IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 7 Issue 1, January 2020
ISSN (Online) 2348 – 7968 | Impact Factor (2019) – 6.248
www.ijiset.com

Using Hadoop Click Stream Analytics in E-Commerce

P.D.S.S.Lakshmi kumari [1], S.Sureshkumar [2]

Assistant professor, Department of IT, SRKR Engineering College, Bhimavaram, A.P, India[1]

Assistant professor, Department of CSE, SRKR Engineering College, Bhimavaram, A.P, India[2]

ABSTRACT: The paper entitled Click Stream analytics on e-commerce data using Hadoop, where the data is high in volume and velocity, store it for analysis in a cost-effective manner for intensified insight and decision making. Basically, Click Streams are the records based on user interaction with website or other application. Each row of the click stream contains a timestamp and an indication of what the user did. Every click performed by the user are stored in some database hence the name "Click Stream" obtained. Today's e-commerce organizations are striving to predict the analysis about their sales and services to find their business status from their beloved customers. The response from the customers based on their activities on the websites play a key role to identify their present market value and it also helps to increase their sales rate in future. These organizations store the information of all customers in their company data bases for future analysis which is referred commonly as big data, because its size increases in a dynamic proportion day by day. One of the main applications of big data intelligence which is ideal for ecommerce websites is Click Stream of data which depends on clicks. A general and old approach is to load these data and processing is by using traditional databases but takes huge time to process and also involves many complexity issues. Here in this paper click stream data is processed, analyzed with the structure of Hadoop using Sqoop,Pig, Hive and many other tools which provides large scale processing performance.

Keywords*:* Hadoop, Click Stream, Big Data, Pig, Hive &Sqoop .

I. INTRODUCTION

Big knowledge could be a assortment of enormous knowledge sets. Since it is in large size it's not economical to method those victimization ancient ways. Several problems have to be faced in processing big data such as capturing the data, storage, search, sharing, transferring, analysis etc. The development to giant data sets is extra information derived from analysis of a single large set of correlated data, as compared to separate smaller sets with the equivalent total amount of data. This helps big data in finding the relations between various fields which may help in various ways, like decision making, understanding the business trends, long term planning, fighting crime, and getting real-time roadway traffic conditions. But thanks to their correlative behavior it becomes troublesome to question them. Professionals try to create results from this vast quantity of knowledge.

This explosion of data is seen each segment within the computing industry. The Internet companies such as Google, Yahoo, Face book etc... deals with large amounts of data generated by the user and this data are in the form of blog posts, photographs, status messages, and audio/video files. There is a also huge amount of data which is indirectly generated by web sites in the form of access log files, click through events etc. Study of this data can bring functional patterns about the activities of user. Most of this data is generated frequently, and the data sets are stored temporarily for a fixed period and once they are used, they are discarded after that. According to Google, the online data on web today is 281 Exabyte, which was 5 Exabyte in 2002.There has been an amazing increase in user generated knowledge since 2005.[1]

II. RELATED WORK

Challengesin massive knowledge analysis embrace knowledge inconsistencyand unity, quantifiability, timeliness, and security before knowledge analysis, knowledge should be made. However, considering the variability of datasets in massive knowledge, the economical illustration, access, and analysis of unstructured or semi structured knowledge are still difficult. Understanding the tactic by that knowledge maybe preprocessed is very important to boost knowledge quality and also the analysis results. Datasets as usually terribly massive at many GB or a lot of, and that they originate from heterogeneous sources. Hence, current real-world databases are extremely prone to in consistent, incomplete, and clattering knowledge. Therefore, various knowledge preprocessing techniques, as well as knowledge cleanup, integration, transformation, and reduction, ought to be applied to get rid of noise and proper inconsistencies. Every sub process faces a special challenge with reference to data-driven applications. Thus, future analysis should address the remaining problems associated with confidentiality. These problems embrace encrypting massive amounts of knowledge, reducing the computation power of secret

writing algorithms, and applying totally different secret writing algorithms to heterogeneous knowledge.

Privacy is major concern in outsourced knowledge. Recently, some controversies have disclosed however some security agencies are victimization knowledge generated by people for his or her own advantages while not permission. Therefore, policies that cowl all user privacy issues ought to be developed. Furthermore, rule violators ought to be known and user knowledge shouldn't be abused or leaked.

