DATA ANALYSIS OF AIR QUALITY INDEX USING BIG DATA TOOLS

SUBMITTED BY
205219001- ABDUL JAMSHEED V
205219013- DIVYA ANWESH SAHU
205219015- M.GOWTHAM BUDHA

ABSTRACT

Urban air pollution management requires an advanced modeling and information analyzing and processing techniques. Designing a system of air quality management must be based on a distributed and adaptive problem-solving approach, which is the aim of this work. In this article, we infer real-time and detailed information regarding the air quality throughout the city of SOFIA based on pollutants and meteorological data (historical and real-time)reported by the existing monitoring stations.All these types of inputs fed the system, which uses an advanced computational module for analyzing and extracting needed information through the use of a set of algorithms and online analytical processing(OLAP)tools.The advanced features of an air quality management system are necessary and expected to be helpful to stakeholders, planners and decision makers, so that they can reliably generate a statement and prevision on the air quality, simulate and analyze more information in the decision making process. Here we use map reduce to retrieve top 5 polluted cities in sofia which is a state in BULGARIA. We will query the the results from HDFS using HIVE and visualise the results in TABLEAU.

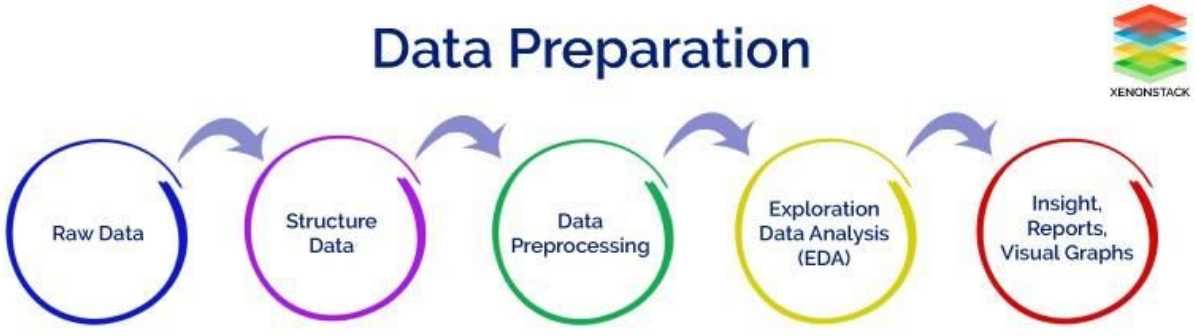
OBJECTIVE

The objective of this project is to develop an air quality management system for air areas within the city of SOFIA, which can allow appropriate measures to specific atmospheric emissions in particular areas. The air pollution data of sofia is analysed from the log files. Different sensors are used across the city to get the pollution data and stored in the file. We will this data to analyse the top 5 cities which have higher pollution. We will visualise this data using real time analysis using tableau. This real time visualization can be used by authorities to plan appropriate measures to cope up with the pollution and improve the air quality.sensor records the air pollutant index based on the two attributes. Ozone particle matter (p1) and sulphur dioxide(p2). We use mapreduce framework to analyse the top 5 places which have maximum pollution on hourly basis from the sofia city. This results are stored back in HDFS. We will visualize this results using tableau. Tableau cannot directly retrieve data from HDFS. So we query data from HDFS using HIVE and pass this data to tableau.

DATA MINING STEPS

Data mining is a five-step process:

1. identifying the source information
2. Picking the data points that need to be analyzed
3. Extracting the relevant information from the data
4. Identifying the key values from the extracted data set
5. Interpreting and reporting the results



1. Identifying the source information

We take the Sofia air quality dataset from kaggle which has been taken from https://airsofia.info/

2. Picking the data points that need to be analyzed

This dataset has many features: we will consider the following features to analyse the data

- sensor_id
- location
- lat
- lon
- timestamp
- P1 (Ozone particle matter)
- P2 (sulphur dioxide)

3. Extracting the relevant information from the data

We will take P1 and P2 and we will take the mean of this. Using this target variable we will process the data and display the results along with the latitude and longitude information.

