LFA - LOGISTIC FREIGHT ASSISTANCE USING HADOOP MAP REDUCE FRAMEWORK

Navya Francis, Computer Science and Engineering
KMEA Engineering College. Ernakulam City, Kerala, India
deepunavya13@gmail.com

***Abstract* — In recent years we are using the logistic freight system for the shipping of goods, and these shipping are done by the freight forwarders/carriers from one point to the alternative point. Shipping charges are the freight bills, which has to be paid by the specific organization after the delivery of freight. The freight forwarders/carriers can make mistakes during the preparation of the freight bills. Due to this accurate auditing of the bills has to be prepared otherwise the organization will have to overpay for the services which they haven't used. Now these auditing is done physically and it require lot of human power and time. Meanwhile there is lot of chance for human and process errors. LFA (Logistic Freight Assistance) is the solution to these problems. It helps to audit the freight bills and using LFA the stakeholders can make informed judgments. It overpowers the human and process errors. The system is developed based on the hadoop mapreduce framework. The data are stored in HDFS and processed using the hadoop map reduce concept.**

Keywords* – *Visual Analytics, Freight, Freight Audit, MapReduce, Logistics, Hadoop

I. INTRODUCTION

Logistics [1] is the emerging area of concern. The movement of goods from one point to alternative point is meant by the cargo. The transportation of freight is done by the Freight forwarders/shippers. Freight transportations are charged by the freight shippers, as the freight charges [2]. The reviewing of these charges has to be done otherwise we will have to pay more for the services which the organization have not incurred.

There are different freight transportation [3] methods. Air, ship and ground are some of them. In this ground transportation is cheaper and mostly used for the logistic freight transportation within the same country.

The auditing of these freight charges can be done in dissimilar ways. In this most of the organization, they do the subcontracting because it provides steady result. The manual handling of these freight audit can take lot of time and human manpower.

Due to this the manual processing [4] is suitable only for the small auditing because in an organization there can be number of products shipped in each and every month. The subcontracting method is the best method, but since it is costlier small concerns cannot swallow the cost of subcontracting.

The logistic freight related details is one of the bigdata application [5]. Bigdata contains large spectrum of data. Data are generated in each and every second of time. The data are generated in different modes in different variety. Since it contain large volumes of data in structured and unstructured format, the handling of these data is the biggest challenge in the recent years.

And the solution to this problem can be formed by using the hadoop [6]. Hadoop Distributed File System (HDFS) and the MapReduce are the two main concepts.

Hadoop Distributed File System (HDFS) [7] acts as the storage mechanism and the MapReduce acts as the programming language in hadoop. Data which has to be managed, has to be stored in the HDFS and these data are retrieved whenever needed for processing.

In mapreduce framework [8], mapper part will gather the details from different clusters and the reducer part will consolidate the results inorder to get reliable production which can be used to make informed decisions.

The visual analytics helps us to make the auditing process far simpler and user approachable manner and thus an unfamiliar individual, with the freight audit can do the auditing within few interval.

II. VISUAL ANALYTICS

Visual analytics [9] is the method of analytical reasoning assisted using visual interfaces. It brings scientific and technical communities together from various areas. It has many goals and techniques related to the information and scientific visualizations. Information visualization provide the direct interface between the machine and the user. These handles the data structures like graph, trees etc. Scientific visualization deals with the geometric structures.

Analytical reasoning [10] helps the user to get the deep knowledge and due to this the correct decisions can be taken by the user. It facilitate high quality of the human judgment. The tasks are done with the combination of the user and the collaborative analysis which is done in the stressful environment.

The complexity of the logistics are analyzed, visualized, optimized and modelled with the help of visual analytics. It helps to understand the past and current situation in quick and easy manner. Thus visual analytics helps the experts to make well-versed decisions which are reliable.

III. HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

HDFS was designed by Apache Nutch project [11] as an infrastructure extension. It is designed for running the computations performed on the cluster. The HDFS has many salient features. Some of them are fault tolerant, reliable, scalable etc. he architecture mainly consist of three parts. They are as follows

- Name node and data node
- The file system namespace
- Data replication

The hadoop contains the master slave architecture the data nodes acts as the slaves and the name node acts as the master. The name node holds and manages the metadata in the HDFS. In this there is no data flow on name node.

