Term Project

Google BigQuery

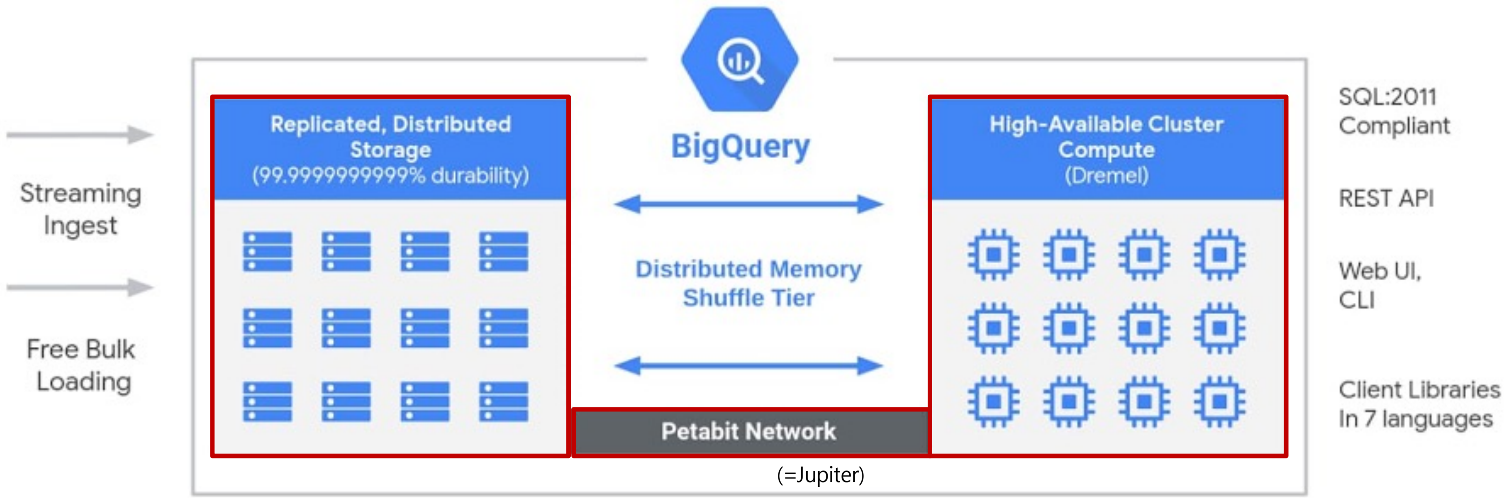
201736055 한지안
201833299 허서윤

Contents

What is BigQuery?

