Sprints

Big Data Masterclass

Project : COVID-19

Ahmed Reda

Amr Saleh

Agenda

Dataset Schema

Business Requirements

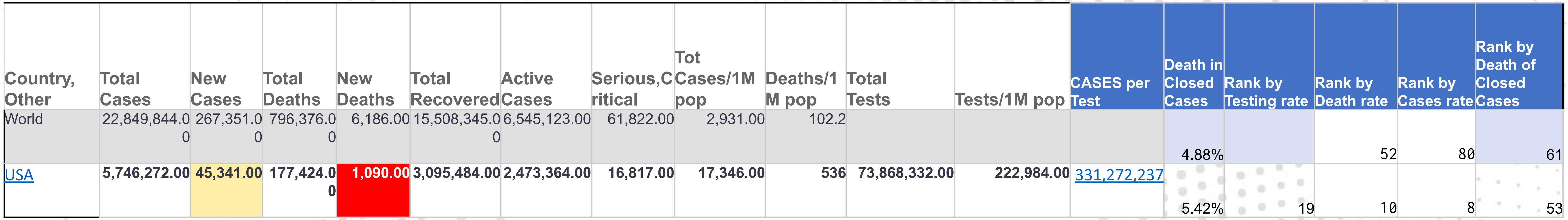
Technical Requirements hints

Sprints

Dataset Schema

Dataset Schema

Dataset schema : covid-19 August.xlsx



Country, Other	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	Active Cases	Serious,Critical	Tot Cases/1M pop	Deaths/1M pop	Total Tests	Tests/1M pop	CASES per Test	Death in Closed Cases	Rank by Testing rate	Rank by Death rate	Rank by Cases rate	Rank by Death of Closed Cases
World	22,849,844.00	267,351.00	796,376.00	6,186.00	15,508,345.00	6,545,123.00	61,822.00	2,931.00	102.2				4.88%		52	80	61
USA	**5,746,272.00**	**45,341.00**	**177,424.00**	**1,090.00**	**3,095,484.00**	**2,473,364.00**	**16,817.00**	**17,346.00**	**536**	**73,868,332.00**	**222,984.00**	331,272,237	5.42%	19	10	8	53

- **Country**
- **Total Cases**
- **New Cases**
- **Total Deaths**
- **New Deaths**
- **Total Recovered**

- **Active Cases**
- **Serious, Critical**
- **Tot Cases/1M pop**
- **Deaths/1M pop**
- **Total Tests**
- **Tests/1M pop**

- **CASES per Test**
- **Death in Closed Cases**
- **Rank by Testing rate**
- **Rank by Death rate**
- **Rank by Cases rate**
- **Rank by Death of Closed Cases**

