Big Data project

Covid-19 Analysis

A Complete pipeline and analysis
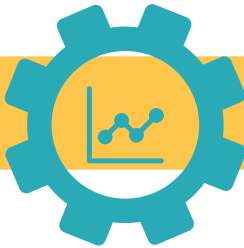

Big Data
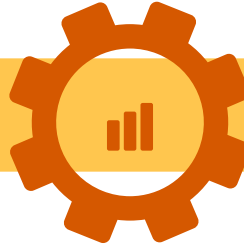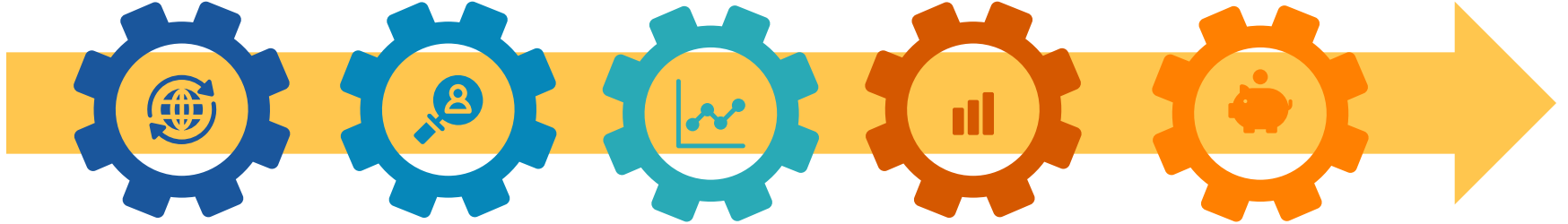
Overview

Ingestion   Storage   Analysis   Visualization   Conclusion
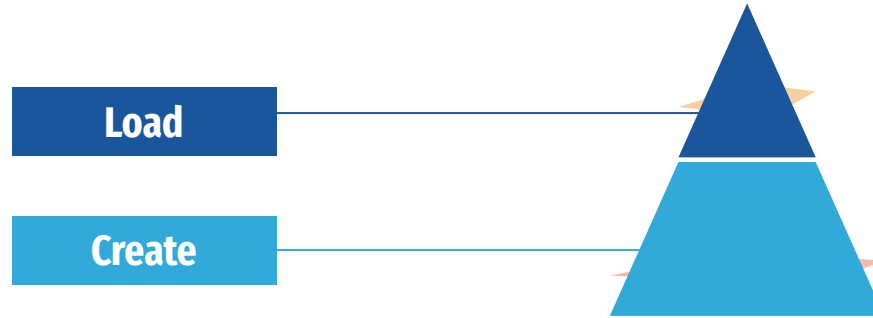
Data ingestion

Use Winscp to transfer data from win to linux

Create directories of data to trigger in data storage.

