BigQuery: Anàlisi de Big Data amb SQL

Què és BigQuery ?

BigQuery és una Base de dades de Google que et permet fer servir consultes SQL (Structured Query Language) sobre conjunts de dades molt grans.

BigQuery és una Base de dades que no requereix cap instal·lació, ja que està pensada per fer-la servir com un servei. Això ens permet centrar-nos en l'anàlisi de dades i oblidar-nos de comprar, instal·lar i gestionar servidors dedicats a la nostra empresa.

Aquesta Base de dades funcionarà al núvol (Cloud) i simplement ens haurem de preocupar d'anar-li enviant les dades que anem recopilant, i com si es tractes d'una base de dades tradicional, fer-li consultes SQL.

El preu que pagarem per emmagatzemar la nostra informació és molt reduït, l'administració ens sortirà de franc, però haurem de pagar una petita quantitat per cada consulta (en el cas de Google, a partir de 1TB al mes), que és un llindar força generós per poder provar si ens interessa un sistema d'aquestes característiques.

Google ens garanteix la confidencialitat i seguretat de les nostres dades segons les polítiques de la Unió Europea. A més, tota la informació viatja sempre encriptada.

A l'inLab estem desenvolupat un projecte d'anàlisi de dades amb una tecnologia molt similar d'Amazon (un altre proveïdor de serveis al Cloud) anomenat Amazon Athena per analitzar dades provinents d'un munt de sensors.

Quina potència té ?

La potència de BigQuery es basa en tres pilars fonamentals:

- Una estructura de dades interna basada en columnes que permet l'emmagatzematge i consulta d'informació d'una forma molt eficient.

- Un ús exhaustiu del núvol, realitzant totes les consultes en paral·lel per obtenir la major velocitat. No és estrany que una única consulta estigui executant-se en paral·lel en centenars de servidors, cadascun processant una petita quantitat d'informació abans de proporcionar-nos la resposta final. D'aquesta manera, Google utilitza la seva enorme infraestructura per aconseguir temps de resposta que ens haguéssin semblat increïbles fa pocs anys.

- Tot i no ser una Base de dades convencional, el llenguatge de consulta és SQL, pel que si ja heu utilitzat SQL en el passat, és molt fàcil utilitzar BigQuery.

Com ho puc provar ?

Una manera senzilla i gratuïta de provar-ho sense haver-nos d'instal·lar ni tant sols un programa per **a** poder realitzar consultes SQL, seria donar-nos d'alta a Kaggle (https://www.kaggle.com/)

Kaggle és un lloc fantàstic per aprendre la ciència de les dades (Data Science), ja que tenim recursos per aprendre i sobretot un munt de Kernels (entorns virtuals, totalment configurats, on podem llançar comandes per analitzar dades i amb les dades ja carregades) amb anàlisi de dades des de punts de vista molt diversos i sobre una gran varietat de DataSets (conjunts de dades) públics.

Un cop ens donem d'alta, podem anar a l'opció de "DataSets" i com a "File Type" triar "BigQuery".

A data d'avui hi ha 40 DataSets on podem trobar per exemple el de Repositoris de GitHub, on les dades comprimides ocupen més de 3TB !

Si seleccionem aquest DataSet i anem a la pestanya de Kernels podem veure les diferents **a**nàlisi que han anat fent els DataScientists i quines comandes han fet servir per extreure informació d'aquestes dades.

Quan treballem amb BigQuery, cal tenir en compte que abans de fer qualsevol consulta, hem de comprovar la mida de la query abans de llançar-la. Recordem que només la BBDD dels repositoris de Github ja ocupa 3TB, i que a Kaggle, com a molt podem consultar 5TB cada 30 dies.

En general els passos per executar una primera query en un nou DataSet serien:

1. Crear un helper
2. Llistar les taules que tenim disponibles
3. Triar una de les taules i veure'n l'estructura
4. Escriure una consulta, **limitant-ne les files resultants**
5. **Estimar la mida** del que ens retornarà la consulta
6. Fer la consulta i guardar-nos-la en un DataFrame (estructura de dades) que ens permeti analitzar / visualitzar la informació