4. Identifying the key values from the extracted data set

we use map reduce framework to extract the top 5 cities which have the highest pollution. Code for the map reduce is given below.

data_visualize.py | test_top5.py

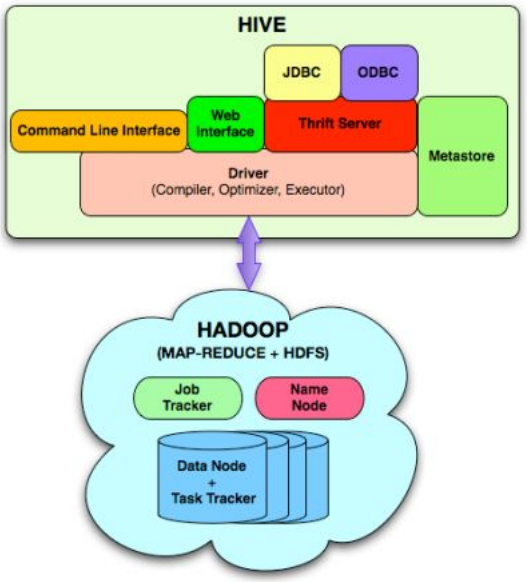
```
@author: bda
"""


#Import Dependencies
from mrjob.job import MRJob
from statistics import median

class MRWordCount(MRJob):

  def mapper(self,_,lines):
    id, sensor_id, location, latitude, longitude, timestamp, p1, p2 = lines.split(',')
    if(sensor_id != 'sensor_id'):
        day = timestamp[0:10]
        pm = p1

        yield day, str(pm) +' ' + latitude + ' ' + longitude

  def reducer(self,key,values):
      list1 = list(values)
      list1.sort(reverse = True)

      yield key, list1[0:5]

if __name__ == '__main__':
    MRWordCount.run()
```

5. Interpreting and reporting the results

We use tableau to visualise the results. Tableau is a real time visualization tool which is popularly used in many companies. Tableau cannot use data directly from the HDFS . So we use HIVE to extract data from HDFS and this data is used to visualise by tableau.

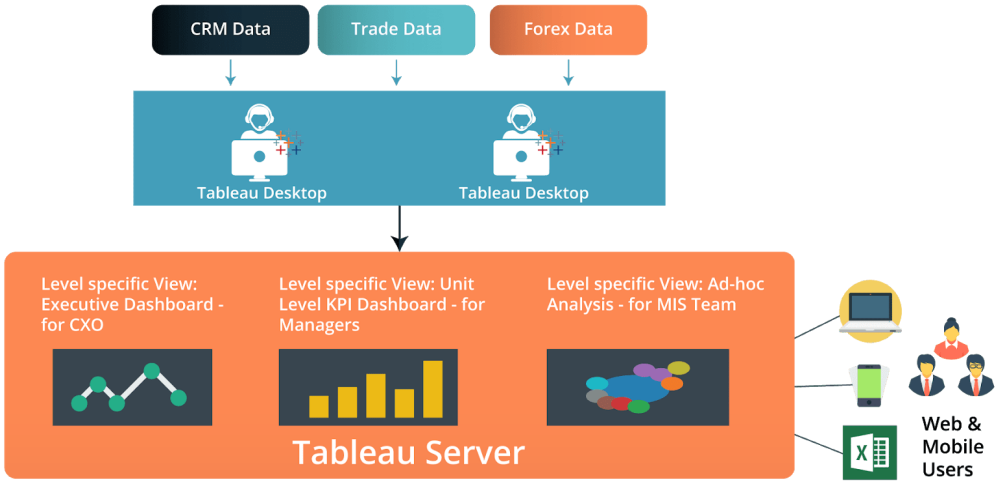
HIVE

Hive is used to query data from HDFS, since tableau cannot process the data from HDFS directly.a data warehouse must be integrated with the Hadoop engine to exploit its MapReduce parallel processing power. Once data is retrieved and stored in the data warehouse of HIVE, it is sent to Tableau for data visualisation



DATA VISUALIZATION

Tableau is used for data visualization



using tableau real time visualization of the data is possible. This visualization can be used by authorities to plan the remedial measures and check the changes occurred in those areas. This will help to improve the data quality of the sofia.