Serverless cloud service for users outside Google
based on the Dremel project

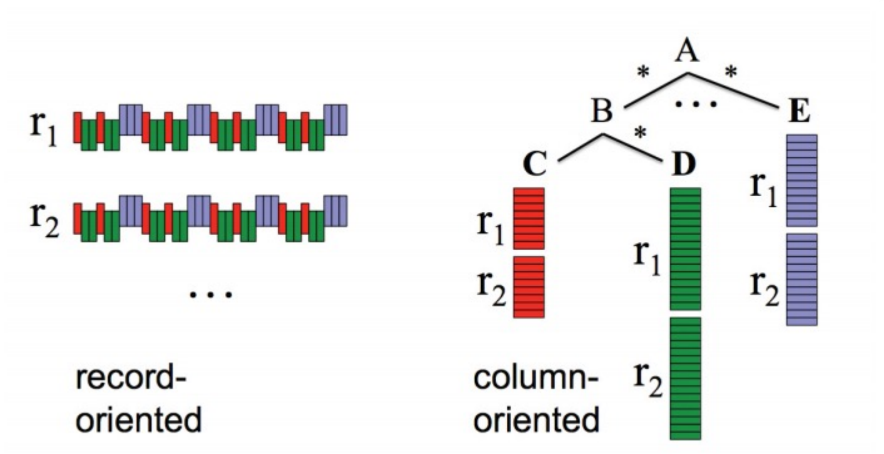
Architecture of BigQuery



Architecture of BigQuery

Architecture of BigQuery

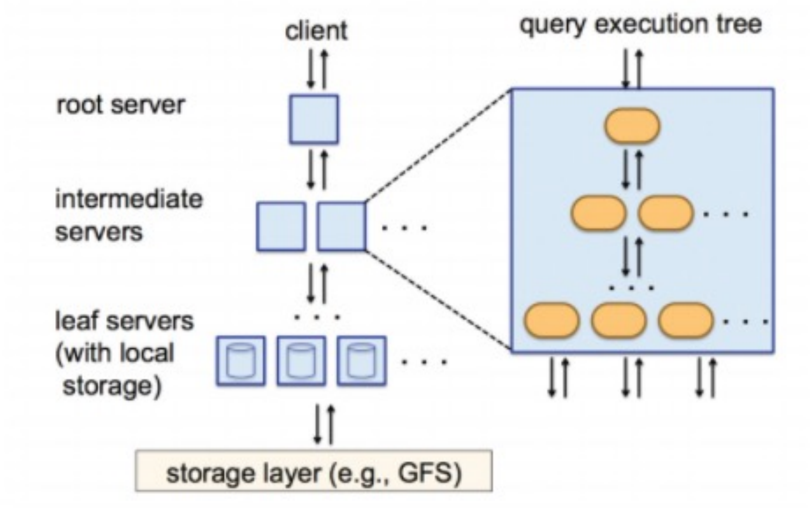
① Columnar Storage



- Save the table in an optimized column format
- Same data type data are gathered and stored
- Each table is compressed and encrypted on a disk
- Storage is durable, and each table replicates between data centers

Architecture of BigQuery

② Tree Architecture Distribution



❑ Root server

- Receives SQL query from the user
- Splits it into a small SQL query

❑ Intermediate server

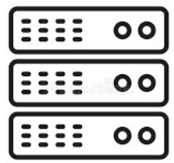
-Query split and sent to leaf node (=slot)
-Gets through columnar storage process.

❑ Leaf server

- Reads data from a storage layer
- Delivers results to the parent node.

Key Features

① Serverless



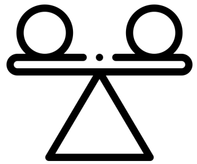
Usually, there is no server, and computer resources are used only for analysis
Only need to pay at this time.

② Cloud

No need to install/operate as a service.
Large capacity support, fast performance support, low price

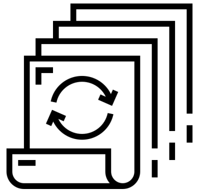
Key Features

③ Stability



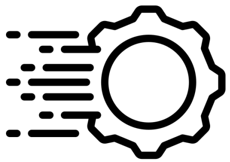
The risk of data loss is low because three copies are distributed and stored in different data centers

④ Batch & Streaming



It provides a batch that loads data at once, and a streaming function that allows you to input data in real time.

