International Journal of Advanced Research Methodology in Engineering & Technology,
Volume 1, Issue 1, December 2016

DATA PREPROCESSING ENVIRONMENT SETUP USING HADOOP ECOSYSTEM HIVE

B.Sasi madhumitha/ III Year – CSE/Knowledge Institute of Technology, Salem

D.Swathi /III Year – CSE /Knowledge Institute of Technology, Salem

Mr.S.S.Aravinth,Asst.Prof/CSE, Knowledge Institute of Technology, Salem

Mail ID: ssacse@kiot.ac.in Mobile: 098944 – 48683

ABSTRACT:

Hbase is the one which has the benefits of distributed storage system in the open-source environment and HDFS (Hadoop Distributed File System) is the highly fault tolerance system which runs on the hardware is the unique function compared to other distributed system. But the Hbase distributed storage run on just above the HDFS as part of the Apache Software Foundation's Apache Hadoop. Hbase is the one of the application of the Hadoop which facilitate the users for read and write random access of large data. In modern times Hbase is used by the Google to manage the huge amount of the structured data. Hbase has the functionalities of the big table and also than it so Hbase is used to support the big table. Hbase is the developed by the Java language where the Hbase is similar to the other non SQL languages. Few features of the Hbase on the column basis are compression, in-memory operation and the Boom-Filters where these features are expect to grab by the Hive in the Hadoop in the future. In the Hbase the Table Input are get in the some Format which facilitate the Map Reduce function to takes place and the output of the function is given to the tables of the Hbase. This paper includes the Introduction to the Hbase, Installation of Hbase, Architecture of Hbase, RDMS, how Hbase overcome the problems of RDMS(Relational Database Management System).

KEYWORDS:

BIG DATA, USER DATA, XML DATA, HIVE TABLES , SQL TABLES & ANALYSIS.

I. INTRODUCTION:

Big data analytic has been emerging as a powerful tool for businesses as today's world is generating huge volumes of data. Big data is an emerging technology that allows extracting useful information from huge volume of structured data such as sensor values, traditional databases, GPS data as well as unstructured data such as social media data, multimedia data which is not possible by traditional technologies. Future predictions can also be made by Big data.

Hadoop is the poster child of Big data. Hadoop is used for data storage. It allows to store as much data as wanted, in whatever form needed by adding more servers to a hadoop cluster. Every server adds extra storage and processing power to the cluster. This allows low cost data storage using hadoop than existing methods.

2 PROPOSED SYSTEM

The browsing history should be collected to know more about the domains and webpages that students

browse while they are surfing internet. This data is a treasure trove to know the browsing behavior and a browsing effectiveness score can be assigned to every student to measure their outcomes from browsing. Good URLs can be suggested to people who don't use the internet effectively. There comes the recommendation part. To collect the browsing logs, there should be some network levers or ISP settings to get the list of webpages being surfed and IP from which the traffic originated. By using all these data the insights have to be found and the browsing behavior of the student community has to be optimized. This would also help the students to get good exposure. It's all about enriching browsing experience

