Data Analyzing Using Map-Join-Reduce in Cloud Storage

Ruchi Bhardwaj
Department of Computer Science and Engineering
Sharda University
Greater Noida, India
ruchi.bhardwaj025@gmail.com

Neetesh Mishra
Department of Computer Science and Engineering
Sharda University
Greater Noida, India
neetesh.mishra30@gmail.com

Rajiv Kumar
Department of Computer Science and Engineering
Sharda University
Greater Noida, India
rajivbec@gmail.com

Abstract— **Data analysis and maintenance in cloud computing is a challenging task which allows large volume of data to be processed in large clusters. Recent days Map Reduce Model have shown great value in processing huge amount of data on very large clusters. Map Reduce paradigm consists of two phases, mapper and reducer. Mapper performs filtering criteria and Reducer performs aggregation task, but Map Reduce supports a homogenous data set that signifies the same filtering logic is applied by mapper function on each tuple in the data set. However these techniques do not performed well in case of complex data analysis that may require the joining of multiple data sets. In order to improve these problems a CloudView framework has been proposed for data storage, processing and analyzing the massive machine data which are collected from cloud environment in which Case Based Reasoning (CBR) approach is used for fault prediction. In this paper, an Enhanced CloudView (ECV) framework has been proposed for data processing, maintenance and analyzing the massive machine data. CloudView is formulated by Map Reduce model whereas ECV framework will use Map-Join-Reduce model. This model will performs mapping-join-reduction task in two successive Map Reduce jobs. First it will filter the logic to all the datasets in parallel, joins the resulted tuple and further reduces for final aggregation and finally, it combines all partial aggregation results and produce the final result. This additional joiner model will perform a fast processing in a heterogeneous data set by using join reduce function, which will improve the efficiency and scalability of the system.**

Keywords—Cloud computing; map reduce; map join reduce; cloudview; extended cloudview.

I. INTRODUCTION

Cloud computing has been an emerging area in the eyes of researchers since last few years. Cloud computing shows the path for an affluent societies. Data maintenance in cloud computing is a challenging issue. It reduces the risk by increasing the accuracy and reliability of data. Recent machines and systems like nuclear power plants , gas turbines and wind turbine farms have number of sensors that gather the data continuously for monitoring the system conditions and for fault predicting and monitoring in cloud computing. A number of methods have been proposed for this purpose. Recently a CBR approach for fault prediction is proposed by [1] but this approach is widely used in industry. By using this method, find the solutions for new problems by using the information of past cases. The information of past cases is assembling in the Case Base. The CBR cycle is based on four "R Process" that is, Retrieve, Reuse, Revise and Retain process. First R process retrieve the similar kind of cases from the case base, second R process, reuse the knowledge from retrieve case to solve the target case problem, third R process, revise the solution and the fourth R process, retaining the new solution information to the case base. This signifies that the CBR is the problem solving and learning approach. Further in order to improve the efficiency of the system CloudView [2] has been proposed. This framework is aimed at supporting wide variety of data analysis algorithm within cloud architecture; it uses the CBR approach for fault predicting. This is based on Map Reduce function. Map reduce is a popular tool for processing large scale data analysis in cloud environment. However there are some problems with this approach. It is mainly designed for performing a filtering-aggregation data analytical task on a single homogenous data set. Expressing join processing function map () and reduce () is not very convenient. Also in certain situations multiway join using map reduce is not efficient. Further to improve the scalability and efficiency of data in the system, we propose a framework called Enhanced Cloud View which provides cloud storage for storage, analyzing, processing the data which are used for predicting the fault for dawning faults. This is based on Map Join Reduce which improve the scalability of data in the system. By using Map Join Reduce, we process the multiple heterogeneous data sets. It has two successive jobs of Map Reduce. The framework

is based on Hadoop [3] which is a framework for running applications on large clusters built of commodity hardware. Hadoop comprises of two major components i.e, HDFS (Hadoop Distributed File System) [4] and Map Reduce.

