Biohouse, Volume 3: Issue 3, May-June 2017

Biohouse Journal of Computer Science

ISSN 2379-1500

Hybrid Approach to Enable Real-time Queries to End-Users

Atheer Musfer Al-hasf, Mona Mushabeb Al-qahtani, Hajer Saad Al-qahtani,
Asma Ali Al-shehri, Maryam Ali Al-shehri, Lalitha Saroja Thota,Mohrah Alalyan
Dept. of Computer Science, King Khalid University, Khamis Mushayt, Abha
Kingdom of Saudi Arabia

Abstract: *Data management is an important asset to any organization. The information grows everyday and it gets difficult to store information in local servers. There have been challenges regarding maintaining huge data sets of information and use it for analytical purpose to predict trends and make strategies accordingly. It helps organization to progress and their data gets secured without being on risk. Hadoop is ideal to process large data information and is designed to assist developers to manage clusters of information. MapReduce is linearly scalable model in which programmer exchanges information easily and can implement desired functionality easily. The document is based on a research and is carried out to enhance and propose different techniques to modify existing framework*

I. INTRODUCTION

Big data refers to huge sets of data that are tough to be managed, analyzed and processed using conventional tools and frameworks as it consumes a lot of time. It has been quite challenging for a period of time to figure up a better solution to cope up with big data. Data that is usually over 30-50 terabyte or more depending upon the growth rate is considered as 'Big data'. Obviously, it gets tough to handle much information and process it according to needs. Many scientists in different decades have come up with solutions to minimize the problem and assist in efficient data handling. Structuring data in an organized form that avoids the maximum redundancy and decomposing it into different layers to learn better flow of the information and processes is one prerequisite to the solution. The processes can be decomposed into three different layers those include application, computational and infrastructural layer.

Hadoop is an open source distributed framework that assists allows to process big data information. It is most flexible in order to store, process and analyze huge sets of information. It is inspired and developed by Google's Map Reduce paradigm which now comes under Apache software.

The fundamental technique to target huge data sets is to break them into multiple parts. Each fraction of information can be processed at the same time as Hadoop runs a parallel computational technique of processing data. A huge data set can be cut down into smaller pieces to process and analyze subsets of data more efficiently and conveniently. Hadoop is comprised of different components that are responsible for performing different tasks and all are available under Apache license

II. SCOPE OF WORK

The document is an illustration of a research that has been carried out in order to make propositions regarding managing huge amount of data. Hadoop sets an ideal approach to process petabytes of data sets and analyze different trends of information. It is designed for batch processing rather than interactive use by users. The scope of the system is to enhance techniques in handling and managing big data

III. PROBLEM STATEMENT

Hadoop possibly serves the best services in order to carry out big data routines. But, there have been certain limitations of the systems such as data Latency, Security Issues, Incapability with small data and In-depth Knowledge

IV. EXISTING SYSTEM

Hadoop works on large number of machines without sharing any memory. Hadoop distributed file system is capable of storing huge sets of data without losing out any information. It stores and sets to different servers for processing data information. It is capable to revive data even if the infrastructure gets disturbed and components fail to execute their task. Hadoop sets a cluster of machines to carry out different work processes; if any of them fails to operate then the load is shifted to other available machines in the cluster. Data is been decomposed into blocks of information and each block is processed in a parallel manner to store and information to different pool of servers. HDFS stores copies of information to multiple servers so no data is lost.

MapReduce serves as processing pillar in the system as it allows shifting information in different parts and process in parallel. MapReduce is responsible for providing results to the queries directed by the client. 'Map' function divides and processes query at the node level where 'Reduce' helps to combine results of the prior function by retrieving it from multiple server ends and determine results to the query. MapReduce is particularly designed for batch processing and not optimized for synchronous processing

