International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 09 | Sep 2020 www.irjet.net p-ISSN: 2395-0072

IDENTIFICATION AND ANALYTICS OF TOURIST DATA USING BIG DATA

Madhuvandhy V[1], Suriya K[2], Swetha S[3], Mohan Raj K R

[1,2,3]Student, [4]Assistant Professor,
[1-4]Department of Information Technology, Velammal Engineering College

---

***Abstract*—** The rapid development of the tourist industry and the insufficient information regarding tourists has placed tremendous pressure on traffic in scenic areas. To ensure their wellbeing a medium is necessary .This medium acts as analysis for tourist identification and preference using city scale transport data. This can be used by tour agencies, transport operators, tourist and also by the ordinary stakeholders using big data technology.

***Keywords*—** Availability of proper information to every user, implementing best source of analysis, reducing time consumption, graph analysis generation, providing data to improve tourist industry.

1. INTRODUCTION

In this project, we are analysing Transportation System data by using hadoop tool along with some hadoop ecosystems like hdfs, map reduce, sqoop, hive and pig. By using these tools we can process no limitation of data, no data lost problem, and can get high throughput, less maintenance cost. It is an open source software and compatible on all the platforms since it is Java based. This application contains separate modules for each section with functionalities such as identification and analysis.

2. EXISTING SYSTEM

Existing concept deals with providing backend by using MySQL. MySQL is a relational database which contains lot of drawbacks i.e. data limitation is that processing time is high when the data is huge and once data is lost it cannot be recovered. It often consumes a lot of unnecessary delays and they don't maintain a centralised source of record. This system lacks in processing large amount of data in a single execution and getting results will take more time and maintenance cost is very high.

3. PROPOSED SYSTEM

The proposed system contains a backend application that can be used by user at any level to plan their trips in a time efficient way. The analysis is viewed in a graph model for the easy understanding. The transport data's are frequently updated by using the big data techniques. Proposed concept deals with providing database by using Hadoop tool where we can analyse no limitation of data and simply add number of machines to the cluster. This provides the users the ability to get results with less time, high throughput and maintenance cost is also very less. This is achieved by using the concepts of joins, partitions and bucketing techniques in Hadoop. Hadoop is an open source framework which has overseen by the apache software foundation and it is used for storing and processing huge datasets with a cluster of commodity hardware. The Hadoop tool contains two things one is hdfs and the other one is map reduce. The map reduce is used to reduce the analysis, helps in quick access by using three stages such as map stage, shuffle stage and reduce stage. The HDFS is Hadoop Distributed File System where files and data's are stored in different systems and any loss of data can be recovered easily. Other Hadoop ecosystems like sqoop, hive and pig are also used.

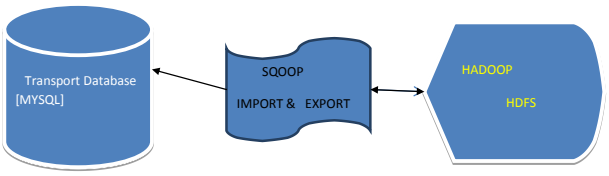
4. PROPOSED WORKFLOW

The project is carried out in following modules:

1. Preprocessing Transport System DB
2. Storage
3. Analyse Query
4. Analyse Latin Script(PIG)
5. Processing(MAP REDUCE)

In Preprocessing Transport System Database module, by using the Microsoft Excel different fields are analysed and then it is converted into comma delimited format which is said to be csv (Comma Separator Value) file and moved to MYSQL backup through Database to avoid loss of data.

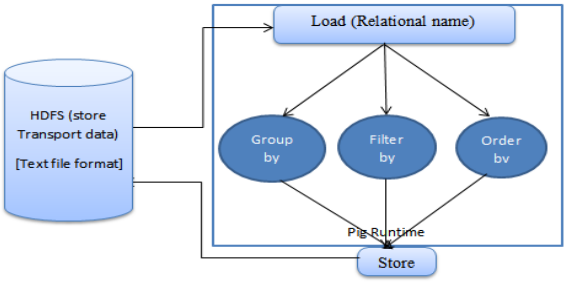
In Storage module, the backup data which has been stored in MYSQL are collected and importing all those data by the using the sqoop commands to HDFS (Hadoop Distributed File System). Now all the data are stored in HDFS were it is ready to get processed buy use of hive.