The rest of the paper is organized as follows: in section II, presents the brief description of the components of Hadoop, section III, presents the approaches used for fault analysis, section IV, and presents the proposed work and finally section V concludes the paper.

II. HADOOP

Hadoop is a framework for running applications on large clusters built of commodity hardware. It comprises of two major components, Distributed File System HDFS and Map Reduce.

A. HDFS:

Hadoop provides a distributed file system and framework for analysis and transformation of huge amount of data from clusters by using Map Reduce paradigm. The important characteristics of Hadoop breaks the data from Hadoop clusters into smaller pieces called as blocks and distribute throughout the clusters. After that, the Map Reduce functions that is, Map and Reduce functions executed on smaller subsets of large datasets, by doing this it provide scalability. HDFS stores file system metadata and application data separately.

B. Map Reduce:

Map Reduce is a parallel data processing paradigm that works in tandem with HDFS. It consists of two phases: the Mapper and the Reducer. Mapper performs map task, they read data from HDFS. Map is an individual task that transform input record to intermediate key/value pairs , the intermediate results is stored in the local buffer. When all the map tasks are completed, then the Reducer task is begin, which aggregate the intermediate result of the same key.

Now days, the continuous streaming data constitutes of important and consist large portion of web data, like weather forecasting data , news , stock data etc. Map Reduce does not support the stream data processing, which decrease the efficiency of stream processing application. Map Reduce also face problem during processing of complex data analytical task. It is designed to perform the filtering aggregation data analytical task on same type of data that is homogenous data set. To solve this problem [5] introduce the extension of Map Reduce paradigm that is Map Join Reduce paradigm.

C. Map Join Reduce

Map Join Reduce is the extension of Map Reduce Paradigm which introduce a filtering-join-aggregation model which is extension of filtering aggregation of Map Reduce paradigm. This model is suitable for processing multiple heterogeneous data sets, they introduce join () that is joiner. The Map Reduce paradigm performs two phases, Mapper phase, which perform filtration logic and Reducer phase that perform aggregation logic, which is mainly suitable for homogenous data sets on which the same filtering logic by mapper is applied on the data set. Map Join Reduce extends to process multiple heterogeneous data set. This model allows pipelining intermediate results between joiners and reducers. Joiner and Reducer runs inside the same reduce task phase.

III. APPROACHES

This section describes the approaches for Machine data analyzing, for fault diagnosis and for predicting the nature of the fault in Cloud storage. Basically Fault analysis approach in cloud storage is broadly categorized into offline, online and hybrid as depicted in Fig. 1.

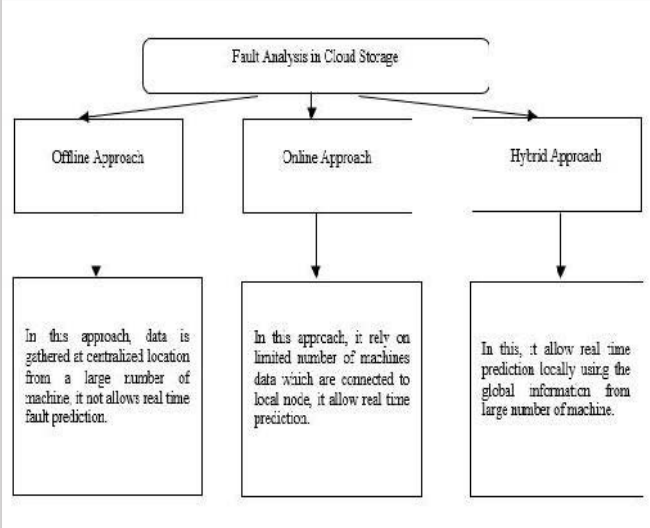

Fig. 1. Fault analysis approach in cloud storage.