Load

Create

Data Storage

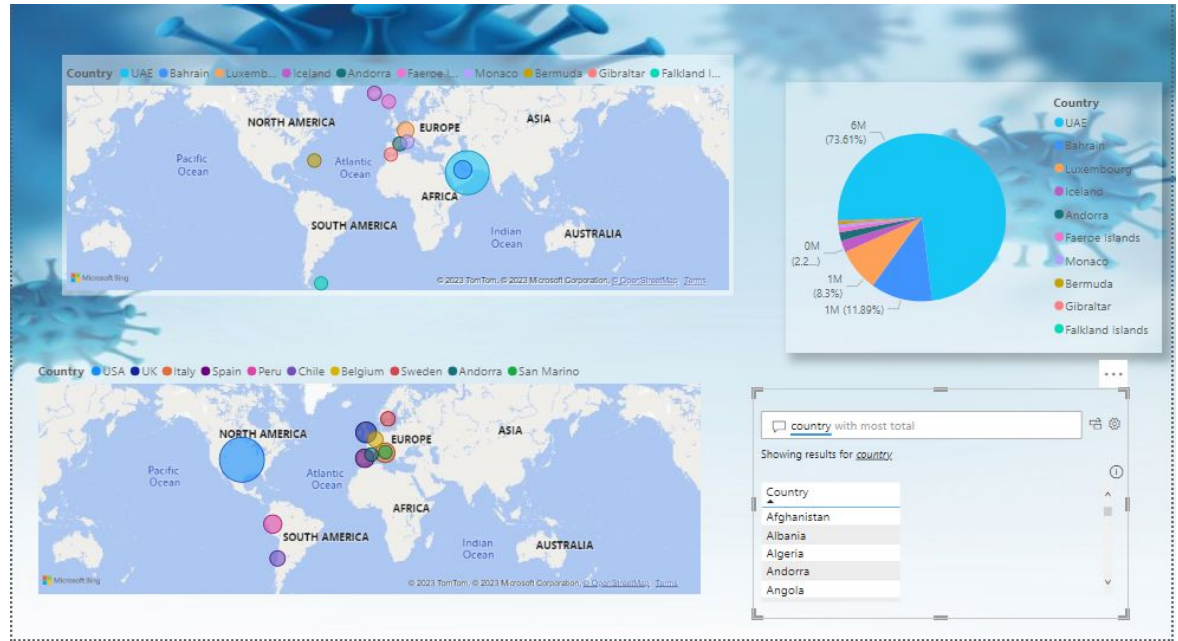Use the LOAD_TO_HDFS shell script to put it in distributed

Use

Data Analysis

We will use at the analysis phase hive to query the data :
1- create database and save the data.
2- create staging table to hold the data
3- create partition table to hold the data into partition to be easy to query
4- create output table to hold the output data .
5- load the top countries in deaths and tests into output table.
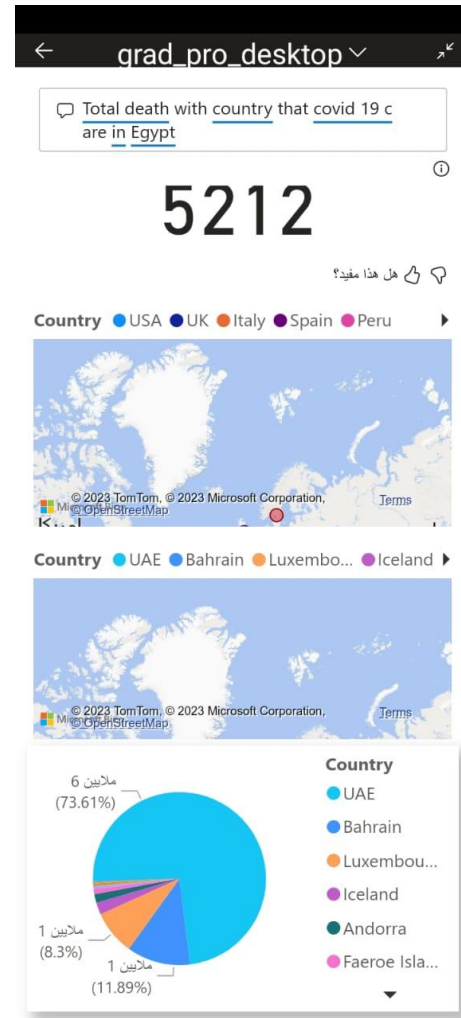6- save into csv file into local linux.
7- delete the database.(update)

Data Visulization

We will use power bi to dashboard the data output

App ver

Mobile Ver

Conclusion

1- we can say uae has better treatment as the number of tests is the hightest but it's not also accurate it has small population it may has good treatment but i can't say best .
2- we can say that usa has the most bad treatment system as it has the most number of deaths but we can't say that as it has over 200 million of people , yes it may have too much deaths but we can't also be accurate it that.
3- data need to be more augmented
4- dashboard need to relate more variables to each other to have more judgement in your insights.

Thank you

Github :   https://github.com/NeTrOn404/Big-data-Masteclass
Linkedin :   https://www.linkedin.com/in/ebrahim-rabea-b01a68215/