**Architect Diagram for Bitcoin ETL pipeline**

Graphical user interface

Description automatically generated

1. **Data Sources():**This layer will collect data from different data sources.  Data sources could be files(Bitcoin data from broker web), real time bitcoin data. That can be downloaded in GCS Bucket location using python code
2. **Data Ingestion.** This layer will import data from the data sources. We can create different plugins to get data from different sources.. Apache Kafka to get real-time reviews data from Data base. Files can be directly copied to GCP Buckets. Custom API’s can be built to copy data to the GCP bucket
3. **Storage.** For Storage purpose GCP Bucket can be used, We can land data to these buckets using Ingestion Layer. And can be directly consumed by Hive Tables, Apache Spark, etc.
4. **Transformation**. For ETL purposes we can use Apache spark in combination with Hive tables. Transformed data from Apache spark can be saved as an external table.

**Using Apache Spark for transformation, because of in memory computation, support of real time steaming**.

1. **Data Exploration.** HiveQL can be used for this purpose. The output from spark will store in hive tables and tables can be used for data exploration purposes. This layer is highly useful for Data-scientists.
2. **Analytical Layer:** In this layer, we will use output data for trading pattern detection using machine learning models.
3. **Data Visualization:**  This layer will be used to get real-time dashboards and BI tools for data analysis. Visualising the price and indicators like std\_dev, moving average on charts, other data dashboards.
4. **Serving Layer:** This layer can be used to serve the output to users like trade will be initialise whenever a trade condition is detected from ML models.
5. **Governance:** It is a transversal component that monitors data manipulations within the data lake and preserves enough information to trace them.

**Note:** Apache Airflow can be used to schedule ETL pipeline. Including python script to download data directly from source url’s.

**GCP:** Google Cloud Platform

**GCS:** Google Cloud Storage