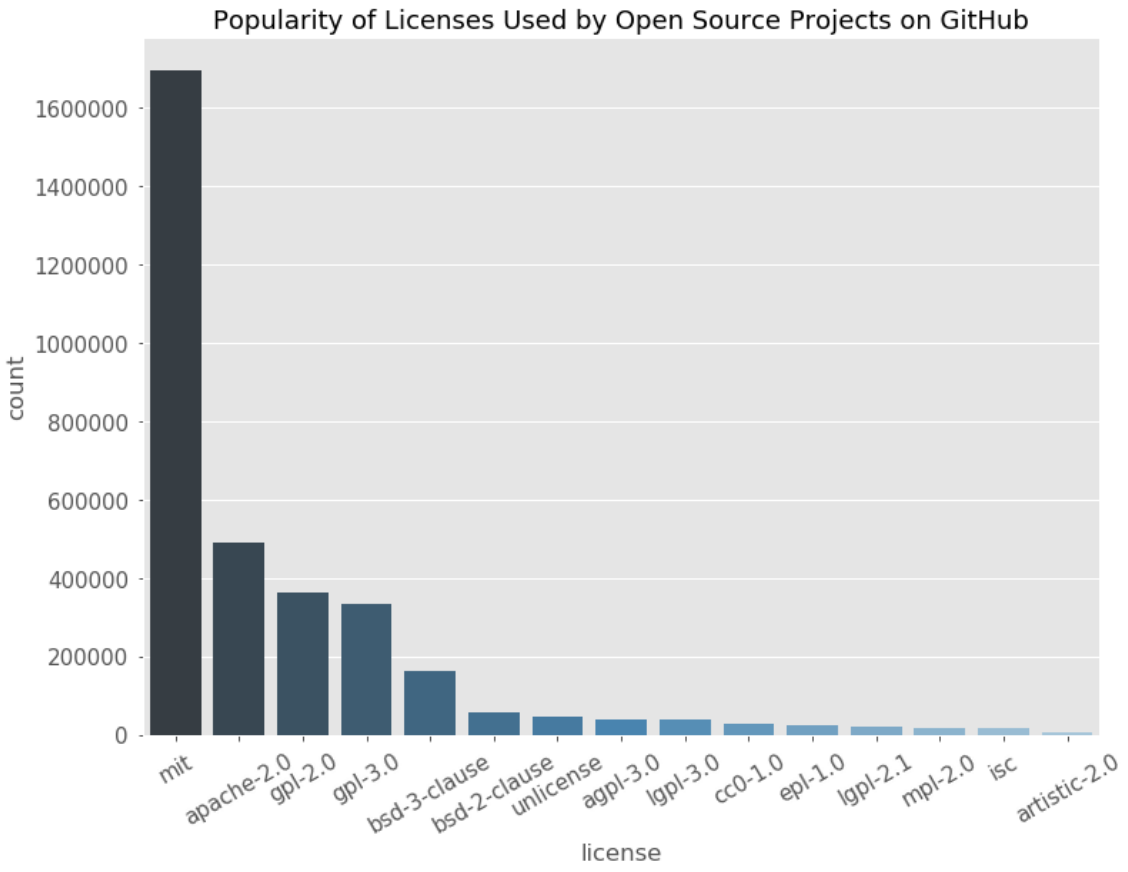
Les comandes serien:

```
# Crear un Helper
bq_assistant = BigQueryHelper("bigquery-public-data",
"github_repos")


# Llistat les taules que tenim al DataSet
bq_assistant.list_tables()

# Mirar quins atributs te una certa taula
bq_assistant.table_schema("licenses")
```

```
# Escriure la Query amb limitació de files

QUERY = """
        SELECT message
        FROM `bigquery-public-data.github_repos.commits`
        WHERE LENGTH(message) > 6 AND LENGTH(message) <= 20
        LIMIT 2000
        """

# Estimar la mida del resultat de la consulta
bq_assistant.estimate_query_size(QUERY)

# Executar la Query i guarder-nos el resultat en un DataFrame
# on podrem explorar els resultats i/o fer-ne una representació
gràfica
df = bq_assistant.query_to_pandas_safe(QUERY)
```

Si voleu aprendre pas per pas com fer anàlisi amb BigQuery, podeu anar al tutorial d'SQL dins de l'apartat "Learn" o si voleu veure un exemple concret que treu una gràfica de barres amb el tipus de llicències dels repositoris de Github, podeu consultar el següent Kernel:

https://www.kaggle.com/mrisdal/safely-analyzing-github-projects-popular-licenses



Referències

- https://cloud.google.com/bigquery/
- https://www.kaggle.com/
- https://bbvaopen4u.com/es/actualidad/analisis-de-big-data-como-servicio-con-google-bigquery
- https://www.kaggle.com/mrisdal/safely-analyzing-github-projects-popular-licenses
- https://aws.amazon.com/athena/