

BigQuery – performance

Cele szkolenia

Szkolenie pozwala uczestnikom:

- optymalizować zapytania,
- właściwie modelować dane poprzez partycjonowanie i klastrowanie,
- analizować plany zapytań,
- wykorzystywać widoki zmaterializowane oraz cache,
- stosować optymalne metody ładowania danych źródłowych,
- umiejętnie monitorować zapytania.

Umiejętności

Dzięki szkoleniu uczestnik będzie:

- rozumieć, jak BigQuery wykonuje zapytania oraz co wpływa na ich czas i koszt,
- dobierać właściwe partycjonowanie i klastrowanie danych,
- pisać zapytania skanujące minimum danych oraz optymalizować złączenia (joins) i funkcje okna,
- diagnozować wolne zapytania oraz umiejętnie szukać wąskich gardeł (bottlenecków),
- przewidywać koszt zapytań przed ich uruchomieniem.

Profil uczestników

Szkolenie jest przeznaczone dla:

- osób odpowiedzialnych za performance (wydajność) zapytań,
- analityków danych i data scientistów,
- programistów pracujących z BigQuery lub planujących migrację danych do tego środowiska,
- specjalistów IT odpowiedzialnych za integrację, raportowanie i analizę danych,
- osób odpowiedzialnych za automatyzację przetwarzania danych.

Przygotowanie uczestników

Uczestnicy powinni:

- znać zaawansowane konstrukcje SQL,
- mieć doświadczenie w pracy z BigQuery,

- znać pojęcia związane z analityką danych i przetwarzaniem zapytań

Szczegółowy program szkolenia

1. Powtórzenie informacji o konstrukcjach SELECT

- Grupowanie danych.
- Złączenia tabel.

2. Optymalizacja zapytań

- Filtracja i ograniczanie skanowanych danych.
- Joiny:
 - kolejność JOIN-ów,
 - broadcast vs shuffle join,
 - kiedy JOIN staje się kosztowny.
- Window functions – najczęstsze pułapki.
- CTE (WITH) – zalety i wady.

3. Modelowanie danych

- Columnar storage – co to oznacza w praktyce
- Partycjonowanie
- Klastrowanie – dobieranie kolumn
- Denormalizacja vs Joiny

4. Analiza planu zapytania

- Czytanie Query Execution Plan.
- Stages, shuffle, slot time.
- Jak znaleźć bottleneck?
- Case study: „Dlaczego to zapytanie jest wolne?”.

5. Materializacja i cache

- Query cache – jak działa i kiedy się używa
- Materialized Views – ograniczenia vs korzyści
- Tabele agregacyjne

- Scheduled queries jako element performance

6. Funkcje przybliżone i skróty

- APPROX_COUNT_DISTINCT i inne APPROX*
- Kiedy dokładność < czas

7. Dane źródłowe i formaty

- Ładowanie danych: batch vs streaming
- Parquet / Avro / JSON – wpływ na performance
- Schematy, NULL-e, typy danych
- Optymalizacja historycznych danych

8. Monitoring i troubleshooting

- Wykorzystanie INFORMATION_SCHEMA.
- Audit logs.
- Wyszukiwanie najdroższych zapytań.
- Performance anti-patterns (realne przykłady z życia).

Metoda realizacji szkolenia

Szkolenie realizowane jest w formie naprzemiennie następujących po sobie mini wykładów oraz ćwiczeń praktycznych. Szkolenie łączy w sobie fachową wiedzę merytoryczną z praktycznymi przykładami jej wykorzystania w środowisku pracy.

Liczba dni, liczba godzin szkoleniowych

2 dni, 16 godzin szkoleniowych

Ścieżka rozwoju po szkoleniu

- *Security w środowisku BigQuery*
- *Analiza kosztów wykorzystania środowiska GCP*

Informacje dodatkowe o szkoleniu:

Poziom szkolenia: ☐ podstawowy ☐ średnio zaawansowany ☒ zaawansowany

Szkolenie w formie: ☒ stacjonarnej ☒ zdalnej

Język szkolenia: ☒ polski ☒ angielski

Liczebność grupy - szkolenie stacjonarne: min: max: 8

Liczebność grupy - szkolenie zdalne: min: max: 8

Wymagania techniczne:

Dostęp do środowiska GCP.