In the file system namespaces, HDFS supports the empirical file structure. The directories can be created and managed by the user or application. Name node handles the file system namespaces. It record the alterations and the associated properties. The number of replica maintained can also be controlled by the application which is maintained by the HDFS replication factor and the details are stored in the name node.

Data replication is done in HDFS to manage large files. The data are stored in same size blocks in sequence. These blocks are replicated to test fault tolerance in which the size of block and replication factor are configurable. The number of copies can be customized by the application.

The heart beat is the communication mechanism used to understand whether a data node is active or not. The heart beat contains the data node id and the block report. Which helps us to understand the functional status of the data node.

IV. HADOOP MAPREDUCE FRAMEWORK

Hadoop is used as the solution to solve the big data problems. Hadoop [12] is open source developed by Apache. It acts as cross platform. It will simplify the implementation of the data.

Data is distributed and replicated across the cluster. It provide large amount of data storage. Since the data is distributed, it provides fast accessing of data. It provides reliability and scalability.

Hadoop has master slave architecture [13]. The name node is the master and data nodes act as the slave nodes. The name node is stored in the in-memory and the data nodes is saved in different machines. The actual data are stored in the data node.

The core components [14] of the hadoop are HDFS, mapreduce, Hbase, Hive, Pig, Sqoop, Flume, Oozie, Zookeeper. HDFS is the storage mechanism. Mapreduce is the framework and the programming language. Map reduce is implemented using Pig and Hive. The workflow is managed by the Oozie and Zookeeper acts as the coordinator. Sqoop and Flume are the enterprise data integration.

Mapreduce [15] is the simple programming method. The map reduce has a pipeline in which the processing is done. The pipeline consist of the input data, mapper part, reducer part, output data and driver.

The driver will control the overall execution of the mapreduce programming. Input data is the input which we give for processing and output data is the result which we get after processing the data.

Mapper will take the input records based on the user requirement and reducer will consolidate the output intermediate results which are produced by the mapper inorder to get the final result of the mapreduce execution. The Mapreduce execution pipeline is shown in Figure 1.

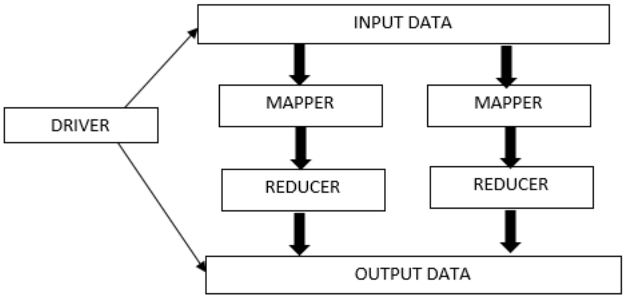

Figure 1. Mapreduce pipeline execution

Distributed cache [16] is an additional option. In this we can store the intermediate results which are the output from the mapper. Usually the intermediate results are left unused. If the cached results are used, execution time can be reduced.

Combiner is another optional part. In this the results produced are combined together inorder to get single valued result.

The mapreduce programing can written in different languages. Some of them are JAVA, Python and Pig [17] etc. These are the basics of the mapreduce programming.

V. LFA (LOGISTIC FREIGHT ASSISTANCE) SYSTEM

It is one of the bigdata application for checking logistic freight. LFA (Logistic Freight Assistance) is planned based on the hadoop mapreduce framework and visual analytics. It consist of four modules and they are as follows:

- Collector Phase
- Transformer Phase
- Output Phase
- Visualization Phase

The figure 2 shows the design of the LFA (Logistic Freight Assistance) system.

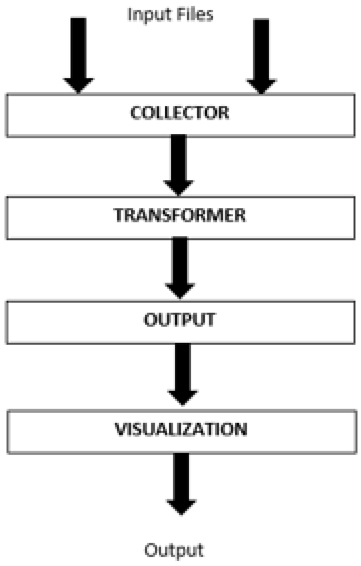

Figure 2 LFA (Logistic Freight Assistance) system design

A Collector Phase

In this phase the data which has to be processed are collected and stored in HDFS. The data collection can be done either from the organization or from the amazon web services (AWS)

B Transformer Phase

In this phase the data which is collected is altered and it is done by the mapreduce framework. The mapper part will collect the relevant data and the reducer part will consolidate the intermediate data which is the output from the mapper part and produce the reliable solution.