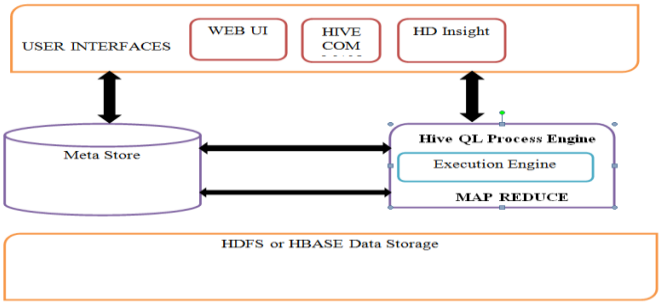


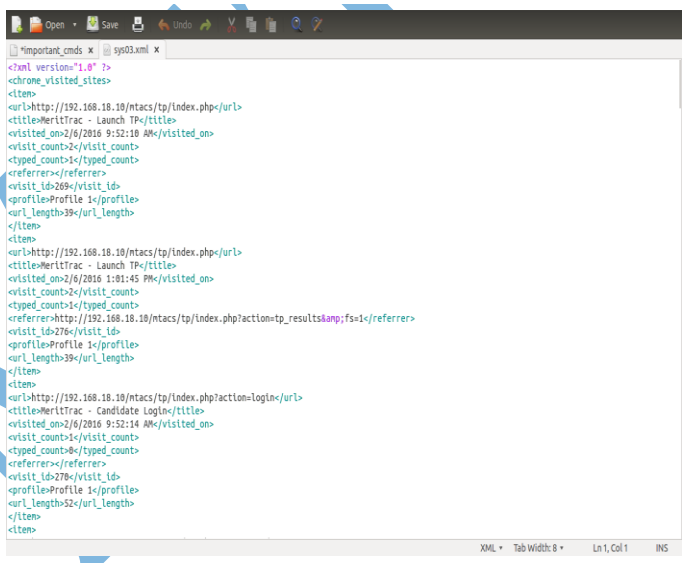
Figure 2.1 Block Replication & Data nodes

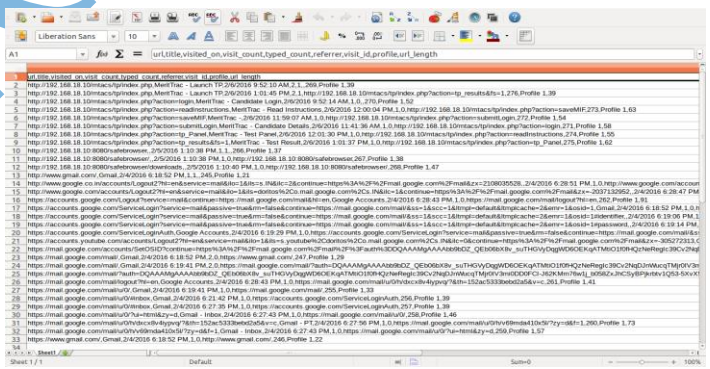
3 MODULES

- Hadoop Cluster Creation
- Browsing history Data Collection
- Creating Tables and Inserting Values in the HIVE tables
- Showing data in HIVE tables and Maintaining Meta data in Tables
- Data Ingestion, Data History Creation and History Table Manipulation

3.1 Hadoop Cluster Creation

A Hadoop cluster is a special type of computational cluster designed specifically for storing and analyzing huge amounts of unstructured data in a distributed computing environment. Such clusters run Hadoop's open source distributed processing software on low-cost commodity computers. Typically one machine in the cluster is designated as





the Name Node and another machine as the Job Tracker, these are the masters. The rest of the machines in the cluster act as both Data Node and Task Tracker, these are the slaves.

Hadoop clusters are often referred to as "shared nothing" systems because the only thing that is shared between nodes is the network that connects

them. Hadoop clusters are known for boosting the speed of data analysis applications. They are also highly scalable: If a cluster's processing power is overwhelmed by growing volumes of data, additional cluster nodes can be added to increase throughput. Hadoop clusters also are highly resistant to failure because each piece of data is copied onto other cluster nodes, which ensures that the data is not lost if one node fails.