A. OFFLINE APPROACH:

In Offline Approach as shown in Fig. 2 the data is collected and stored at a centralized location from a large number of machines, which is used to recognize the nature of the fault. Devaney and Cheetham [6] used case based reasoning approach for Gas turbine diagnostics used by general electric, in which the data sets from a large number of machines are gathered at a centralized location that is maintenance center. The disadvantage of Offline approach is, it take a time scale of weeks to months for updating the information of new cases so that the CBR system albeit slowly, in the dynamic environment. The Offline Approach does not allow real time sensor data analytics task for predicting faults in dawning faults.

B. ONLINE APPROCH:

The Online approach as shown in Fig. 3, allows the real time sensor data analytics task from machines to predict the fault of the nature for dawning faults. SKF WindCon [7] is a condition monitoring solution for Wind turbines developed by SKF, which is based on a local approach. Fig. 3 shows in online approach. There is a local node in each cluster or in each plant, which perform the function of recognizing the nature of dawning fault and store the data of all the machines of that plant, which is directly connected to the Internet service which is directly access by a unique url, to view and monitoring the

machines which are connected to the local node. The limitation of Online Approach is that it contains a limited number of cases for fault prediction of dawning faults unlike the offline Approach in which the case base are more elaborate.

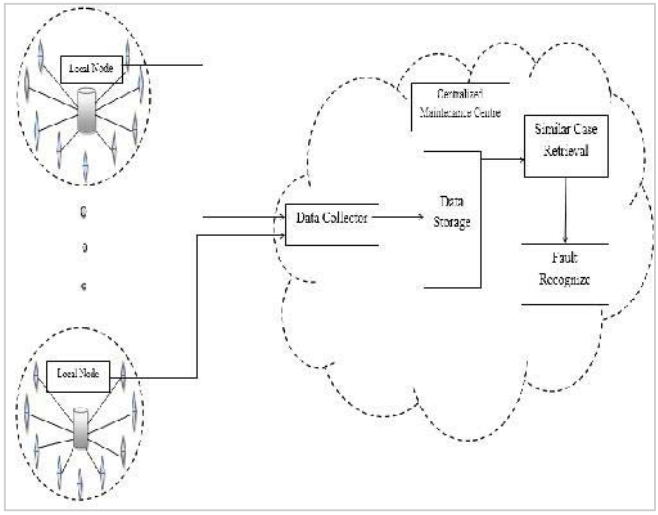

Fig. 2. Offline approach used in cloud storage.

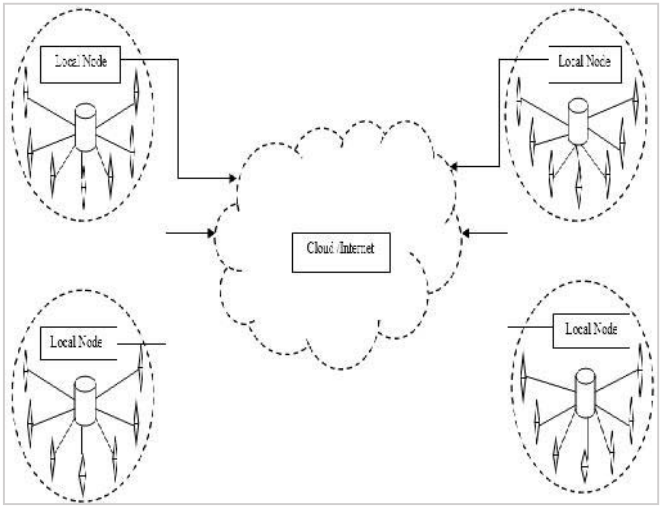

Fig. 3. Online approach used in cloud storage.

