ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering

Vol. 1, Issue 5, July 2013

A Survey on Hadoop Technology to Develop ETL for Efficient Datawarehouse

M.Murali Krishna , B. Ramakantha Reddy, Y.K. Alluraiah

Professor, Dept. of CSE, S V College of Engineering, Tirupati, A.P, India

Assistant Professor, Dept. of CSE, S V College of Engineering, Tirupati, A.P, India

Professor, Dept. of CSE, S V College of Engineering, Tirupati, A.P, India

ABSTRACT: In recent years, there is a tremendous growth in data volume. Also many other sources than traditional structured data ex. log files, web data, stream data and sensor data need to be stored in the OLTP. It is not suitable for an organization to neglect valuable information from these sources. ETL tools extract meaningful information from various data sources, various transformations of data are carried out in transformation phase and then load into the data warehouse. Traditionally, commercial ETL (Extract –transform-load) tools ex. Informatica, Micro strategy ,Pentho etc. used to transfer OLTP data to other database known as data warehouse . MapReduce technology is becoming popular among people with specialty of ETL task compelling organization to gain benefits from it. Hadoop, an open source MapReduce framework, is capable of handling massive data, provide cheap storage, process structured as well as unstructured data and has massive scalability. It can be seen as viable alternative for migrating ETL job. Although Hadoop is beneficial for large scale industries, many small organizations having small amount of data is also looking for leveraging their business intelligence on it. In this paper, we will explore the new opportunities of utilizing Hadoop for performing business intelligence with specifically ETL phase of datawarehouse.

KEYWORDS: Datawarehouse, ETL, Hadoop, Hive, MapReduce

1. INTRODUCTION

In recent years, MapReduce technology is getting popular with many organizations. Also, several papers have suggested the advantages of MapReduce technology. MapReduce designed for Extract-Transform –load task complements DBMSs [1] .Even though DBMS are able to perform same workload as MapReduce model, several characteristics of MapReduce like ETL and “read once” datasets, complex analytics, suited for limited-budget operation are

attracting many organization to towards MapReduce applications. In addition to this, many large scale organizations like Yahoo, Facebook have been using MR for processing Petabyte and terabyte of data [2]. ETL which takes 80% time for development of datawarehouse [2] gathers information from unstructured and semi-structured nature of source data. As these sources are difficult to process, they are adding complexity in carrying out ETL task compelling organization to migrate on another platform [1].Recently; researchers are studying the MapReduce to explore this upcoming Technology. There are various characteristics of MapReduce technology such as it is not an Extract-Transform-Load (ETL) tool. It is a platform that supports running ETL [2]. Migration of ETL processing to Hadoop, an open source framework of MapReduce can achieve advantage of reduced processing time

.It will tremendously put down the cost associated with it. There are various advantages of Hadoop technology such as its massive scalability, ability to handle complex logic and manage unstructured data with MapReduce, cheap storage. Due to these reasons Hadoop can be proved as an ideal ETL platform [1].In this paper we will analyze the compatibility MR for performing ETL task, study data warehousing solutions using upcoming technique. This proposes Hadoop as an next generation ETL platform for building datawarehouse.

This work is organized as follows. In section 2 we define ETL system along with its role in building datawarehouse. Fundamentals of Hadoop and its assessment as a viable alternative platform for ETL processing have been described in section 3. Subsequent sections discuss the various cases claiming the Hadoop as a relevant platform for migration of ETL jobs followed by the conclusion.

II. EXTRACT-TRANSFORM-LOAD(ETL)

Development of datawarehouse involves the ETL process. It is complex combination of process and technology. This system consists of three functional entities: Extract, Transform and Load. Extract function extracts relevant

DOI: 10.15680/IJIRCCE.2013. 0105017

information from source data for decision making which then needed to be transformed into different schema to match the datawarehouse schema .The final function load the data into datawarehouse [3]

III. UPCOMING TRENDS IN DATA WAREHOUSING PLATFORM AND REQUIREMENT

The datawarehouse platform is categorized on the basis of their operating system, hardware server ,storage system .By considering the above cloud and Hadoop based data warehousing platform are the most reliable to meet today's requirement .ETL process also need to be develop on the same platform to gain the maximum profit [4] In [5], organizers have listed out their demands in order to process Petabyte of data daily. They have demanded the requirement develop the efficient ETL, capable of handling all kinds of data.