Sprints

Business Requirements

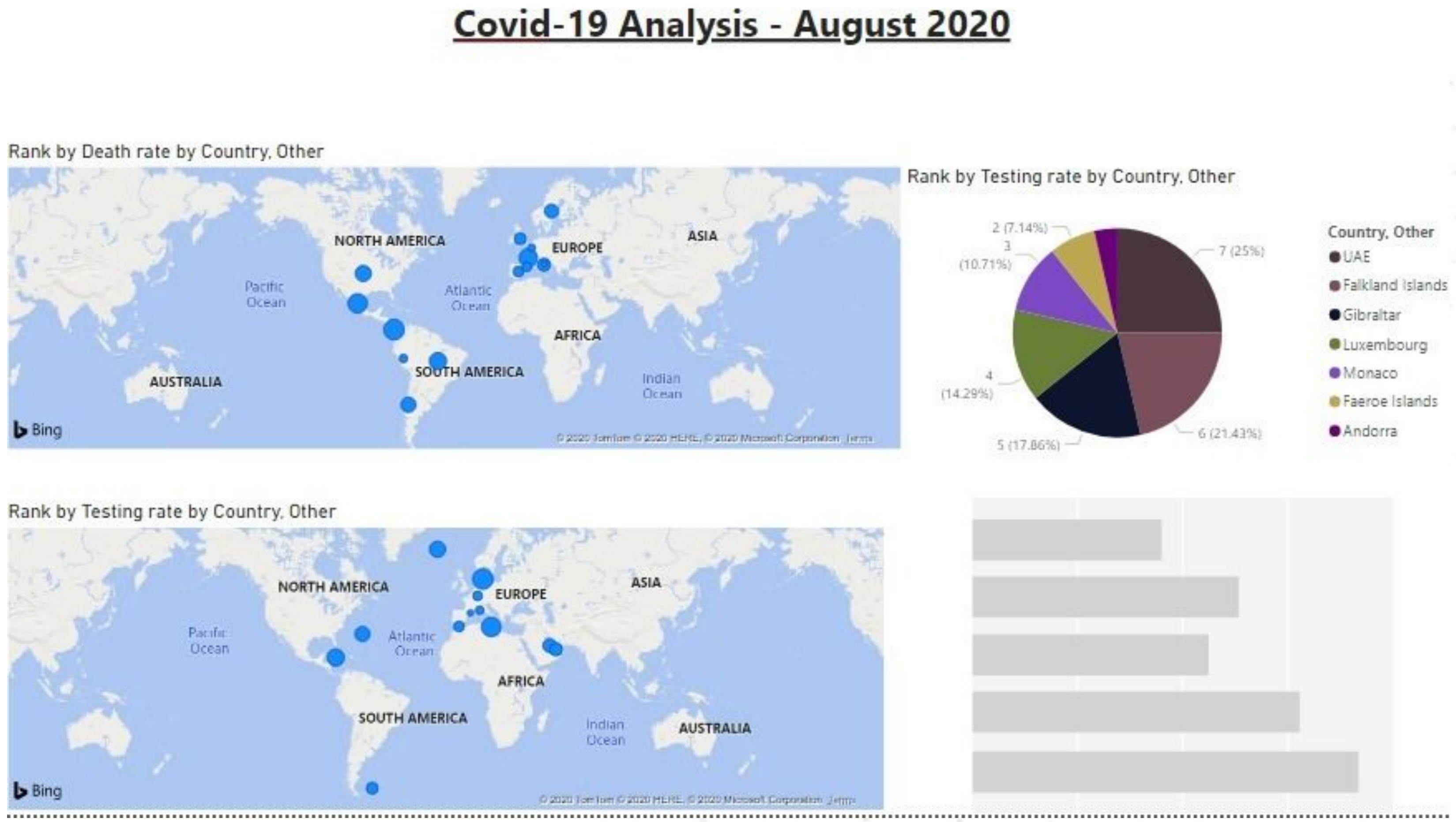
Business Requirements

Create an automated pipeline workflow from ingestion till visualization for COVID dataset

1. show on a map the top 10 ranking countries in death rate
2. show on a map the top 10 ranking countries in testing rate
3. show the top 10 ranking countries in testing rate on a pie chart
4. Add a custom chart of your choice in the empty section of the dashboard

Hint:

Rate means Count per Million Population
The final result should look like the shown visualization



Covid-19 Analysis - August 2020

Rank by Death rate by Country, Other

Rank by Testing rate by Country, Other

Rank by Testing rate by Country, Other

Sprints

Technical Requirements Hints

Business Requirements Hints

Sprints

1. Create Folder on the Virtual Machine nam “/home/cloudera/**covid_project**”

2. Create folders under “**covid_project**” **(landing_zone and scripts)**

3. Upload dataset **“covid-19.csv”** into VM using WinSCP into landing zone name it **“/home/cloudera/covid_project/landing_zone/COVID_SRC_LZ”**

4. Load the dataset from **“COVID_SRC_LZ”** to HDFS directory name it **“/user/cloudera/ds/COVID_HDFS_LZ”** using **HDFS cli** commands in a shell script

5. Sample shell script is added on google drive shared folder

Business Requirements Hints

6. Create database on Hive and create schema for each Hive loading stage

   I. 1st Hive staging table for pointing to dataset location to select data from
   II. 2nd Hive ORC table is partitioned by Country and data are loaded dynamically into it to speed query
   III. 3rd Final hive table to generate the final report which will generate output file to be visualized

7. Create an Oozie workflow actions from (Cloudera HUE in VM) to run the HDFS shell script and execute the Hive queries (HDFS and Hive actions)

Business Requirements Hints

Sprints

8. Run the Oozie workflow Job manually from the HUE to get final output

9. Pick the generated final output file from HDFS file location of the last Hive table “/user/cloudera/ds/COVID_FINAL_OUTPUT”

10. Download the final output report file and visualize it on Power BI

Resource Location

Download all resources using the following link :

https://drive.google.com/drive/folders/1AuDHNCgHN9-b7Lq8ubZukexNSuS84Hw8?usp=sharing

1. Dataset : covid-19.csv, covid-19 August.xlsx
2. HDFS : Linux shell script
3. Hive : Hql scripts
4. Oozie : sample Oozie script : workflow, job.properties and run.sh;

This part need to be implemented on Cloudera HUE using workflow tab and drag and drop actions to create a workflow pipeline

Sprints

THANK

YOU

www.sprints.ai

Sprints.ai