C. HYBRID APPROACH:

The Fig. 4 shows the architecture of proposed Hybrid approach. It is the combination of Offline approach and Online approach .The proposed framework called ECV which is a extension of the CloudView framework that allow massive data analyzing and maintenance in a cloud storage, with the benefit of real time sensor data analysis for fault prediction the dawning fault locally with the offline approach. ECV provides the ability to maintain the data which are achieved by performing several steps that is, it gathered the data from the sensors of the large number of machines, further it filters and aggregate the data, then collect the data from the past failure cases from various machine in a case base that step is termed as the creation of case base and finally, in case base the new cases are updated at computing node in online mode which are directly connected to the cloud i.e case base updation. The benefit of proposed hybrid approach is that it is able to analyze the large amount of data in real time sensor and perform the fault prediction in machines by using the information of previous faults from a large number of machines. In this approach the HDFS and Map Reduce are used to maintain the data at cloud storage. The ECV consists some components like Data Collector, HDFS, Case Base Creation and Case Base Updation. The case base creation and case base updation are formulated by using Map Join Reduce model in ECV.

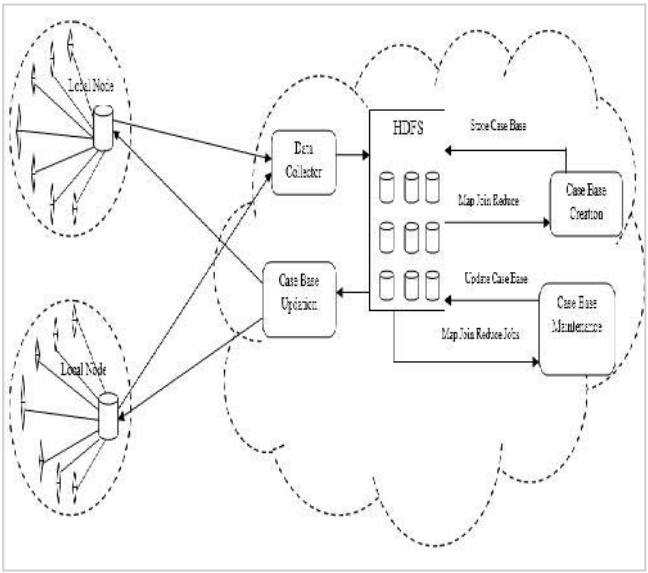

Fig. 4. Proposed system using hybrid approach

IV. RELATED WORK

There are two categories of the system which are able to perform large scale data analytical tasks on a large cluster that is Parallel Database System and Map Reduce based system. The research on Parallel Database System in the late 1980s [8]. The differences between both these systems are express in terms of their performance and scalability is explained in [9] and [10]. The scalability issues of these are further studies in [11]. Map Reduce is a parallel processing paradigm which is popularly used to handle large data from a large cluster. In [12] provides a simple and powerful interface that enable automatic parallelization and distribution of large scale computations, combined with an implementation of this interface that achieves high performance on large clusters of commodity PCs. [1] They presents description of CBR principles, methods and systems. CBR is a recent approach to problem solving and learning. They explain a task-method decomposition of CBR; representation of cases, or collection of cases on which the CBR is heavily dependent. In this paper we are proposing the extension of CloudView framework, which used CBR approach for machine fault prediction by using the case base, in which the past cases of failure information is stored, they used Map Reduce model

for creating its some components. In proposed work we are going to use Map Join Reduce model [5] for creation of the component of proposed framework that is ECV framework, which will allow to handle the heterogeneous data set, by using this technique we are going to improve the efficiency and scalability of data in the proposed framework. In this paper we are proposing the ECV Framework for storage, processing, analysis and maintenance the massive machine data, which are collected from a large number of machines in cloud storage, which is extension of CloudView framework, which used CBR approach for machine fault prediction by using the Case base, in which the past cases of failure information is stored. They used map reduce model which is suitable for homogenous data sets, in our proposed work we are going to use Map Join Reduce model for creation of the component of proposed framework, its functionality is to allow to handle multiple heterogeneous data set. There are some component of ECV they are, Data collector, Case base maintenance and Case base updation. In data collector of ECV, the stream data coming from the large number of machines which are aggregate in the data aggregator, it produce unstructured file , which are processed under the operation of Map-Join-Reduce producing the sequence structured file which are stored in the Hadoop distribution file system. By using Map Join Reduce improve the performance, efficiency and scalability of data in the proposed framework.

