Analysing the Crime rates in Republic of Ireland using HDFS

.

*Vineeth Mathangi,*

B00828528,

[Mathangi-V@ulster.ac.uk](mailto:Mathangi-V@ulster.ac.uk)

*Msc Internet Of Things FT,*

University of Ulster.

*Abstract*— At present, the current rise in data size in the era of internet, big data storage, analysis, and processing are becoming hot subjects in research and business. To accomplish application processing, most large data processing platforms use the MapReduce programming model. Hadoop's implementation and analysis methods, for example, are as continues to follow HDFS gathers data and stores it in distributed storage systems, which are clustered storage nodes. The high computational nodes then perform map operations on the data read from the storage nodes. Furthermore, the compute nodes also communicate with one another and perform reduction operations in order to obtain computation results. In the process of acquiring and storing data, storage nodes largely perform I/O optimization tasks, hence their computer resources aren't fully utilized. This piece of paper shows that processing of the bigdata analysis based on the HDFS (Hadoop distributed file system) and the important visualizations can be seen in Zeppelin . This concept can help reduce disc transfer and network connection data sizes, as well as program runtime. performances.

Keywords—Crime related datasets, HDFS (Hadoop distributed file system), Ambari, Putty ,Zeppelin.

# Introduction (*Heading 1*)

**BIGDATA**: The digital traces or data that arise in this digital era is referred to as big data. Small data, such as sports scores and employee shift schedules, is available in small quantities and can be readily absorbed by those with little or no computer processing ability. Small data is usually simple to keep because information is usually available in an organized format, like first and last identity, address, gender, and so on, and can be retained within an enterprise's own technology in a data center. Big data, on the other hand, is created in vast amounts and arrives with little or no organization [1]. Let us imagine that structured data is comes under all kind of files and documents, however audio and medical records could be unstructured. Data's difficult and interpreting and presenting it for human consumption necessitates the use of specialized tools. Big data is not only vast, but it is also continuously collected and gets bigger and bigger in a short amount of time. Finally, due to its size, big data is stored in the cloud or on dedicated server farms.

**IRELAND CRIME RATES**: Due to people involved or their long-term consequences on society or legislation, a few important crimes in Ireland have gained renown. Ireland had the 11th lowest violent crime rate in Europe and the 23rd lowest percentage worldwide as of 2019[2].

**HADOOP:** Apache Hadoop is a platform that is both free and open source. Hadoop is becoming more popular in the analytics business, as well as in several fields of data storage and processing, due to the simplicity and cost-effectiveness. The Apache software foundation released Hadoop 1.0, the first version of Hadoop. The Apache Hadoop framework uses a distributed processing paradigm to process the large amount of data sets across a computer cluster. This framework has its own and simple computational model which is extremely scalable and fault tolerant. Meanwhile, The Hadoop-2 framework is made up of four essential components they are Hadoop common is a utility library for other modules, Hadoop distributed file system (HDFS) is used to store massive amounts of data and provides high throughput access to application data, Hadoop YARN is a framework for job scheduling and cluster Resource Management. secondly, Hadoop MapReduce is a YARN-based system for parallel processing of large datasets [3]. As keeping all this into There are classified into two parts to MapReduce Map and Reduce. However, The Hadoop system has some characteristics, according to the capabilities, such as volume, velocity and veracity [4].

**HADOOP ECOSYSTEM:** Back then 2004, MapReduce and some GFS techniques were introduced by the Google and then later an open source platform was introduced which is said to be Apache Hadoop, scripted in java language. Moreover, Hadoop is everywhere, it was used by large companies such as Facebook, twitter and LinkedIn etc. Hadoop particularly used for large sets amount of data and is to overcome with the storage, processing issues [5].

*Components of Hadoop Eco-system: Hadoop* ecosystem was particularly based on the two strong tools such as HDFS & MapReduce and other different components are Pig, HBase, Hive, Sqoop. However Hadoop distributed file system (HDFS) [6]are inspiration of the Google systems and also MapReduce is best for processing the tasks which can handle parallelly because of engine can permits the parallel processing of the nodes . following figure illustrates each components in eco system [7].