Fast data loading out
Fast query processing
Highly efficient storage utilization
Strong adaptivity to highly dynamic workload patterns

A. NEED TO MIGRATE FROM EXISTING RDBMS SYSTEM TO DDMS SYSTEM

Traditionally, RDBMS system have been in use to perform various task .It provides
security and maintenance but lacks in scalability. Also [1] thoroughly discuss the difference between them. According to paper both the system complements each other. But there are several applications where MR system may prove better solution over the parallel DBMS like ETL task and complex analytics. As MR system stores the data into the storage system which is analogous to the ETL processing. It extracts logs of information from different sources, parse the whole data and clean it. Afterwards, various transformation like' sessionalization', data cleaning ,filtering ,lookups is performed on it to load into a storage system .Another advantage of MR system is that it performs the complex analytics . Complex data flow program with output of one application should be input to another to make multiple passes over data .This application can be easily develop with MR system.

Hence Hadoop is open source framework of MapReduce it can be proved as a viable alternative for performing ETL task for building datawarehouse.

In [6] TDWI conducted the survey of 263 user organization and their responses were recorded. There are number of questions and one questions which forcefully demanding for different platform is that "In your perception, what would be useful applications of Hadoop in your organization?" About 41% claims data staging area for datawarehouse is possible usage of Hadoop

IV. AN INTRODUCTION TO EMERGING TECHNOLOGY WITH EXAMPLE

In this section, fundamentals of Hadoop technology has been discussed

A. HADOOP

Hadoop is open source framework based works in a distributed environment [7].It is
built on java programming model. According to [8], Hadoop shows MAD characteristics
.The 'M' stands for magnetic i.e. it can store all kind of data sources and attracts them towards itself. The 'A' refer to the agility as various operations on big data easily can be easily performed on it. The 'D' stands for Deep. It is capable of performing ad-hoc and complex analytics. To perform depth analytics over the big data, Hadoop provide desired result. [9].

B. *HDFS*

The Hadoop Distributed File System (HDFS) is a distributed file system designed to
run on commodity hardware[10] .HDFS is designed to be deployed on low-cost hardware. It is highly fault tolerent.HDFS is suitable for applications that have large data sets [11].

C. MAPREDUCE PROGRAMMING MODEL

To process the large amount of data one relevant programming paradigm is MapReduce.

In this model, Map function process Key/Value pair to generate intermediate key/value pairs. Reduce function later collect these values to produce the output file [12, 13].

D. AN RELEVANT EXAMPLE FOR INTEGRATION WITH HADOOP

Reference [14] explains when one should think for another platform for performing business intelligence on

DOI: 10.15680/IJIRCCE.2013. 0105017

Hadoop. It has described it by giving a relevant example. Nowadays traditional meters to collect electricity readings had been replaced to smart meter. CostCutter utility company which serves 10 million houses. They had decided to collect meter reading quarterly instead of monthly due to increase in labour cost Number of readings are getting increasing .Apparently they have started collecting about 10 million bills by quarterly collecting 10 million readings. In addition to it, government law impose compelled CostCutting Company to deploy Smart meter as there has also been increase in oil prices. Now, it is compelling for them to collect hourly reading from every house. As a result of it, it has collected 21.6 billion sensor reading per quarter reading. Now, CostCutting Company face scalability problem as well as maintenance problem. Billing plans require the analysis of data which should be correlated with the weather conditions and local events.

