Processing Cassandra or Mongodb dataset with Hadoop-Streaming Based Approaches

[1]Gopal R. Chandangole, [2]Shweta A. Joshi
[1] Student ME (Computer),[2]Prof. ME(Computer)
[1]Department of Computer Engineering
[1]Flora Institute of Technology, Pune,India.

***Abstract*– Now a days Bulk of data generating on the system. and today's user accessing, searching and sorting the data from database is very difficult. To overcome this problem, data is distributed in different node using Hadoop technology. A system is proposed in which the collected data is to be distributed using map reduce technique for sorting the data is very easily on Hadoop environment. In this case used Cassandra and mongodb tools to storing large amount of data on Hadoop Framework. NoSQL data is stores in unstructured data format which is a key focus area for "Big Data" research. The Hadoop Framework used to large amount of data on a different nodes in a cluster data. NoSQL databases due to structure and unstructured data of high scalability for getting high performance of system. A Cassandra is to provide the platform for the fast and efficient data queries. and the mongodb is storing large amount of data and connected between different node with the Hadoop MapReduce engine.**

***IndexTerms*– Index Terms Hadoop Streaming, Cassandra, Mongodb,MapReduce.**

I. INTRODUCTION

The large amount of data collection taking place the result of social media interaction, scientific experiments, data generation from many different sources[1] "new generation" data generation from many different sources data presents challenges as is not all relational and lacks predefined structures. The Data is formatted in different way is originates from different sources. To use the NoSQL data Model is storing data in differently structured format in separate databases. The NoSQL is allowing datasets of sharing records but it is in the different structure While Map Reduce et al.[11] frameworks provide the large storage and capabilities for data in any form like structured or unstructured format the Map Reduce written in non-java to be used within such "Big Data" processing When the data is generated in NoSQL. This paper to present data processing to allow not only in Hadoop Map Reduce programs written in Java with NoSQL (not only SQL) structure format.The storage systems for improving Performance. NoSQL is used to combine with MapReduce framework. The Cassandra data set is an open source for storing large amount of data on a Hadoop Framework and also using a MARISSA It is does not require Data node and Name Node such optimizations to the shared file system to base on Master Node. The process the data from NoSQL stores and the performance impact of first downloading to a MapReduce framework. The Hadoop Framework is mostly use for generating large amount of data like Facebook and Twitter. Hadoop Framework contain different node like Name Node, Data Node, and Master Node. It is used for generating large amount of data and distributed on the different node is very easily.

The contributions of this paper are as follows:

- The introduce a MapReduce streaming with a MapReduce framework datasets use to downloaded from the Cassandra data set.
- To comparison between the data directly from Cassandra dataset using the Map Reduce for performance of processing data columns under in various application.
- System proposes an alternative approach using Hadoop for reading Cassandra dataset records directly from the database.

II. BACKGROUND

A. CASSANDRA Data Set

Cassandra [12] is an open source and non-relational column oriented distributed database developed by Facebook. It is designed to store the large amount datasets for used in a peer-to-peer structure to promote horizontal scalability to improve the performance of system. It is now an open source Apache project. Interesting aspects of the Cassandra framework include independence from any additional file systems like HDFS, scalability to support for balanced data partitioning.

B. MapReduce Process

The Map Reduce starts with the splitting data an input dataset over a set of System user and processing these data splits in parallel processing with user-defined map and reduce functions. The model process on input distribution, parallelization, and scheduling. The Apache Hadoop is the open source Map Reduce implementation on two fundamental components the Hadoop Distributed File System (HDFS) and the Hadoop Map Reduce Framework for the data management.

C. MARISSA Tools

It is does not require Data node and Name Node such optimizations to the shared file system to base on Master Node. Each user node points to the executable to input splits data it is responsible for monitors the status of the local job and informs to the master node when the local tasks are completed. Support to some features is implemented within in MARISSA tools that are considered in Hadoop. MARISSA does not require processes like Task Trackers and Data Nodes for execution of Map Reduce operations. Because it is process on Master node. it is does not require processes like Task Trackers and Data Nodes for execution of Map Reduce jobs.