There are two levels of click stream analysis, traffic analytics and e-commerce analytics. Traffic analytics operates at the server level and tracks how many pages are served to the user, how long it takes each page to load, how often the user hits the browser's back or stop button and how much data is transmitted before the user moves on. E-commerce-based analysis uses clickstream data to determine the effectiveness of the site as a channel-to-market. It's concerned with what pages the shopper lingers on, what the shopper puts in or takes out of a shopping cart, what items the shopper purchases, whether or not the shopper belongs to a loyalty program and uses a coupon code and the shopper's preferred method of payment.
Because an extremely large volume of data can be gathered through clickstream analysis, many e-businesses rely on big data analytics and related tools such as Hadoop to help interpret the data and generate reports for specific areas of interest. Clickstream analysis is considered to be most effective when used in conjunction with other, more traditional, market evaluation resources.

III. **PROPOSED MODEL**

Step-1

- Loading data from click stream database to HDFS & performing preprocessing using Pig.

Step-2

- Performing analytics & finding insights on click stream data using Hive.

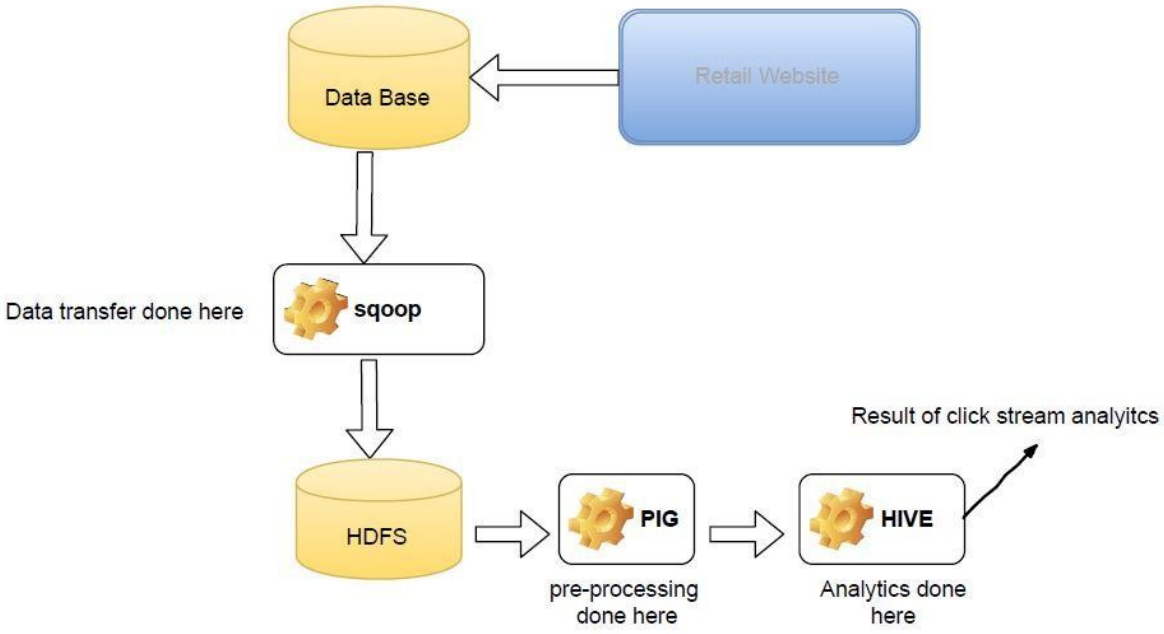
It includes:

1) Finding top most clicking items.
2) Finding top most selling items.
3) Finding out common items in top selling and also top clicking
4) Giving Day wise sales report.

5) Finding hourly sales report. etc.

Step-3

Visualizing Clickstream results using R



IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 7 Issue 1, January 2020
ISSN (Online) 2348 – 7968 | Impact Factor (2019) – 6.248
www.ijiset.com

IV. EXPERIMENTAL RESULTS

At first User provides complete click stream clicks & buys data sets and those files are huge in content and contains nearly about 1lakh rows and 4 columns. These below are the example contents for clicks and buy data sets.

1	2014-04-07T10:51:09.277Z	214536502	0
1	2014-04-07T10:54:09.868Z	214536500	0
1	2014-04-07T10:54:46.998Z	214536506	0
1	2014-04-07T10:57:00.306Z	214577561	0
2	2014-04-07T13:56:37.614Z	214662742	0
2	2014-04-07T13:57:19.373Z	214662742	0

fig 4.1 clickstream clicks input data (session ID, Timestamp ,Itemid , category)

420374	2014-04-06T18:44:58.314Z	214537888	12462	1
420374	2014-04-06T18:44:58.325Z	214537850	10471	1
281626	2014-04-06T09:40:13.032Z	214535653	1883	1
420368	2014-04-04T06:13:28.848Z	214530572	6073	1
420368	2014-04-04T06:13:28.858Z	214835025	2617	1
140806	2014-04-07T09:22:28.132Z	214668193	523	1

(fig 4.2 clickstream buys input data(itemid,TimeStamp,Sessionid,price,Quantity)) Here we use Hadoop Framework for processing clickstream data and finding out some interesting insights about that data.