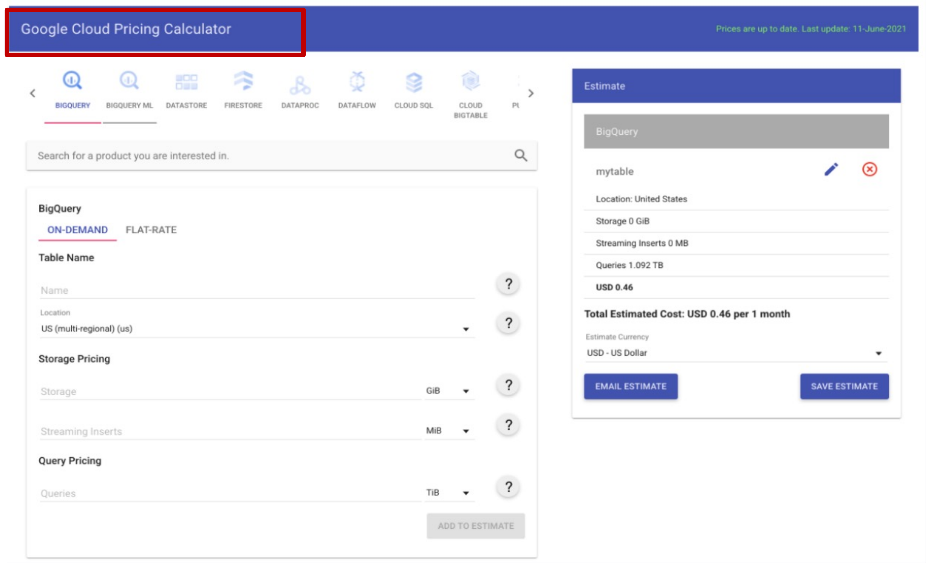
⑤ Dermel project



Structured data can be distributed and analyzed quickly

Pros of BigQuery

① Cost-effective because it is possible to estimate query costs

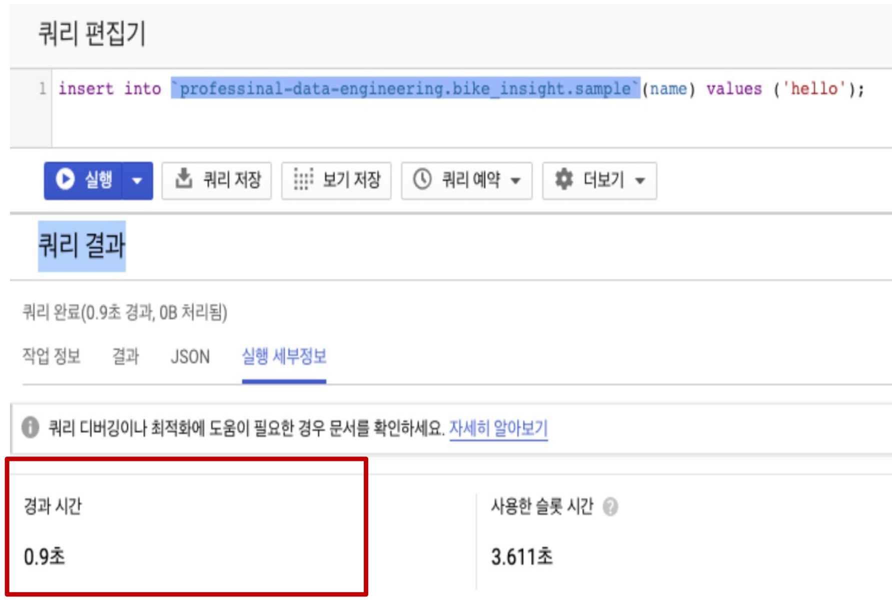


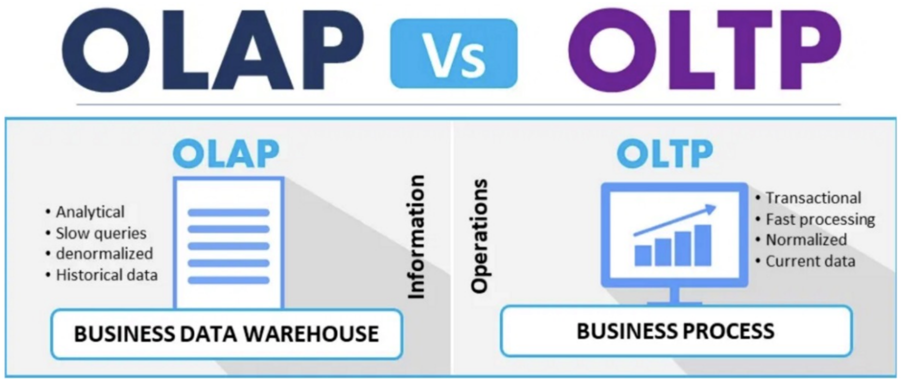
② Models can be created & tested using SQL queries