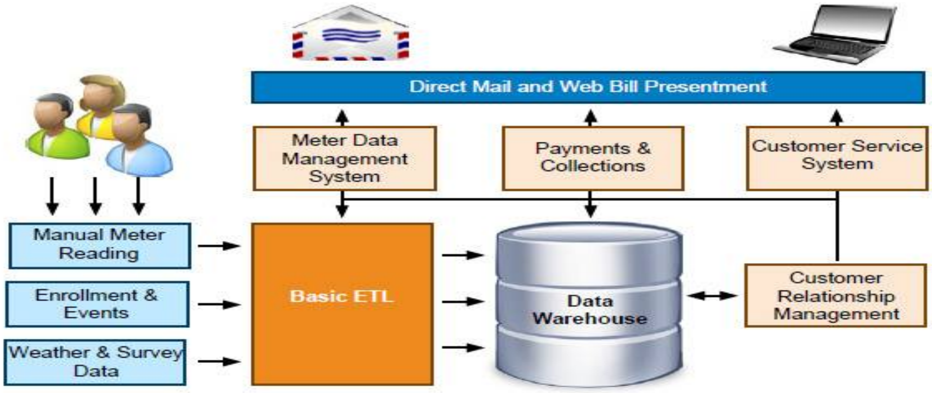

Fig.1.before: data flow of meter reading done manually

To get solution over the above problem, Company urgently requires another platform to replace their whole system. As a result, they have decided to use Hadoop for processing the data. The processing of data in Hadoop is analogous to the extraction and transformation operation. It also provides massive scalability by joining number of nodes as per the requirement to the system at reduced cost. It is possible for the CostCutting Company to collect reading at every 5 or 60 minutes after migration their system on Hadoop.

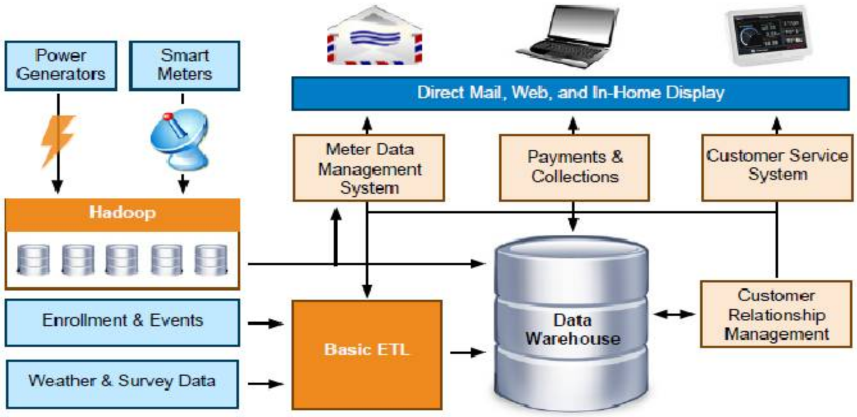

Fig.2 After meter reading every 5 or 60 minutes via Smart meters

According to [4] Hadoop can be reasonably considered as the evolution of next- generation Data Warehousing systems, with particular regards to the ETL phase of such systems. In [11] paper, key points of Hadoop based technology has been mention .According to it, In Hadoop file system, once data has been loaded, no alteration can be made on it. It is just like once-write-read- many. It also mentions that, Hadoop is not a ETL tool. It supports the ETL environment .Once data has been loaded into HDFS; it is required to write transformation code.

V. DATAWAREHOUSE AND HADOOP : RELATED WORK

In [15], ETLMR framework has developed using the Hadoop .This framework can handle the scalability issue as well as the slowly changing dimensions of data. The processing of snowflakes and start schema has been described in this framework. In [16], the effective data modelling technique has been suggested. The data from the star schema has been transferred using the SPOOP and performed various queries ad-hoc queries on it. The experimental results shows that star schema quickly response to the query rather than the flat file. This study also motivates to perform the transformation itself on Hadoop and store the star schema in hdfs.

In many large companies, Hadoop has been used as datawarehouse .The Facebook has developed the data warehousing platform and been named as Hive. It is an open source data warehousing solution build on top of Hadoop and developed by team in 2007 [2].As per [2], Hive at Facebook has been handling large amount of data which contains tens of thousands of table and it has 700 TB data with 200 users. It has been specifically developed to perform ad-analysis and reporting. The HiveQl is replica of SQL used to query HIVE, analysis of data in Hadoop has been relatively easy for SQL users. It does not require the extensive knowledge of MapReduce.