III. LITERATURE SURVEY

This system proposed will provide a new approach analysis big data mining [3] with Cassandra and mongodb which is based on MapReduce Paradigm on Hadoop. This new approach will try to improve the computational time more fault tolerance of system and will handle or deal with Big data analysis. Data Mining is a process to generate pattern and rules from the various types of data marts and data warehouses in this process there are several steps which contains data cleaning and data anomaly detection then clean data is mined with various approaches. In this research have to discussed data mining on large datasets (Big Data) with this large data set major issues are scalability and security to improve the performance. In the Proposed system we are going to mine the big data on Hadoop environments and mongodb and we will try to mine the data with sorted or double sorted key value pair for and analyze the outcome of system. the Bulk amount of data collection taking place the result of social media interaction scientific experiments data generation from many different sources and lacks of predefined structures.

Existing System: To Proposed different tools Like Cassandra and MARISSA are as given below. B and C.

A. MapReduce Process

MapReduce [10] starts with the idea of splitting the input dataset over a set of machines called user and processes these data splits in parallel with user-defined map and reduce functions. and Hadoop Streaming is a MapReduce Framework for data management and job execution. A JobTracker running on the master node is responsible for resolving the job details. Hadoop is implemented in Java language and requires map and reduce operations and use the Hadoop API.

B. CASSANDRA Data Set

Cassandra [12] is an open source and non-relational column oriented distributed database developed by Facebook. It is designed to store the large amount datasets for used in a peer-to-peer structure used in horizontal scalability to improve the performance of system. It is an open source Apache project. Cassandra Framework used additional file system like HDFS, Scalability and replication for balance data partitioning. Cassandra is a column based NoSQL database able to store large amount of data. Each row consists of (key, value) pairs. Values may contain an additional level and offset of data value but these keys are not indexed.

C. MARISSA Tools

MARISSA [16] it is does not require processes like *Task Trackers* and Data *Nodes* for execution process of Map Reduce operations. Because it is process work on Master node the input data is split by the master node using the Splitter module and the data placed into the shared file system each user node has access to input data.

Proposed System: To Proposed different tools Like Cassandra and Mongodb are as given below.

D. CASSANDRA *Data Set*

Cassandra [12] is an open source and non-relational column oriented distributed database developed by Facebook. It is designed to store the large amount datasets for used in a peer-to-peer structure to promote horizontal scalability to improve the performance of system. It is now an open source Apache project. Cassandra Framework used additional file system like HDFS, Scalability and replication for balance data partitioning.

E. Mongodb

Mongodb tools is Open Source used to connect many node and distributed data in different node at same time so the system will getting high Performance for generating large amount of data. In the Proposed system we are going to mine the big data on Hadoop environments and mongodb and we will try to mine the data with sorted or double sorted key value pair for and analyze the outcome of system. the Bulk amount of data collection taking place the result of social media interaction scientific experiments data generation from many different sources and lacks of predefined structures.

IV. MAPREDUCE WITH CASSANDRA DATASET

A. Map Reduce Streaming For the Cassandra Datasets

Map Reduce Streaming for the Cassandra Datasets executable. This pipeline shown in Figur.1, has three main stages: 1) Data Preparation 2) Data Transformation (MR1) and 3) Data Processing (MR2*).*

1) Data Preparation Process: Data Preparation, figur.1a, this Step used of downloading the data from Cassandra servers to the related file systems like HDFS for Hadoop Streaming and shared file system for MARISSA.[1] The Cassandra used to exporting the records of a dataset in JSON formatted files using each node to download the data from the local Cassandra server to the file system. Every user connects to the local database and starts the operation to export the records in unique file on a shared file system.

