International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Comparative Study of MapReduce and Pig in Big Data

Radhiya A. Arsekar†*, Ankita V. Chikhale†, Vaibhav T. Kamble† and Vinayak N. Malavade†

†Department of Computer Engineering, Mumbai University, Maharashtra, India

Accepted 10 March 2015, Available online 15 March 2015, **Vol.5, No.2 (April 2015)**

Abstract

In today's world of economic and social industry there exists an over demand and rising need of information. Due to which a problem of storage and maintaining big data comes into picture. It is challenging task to manage and retrieve relevant information from big data. This data basically has storage memory in terabyte and petabyte hence becoming it difficult to process and analyze. The specified project basically uses Hadoop, a tool specified by Apache Server, for information retrieval. Hadoop is a Java Software Framework that supports data intensive distributed applications and is developed under open source license. Many websites including Facebook and Twitter rely on Hadoop. The two major pieces of Hadoop are HDFS and MapReduce. In this paper we are focusing on MapReduce technique, one of the most common techniques used for retrieval of information.

Keywords: *Big data, Hadoop, HDFS, MapReduce, Pig.*

1. Introduction

In today's electronic age, increasing number of organizations are facing the problem of explosion of data and the size of the databases used in today's enterprises has been growing at exponential rates. Data generated through sources like business processes, transactions, social networking sites, web servers, etc. remains in structured as well as unstructured form. For processing and analyzing this huge amount of data or for extracting meaningful information from them, Hadoop is brought to use. Hadoop is an open source software framework created by Doug cutting and Michael J. Cafarella (Apache Hadoop). 'Hadoop' name was inspired by Doug cutting son's toy elephant. Hadoop was designed for handling large data sets. Google published two papers in 2003 that described the architecture of Google's distributed file system, called GFS, which was would solve their storage needs for the very large files generated as a part of the web crawl and indexing process . Hadoop was written in java programming language and could handle all type of data including audio files communication records e-mails, multimedia, picture and log files. Using Hadoop there exists unlimited storage and processing of data. Hadoop consists of sub projects namely MapReduce, Pig, Hive, Zookeeper and Mahout. Among these, MapReduce tends to be a popular technique, in which application is divided into many small fragment each of them is executed on various nodes in cluster. Hadoop uses HDFS for storage purpose and file handling providing fault tolerance as well as throughput access to large data set. Also one of the most important benefits of Hadoop is to limit the communication between the nodes and makes the system more reliable (Shivaraman Janakiraman).

For storing more and more data, people basically opt for two techniques, that is, either scale up or scale out. In scale up, for a bigger database containing more data, a smaller machine is replaced by a bigger machine whereas in scale out, more resources are added rather than replacing the old one. Hadoop hence uses the scale out technique with a Hadoop clusters having ten to hundreds of machines is standard.

Big Data

The term Big data is used for data sets so large and complicated that it becomes difficult to process, capture and manage data within a tolerable elapsed time. In today's world, we're surrounded by data. People and Machines, both, are generating and keeping more and more data. Popular sites such as Google, Yahoo, Amazon, and Microsoft needed to go through terabytes and petabytes of data to figure out which websites were popular, what books were in demand, and what kinds of ads appealed to people. The main problem they faced was that existing tools were becoming inadequate to process such large sets of data. Big Data impacts include Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data - the equivalent of 167 times the information contained in all the books in the US Library of Congress, Facebook handles 40 billion photos from its user base and so on (Aditya B *et al)*,

*Corresponding author: **Radhiya A. Arsekar**

2012).Example of big data includes social data from social network or large scale ecommerce where data size is constantly ranging from terabytes to petabytes in a single data base. The problem faced in this electronic age is that, using relational databases and desktop statistics/visualization packages it is difficult to process data on thousands of servers. Various problems which arise while processing big data include scalability, unstructured data, accessibility, real time analytics, fault tolerance and many more. These problems are not only limited to structuring of data but also the type of data and storage of data i.e., whether the data encodes video, images, audio, or text/numeric information.

Big Data consist of following terminologies

A.Hadoop
B.HDFS

A. Hadoop: Hadoop is an open source platform that provides implementations of both the MapReduce and GFS technologies and allows the processing of very large data sets across clusters of low-cost commodity hardware (Garry Turkington).

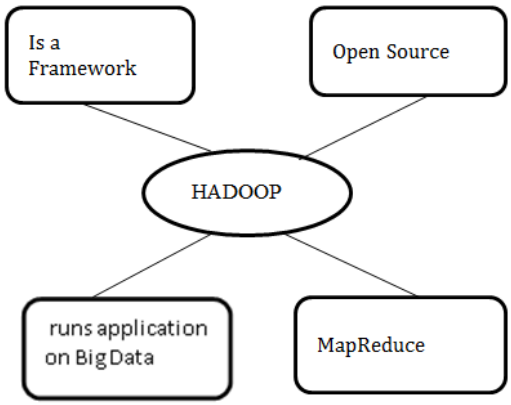

Fig 1.1 Hadoop architecture

Figure 1.1 gives us idea about Hadoop architecture. In here, different clients provide information to cluster of commodity machines networked together in one location. Storage and processing of data is the main job of this Hadoop cluster of servers. Users, which act as clients, submit their jobs to Hadoop cluster which can be stored in their individual desktop at remote location. Hadoop consists of two main sub projects namely HDFS (Hadoop Distributed File System) and MapReduce.HDFS basically caries out replication of data which achieves higher throughput and availability of data. MapReduce tasks run directly on the HDFS to hold the required data hence integrating tightly with HDFS.