A data warehousing framework developed with MapReduce in [17]. It has been built for online advertising application allows custom optimization. The processing of data has been carried out by MapReduce paradigm. The demonstration Cheetah framework can has been given in detail in this paper. This framework suggests pre-joined of fact and dimension table so that end users do not require to fire join query. In [18] also, a discussion has been carried out on strong need of migration of ETL processing to Hadoop for future requirement. It also claims the same outcome migration as reduce cost and processing time. Hadoop mainly exhibits the characteristic quality like scalability, process unstructured data with MapReduce ,cheap storage and ability to perform complex analytics makes it an attractive alternative for ETL platform .there is a another interface known as a pig which can be used to perform the ETL script[19].

As all above examples related to the large companies where daily need to process data is in Petabyte. Various research has been conducted to find the answer whether Hadoop is suitable for small organisation or not. In these companies data is not large but they require cost effective solution for business intelligence .It can reduce their data processing time for developing datawarehouse. The paper [20] conducted the extensive analysis of suitability of Hadoop for small scale data for midsized organisation. In this study, comparative analysis among MYSQL, Hadoop+MapReduce and hive has been performed. The data set are ranging from 200 MB to 10 GB are used to conduct experiments. Result of this study draws various conclusions. This study claims that for data up to 1GB, MYSQL outperforms than Hadoop and Hive. Second experiment shows data ranging from 1 to 2GB, Hadoop+MapReduce architecture gives accurate result than MYSQL and Hive .Beyond 2 GB data, Hive tremendously outperform than other two. This study explores research opportunities in future .In this study, limited analysis has been done and does not run extensive analytics to confirm the result. Instead of this various issues, this study explores new opportunity for small and medium industries so that they can also avail the advantages from usage of Hadoop Data pre-processing is carried in data staging area for the data to be loaded into a datawarehouse. It required the lot of pre-processing .Paper [21] claims that size of data staging area is increasing day by day than datawarehouse. By considering the advantages of Hadoop technology as, massive storage at cheaper rate, scalability; data storing in the form of file and processing of unstructured data Author suggested that data pre-processing and staging can be performed on and then data can be loaded to datawarehouse as it has advantage of. This concept has another advantage. This concept has lot of advantages as only cleaned data is loaded into DW for end user to access it. Data processing in staging area has been detached from source system so that original system does not get affected by data analysis process.

The storage capacity of Hadoop system is massive attracting organizations to store their historical and raw data. Datawarehouse are now a day has been observed with vast of data collected in it over a period of time. It does not require all the data for analysis and reporting .Also, it cannot totally be pruned. This is also causing headache for an organization. Storing and maintaining older datawarehouse data and raw source data for long period has becoming expensive [21].The relevant solution over this issue is that storing older and raw can also be migrated to Hadoop so that data can be retained as long as it is required.

VI. FUTURE RESEARCH

As stated above, there is a requirement of future research in the specific area and these are

- SQL engine: the Hive does not support the concept of mature SQL engine. This makes it difficult to apply traditional dimensional modeling technique in Hive.
- Limitation on join : joins on only equality has been allowed in the hive .This makes it difficult to join various tables and prune data while joining
- No study has performed the transformations on hive itself and measures its performance. The various transformations such as filtering, cleaning, lookups can performed on hive itself and load data in various schema.

VII. CONCLUSION

The main aim of this paper is to assess the capacity of Hadoop technology for building datawarehouse specifically the ETL part of it. In this paper we have discussed the various advantages of Hadoop based on MapReduce framework works. Also various case studies have been studied for different usage of Hadoop other than ETL such as data staging area. An details assessment Hadoop as alternate ETL platform has also been carried out .The relevant example has been discussed when one should think for Hadoop platform.

