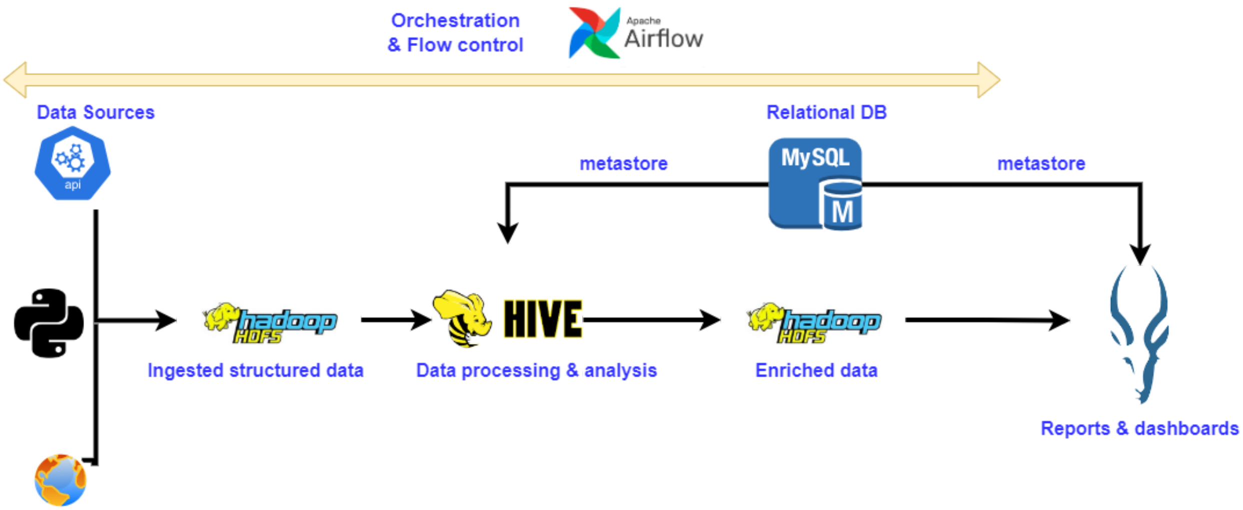
Flow and scheduling:



The End of the day trading data will be ingested from external Financial API's (e.g. Yfinance) at market closure time + 1 hour, daily, on the days the market was open. The structured data will be stored in HADOOP cluster (hdfs) using Pyarrow package, in Parquet format. This data will be processed and enriched using Hive, and stored again, becoming available for users, to query it using Impala. Relational DB(e.g. MySQL) is required for storing the metadata used by Hive and Impala.

Orchestration: Apache Airflow

The Airflow platform is an open-source tool for describing, executing, and monitoring workflows, used as an ecosystem for triggering API calls every minute. This tool had been used in order to utilize the convenience of creating a custom python code (DAG) that schedules the tasks order and runs timeslot. It's also easy to integrate the tasks python code in the DAG python with minimum configuration. Airflow can be easily integrated and used in Hadoop-Claudera environment.

Storage: Hadoop HDFS

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS is part of the Cloudera environment solution, thus it is convenient to use it as a storage component, for data analysis over parquet files, via Hive and Impala.

Compute engine: Hadoop cluster, batch data processing with Hive

The Apache Hive data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL. Structure can be projected onto data already in storage. A command line tool and JDBC driver are provided to connect users to Hive. Hive can be integrated with HDFS and its part of the Hadoop-Cloudera solution.

Visualization Layer: Impala

Apache Impala is an open source, native analytic database for Apache Hadoop.Impala provides low latency and high concurrency for BI/analytic queries on Hadoop (not delivered by batch frameworks such as Apache Hive). Impala also scales linearly, even in multi-tenant environments. Impala can be integrated with HDFS and its part of the Hadoop-Cloudera solution.