Dataproc overview  
Dataproc is a managed Spark and Hadoop service that lets you take advantage of open source data tools for batch processing, querying, streaming, and machine learning. Dataproc automation helps you create clusters quickly, manage them easily, and save money by turning clusters off when you don't need them. With less time and money spent on administration, you can focus on your jobs and your data.

Dataproc Architecture

Google Cloud Dataproc architecture is designed for scalable, efficient, and cost-effective processing of large datasets using Apache Spark, Apache Hadoop, and other open-source big data tools. Below is an outline of the key components and layers involved in a typical Dataproc architecture:

**1. Data Sources**

* **Storage Systems**:
  + Google Cloud Storage (GCS) for raw, processed, and analytical data (Bronze, Silver, Gold layers).
  + BigQuery for queryable datasets and reporting.
  + External databases (e.g., MySQL, PostgreSQL, or MongoDB).
* **Streaming Sources**:
  + Pub/Sub for real-time event streaming.
  + Kafka or other message brokers.

**2. Dataproc Cluster**

* **Cluster Components**:
  + **Master Node(s)**: Manages cluster resources, job execution, and coordination.
  + **Worker Node(s)**: Perform distributed data processing and computation.
  + **Optional Preemptible Workers**: Cost-saving nodes for non-critical workloads.
* **Cluster Configuration**:
  + Choose the machine type (e.g., n1-standard-4) and number of nodes based on workload.
  + Custom initialization actions (e.g., installing libraries, tuning Spark).
  + Autoscaling to dynamically adjust resources based on workload demand.

**3. ETL Layers**

* **Bronze Layer**:
  + Ingest raw data from sources into GCS or HDFS.
  + Focus on minimal transformation and retention of raw formats (e.g., JSON, CSV, Avro).
* **Silver Layer**:
  + Transform and clean raw data using Spark jobs.
  + Standardize formats (e.g., Parquet) and enrich datasets.
  + Store processed data in GCS or BigQuery for further analysis.
* **Gold Layer**:
  + Aggregate and optimize data for analytics or machine learning.
  + Create dimensionally modeled datasets or feature stores.

**4. Workflow Orchestration**

* **Apache Airflow** (or Cloud Composer):
  + Manage ETL workflows for Dataproc.
  + Schedule jobs (e.g., PySpark, HiveQL, Pig scripts).
  + Monitor task dependencies and retry logic.
* **Optional Alternatives**:
  + Dataproc Workflow Templates for simpler orchestration.
  + Vertex AI Pipelines for ML-focused workflows.

**5. Data Access and Analytics**

* **BigQuery**:
  + Direct access to Gold layer datasets for SQL-based analytics.
* **BI Tools**:
  + Connect Power BI, Tableau, or Looker to Dataproc or BigQuery.
* **ML Tools**:
  + Export prepared datasets to Vertex AI or TensorFlow for machine learning.

**6. Monitoring and Logging**

* **Cloud Monitoring**:
  + Monitor resource usage, cluster health, and job performance.
* **Cloud Logging**:
  + Centralized logging for debugging and tracking workflows.
* **Dataproc Job Logs**:
  + Access detailed Spark/Hadoop logs for troubleshooting.

**7. Security and IAM**

* **IAM Roles**:
  + Restrict access to Dataproc, GCS, and BigQuery by roles (e.g., Data Engineer, Data Analyst).
* **VPC and Subnets**:
  + Ensure private communication within Dataproc clusters.
* **Encryption**:
  + Enable encryption for data at rest (GCS) and in transit (TLS).

**Sample Workflow**

1. **Ingest Data**:
   * Pull data from Pub/Sub into GCS (Bronze layer).
2. **Transform Data**:
   * Run PySpark jobs in Dataproc to clean and enrich data (Silver layer).
3. **Load Data**:
   * Save transformed data to BigQuery (Gold layer).
4. **Analyze Data**:
   * Use Looker for reporting or Vertex AI for machine learning.

This modular design allows Dataproc to efficiently process large-scale data, providing flexibility for various big data and analytics tasks.

Basic ETL Architecture using Dataproc