Google BigQuery ML

Cons of BigQuery

Specialized in analysis and OLAP
Not suitable for OLTP





Establishing and Using Classification Model in Census Data

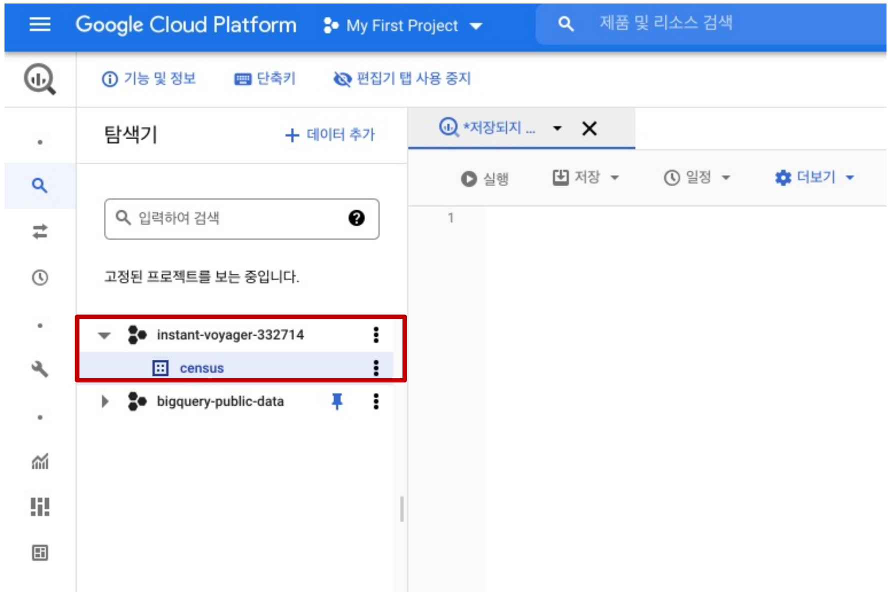
US Census Data

United States Census Bureau

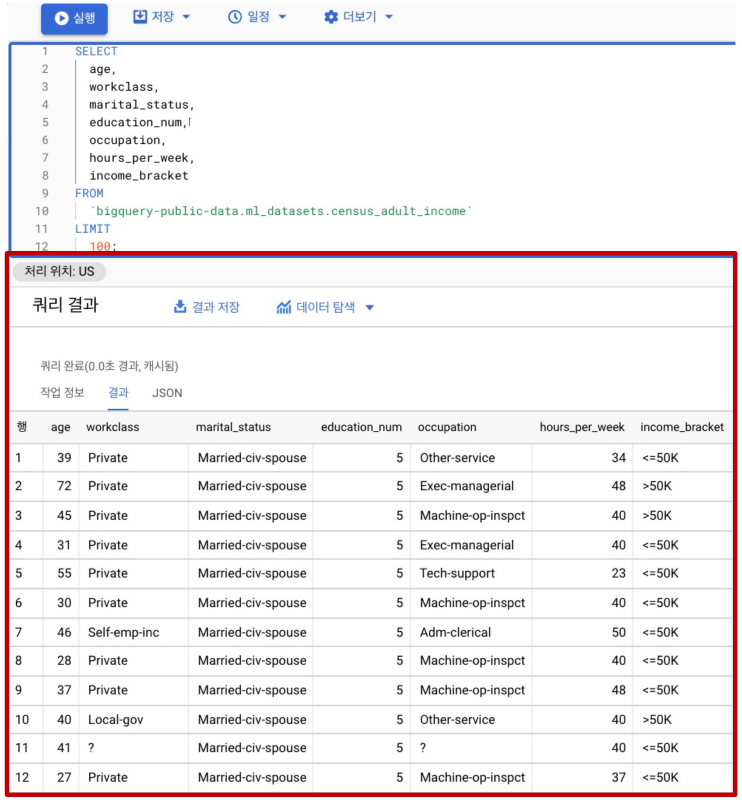
2000 and 2010 US Census data

Data Information
Age
Workclass
Marital_Status,
Education_num
Occupation
Hours_per_week
Income_bracket

1. Big query data sets are created to store models.



2. Returns 100 rows from a dataset



Implementation of BigQuery

3. Create a view to compile training data

실행 저장 일정 더보기

```sql
CREATE OR REPLACE VIEW
  `census.input_view` AS
SELECT
  age,
  workclass,
  marital_status,
  education_num,
  occupation,
  hours_per_week,
  income_bracket,
  CASE
    WHEN MOD(functional_weight, 10) < 8 THEN 'training'
    WHEN MOD(functional_weight, 10) = 8 THEN 'evaluation'
    WHEN MOD(functional_weight, 10) = 9 THEN 'prediction'
  END AS dataframe
FROM
  `bigquery-public-data.ml_datasets.census_adult_income`
```