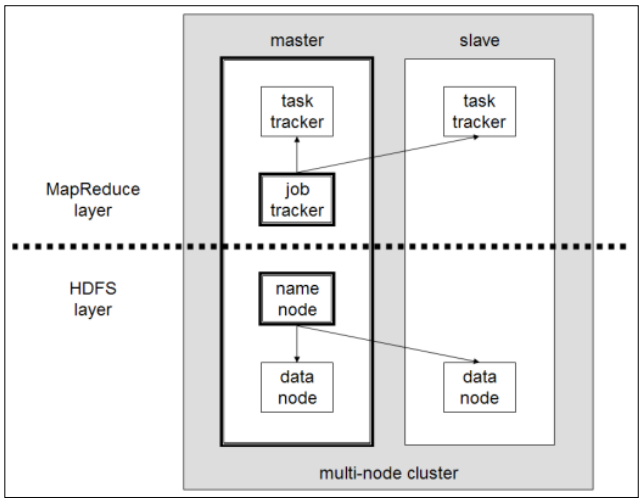

Fig 1: HADOOP NODE

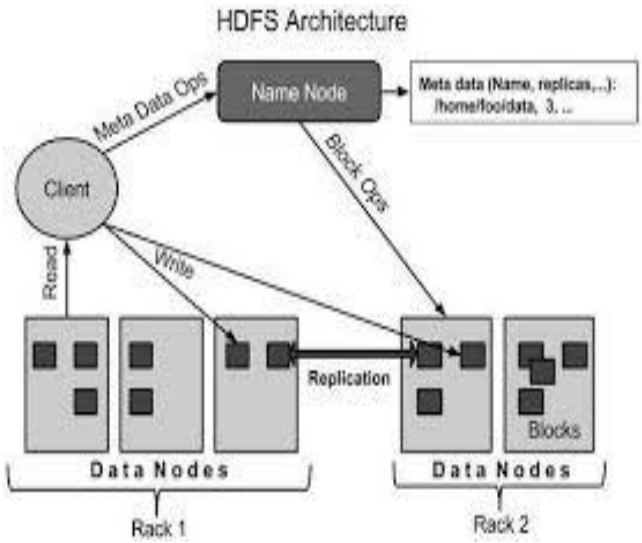

Fig 2: HDFS ARCHITECHTURE

MapReduce is a framework using which we can write applications to process huge amounts of data, in parallel, on large clusters of commodity hardware in a reliable manner. MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job

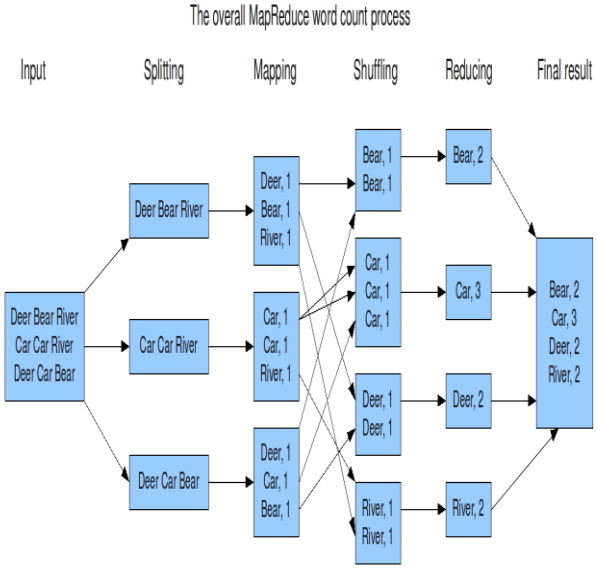

Fig 3 MapReduce Architechture

V. PROPOSED SYSTEM

The purpose of the system is to enhance techniques of storing and processing huge set of data information. Hadoop is a distributed file system that takes structured and non-structured data from different sources and processes it following the conventional data management methods. The system is designed to exchange information from a single server to thousands of machines with parallel processing. The information is broken into different parts and then processed in parallel manner which has much room for improvement. The research identifies parallel execution of processes as a problem and proposes to enhance it with batch processing. The information is huge and can be collected from various sources and can grow with time. Enhancements that the system has to offer are Data Latency, Real time queries, Easy Interaction, Synchronous processing and Real time search

VI. RESEARCH OBJECTIVES

Primary objective of the research is to find an optimal approach to support Hadoop with a framework that is enhanced with capabilities to capture, process and provide analytics in an efficient manner. The platform aims to work out on volume, velocity and handling different variety of data. Velocity is targeted in order to make processing more efficient so that data leads to perfect analysis. The Objectives also include Data Freshness, Timeliness, Search Enhancements, Real-Time Querying

VII. SYSTEM DESIGN

MapReduce framework can be illustrated though a sequence diagram. It is illustrated below:

