Covid-19 Project

Overview

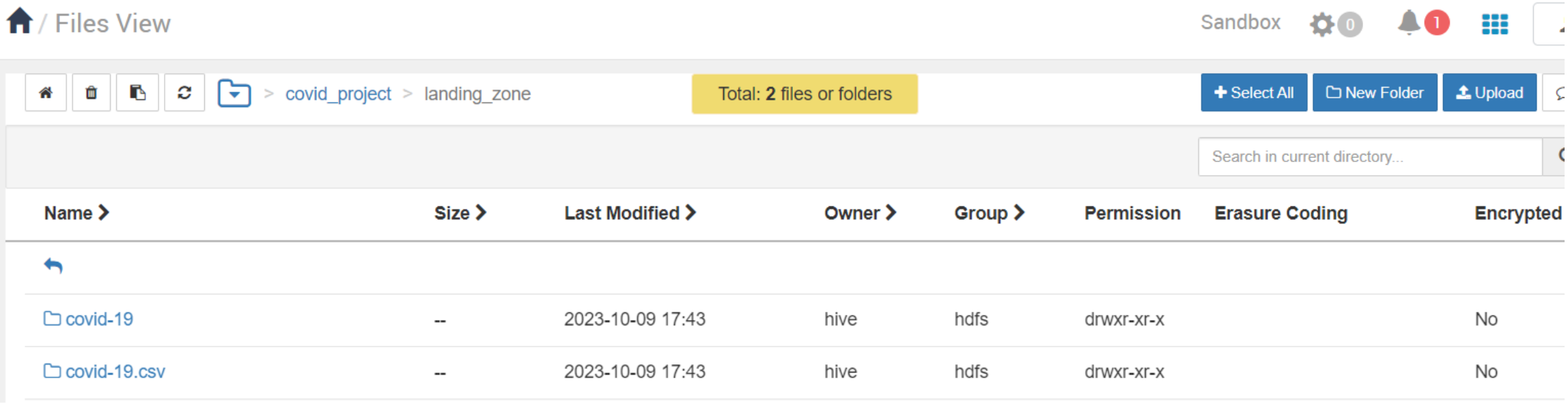
- Introduction
- Data Ingestion
- Data Transformation
- Work Flow
- Data Visualization

Introduction

- This is the graduation project for the Data Engineering Masterclass presented by Sprints.ai
- In this presentation, I will explain all the steps taken to develop and visualize some COVID-19 data set.

Data Ingestion

- To ingest the data used for this project, I used HDFS as a storage capacity and created required tables accordingly.
- The path used to store raw data is /user/raj_ops/home/cloudera/covid_project/landing_zone



Name	Size	Last Modified	Owner	Group	Permission	Erasure Coding	Encrypted
covid-19	--	2023-10-09 17:43	hive	hdfs	drwxr-xr-x		No
covid-19.csv	--	2023-10-09 17:43	hive	hdfs	drwxr-xr-x		No

Transformation/Processing

- Next step is to perform some transformations/processing on the raw data to obtain the required output.
- Firstly, the raw data is partitioned on the Country_name column to provide some efficiency and performance enhancment.

country_name=Afghanistan	--	2023-10-09 20:53	hive	hadoop	drwxrwxrwx
country_name=Albania	--	2023-10-09 20:54	hive	hadoop	drwxrwxrwx
country_name=Algeria	--	2023-10-09 20:54	hive	hadoop	drwxrwxrwx
country_name=Andorra	--	2023-10-09 20:54	hive	hadoop	drwxrwxrwx
country_name=Angola	--	2023-10-09 20:53	hive	hadoop	drwxrwxrwx
country_name=Anguilla	--	2023-10-09 20:54	hive	hadoop	drwxrwxrwx
country_name=Antigua and Barbuda	--	2023-10-09 20:53	hive	hadoop	drwxrwxrwx
country_name=Argentina	--	2023-10-09 20:54	hive	hadoop	drwxrwxrwx
country_name=Armenia	--	2023-10-09 20:54	hive	hadoop	drwxrwxrwx
country_name=Aruba	--	2023-10-09 20:54	hive	hadoop	drwxrwxrwx
country_name=Australia	--	2023-10-09 20:54	hive	hadoop	drwxrwxrwx
country_name=Austria	--	2023-10-09 20:54	hive	hadoop	drwxrwxrwx
country_name=Azerbaijan	--	2023-10-09 20:54	hive	hadoop	drwxrwxrwx
country_name=Bahamas	--	2023-10-09 20:54	hive	hadoop	drwxrwxrwx