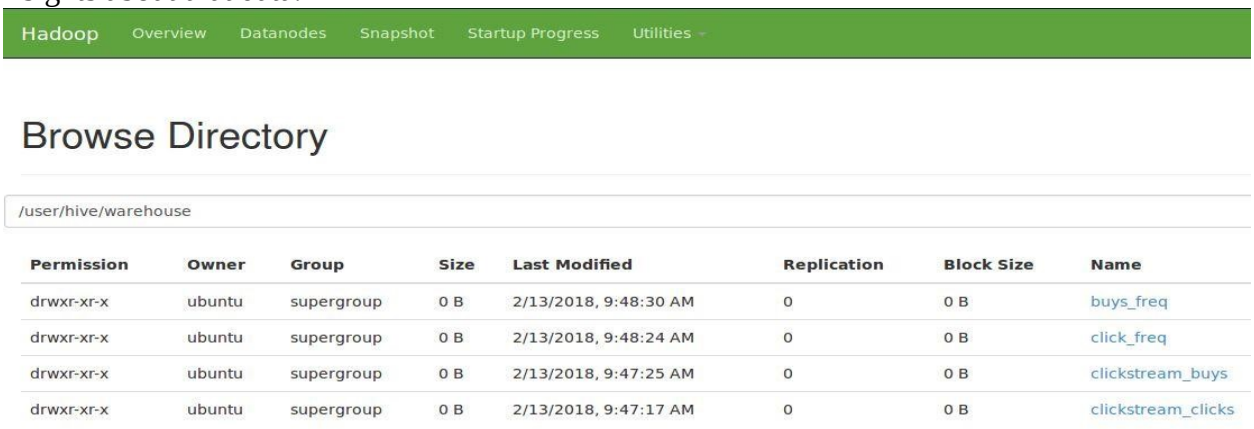


Fig 4.3 Hive implementation output [STORAGE VIEW]

Fig 4.3 shows the file structure the showcase the hive output. Simply it contains the data related to top sold items and top clicked items list of an e-commerce data set that we have taken as a reference source for clickstream data. The below Fig 5.4 indicates the same output in through CLI(command line interface).

```
2018-02-13 09:47:56,960 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local873313990_0004
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 263920 HDFS Write: 0 SUCCESS
Stage-Stage-2:  HDFS Read: 263920 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
214821277       16
214826801       14
214826627       11
214821290       10
214826955       10
214826606       9
214839313       9
214821285       9
214826705       7
214716930       7
214826803       6
214821272       5
214753507       5
214820231       5
214716932       5
Time taken: 14.497 seconds, Fetched: 15 row(s)
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different
spark, tez) or using Hive 1.X releases.
Query ID = ubuntu_20180213094757_e6a25846-d4a2-45b0-8cee-cb48125463e6
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
```

Fig 4.4 Hive job running [CLI - view]

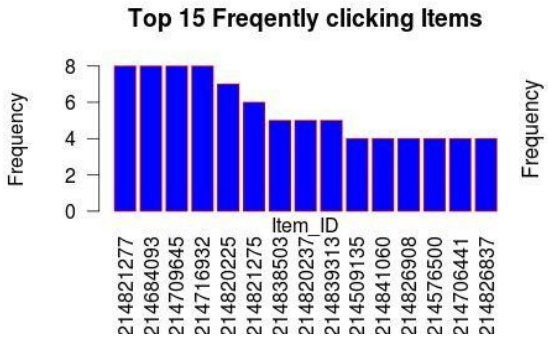


Fig 4.5 frequently clicking items list of Top15

Fig 4.6 frequently buying items list of Top 15

Fig 4.5, 4.6 and 4.7 are obtained from Hadoop streaming using R, which provided effective visualization.

- Fig 4.5 showcases frequently clicked items list of Top 15 where
- Fig 4.6 showcases frequently buying items list of Top15
- Fig 4.7 sales rate based on hour .

These results would be useful in case of recommending those top sold items to the existing users in real time. Consider a case where a particular new item has to be introduced in to the online market, then the company will absolutely query about the sales rate based on time constraints. Fig 4.7 provides the sales rate based on hour & it indicates the most happening hour which the sales rate would be high. These kinds of many real time and value based insights could be found using this method.

