21CSA523A Data Engineering for AI

Mid Review Report

Project Title

Covid-19 Data Warehouse

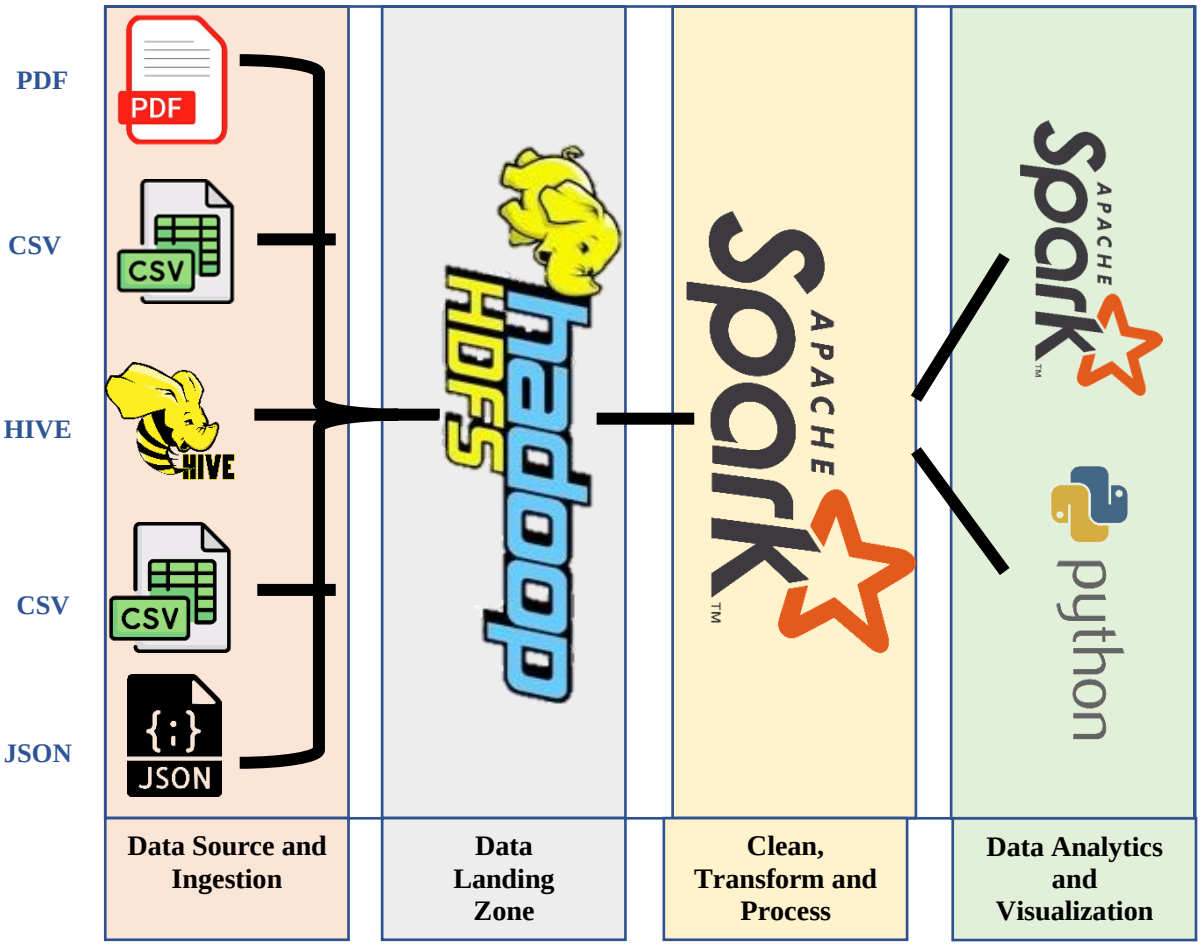
Prepared by
Name of the student: Manav Patadia
Roll Number: AA.SC.P2MCA2107423
Degree and Semester: MCA – 3RD Semester
Email: manavnp_mca2107423@ahead.students.amrita.edu

March 2023

Objective: Objective of this project is as below:

1. Extract data from various websites about covid-19.
2. Process, clean and transform extracted data and form data marts and data warehouse.
3. Analyze its impact on health, economics and environmental aspects of human lives across India. I have stock data to measure impact of covid19 on Indian economics. I have pollution data to measure impact of covid19 on environment. I have death data to measure impact of covid19 on Indian health.
4. Visualize analyzed data in python

Block Diagram:



Stakeholders of this project:

Government, Financial advisors, Stock analysts, Students

Tools Used:

Py-Spark (Data Processing), Hive (Data Storage), HDFS (Data Storage) and Python (Visualization)

Region Selected for this Case study: India

Dataset details:

Dataset 1: Daily Covid 19 Case Details

Columns: Date, State, Confirmed_Indian_Cases, Confirmed_Foreign_cases, Cured, Deaths

Source:

1. https://raw.githubusercontent.com/datameet/covid19/master/data/mohfw.json

Format: JSON

Record Count: 34991

Dataset 2: State_codes

Columns: Subdivision category, 3166-2 code, Subdivision name

Source https://en.wikipedia.org/wiki/ISO_3166-2:IN

Format: CSV

Record Count: 37

Dataset 3: Daily Covid Vaccination details

Columns: Date, State, dose_1, dose_2, 15_18_years_dose_1, 15_18_years_dose_2, 12_14_years_dose_1, 12_14_years_dose_2, precaution_dose, total_doses

Source: https://thejeshgn.com/2020/03/16/novel-corona-virus-covid19-archive-api-india-data/

Format: PDF

Record Count: 589 PDF files and Total Rows: 22447

Dataset 4: Pollution Data

Source: https://api.data.gov.in/resource/3b01bcb8-0b14-4abf-b6f2-c1bfd384ba69

Big Query Public Table: bigquery-public-data.openaq.global_air_quality

Columns: location, city, country, pollutant, value, timestamp, unit, source_name, latitude, longitude, averaged_over_in_hours, location_geom

Format: Big Query Google Table

Record Count: 5594614

Dataset 5: List_of_cities_and_towns_in_india

Columns: Subdivision category, 3166-2 code, Subdivision name

Source:https://www.downloadexcelfiles.com/sites/default/files/docs/list_of_cities_and_towns_in_india-834j.csv

Format: CSV

Record Count: 1318

Dataset 6: Daily Nifty-50 Details

Columns: Date, Open, High, Low, Close

Source: https://www.niftyindices.com/reports/historical-data

Format: Hive Table

Record Count: 927

Plan to Execute:

Step 1: Load data from 6 different sources into spark DataFrames

Step 2: Prepare data pipeline for the datasets using Spark

Step 3: Clean all the dataset thoroughly using Spark

Step 4: Transform and process the data to make it ready for analysis using Spark

Step 5: Analyze the datasets to draw conclusions from them using Spark

Step 6: Explain conclusions using visualizations using Python