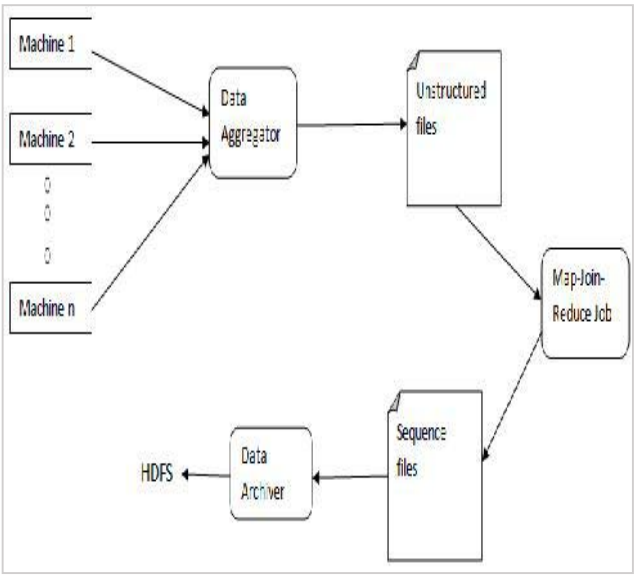

Fig. 5. Data collector of proposed system.

V. CONCLUSION

In this paper, we proposed a Map-Join-Reduce in fault recognition and their analysis in Cloud storage. The proposed system will improve the Map Reduce runtime with more efficient manner. The novelty of system lies in a filtering-join-aggregation model. This will allow a user to specify data analytical task that requires multiple dataset for aggregation. The system will offer three functions: map (), join () and reduce (). Further the proposed system will be tested for benchmark study against Hive on Amazon EC2.

REFERENCES

[1] A. Aamodt and E. Plaza, "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches,: AI Comm., vol. 7, pp.39-59,1994.

[2] A. Bagha and V.K. Madisetti, "Analyzing Massive Machine Maintenance Data in a Computing Cloud", IEEE Transaction, vol. 23, October 2012.

[3] Apache Hadoop, http://hadoop.apache.org, 2012.

[4] HDFS, http://www.aosabool.org/en/hdfs.html

[5] D. Jiang, A.K.H Tung and G. Chen, "Map-Join-Reduce: Towards Scalable and Efficient Data Analysis on Large Cluster", IEEE Transaction, vol. 23, September 2011.

[6] M. Devaney and B. Cheetham, "Case-Based Reasoning for Gas Turbine Diagnostics," Proc. 18th Int'l FLAIRS Conf., 2005.

[7] H. Timmerman, "SKF WindCon Condition Monitoring System for Wind Turbines," Proc, New Zealand Wind Energy Conf., 2009.

[8] D. DeWitt and J. Gray, "Parallel Database Systems: The Future of High Performance Database Systems," Comm. ACM, vol. 35, no. 6, pp. 85-98, 1992.

[9] A. Pavlo, E. Paulson, A. Rasin, D.J. Abadi, D.J. Dewitt, S. Madden, and M. Stonebraker , "A Comparison of Approaches to Large-Scale Data Analysis," Proc. 35th SIGMOD Int'l Conf. Management of Data (SIGMOD'09).

[10] J M. Stonebraker, D. Abadi, D.J. DeWitt, S. Madden, E. Paulson, A. Pavlo, and A. Rasin, "Mapreduce and ParallelDBMSs:Friends or Foes?" Comm.ACM, vol.53, no. 1, pp. 64-71, 2010.

[11] A. Abouzeis, K. Bajda-Pawlikowski, D.J. Abadi, A. Sileberschatz, and A. Rain, "HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads," Proc. VLDB Endowment, vol.2, no. 1, pp. 922-933, 2009.

[12] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters,"Proc. Operating Systems Design and Implementation (OSDI), pp. 137-150, 2004.