B. HDFS: HDFS is a distributed file system, unlike the regular file system which spreads storage across multiple nodes. It is designed for storing very large files. Very large files in this context mean the files that are hundreds of megabytes, gigabytes, or terabytes in size. The key features are: Most file systems store around 4-32 KB whereas HDFS stores files in blocks typically at least 64 MB in size.

The main disadvantage of HDFS being that it is poor at seeking requests for many small files. HDFS handles workloads of write-once and read-many type. Each storage node runs a process called a DataNode that manages the blocks on that host, and these are coordinated by a master NameNode process running on a separate host.(Garry Turkington) NameNode are termed as master nodes whereas DataNode are termed as worker nodes. The file system namespace is managed by NameNode, which maintains the file system tree as well as the metadata for all the files and directories in the tree. The output is stored as the namespace image and the edit log. The NameNode knows the DataNode but does not store the block locations persistently and when the system starts, the information is reconstructed. A file system is accessed by client in order to communicate with the NameNode and DataNode. DataNode also termed as 'the workhouses of the file system stores and retrieves blocks when they are told to. DataNode also keep a track of NameNode by periodically providing it with lists of blocks that they are storing.

HDFS uses replication, unlike other file systems that handle disk failures by having physical redundancies in disk arrays or similar strategies HDFS.HDFS is a distributed file system, unlike the regular file system which spreads storage across multiple nodes. It is designed for storing very large files. Very large files in this context mean the files that are hundreds of megabytes, gigabytes, or terabytes in size. The key features are: Most file systems store around 4-32 KB whereas HDFS stores files in blocks typically at least 64 MB in size. The main disadvantage of HDFS being that it is poor at seeking requests for many small files. HDFS handles workloads of write-once and read-many type. ‹Each storage node runs a process called a Data Node that manages the blocks on that host, and these are coordinated by a master NameNode process running on a separate host.(Garry Turkington)NameNode are termed as master nodes whereas DataNode are termed as worker nodes. The file system namespace is managed by NameNode, which maintains the file system tree as well as the metadata for all the files and directories in the tree. The output is stored as the namespace image and the edit log. The NameNode knows the DataNode but does not store the block locations persistently and when the system starts, the information is reconstructed. A file system is accessed by client in order to communicate with the NameNode and DataNode. DataNode also termed as 'the workhouses of the file system store and retrieve blocks when they are told to. DataNode also keep a track of NameNode by periodically providing it with lists of blocks that they are storing.HDFS uses replication, unlike other file systems that handle disk failures by

having physical redundancies in disk arrays or similar strategies.

2. Literature Survey

Hadoop consist of many techniques for parallel processing out of those techniques, in this paper, we are comparing two of the most widely techniques namely MapReduce and Pig.

A. Mapreduce: MapReduce programming model, by map and reduce function realize the mapper and reducer interfaces (ZHOU *et al*). It consists of mapper and reducer.

1. Mapper

Map function takes the input as a pair of key value and produces a group of intermediate <key, value> pairs which consists of two parts. The value is nothing but the data related to the task and the key is the group number of the value. These intermediate values with same key are combined and then forwarded to the reduce function.

2. Reducer

Reduce handles the intermediate <key, value> pairs provided by the Mapper. Reduce function carries out the 'merge' process, wherein it merges these values to get a small set of values. These <key, value> pairs can be seen as a letter, key being the letter's posting address and value being the letter's content. Hence we can imply that letters having the same address letters will be delivered to the same place. Programmer's job is to set up correctly these <key, value> pairs, MapReduce framework will do the rest.

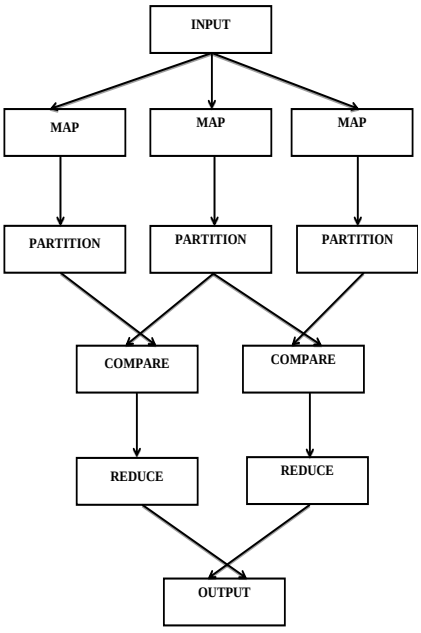

Fig 1.2 MapReduce Functional Diagram

This figure illustrates the information given below:

- Input Data:

The initial progression of MapReduce Task starts here and obtained data can reside in HDFS, HBase or other stages.