뷰 스키마

필터 속성 이름 또는 값 입력

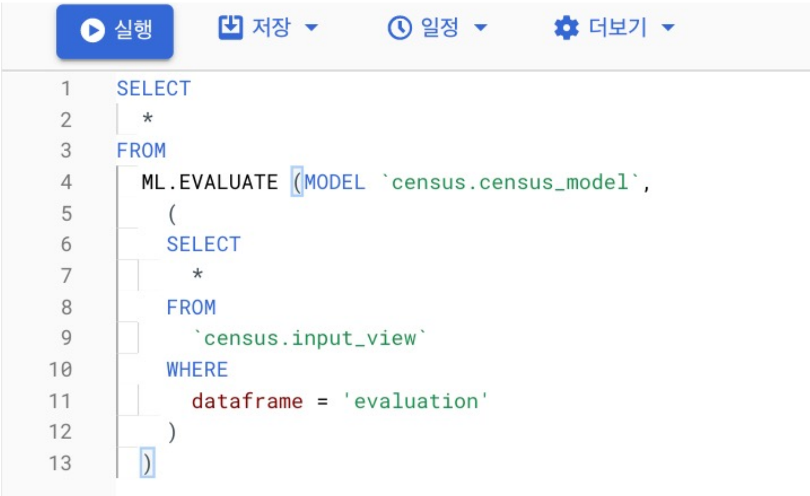
필드 이름	유형	모드
age	INTEGER	NULLABLE
workclass	STRING	NULLABLE
marital_status	STRING	NULLABLE
education_num	INTEGER	NULLABLE
occupation	STRING	NULLABLE
hours_per_week	INTEGER	NULLABLE
income_bracket	STRING	NULLABLE
dataframe	STRING	NULLABLE

4. Create a logistic regression model

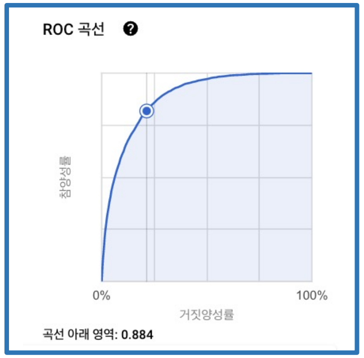
실행 저장 일정 더보기

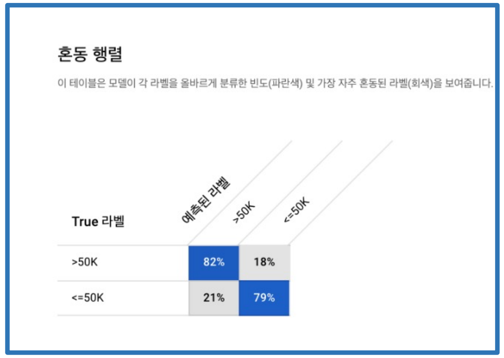
```sql
CREATE OR REPLACE MODEL
  `census.census_model`
OPTIONS
  ( model_type='LOGISTIC_REG',
    auto_class_weights=TRUE,
    input_label_cols=['income_bracket']
  ) AS
SELECT
  * EXCEPT(dataframe)
FROM
  `census.input_view`
WHERE
  dataframe = 'training'
```

5. Use the ML.EVALUATE function to evaluate the performance of the model




```
SELECT
  *
FROM
  ML.EVALUATE (MODEL `census.census_model`,
    (
    SELECT
      *
    FROM
      `census.input_view`
    WHERE
      dataframe = 'evaluation'
    )
  )
```





쿼리 결과 결과 저장 데이터 탐색

쿼리 완료(0.9초 경과, 2.6MB 처리됨)

작업 정보 결과 JSON 실행 세부정보

행	precision	recall	accuracy	f1_score	log_loss	roc_auc
1	0.5740551583248212	0.7635869565217391	0.8168009919404836	0.6553935860058309	0.3944068834574654	0.8827432567432567