select * from covid_db.covid_output_tests | Enter a SQL expression

	total_tests	testing_rate	country
1	137,457	1,778,642	Andorra
2	80,312	1,642,742	Faeroe Islands
3	51,953	1,322,632	Monaco
4	706,629	1,126,386	Luxembourg
5	28,366	841,971	Gibraltar
6	2,256	645,863	Falkland Islands
7	6,265,918	632,496	UAE
8	1,011,805	592,064	Bahrain
9	36,203	581,621	Bermuda
10	191,690	561,236	Iceland

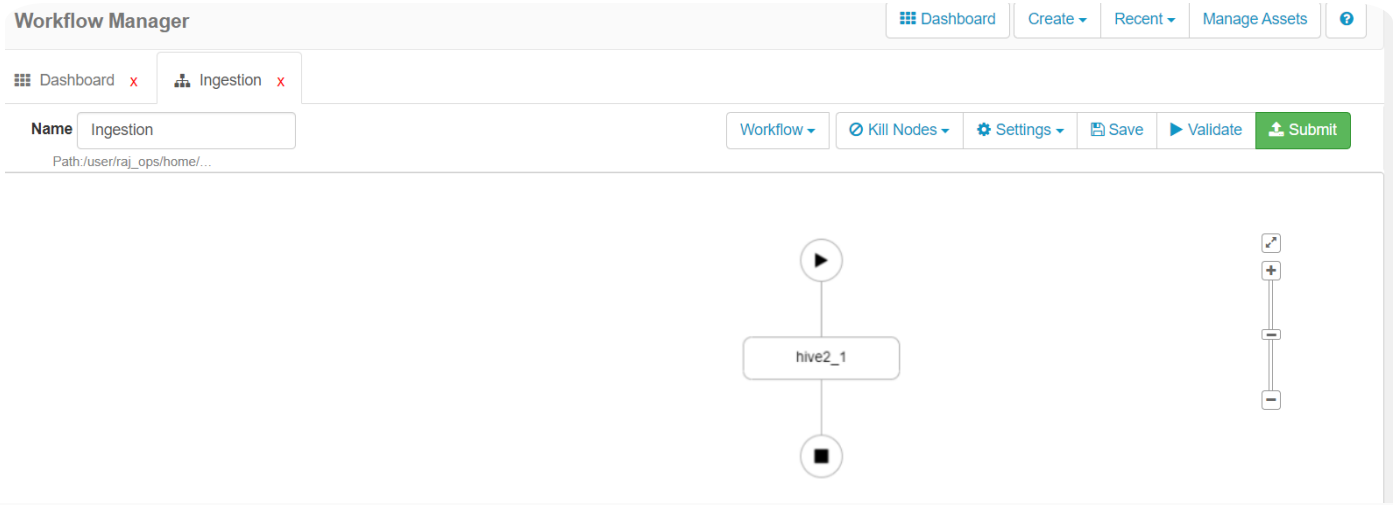
country	deaths	total_deaths
San Marino	1,237	42
Belgium	860	9,969
Peru	818	27,034
Andorra	686	53
Spain	616	28,813
UK	609	41,403
Italy	586	35,418
Sweden	574	5,805
Chile	558	10,671
USA	536	177,424