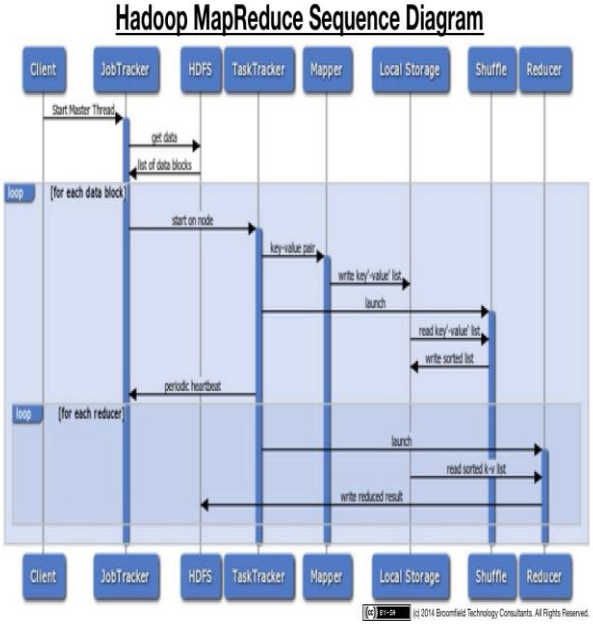

Fig 4 Sequence Diagram

VIII. SYSTEM IMPLEMENTATION AND TESTING

The project is implemented using Java

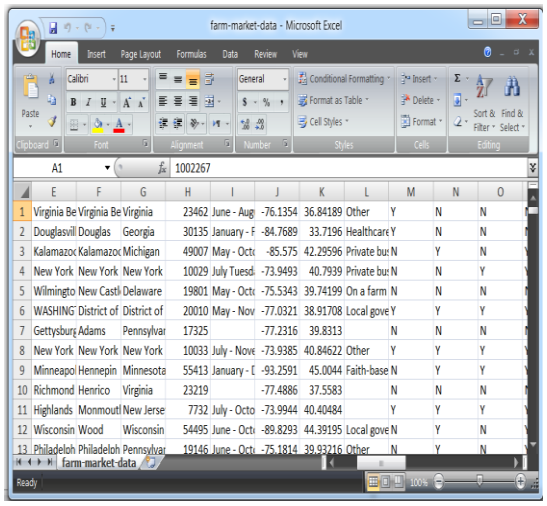

Fig 5 Database screens

We are working on very large data and this means that we have in the databases more than Terabytes and in this project was the use of a database published by the US Department of Agriculture was designed on Microsoft Excel and this does not mean that the implementation of the databases only Excel, Different databases, processing and analysis

The system is tested for Unit Test, Integration Test, System Testing and Acceptance Test and is working good.

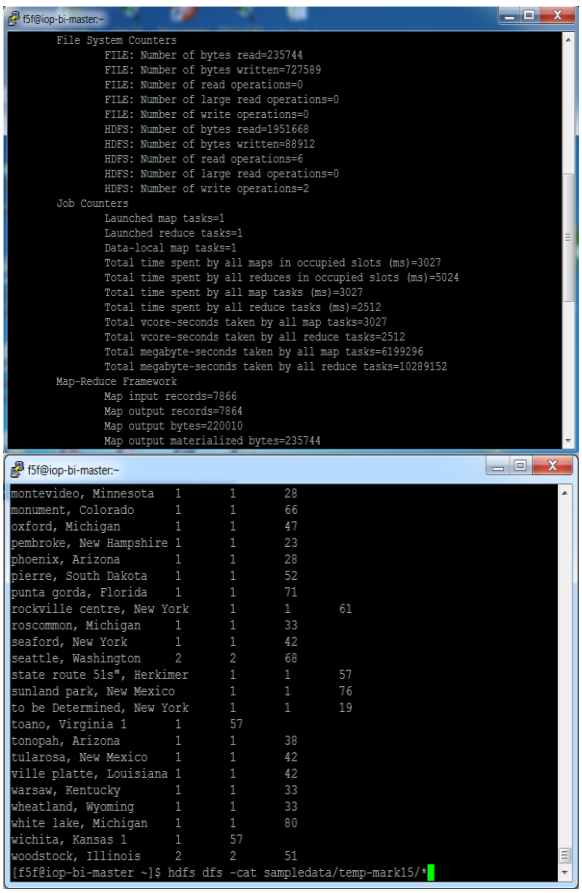

Fig 6 Design screens

IX. CONCLUSION

A research study has been carried out in order to make enhancements in the frameworks of Hadoop. Hadoop is ideal for big data processing but there is a room of improvement which is led by the problem statement. Processing rate is directly proportional to the growth of data which is quite challenging for developers to rethink their modeling perspectives. A literature survey has been carried out that how the system has improved over the years. Batch processing is slightly shifting to stream and interactive processing purposefully. Real time queries and search has been suggested using SQL which is much simple in order to fetch data information. The research fulfilled its purpose and stated propositions on the existing system.

X. FUTURE WORK

We look forward to applying the query system in real time in major organizations, ministries, and agencies that deal with huge amounts of data, which improves the functional output of the organization, and easy to deal with massive data, and saves time and effort, and the reduction of the expected errors

XI. REFERENCES

1. Big Data A Revolution that will Transform how we Live, Work and Think By: Viktor Mayer-Schonberger and Kenneth Cukier .
2. Hadoop The Definitive Guide, 4th Edition By: Tom White.
3. Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today′s Businesses by Michael Minelli, Michele Chambers, Ambiga Dhira
4. Online: https://www.edrawsoft.com/ uml-introduction.php
5. Online: http://www.ijsrp.org/research-paper-1014/ijsrp-p34125.pdf

International Journal Series

ISSN 2379-1500