C Output Phase

In this phase the output produced from the transformer phase can be viewed.

D Visualization Phase

The output phase results can be visualized using the visualization tool which will be give more clear and understandable by the user.

LFA (Logistic Freight Assistance) system will provide the organiszations with the past and current freight audit details of the organizations within few interval of time. Using this system the experts can take many decisions, which benefit the organization.

VI. MODELLING OF LFA (LOGISTIC FREIGHT ASSISTANCE) SYSTEM

The data which is required to process the LSA are stored into HDFS and the contents are processed with the help of mapreduce concept. It is implemented using ClouderaVM. The mapreduce algorithm is written using the PIGLATIN. The counters after processing the LFA (Logistic freight assistance) is shown in table 1.

TABLE 1
Mapreduce Execution Parameters

Parameters Received	Map (Bytes)	Reduce (Bytes)
File Bytes Read	0	115255
File Bytes Written	1549129	1558808
HDFS Bytes Read	12140698	0
HDFS Bytes Written	0	811789

VII. CONCLUSION

Logistic freight audit can be done in simple way using the LFA (Logistic Freight Assistance) system. It is a visual analytics method for solving the issues regarding the freight audit. LFA can be used by any individual inorder to perform the freight audit of the logistic system. It will provide a user friendly environment. LFA is implemented using the concept of the hadoop mapreduce framework. This helps the organization to view the past and current status of the freight they have incur. Using LFA (Logistic Freight Assistance) system the expert can make informed decisions.

As a future work, the visualization of the complete freight details can be done with the help of graph, dashboards etc.

REFERENCES

[1] http://en.wikipedia.org/wiki/Logistics Retrieved 2015-02-27.
[2] http://en.wikipedia.org/wiki/Freight_rate Retrieved 2015-02-27.
[3] http://en.wikipedia.org/wiki/Freight_transport Retrieved 2015-02-27
[4] http://en.wikipedia.org/wiki/Freight_audit Retrieved 2015-02-27.
[5] http://en.wikipedia.org/wiki/Big_data Retrieved 2015-02-26.
[6] Ramesh Kumar, Dr.Vijay Singh Rathore "Efficient Capabilities of Processing of Big data using Hadoop Map Reduce" International Journal of Advanced Research in Computer and Communication Engineering, Volume 03 Issue 06, June 2014, Pages 7123-7126.
[7] http://wiki.apache.org/nutch/NutchTutorial Retrieved 2015-03-12.
[8] http://en.wikipedia.org/wiki/MapReduce Retrieved 2015-02-27.
[9] https://en.wikipedia.org/wiki/Visual_analytics Retrieved 15.03.2015.
[10] http://www.purdue.edu/discoverypark/vaccine/assets/pdfs/publications/pdf/Science%20of%20Analytical%20Reasoning.pdf
[11] http://wiki.apache.org/nutch/NutchTutorial Retrieved 2015-03-12.
[12] http://en.wikipedia.org/wiki/Apache_Hadoop Retrieved 2015-03-12.
[13] Shital Suryawanshi, Prof.V.S.Wadne "Big Data Mining using Map Reduce: A Survey Paper" IOSR International Journal Computer Engineering, Volume 16 Issue 06, Nov- Dec 2014, Pages 37-40.
[14] Boris Lublinsky, Kevin T Smith, Alexey Yakubovich "Professional Hadoop Solutions" Proc.WROX, Pages 1-96.

[15] Jefffrey Dean, Sanjay Ghemawat “MapReduce: Simplified Data Processing on Large Clusters” Communications of the ACM, Volume 51, Number 1, Pages 107-113

[16] Yaxiong Zhao, Jie Wu “Dache: A Data Aware Caching for Big-Data Applications Using The Map Reduce Framework” International Journal of Tsinghua Science And Technology, Volume 19 Number 1, February 2014, Pages 39-49.

[17] Alan Gates “Programming Pig” Proc. O’Reilly Media. Pages 1-170.

[18] http://www.bbc.com/news/business-26383058 Retrieved 2015-03-12.