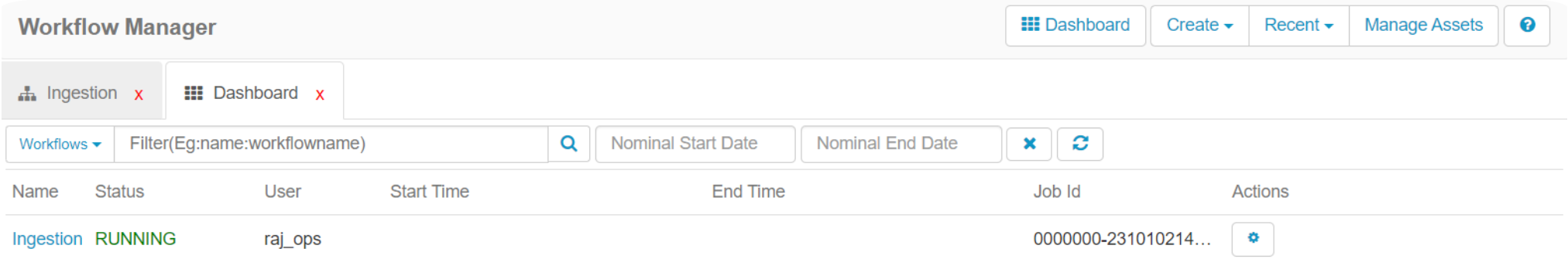
Transformation/Processing

- Next step is to create the tables need to be visualized:
  - Top 10 ranking countries in Death rate
  - Top 10 ranking countries in Test rate.

WORK FLOW

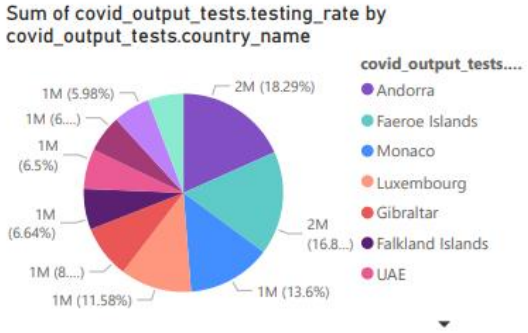
- To automate the creation and execution of the output tables, a work flow is developed using Apache oozie.

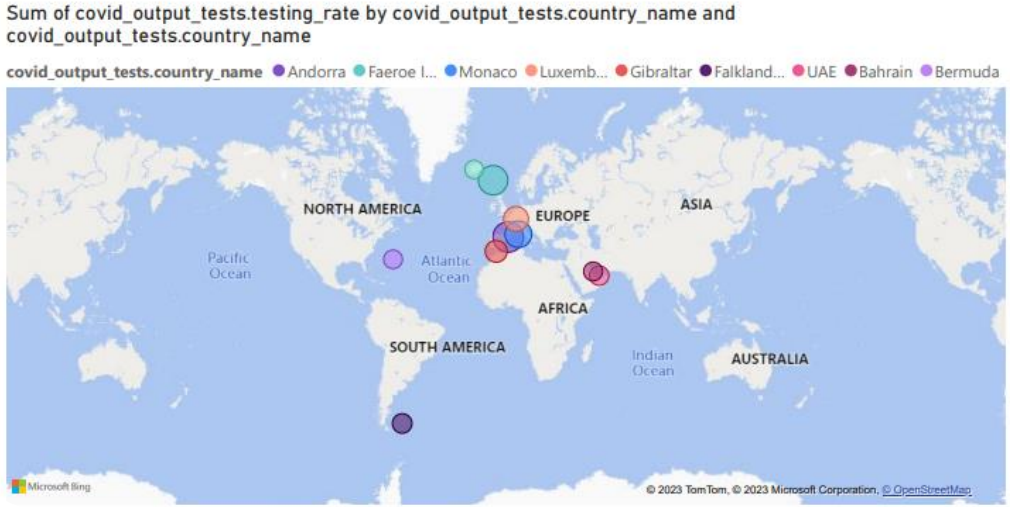


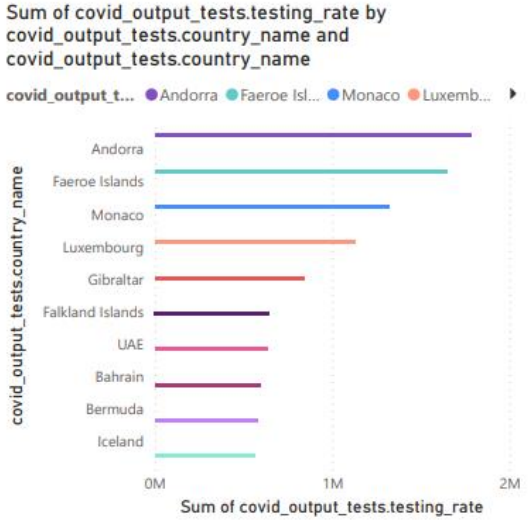


Data Visualization

- Power BI is used to perform the data visualization step, where some data analysis is done using charts and maps.







References

- GITHUB REPO:
  - https://github.com/Zeyad-Abady/Sprints_MasterClass/tree/main/COVID_PROJECT