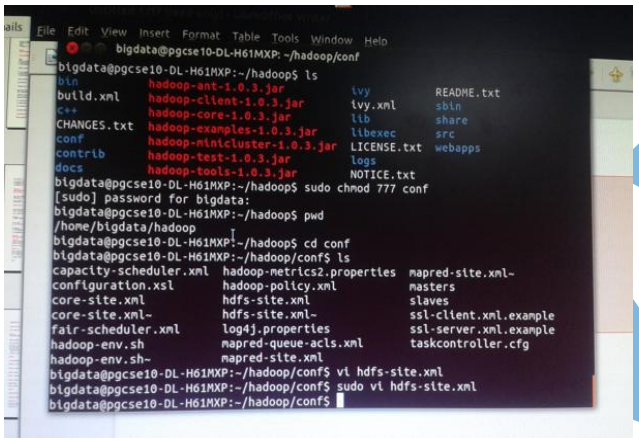


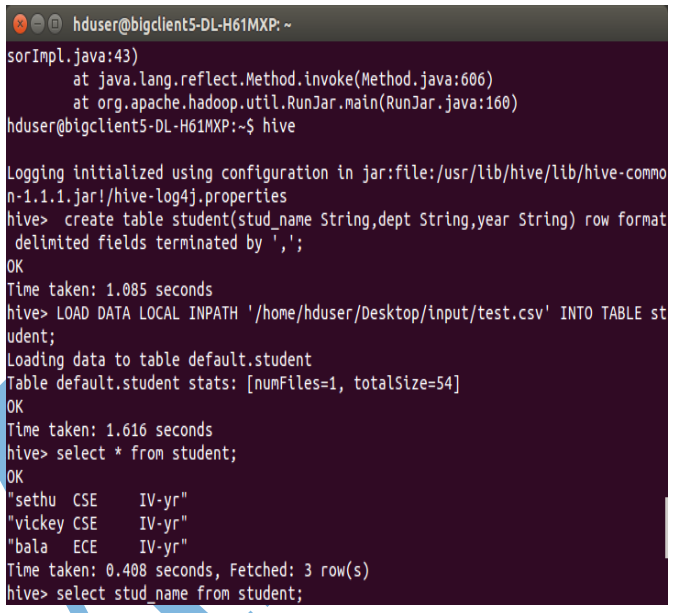
Figure 3.1.1 Block Replication & Data nodes

3.2 Browsing history Data Collection

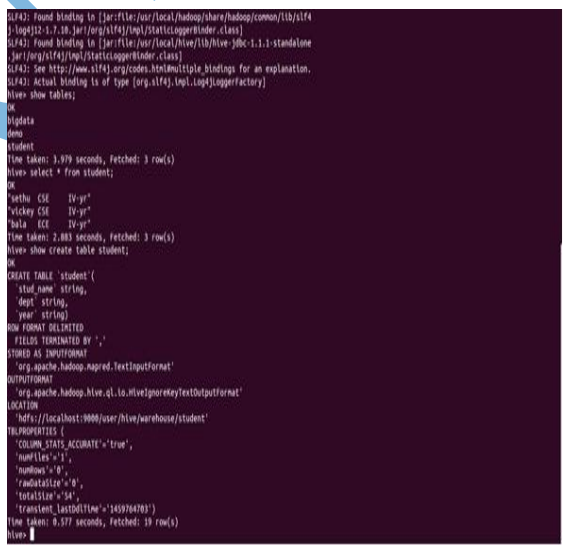
This module is used for preprocessing the data and predict what the user needs. The data have been collected using browser history analyzer tool. It gives the output in the expected format. It reveals URL, time length, source of the site etc.., which will be easier to analyze the data.

