Google Cloud Platform Data Analytics Services

Janakiram MSV

Learning Objectives

- Overview of GCP Data Analytics Services
- Cloud Pub/Sub
- Cloud Dataflow
- Cloud Dataproc
- Cloud Datalab
- BigQuery

Demo: Analyzing data with BigQuery

- Use Cases of Data and Analytics Services

GCP Data & Analytics Services

Overview of GCP Data Analytics Services

- Data analytics include ingestion, collection, processing, analyzing, visualizing data
- GCP has a comprehensive set of analytics services
- Cloud Pub/Sub is used for ingesting data at scale
- Cloud Dataflow can process data in real-time or batch mode
- Cloud Dataproc is a Big Data service for running Hadoop and Spark jobs
- BigQuery is the data warehouse in the cloud
- Cloud Datalab is used for analyzing and visualizing data

Google Cloud Pub/Sub

Google Cloud Pub/Sub

- Managed service to ingest data at scale
- Based on the publishing/subscription pattern
- Global entry point to GCP-based analytics services
- Acts as a simple and reliable staging location for data
- Tightly integrated with services such as Cloud Storage and Cloud Dataflow
- Supports at-least-once delivery with synchronous, cross-zone message replication
- Comes with end-to-end encryption, IAM, and audit logging

Google Cloud Dataflow

Google Cloud Dataflow

- Managed service for transforming and enhancing data in stream and batch modes
- Based on Apache Beam open source project
- Serverless approach automates provisioning and management
- Inbound data can be queried, processed, and extracted for target environment
- Tightly integrated with Cloud Pub/Sub, BigQuery, and Cloud Machine Learning
- Cloud Dataflow connector for Kafka makes it easy to integrate Apache Kafka

Google Cloud Dataproc

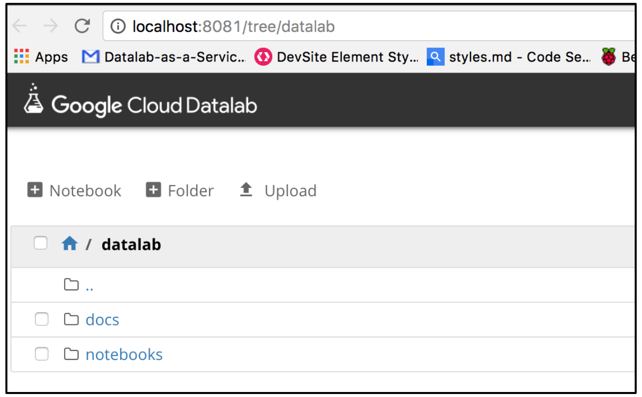
Google Cloud Dataproc

- Managed Apache Hadoop and Apache Spark cluster environments
- Automated cluster management
- Clusters can be quickly created and resized from three to hundreds of node
- Move existing Big Data projects to GCP without redevelopment
- Frequent updates to Spark, Hadoop, Pig, and Hive
- Integrates with other GCP services like Cloud Dataflow and BigQuery

Google Cloud Datalab

Google Cloud Datalab

- Interactive tool for data exploration, analysis, visualization, and machine learning
- Runs on Compute Engine and may connect to multiple cloud services
- Built on open source Jupyter Notebooks platform
- Enables analysis data on BigQuery, Cloud ML Engine, and Cloud Storage
- Supports Python, SQL, and JavaScript languages



Google BigQuery

BigQuery

- Serverless, scalable cloud data warehouse
- Has an in-memory BI Engine and machine learning built in
- Supports standard ANSI:2011 SQL dialect for querying
- Federated queries can process external data sources
  - Cloud Storage
  - Cloud Bigtable
  - Spreadsheets (Google Drive)
- Automatically replicates data to keep a seven-day history of changes
- Supports data integration tools like Informatica and Talend

Google Cloud Platform Fundamentals

Lab Guide for Google BigQuery

```
# Run the below SQL statement in BigQuery

SELECT badge_name AS First_Gold_Badge,
       COUNT(1) AS Num_Users,
       ROUND(AVG(tenure_in_days)) AS Avg_Num_Days
FROM
(
  SELECT
    badges.user_id AS user_id,
    badges.name AS badge_name,
    TIMESTAMP_DIFF(badges.date, users.creation_date, DAY) AS tenure_in_days,
    ROW_NUMBER() OVER (PARTITION BY badges.user_id
                       ORDER BY badges.date) AS row_number
  FROM
    `bigquery-public-data.stackoverflow.badges` badges
  JOIN
    `bigquery-public-data.stackoverflow.users` users
  ON badges.user_id = users.id
  WHERE badges.class = 1
)
WHERE row_number = 1
GROUP BY First_Gold_Badge
ORDER BY Num_Users DESC
LIMIT 10
```

GCP Data & Analytics Service – Use Cases

Use Cases

Product	Service Type	Key Feature	Use Case
Google Cloud Pub/Sub	Ingestion	High-speed ingestion of data	Sensor data, telemetry, and logs
Google Cloud Dataflow	Stream and batch processing	Process data coming from Pub/Sub and data in GCS	ETL for business intelligence and machine learning
Google Cloud Dataproc	MapReduce jobs	Big Data processing based on Apache Hadoop and Spark	MapReduce jobs
Google Cloud Datalab	Visualization	Jupyter Notebooks for interactive analysis	Data exploration and visualization
BigQuery	Data warehouse	Query large datasets in ANSI SQL	Business intelligence

Google Cloud Platform Fundamentals

Resources for Google Cloud Data & Analytics

Key Links

- Cloud Pub/Sub
- Cloud Dataflow
- Cloud Dataproc
- BigQuery

References

- GCP Big Data Products
- BigQuery Quickstart
- GCP Data Analytics Blog