2) Data Transformation Process (MR1): Cassandra allows users to download datasets as JSON formatted files. [1] For the Map Reduce applications to run which are either difficult to be modified and the input data needs to be converted into a JSON format that is expected by Map Reduce Framework. So the software pipeline contains a Map Reduce stage Figur.1b, where JSON data file can be transformed into the other formats. In this stage each input data is converted into another format and stored in output files.

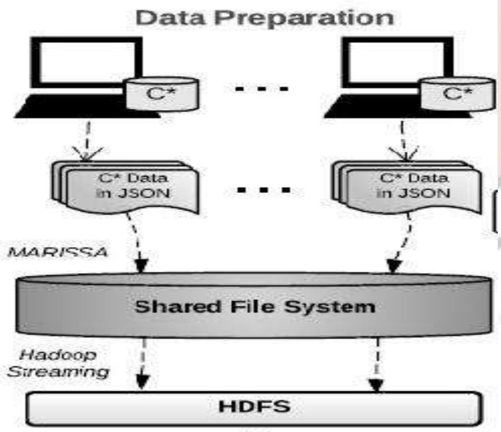
3) *Data Processing Process (MR2):* This is the last step of the Map Reduce Streaming shown in Figur.1. In Figur.1c To run the non-Java executable which is the initial applications over the output data of MR1 the data is in JSON format can be processed by this application to use the MARISSA and the Hadoop Streaming to run executable map and reduce operations.

B. Mapreduce Streaming With Marissa Tool

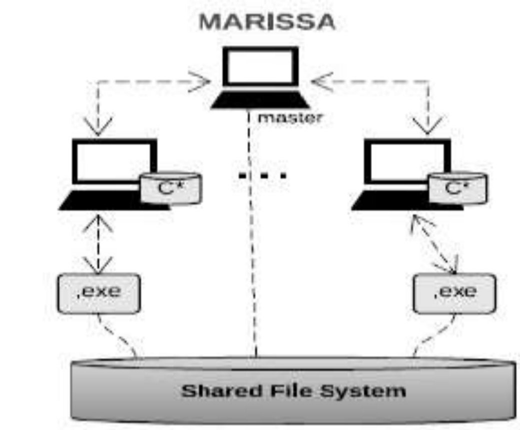
The Splitter module of MARISSA [11] has been modified such that each user connects to the local database server to take the input dataset in JSON format and place to share file system. After the Data Preparation stage shown in Figure. 1a, the input is split and ready for Data Transformation. Figure.2a, shows the architecture of MARISSA. It allows each non-Java Program to interact with the corresponding input splits In the level of Data Transformation each MARISSA mapper runs an executable to convert the JSON data files to the user specified input format. These converted files into the shared file system. The executable to create the next MapReduce stage which is calls to the Data Processing and do not have immediate access to the input splits. So Hadoop Task Trackers read input from HDFS and store into the executable for processing and collect the results to write the HDFS. In the Data Transformation step shown in Figure 1b. MARISSA runs on the user given executable to create the next MapReduce stage, which is use Data Processing. There is no re-distribution or re-creation of splits required since MARISSA is designed to allow iteration of MapReduce operations.