Figure 3.2.2 CSV Format

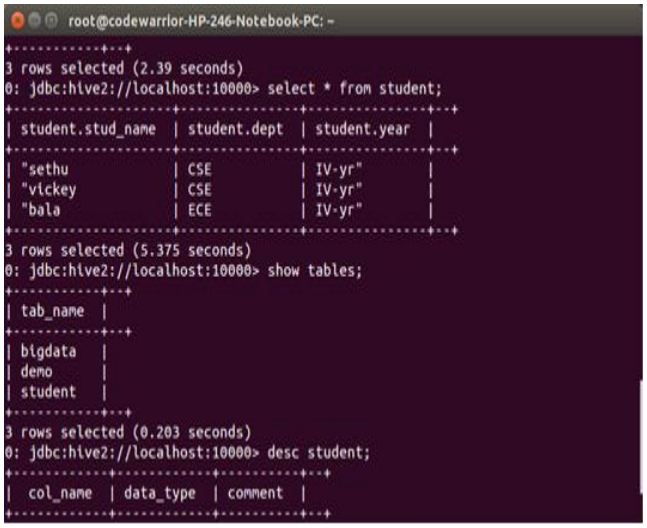
3.3 Creating Table in HIVE



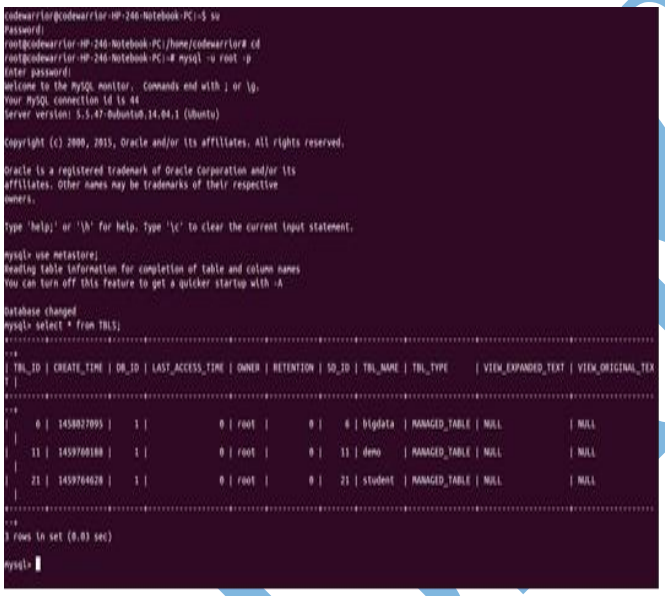
3.4 Inserting Values into the table



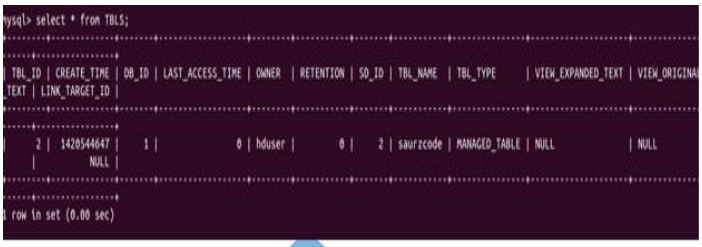
3.5 Showing Values in the table



3.6 Handling Hive Metadata with my SQL



On MySQL database the names of Hive tables

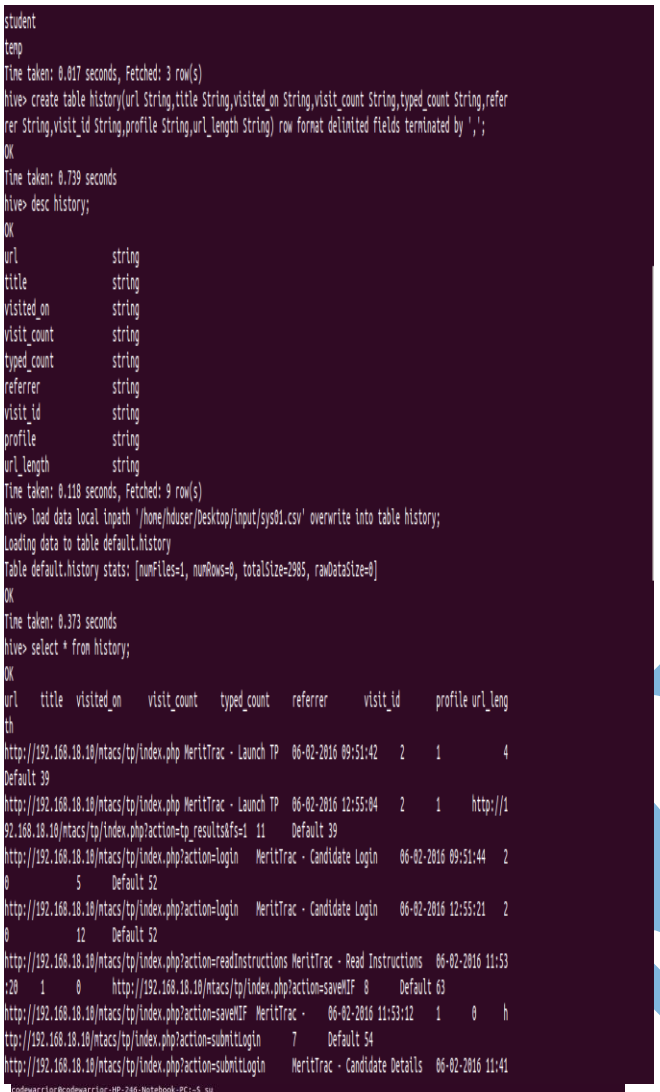


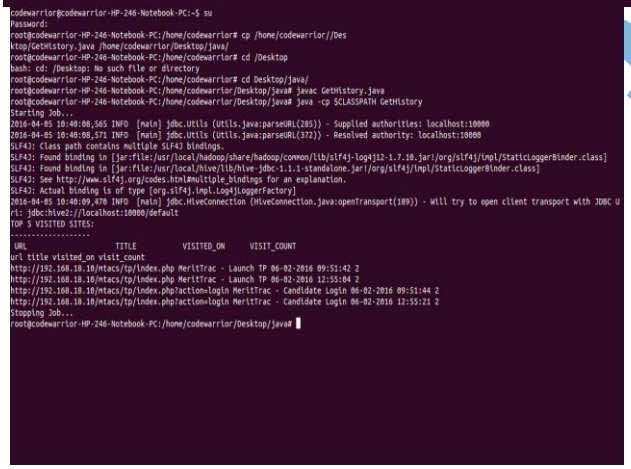
3.7 Data Ingestion, Data History and History Table Generation

The **Hadoop ecosystem** includes other tools to address particular needs. Hive is a SQL dialect and Pig is a dataflow language for that hide the tedium of creating Map Reduce jobs behind higher-level abstractions more appropriate for user goals. Zookeeper is used for federating services and Oozie is a scheduling system.

Hive is an open-source data warehouse system for querying and analyzing large datasets stored in Hadoop files. Hadoop is a framework for handling large datasets in a distributed computing environment. Hive has three main functions: data summarization, query and analysis. It supports queries expressed in a language called Hive QL, which automatically translates SQL-like queries into Map Reduce jobs executed on Hadoop. In addition, Hive QL supports custom Map Reduce scripts to be plugged into queries. Hive also enables data serialization/deserialization and increases flexibility in schema design by including a system catalog called Hive-Meta store.

3.8 Getting History





3.9 History Table Manipulation

4 CONCLUSION AND FUTURE ENHANCEMENT

4.1 CONCLUSION