- Mapper:

Input data is passed on to Mapper where series of <key, value>pairs, which are used for individual map execution, are generated.

- Partition:

The intermediate <key, value> (k2, v2) called as subsets or partitions are input to reduce task. The values of same key are reduced together.

- Shuffle:

Since Hadoop executes several map tasks for a given job shuffling takes place. Shuffling is a process of moving nap output to reducer.

- Sort:

Hadoop sorts out the set of intermediate <key, value> pair to form <key, value> (k2,{v2,v2….}) before they are presented to reducer.

- Reducer:

The reducer's reduce() method is called once for each key assigned to a given reducer. This method receives value associated with which are then returned in undefined order. The output is obtained in (k3, v3) pair.

- Output Format:

The output format defines a location of output data and record written for storing data.

B.PIG

Pig and Hadoop both being open-source software projects are developed under Apache Software Foundation. Pig acts as joining key for SQL and Map-Reduce.SQL-style high-level data manipulation constructs Map- and Reduce-style functions or executable are assembled by pig. Pig is described as a data flow engine used to process large data sets. Companies like Yahoo use pig to deal with their data. The language used by Pig is Pig Latin which handles, one or more physical dataflow jobs, and then also carries out execution of these jobs. Pig currently uses the Hadoop open-source Map-Reduce implementation as its physical dataflow engine (Christopher Olston).

Pig allows three modes of user interaction

1. Interactive mode: In this mode an interactive shell, called Grunt, accepts Pig commands and is triggered only when the user asks for output through the STORE command.

2. Batch mode: In this mode a series of Pig commands, typically ending with STORE are submitted by user as a prewritten script.The semantics are identical to interactive mode (Alan F. Gates *et al*).
3. Embedded mode: Pig Latin commands can be written using java program via method invocations which in turn permits dynamic construction of Pig Latin programs, as well as dynamic control flow.

In paper(Djoerd Hiemstra *et al*) it is concluded that Using 15 machines to search a web crawl of 0.5 billion pages, the proposed MapReduce approach is less than 10 times slower than a single node of a distributed inverted index search system on a set of 50 queries.
If more queries are processed per experiment, the processing times of the two systems get even closer.

In paper (Christopher Olston) it is stated that the current Pig implementation incorporates two simple but critical optimizations: (1) Pig automatically forms efficient pipelines out of sequences of per-record processing steps. (2) Pig exploits the distributive and algebraic properties of certain aggregation functions, such as COUNT, SUM, AVERAGE and some user defined functions, and automatically performs partial aggregation early (known as combining in the Map-Reduce framework), to reduce data sizes prior to the expensive data partitioning operation.

3. Comparison between Mapreduce and Pig

Both PIG and MapReduce are used for parallel processing but MapReduce technique is highly efficient and widely used compared to Pig.

Table 1.1: Comparison between MapReduce and Pig

Pig	MapReduce
Since the entire program is based on PIG transformations, there is no much need of programming skills.	Expertise knowledge of programming is required to write a MapReduce program.
As compared to MapReduce program line of code is less. Equivalence Ratio of MapReduce to pig is 20:1.	Huge line of code is required.
Internally Pig gets converted into MapReduce program and gets executed.	MapReduce program compiles and executes on its own.
It is simple to write and execute Pig script.	It is slightly complex process to write and execute MapReduce program.

Future Work

Hadoop has many new technologies such as Zookeeper and Hive. We would focus on comparing these techniques with MapReduce technique and would note down which is more efficient among these. Future work will focus on performance of Hadoop on cloud platforms.

Conclusions

As Big Data becomes difficult to process using on-hand data management tools or traditional data processing applications there exists use of Hadoop tool to manage big data. The challenges include capture, storage, search, sharing, transfer, analysis and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data. Hadoop basically resolves this problem by managing the big data and resolving the much needed problems faced by big data. .

References

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

Aditya B. Patel, Manashvi Birla, Ushma Nair (Dec 2012) Addressing Big Data Problem Using Hadoop and Map Reducein 2012 Nirma University International Conference on Engineering, Nuicone-2012, 06-08

Shivaraman Janakiraman ‘’Fault Tolerance in Hadoop ForWorkMigration

Hadoop Beginners Guide by Garry Turkington.

ZHOU , Jingsheng LEI, Wenjun YE Large-Scale Data Sets Clustering Based on MapReduce and Hadoop Ping School of Computer and Information Engineering, Shanghai University of Electric Power, Shanghai 200090,China.p

Christopher Olston, Benjamin Reed, Adam Silberstein, Utkarsh Srivastava Automatic Optimization of Parallel Dataflow Programs at Yahoo! Research

Alan F. Gates, Olga Natkovich, Shubham Chopra, Pradeep Kamath,Shravan M. Narayanamurthy, Christopher Olston, Benjamin Reed,Santhosh Srinivasan, Utkarsh Srivastava Building a HighLevel Dataflow System on top of MapReduce:The Pig Experience Yahoo!, Inc

Djoerd Hiemstra and Claudia Ha MapReduce Information Retrieval Experiments, University of Twente..