V. RELATED WORK

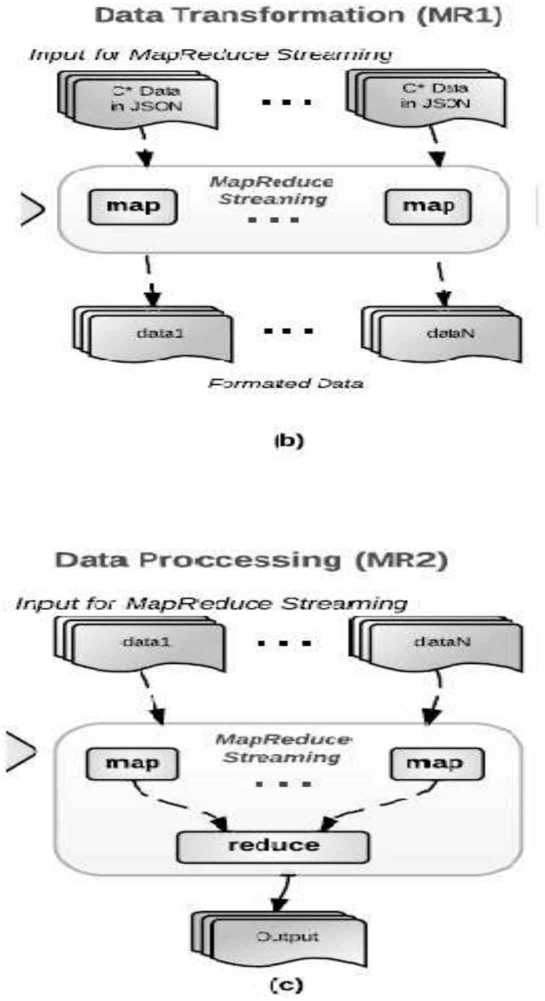
The system will provide to store large amount of data and used in Facebook, Twitter and various approaches use and method proposed. [3] using the NoSQL technologies with the Map Reduce process. The map Reduce Model is to provide a scalable storage and processing framework. Twitter [14] to Process in map Reduce platform to generate the large amount of data. A Hadoop is to perform parallel database systems while exploiting scalability and flexibility of Hadoop MapReduce. Cassandra Framework used additional file system like HDFS, Scalability for balance data partitioning. Cassandra is a column based NoSQL database able to store large amount of data. MongoDB is used to connect different node and also distribute the data on using various operations. Dede et al. [11] provide the performance analysis of using MongoDB storage for Hadoop. To use HBase [2] as a backend to store the data and the web query interface to allow access to the datasets. The MapReduce model provides a distributed framework for scalable storage, querying and processing framework on Hadoop.



(a)



(a)



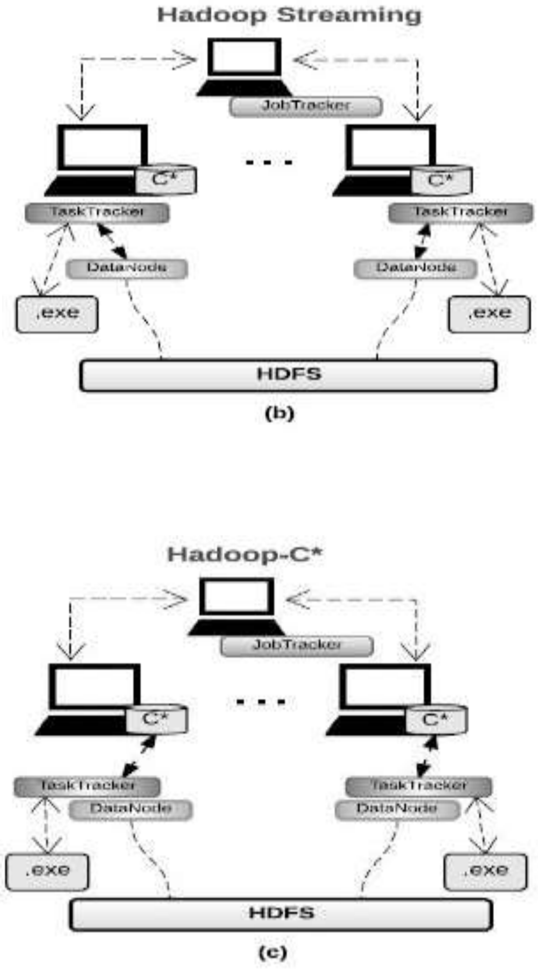
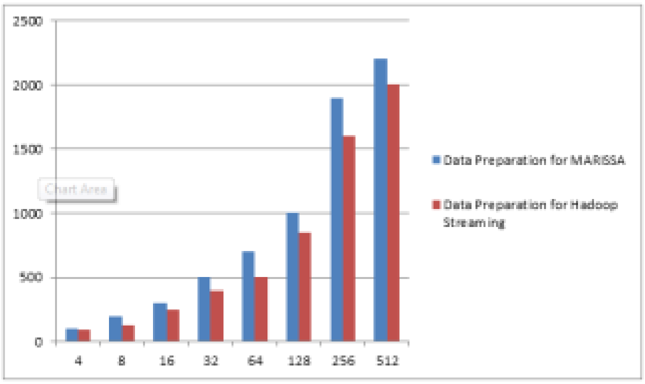