This main objective is to analyze the student's browsing and to guide them with the analyzed source. By the analyzed source the user demand can be predicted; the most popular sites will be recommended to the students from their browsing history. Rating will be also given to the students, who are all browsing for their knowledge acquisition.

4.2 FUTURE ENHANCEMENT

In current work only minimum number of systems has been implemented. In future it will be extended to the whole college. By this way it can help all the students from various departments. In this method all the students can be monitored and will be guided to reach a successful path. In other perspective this can be used by Internet Service Provider (ISP) also. For them, they can easily trace out which users are using banded sides.

REFERENCE:

[1] Jefry Dean and Sanjay Ghemwat, MapReduce:A Flexible Data Processing Tool, Communications of the ACM, Volume 53, Issuse.1,January 2010.

[2] Jefry Dean and Sanjay Ghemwat,.MapReduce: Simplified data processing on large clusters, Communications of the ACM, Volume 51 pp. 107–113, 2008

[3] Brad Brown, Michael Chui, and James Manyika, Are you ready for the era of „big data"?,McKinseyQuaterly,Mckinsey Global Institute, October 2011.

[4] DunrenChe, MejdlSafran, and ZhiyongPeng, From Big Data to Big Data Mining: Challenges, Issues, and Opportunities, DASFAA Workshops 2013, LNCS 7827, 2013.

[5] MarcinJedyk, MAKING BIG DATA, SMALL, Using distributed systems for processing, analysing and managing large huge data sets, Software Professional"s Network, Cheshire Data systems Ltd.