In Analyze Query module, the data from HDFS to HIVE, by the use of sqoop import command where hive is ready to analyse. HIVE can process only structured data to analyse.

By extracting only the meaningful data and neglecting unclenched data, the data's are analysed in more efficient manner.

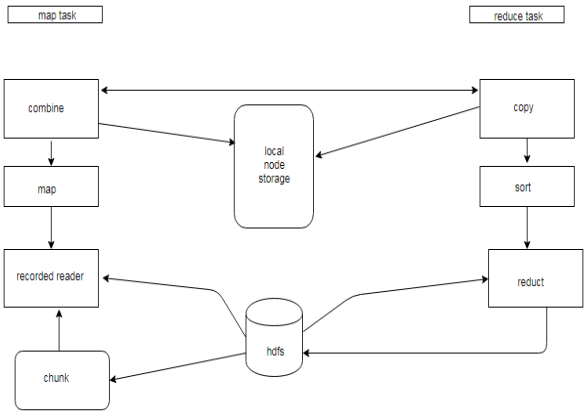
The Analysis Latin Script module are used for analyse of user transport system using Pig The programmers need to write scripts using Pig Latin Language and execute them in interactive mode using the Grunt Shell. The scripts are converted to Map and Reduce tasks internally. By invoking the Grunt Shell, the Pig scripts in the shell are runned.



Pig Latin Statement take a relation as input and produce another relation as output. Whenever a load statement is entered in a grunt shell, its semantic checking will be carried out. The contents of the schema are viewed by using the dump operator. After the dump operation, the MapReduce job for loading the data into the file system will be carried out.

Pig provides many build-in operators to support data operations like grouping, filters, ordering, etc.

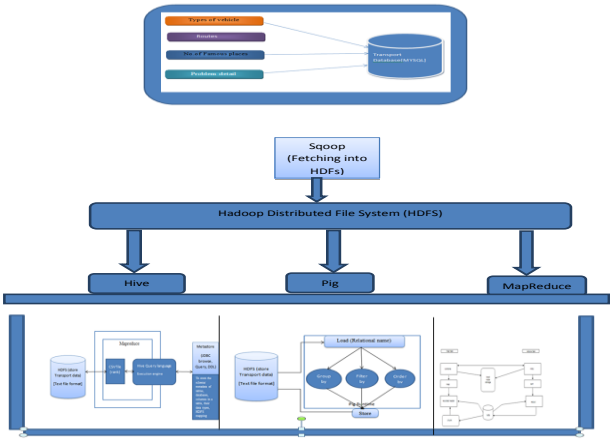
In processing module, Map Reduce algorithm are used. The program executes in three stages. The map or mapper's job is to process the input data and the data's are stored in Hadoop File System(HDFS).The inputs are passed line by line. It processes the data and creates several small chunks of data. The next stage is the combination of both the shuffle and the reduce stage. The reducer is to process the data that comes from the mapper stage. Finally produces a new set of output, which will be stored in HDFS.



5. USER INTERFACE DESIGN

Using this application the user is provided with user friendly portal that can be used to provide data's of transportation for the tourist spots. It also contains graphical representation of data based on the tourist data using big data. The end user can make use of this for visiting more spots in a time efficient manner.

6. TECHNOLOGY AND SYSTEM ARCHITECTURE



The system architecture defines all the techniques used for the identification and the analyse of the data related to tourism using big data. The various hardware processor like Pentium IV 2.6 GHz and Inter Core 2 Duo are used. The software like Hadoop framework and the Cent OS are used as the Operating System. Along with the Database support of MYSQL the entire system architecture are build in.

7. CONCLUSION AND FUTURE WORK

In this world of developing technologies everything has been computerized. With a large number of work opportunities the Human workforce has increased and the stress level has also increased. Thus there is a need for the human to be relaxed and that's the reason why people prefer to visit certain places i.e. tour.

In future this can be enhanced in Apache Spark. Apache Spark gets result hundred times faster than Hadoop. It also supports Java, Python API for the development. It is a open source processing engine built for analytics.

REFERENCES

[1] Zhen Zhang, IEEE, Mianzhi Wang, "Optimal Transport in Reproducing Kernel Hilbert Spaces: Theory and Applications".

[2] N. Mohamed and J. AI- Jarood, "Real-time big data analytics: Applications and challenges".

[3] Chang YU , Zhao-Cheng HE, "Spatial-temporal daily frequent trip pattern of public transport passengers using smart card data.