Fig.1. MapReduce streaming pipeline showing each stage. In **Fig. a. Data Preparation:** each user node exports dataset from local Cassandra servers to shared file system. put the data from shared file system to HDFS. **Fig.b. Data Transformations:** The dataset is converted into the user specified format using MapReduce and **Fig.c. Data Processing:** user set non-Java executable are used as map and reduce operations to process reformatted data.

Fig.2. Three different streaming approaches to process Cassandra datasets with map Reduce. **Fig. a. MARISSA:** The data is first downloaded from database servers to shared file system pre-processed for the application **Fig.b. Hadoop Streamings:** To show the layout of using Hadoop Streaming where the dataset is also placed into the HDFS. **Fig. c. Hadoop C*:** Shows the structure of Hadoop-C* to use to process Cassandra data directly from the local database servers using Hadoop

VI. RESULT

A. Data Preparation



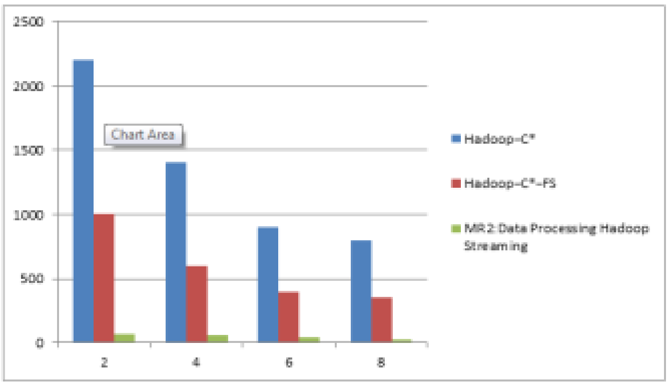


Fig.3. The overhead of moving data from Cassandra into the alternative streaming models for Java applications

Fig.4. Running read intensive workloads over Cassandra data using File system for Map Reduce processing.

VII. FUTURE SCOPE

To proposed two different tools one is mongodb and second is Cassandra data set for improving the Performance up to hug amount of data on the hadoop streaming.

VIII. CONCLUSION

Now a day data increases day by day the storage, retrieval and analysis of big data in structured databases like Oracle and Mysql is not possible to managed the so have to presented many Nosql system among them mongodb is preferable for as an alternate for Mysql , still it is an active search are for data mining to mine knowledge from big data. This Paper Present on "Big Data" used software pipeline to combine the data stores such as Cassandra with distributed programming models such as Map Reduce. In this paper to show two different tools one is Cassandra data set used directly to perform on MapReduce Process and second is Mongodb is used to connect different node and also distributed data in different format file like JSON.

REFERENCES

1] Apache Hadoop. http://hadoop.apache.org.

[2] Apache HBase. http://hbase.apache.org.

[3]Cassandrawiki,operations.
http://wiki.apache.org/cassandra/Operations.

[4] Datastax. http://www.datastax.com/.

[5] National Energy Research Scientific Computing Center.
http://www.nersc.gov.

[6]Projectvoldemort.http://www.project-voldemort.com/voldemort/.

