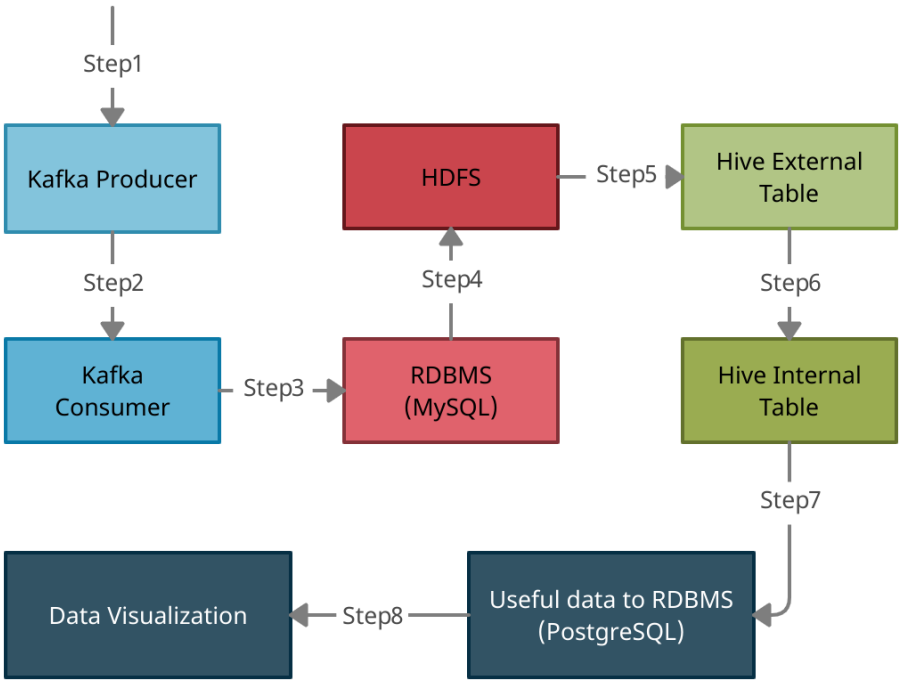
CAPSTONE PROJECT

Description:

In this project, we will get latest datasets to prepare them for an Analysis. In this project, we are going to work with the COVID19 dataset, published by John Hopkins University, which consists of the data related to the cumulative number of total confirmed cases, deaths, recovers, critical, tests in each Country and timeseries datasets of each Country. Also, we have another dataset consist of various states in the USA and store them in our RDBMS. We are going to separately achieve top most confirmed cases, deaths, recovers, critical, tests country wise and state wise from the USA country with timeseries.

Basic Design:



Steps:

Step 1: Kafka producer will fetch three datasets from an API and filter columns of datasets as needed.

Step 2: Kafka producer will make filtered datasets available for Kafka consumer to consume those datasets.

Step 3: Kafka consumer will consume datasets from Kafka producer and save them into RDBMS (MySQL).

Step 4: Using PySpark, all the datasets from RDBMS will be ingested to the HDFS as a CSV file.

Step 5: By enabling support of Hive to PySpark, we will create Hive External table from HDFS data lakes.

Step 6: From Hive External table we will get data, aggregate those data as our needs and create ten new Internal tables within Hive.

Step 7: Store all tables of useful data to RDBMS (PostgreSQL) for further analysis.

Step 8: Visualizing all useful data from PostgreSQL database using Django Application.

Benefits:

- Achieving top most confirmed cases, deaths, recovers, critical, tests of each country, we can have good measures and will be useful for further analysis to make decision in each country.
- Get ranks for all the countries based on confirmed cases of each country.
- For each state from the USA country, by achieving top most confirmed cases, deaths, recovers, tests, we can have good measures and will be useful for further analysis to make decision in each state of the USA country.
- By achieving COVID-19 timeseries of the USA country, we can have comparison on first five months and last five months of confirmed cases increased day by day and how many are getting recovered daily.
- Visualising (On Django Application) all the achieved data with automated pipeline within Apache Airflow, we will have good visual understanding with Bar chart, Line chart and map plots.