*Graphical user interface

Description automatically generated with medium confidence*

In Map Reduce framework there are two set of key/values at first input values are perfectly mapped into output as a key value. below architecture design gives an brief example[8].

Diagram

Description automatically generated

## Methods are to be used :

1) Dockers

2) Putty

3) Gitbash

4) Ambari

5) HIVE 2.0

6) Zeppelin(visualizations)

*Execution process:* Docker is a platform that certainly provides and satisfies the software in containers by using OS virtualizations. Moreover, containers are different from one another build with own libraries and security configurations.

*Prerequisites*: Minimum 16 Gb of RAM needed to support the dockers. With level of intel 5 cores.

*Process* :

*Graphical user interface, application

Description automatically generated*

Above figure provides the information about docker container home. Well container is like a unit of software package that carries code and other dependencies to the applications and runs easily from the computing environment to another environment which is secure .it has the capability of saving all sort of data.

However, docker container runs the sandbox which secures the data in a safe mode and need to start the engine in order to analyze the data.

Now the interesting part begins with putty, where destination port is said to be 2222, similarly the hostname destination address can be sandbox.hortonworks.com .then by setting SSH connection type with the default address 192.168.1.10 putty starts a connection with Horton works. However, background connection going on with the http://localhost:1080 server that complete page redirects to the Ambari.

Graphical user interface

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Above figures shows the evidence of hostname and port.

Ambari is a interface which helps for Apache Hadoop system to simplify the management and monitoring. And the simplification is easy by use of web UI. There are some particular steps need to be followed, when server redirects to the Ambari homepage first steps is to make sure every server is perfectly running and alerts to be zero in order to work server properly, and then by viewing the file view system in Hive 2.0, chosen new datasets need to be uploaded in the admin folder.

Moreover, after uploading the datasets better to preview the tables whether uploaded or not. below figure illustrates the query which is used to open the table.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Zeppelin is used to show the visualizations with query commands. Now already uploaded the datasets.

Command that I have used to show the output

% jdfc (HIVE)

SELECT \* FROM IRECR ;

Figure shows the both the divisions and offenses in Ireland in the year 2018,each region has a different offenses let us see more charts with comparison.

Text

Description automatically generated

Graphical user interface, application

Description automatically generated

Crime rates in Ireland in the year 2018 Attempts/threats, Dangerous and Sexual offences are more when compared to all .

Chart

Description automatically generated

Chart in the year 2018 gives the same information but in different prospective.

A screenshot of a computer

Description automatically generated with medium confidence

By comparing the 2018 and 2019, the offenses in Ireland attempts/threats and sexual offenses is less in 2019 .however, offenses against has become more and more.

**Conclusion:**

In conclusion, by noting all the datasets and their visualizations have been stored in order to use for secondary purpose. However, more visualizations can be done by using some Machine learning techniques.

##### References

1. Big data: The next frontier for innovation, competition, and productivity. [online]. Available: http://www.mckinsey.com/insights/business\_technology/big\_data\_the \_next\_frontier\_for\_innovation .
2. https://www.nationmaster.com/country-info/compare/Ireland/United-Kingdom/Crime
3. T. P. Gondaliya., “Big Data challenges and Hadoop as one of the solution of big data with its Modules”, International Journal of Scientific & Engineering Research, Volume 5, Issue 6, June2014.
4. V. B. Bobade., “Survey Paper on Big Data and Hadoop”, International Research Journal of Engineering and Technology (IRJET), Volume: 03 Issue: 01, Jan-2016.
5. Ghemawat S, Gobioff H, Leung S-T: The Google file system. 19th ACM Symposium on Operating Systems Principles Lake George, NY: ACM Press; 2003.
6. S. Kaisler et al., “Big Data: Issues and Challenges Moving Forward”, DOI 10.1109/HICSS.2013.645, 46th Hawaii International Conference on System Sciences, 2012.
7. Cuzzocrea, A., Song, I., Davis, K.C.: Analytics over LargeScale Multidimensional Data: The Big Data Revolution! In: Proceedings of the ACM International Workshop on Data Warehousing and OLAP, pp. 101–104 (2011)
8. EMC: Data Science and Big Data Analytics. In: EMC Education Services, pp. 1–508 (2012)

**.**