[7] A. Abouzeid, K. Bajda-Pawlikowski, D. Abadi, A. Silberschatz, and A. Rasin. "*Hadoopdb: "an architectural hybrid of mapreduce and dbms technologies for analytical workloads*". Proceedings of the VLDB Endowment, 2(1):922–933, 2009.

[8] G. Ball, V. Kuznetsov, D. Evans, and S. Metson. "*Data aggregation system-a system for information retrieval on demand over relational and non-relational distributed data sources".* In Journal of Physics: Conference Series, volume 331, page 042029. IOP Publishing, 2010.

[9] B. F. Cooper, R. Ramakrishnan, U. Srivastava, A. Silberstein, P. Bohannon, H.-A. Jacobsen, N. Puz, D.Weaver, and R. Yerneni. Pnuts: Yahoo!*'s "hosted data serving platform. Proceedings of the VLDB Endowment*" 1(2):1277–1288, 2008.

[10] J. Dean and S. Ghemawat. Mapreduce: "*simplified data processing on large clusters. Commun. "ACM,* 51(1):107–113, Jan. 2008.

[11] A. Silberstein, B. F. Cooper, U. Srivastava, E. Vee, R. Yerneni, and R. Ramakrishnan." *Efficient bulk insertion into a distributed ordered table."* In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pages 765–778. ACM, 2008.

[12] A. Lakshman and P. Malik. "*Cassandra: structured storage system on a p2p network.*"In Proceedings of the 28th ACM symposium on Principles of distributed computing, PODC '09, pages 5–5, New York, NY, USA, 2009. ACM.

[11] E. Dede, Z. Fadika, J. Hartog, M. Govindaraju, L. Ramakrishnan, D. Gunter, and R. Canon. "*Marissa: Mapreduce implementation for streaming science applications*." In E-Science (e-Science), 2012 IEEE 8th International Conference on, pages 1–8, 2012.

[12] E. Dede, B. Sendir, P. Kuzlu, J. Hartog, and M. Govindaraju. "*An evaluation of cassandra for hadoop."* In Proceedings of the 2013 IEEE Sixth International Conference on Cloud Computing, CLOUD'13,pages494–501, Washington, DC, USA, 2013. IEEE Computer Society .

[13] E. Dede, B. Sendir, P. Kuzlu, J. Weachock, M. Govindaraju, and L. Ramakrishnan. "*A processing pipeline for cassandra datasets based on hadoop streaming".*In Proceedings of the 2013 IEEE Big Data 2014 Conference, Research Track, BigData '14, Anchorage, AL, USA, 2014.

[14] J. Ekanayake, H. Li, B. Zhang, T. Gunarathne, S.-H. Bae, J. Qiu, and G. Fox. "*Twister: a runtime for iterative mapreduce." In HPDC,* pages 810–818, 2010.

[15]C. Yang, C. Yen, C. Tan, and S. R. Madden. Osprey: Implementing mapreduce-style fault tolerance in a shared-nothing distributed database. In Data Engineering (ICDE), 2010 IEEE 26th International Conference on, pages 657–668. IEEE, 2010.

[16]R. Taylor. An overview of the hadoop/mapreduce/hbase framework and its current applications in bioinformatics. BMC bioinformatics, 11(Suppl12):S1, 2010.

[17] K. Shvachko, H. Kuang, S. Radia, and R. Chansler. The hadoop distributed file system. In Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on, pages 1 –10, May 2010.

[18] Z. Fadika and M. Govindaraju. LEMO-MR: Low Overhead and Elastic MapReduce Implementation Optimized for Memory and CPUIntensive Applications. Cloud Computing Technology and Science,IEEE International Conference on, 0:1–8, 2010.

[19] R. Sumbaly, J. Kreps, L. Gao, A. Feinberg, C. Soman, and S. Shah.Serving large-scale batch computed data with project voldemort. In Proceedings of the 10th USENIX conference on File and Storage Technologies, pages 18–18. USENIX Association, 2012.