Also survey conducted by renowned TDWI organization shows strong willingness of organization to use Hadoop for ETL platform. The star schema is suitable for getting faster results also promote to perform ETL on Hadoop. It can be concluded that Hadoop can reduce processing time as well as storage cost in building datawarehouse for an organization require In addition to this ,Case study exhibits the advantage of Hadoop for small scale industries explores new opportunities of research.

VIII. ACKNOWLEDGMENT

I would also like thank Mr. Kuldeep Deshpande, (Head, Ellicium Solutions Pune), for his valuable guidance. I would like to thank, Prof. Leena A. Deshpande, for the guidance and dedication she gave during my project work.

REFERENCES

[1] M. Stonebraker, D. Abadi, D.J. DeWitt, S. Madden, E.Paulson, A. Pavlo, and A. Rasin. MapReduce and parallel DBMSs: friends or foes? Communications of the ACM, 53(1):64–71,2010

[2] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Ning Zhang, Suresh Antony, Hao Liu and Raghotham urthy, Hive – A Petabyte Scale Data Warehouse Using Hadoop, IEEE, 2010.

[3] Merinela Mircea, Business Intelligence--Solution for Business Development, Intech Publisher,2012

[4] kuldeep deshpande, and dr. Bhimappa desai **,**limitations of datawarehouse platforms and Assessment of hadoop as an alternative, Volume 5, Issue 2, pp. 51-58, IJITMIS ,2014

[5] Yongqiang He et all RCFile: A Fast and Space-efficient Data Placement Structure in MapReduce-based Warehouse Systems, ICDE, 2011

[6] Philip Russom, Integrating Hadoop into business intelligence and data warehousing, TDWI best practices report Nov 2013

[7] Hadoop. http://hadoop.apache.org.

[8] Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., and Welton, C, MAD Skills: New Analysis Practices for Big Data, PVLDB 2(2), 2009.

[9] Song .Y, Davis Karen C, Analytics over large scale Multidimensional Data: The Big Data Revolution, Communications of ACM, 2011

[10] Hadoop distributed file system (hdfs). http://hadoop.apache.org/hdfs

[11] T.K.Das and Arati Mohapatro , A Study on Big Data Integration with Data Warehouse, International Journal of Computer Trends and Technology (IJCTT) – volume 9 number 4– Mar 2014

[12] J. Dean and S. Ghemawat, MapReduce: simplified data processing on large clusters, Communications of the ACM, 51(1):107–113, 2008.

[13] Hadoop MapReduce. http://hadoop.apache.org/mapreduce

[14] Awadallah Amar,Graham.Dan, Hadoop and the Data Warehouse-When to use which , Cloudera Inc and Teradata Corporation,2011

[15] Xiufeng Liu, Christian Thomsen, and Torben Bach Pedersen, ETLMR: A Highly Scalable Dimensional ETL Framework Based on MapReduce, pp. 1–31, springer, 2013.

[16] Clark Bradley, Ralph Hollinshead, Scott Kraus, Jason Lefler, Roshan Taheri "Data Modeling Considerations in Hadoop and Hive", Technical paper, SAS,2013

[17] Chen Songting, Cheetah – A high performance custom Datawarehouse on top of MapReduce, Proceedings of VLDB ,Vol 3,No . 2, 2010

[18] Offload your Datawarehouse with Hadoop, Syncsort publication, 2014 [19] Hadoop Pig. Available at http://hadoop.apache.org/pig

[20] Marissa Rae Hollingsworth, Hadoop and Hive as Scalable Alternatives to RDBMS: A Case Study, Boise State University, 2012.

[21] Dhruba Borthakur, The Hadoop Distributed File System: Architecture and Design, Apache foundation, 2007

[22] Gandhali Upadhye and Astt. Prof. Trupti Dange, "Nephele: Efficient Data Processing Using Hadoop" International Journal of Marketing & Human Resource Management (IJMHRM), Volume 5, Issue 7, 2014, pp. 11 - 16, ISSN Print: 0976 – 6421, ISSN Online: 0976-643X.