6. Predict the income class of all respondents

실행 저장 일정 더보기

```sql
SELECT
  *
FROM
  ML.PREDICT (MODEL `census.census_model`,
    (
    SELECT
      *
    FROM
      `census.input_view`
    WHERE
      dataframe = 'prediction'
    )
  )
```

7. More detailed analysis with Explainable AI method

실행 저장 일정 더보기

```sql
#standardSQL
SELECT
*
FROM
ML.EXPLAIN_PREDICT(MODEL `census.census_model`,
  (
  SELECT
    *
  FROM
    `census.input_view`
  WHERE
    dataframe = 'evaluation'),
  STRUCT(3 as top_k_features))
```

작업 정보 결과 JSON 실행 세부정보

행	predicted_income_bracket	predicted_income_bracket_probs.label	predicted_income_bracket_probs.prob	age	workclass	marital_status	education_num	occupation	hours_per_week	income_bracket	dataframe
1	<=50K	>50K	0.05639289147442344	34	?	Married-civ-spouse	7	?	8	<=50K	prediction
		<=50K	0.9436071085255766								
2	<=50K	>50K	0.1311398392666556	21	?	Married-civ-spouse	9	?	30	<=50K	prediction
		<=50K	0.8688601607333444								
3	<=50K	>50K	0.06491298402568849	25	?	Married-civ-spouse	9	?	4	<=50K	prediction
		<=50K	0.9350870159743115								

6. Predict the income class of all respondents

실행 저장 일정 더보기

```sql
SELECT
  *
FROM
  ML.PREDICT (MODEL `census.census_model`,
    (
    SELECT
      *
    FROM
      `census.input_view`
    WHERE
      dataframe = 'prediction'
     )
  )
```

7. More detailed analysis with Explainable AI method

실행 저장 일정 더보기

```sql
#standardSQL
SELECT
*
FROM
ML.EXPLAIN_PREDICT(MODEL `census.census_model`,
  (
  SELECT
    *
  FROM
    `census.input_view`
  WHERE
    dataframe = 'evaluation'),
  STRUCT(3 as top_k_features))
```

행	predicted_income_bracket	probability	top_feature_attributions.feature	top_feature_attributions.attribution	baseline_prediction_value	prediction_value	approximation_error	age	workclass	marital_status	education_num	occupation	hours_per_week	income_bracket	dataframe
1	>50K	0.5211861694930122	hours_per_week	2.0131952297565316	-0.29878674958313434	0.0847954499691751	0.0	59	Private	Married-civ-spouse	4	Other-service	99	<=50K	evaluation
			education_num	-1.8421287256458136											
			occupation	-1.2016119213892107											
2	>50K	0.5322760998839042	occupation	-1.2016119213892107	-0.29878674958313434	0.12928417465217856	0.0	47	Self-emp-not-inc	Married-civ-spouse	9	Other-service	75	<=50K	evaluation
			hours_per_week	1.1871782572907503											
			marital_status	0.9439379268252633											

Reference

Specs of BigQuery
https://syujisu.tistory.com/190?category=907377
https://youtu.be/LhksTFvVriU

Architecture of BigQuery
https://cloud.google.com/bigquery
https://velog.io/@jch9537/%ED%95%9C-%EC%A4%84-%EC%9A%A9%EC%96%B4%EB%B0%B0%EC%B9%98Batch%EB%9E%80

Pros of BigQuery
https://xo.xello.com.au/blog/google-bigquery-5-benefits-cloud-data-warehouse
https://www.quora.com/What-are-the-pros-and-cons-of-using-Google-BigQuery-as-a-database
https://www.xplenty.com/blog/snowflake-vs-bigquery/

Cons of BigQuery
https://www3.technologyevaluation.com/solutions/53566/google-bigquery
https://dzone.com/articles/introduction-to-google-bigquery

Big Query Implementation
https://cloud.google.com/bigquery-ml/docs/logistic-regression-prediction

Data Source Site
https://console.cloud.google.com/marketplace/product/united-states-census-bureau/us-census-data?project=instant-voyager-332714

Thank You