- довіра до постачальника даних - оцінка одержувачем ділових якостей постачальника публічних даних як відповідального, авторитетного, організованого і відносно незалежного видавця цифрової інформації високої якості.

Виходячи з усього вищесказаного, можна зробити висновок, що Data Quality є надзвичайно важливим аспектом прикладних проєктів Big Data. Саме використання якісних даних дозволить добитися найкращих результатів роботи.

Список використаної літератури

1. Redman T. C. Data Driven: Profiting from Your Most Important Business Asset / Thomas Redman.
2. Protocol for a systematic review and qualitative synthesis of information quality frameworks in eHealth [Електронний ресурс] – Режим доступу до ресурсу: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6429947/
3. Fürber C. Data Quality Management with Semantic Technologies / Christian Fürber.
4. Показники якості публічних даних [Електронний ресурс] – Режим доступу до ресурсу: https://habr.com/ru/post/321406/.

УДК 519.6

Шпаченко Н.О., *магістр спеціальності 124 «Системний аналіз»*
Шевченко Н.Ю., *к.е.н., доцент, доцент кафедри інтелектуальних систем прийняття рішень*