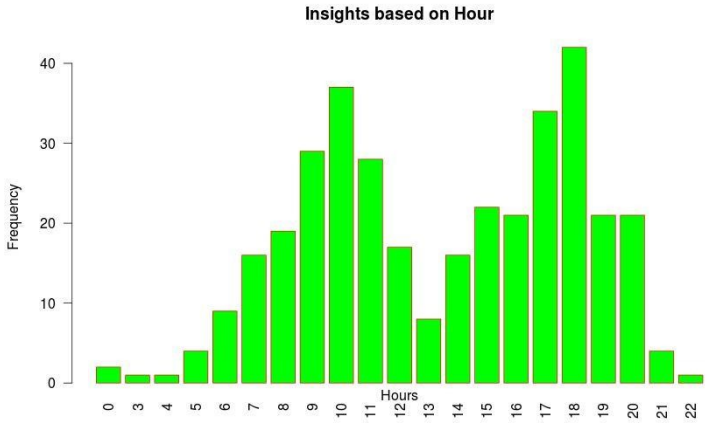


Fig 4.7 Hour based Insights on clickstream data

V. CONCLUSION

The whole information is quickly growing on a social website, e-commerce and many other sites. With the development in web technology and ease of use of raw data infrastructure, the demand for analyze of Click stream information over web has increased a lot. Click stream information Analytics play an vital role in a wide diversity of applications such as decision support systems, profile-based marketing, to know about the visitor and path where he comes from .A group of frameworks and techniques are there to handle Click stream information. In this report a method to tackle the behavior of the visitor has been derived. There are already existing methods to find the behavior of use but they were somehow not so accurate and efficient according to this problem. So, a more refined method has been presented which uses Map Reduce method to refine the raw data to find the behavior of the visitor. The size of the data in this becomes too big. It would be ineffective to access that data sequentially. So, the Map Reduce has been accustomed execute the information parallelly.It improves the solution in terms of time complexity to a great extends because the data is being processed parallel.

A range of clickstream knowledge expeditiously sculptured victimization MapReduce.

Map Reduce could be a programming model that lets developers specialize in the writing code that processes their knowledge while not having to fret regarding the small print of parallel execution. A Map Reduce job sometimes splits the input data-set into freelance chunks that are processed by the map tasks in an exceedingly utterly parallel manner. The framework types the outputs of the maps, which are then input to the reduce tasks. There Map Reduce is incredibly like minded for the advanced click stream knowledge.

REFERENCES

[1] P. Mohan Anand, G. Sai Vamsi, P. Ravi Kumar. "A Novel Approach for Insight Finding Mechanism on ClickStream Data Using Hadoop", 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018

[2] Sharma, P., Mahajan, K., Bhatnagar, V. (2016, February). Analyzing Click Stream Data Using Hadoop. In Computational Intelligence and Communication Technology (CICT), 2016 Second International Conference on (pp. 102-105). IEEE.

[3] Sukhwani, Sumit, Satish Garla, and Goutam Chakraborty. ” Analysis of Clickstream Data Using SAS.” SAS Global Forum 2012.

[4] Makhecha, H., Singh, D., Prajapati, B., Puvar, P. Clickstream Analysis using Hadoop. In International Journal of Computer Trends and Technology (IJCTT) Volume 34 Number 2 - April 2016

[5] Apache-Hadoop, http://Hadoop.apache.org

[6] Shvachko, Konstantin, et al. ”Thehadoop distributed file system.” Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on. IEEE, 2010.

[7] Tom White, (2015) Hadoop: The Definitive Guide. OReilly, Scbastopol, California.

[8] Zhang, Rui, Min Li, and Dean Hildebrand. ”Finding the big data sweet spot: Towards automatically recommending configurations for hadoop clusters on docker containers.” Cloud Engineering (IC2E), 2015 IEEE International Conference on. IEEE, 2015.

[9] Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Ghodsi, A. (2016). Apache spark: a unified engine for big data processing. Communications of the ACM, 59(11), pp.56-65.

[10] Belcastro, L., Marozzo, F., Talia, D. (2018). Programming modelsand systems for Big Data analysis. International Journal of Parallel,Emergent and Distributed Systems, pp.1-21.

[11] Fan, W., Bifet, A. (2013). Mining big data: current status, and forecast to the future. ACM sIGKDD Explorations Newsletter, 14(2), pp.1-5.