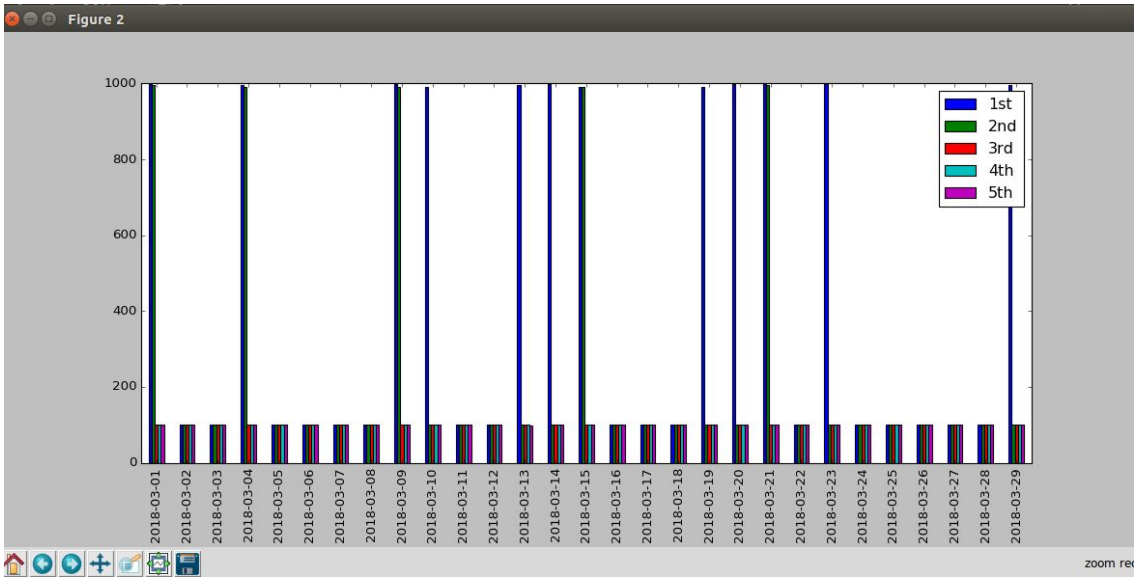
Since tableau is not open source and it requires high performance system we visualize the results using python.

```
(base) bda@hp-HP-ProDesk-400-G3-SFF:~$ sudo gdebi -n tableau-server-2019-4-0_amd64.deb
[sudo] password for bda:
Reading package lists... Done
Building dependency tree
Reading state information... Done
Reading state information... Done
(Reading database ... 287851 files and directories currently installed.)
Preparing to unpack tableau-server-2019-4-0_amd64.deb ...

--Your system does NOT meet the minimum system requirements for Tableau Server--
Tableau Server requires these minimum hardware requirements: http://www.tableau.com/products/server/specs
Either try the install on a different computer, or explore our other options for running a trial of Tableau Server: http://www.tabl
u.com/products/server/download
Tableau Server requires at least 16 GB memory to run, but found only 8 GB of memory.
Tableau Server runs best with at least 8 cores, but found only 4 core(s).
To run a Tableau Server cluster, you must disable temporary IPv6 addresses on all nodes in the cluster. For details, see:
http://kb.tableau.com/articles/knowledgebase/temporary-ipv6 (Disabling temporary IPv6 addresses)
```

visualization using python

data_visualize.py | mrjob_sal.py

```
"""
#hadoop fs -get hdfs:///dibya/output_top5_1/part-00000 output_top5.txt


import pandas as pd
import re
top5_df = pd.read_csv('output_top5.txt', sep="\t", header=None)

#data.columns = ["a", "b", "c", "etc."]
years = [i for i in top5_df[0]]
perday_pmvalues = []
for i in range(len(top5_df)):
    temp = str(top5_df[1][i]).strip("[]")
    temp = re.sub('"','',temp)
    temp = temp.split(',')
    pm_values = []
    for j in temp:
        j = j.strip().split()
        pm_values.append(float(j[0]))
    perday_pmvalues.append(pm_values)

df = pd.DataFrame(perday_pmvalues, columns = ['1st', '2nd', '3rd', '4th', '5th'], index = years)

df.plot.bar()


```



CONCLUSION

The growth in the volume of information exchange in the world engages huge quantity of data processing. Most of the organization, including millions of customers needs a system to process a big amount of data daily. Such system must meet the following requirement
1. Fast data loading
2. Fast query processing
3. Highly efficient storage utilization

In order to satisfy this big data challenges in terms of fast processing and managing highly varying and big amount of data, a data warehouse must be integrated with the Hadoop engine to exploit its MapReduce parallel processing power.We have, through this paper discussed the end to end data mining steps using different data analytics tools and learning the working of it. We have visualised the output which can be used to analyse and suggest suitable measures to cope up with the increasing air pollution all over the world.