[19] http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html Retrieved 2015-03-12.

[20] Jefffrey Dean, Sanjay Ghemawat “MapReduce: Simplified Data Processing on Large Clusters” Communications of the ACM, Volume 51, Number 1, Pages 107-113.

[21] Sasiniveda.G, Revathi.N “Data Analysis using Mapper and Reducer with Optimal Configuration in Hadoop” International Journal of Computer Trends and Technology (IJCTT), Volume 04 Number 03, February 2013, Pages 264-268.

[22] Karan B.Maniar, Chintan B.Khatri “Data Science: Bigtable, Mapreduce and Google File System” International Journal of Computer Trends and Technology (IJCTT), Volume 16 Number 03, October 2014, Pages 115-118.

[23] Tom White “Hadoop the definitive guide” Proc. O’Reilly Media, Edition 3, May 2012.

[24] Chuck Lam “Hadoop in Action” Proc. Manning Publication, Edition 1, December 2012.

[25] http://en.wikipedia.org/wiki/Pig (programming_tool) Retrieved 2015-02-27.

[26] http://pig.apache.org/docs/r0.8.1/udf.html Retrieved 2015-02-26.

[27] Donald Miner, Adam Shook “Mapreduce Design Patterns” Proc. O’Reilly Media, November 2012.

[28] http://hortonworks.com/hadoop-tutorial/how-to-use-basic-pig-commands/ Retrieved 2015-02-26.

[29] https://www.controlpay.com/services/logistics-visibility Retrieved 2015-02-26.

[30] http://www-01.ibm.com/software/in/data/bigdata/ Retrieved 2015-02-26.

[31] http://bigdatauniversity.com/bdu-wp/bdu-course/big-data-fundamentals/ Retrieved 2015-02-26.

[32] Sangeeta Bansal, Dr. Ajay Rana “Transitioning from Relational Databases to Big Data” International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1, January 2014.

[33] Han Hu, Yonggang Wen, Xuelong Li “Toward Scalable Systems for Big Data Analytics: A Technology Tutorial” IEEE access practical innovation: open solution, Volume 2, July 2014, Pages 652-687.

[34] Kyongha, yoonjoon, Hyunsik, Yondohn, Bongki “Parallel Data Processing with MapReduce: A Survey” ACM SIGMOD Record, Volume 40 Issue 4, December 2011 Pages 11-20.

[35] HDFS Users Guide – Rack Awareness". Hadoop.apache.org. Retrieved 2015-02-17.

[36] "Improving MapReduce performance through data placement in heterogeneous Hadoop Clusters" (PDF). Eng.auburn.ed. April 2010.

[37] "Cloud analytics: Do we really need to reinvent the storage stack?". IBM. June 2009.

[38] "HADOOP-6330: Integrating IBM General Parallel File System implementation of Hadoop File system interface". IBM. 2009-10-23.

[39] "Refactor the scheduler out of the JobTracker". Hadoop Common. Apache Software Foundation. Retrieved 9 April 2015.

[40] M. Tim Jones (6 December 2011). "Scheduling in Hadoop". ibm.com. IBM. Retrieved 20 November 2013. [11]"Under the Hood: Hadoop Distributed File system reliability with Namenode and Avatarnode". Facebook. Retrieved 2015-03-13.

[41] "Under the Hood: Scheduling MapReduce jobs more efficiently with Corona". Facebook. Retrieved 2015-03-9.

[42] "Zettaset Launches Version 4 of Big Data Management Solution, Delivering New Stability for Hadoop Systems and Productivity Boosting Features | | Zettaset.comZettaset.com". Zettaset.com. 2011-12-06. Retrieved 2052-02-23.

[43] Curt Monash. "More patent nonsense — Google MapReduce". dbms2.com. Retrieved 2015-03-07.

[44] D. Wegener, M. Mock, D. Adranale, and S. Wrobel, “Toolkit-Based High-Performance Data Mining of Large Data on MapReduce Clusters,” Proc. Int’l Conf. Data Mining Workshops (ICDMW ’09), pp. 296-301, 2009

[45] J. Dean and S. Ghemawat, “Mapreduce: simplified data processing on large clusters,” in Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation - Volume 6, ser. OSDI’04.

[46] http://www.slideshare.net/mcsrivas/design-scale-andperformance-of-maprs-distribution-for-hadoop.