Distributed Computing and Hadoop in Statistics

Xiaoling Lu and Bing Zheng
Center For Applied Statistics, Renmin University of China, Beijing, China
Corresponding author: Xiaoling Lu, e-mail: xiaolinglu@ruc.edu.cn

Abstract

Big data is ubiquitous today. The big data challenges current numeric statistical and machine learning methods, visualization methods, computational methods and computational environments. Distributed Computing and Hadoop help solve these problems. Hadoop is an open source framework for writing and running distributed applications. It consists of the MapReduce distributed compute engine and the Hadoop Distributed File System (HDFS). Mahout produces machine-learning algorithms on the Hadoop platform. Rhadoop, Rhive and RHipe are the merger of R and Hadoop. Finally we show two examples of how enterprises use Hadoop to their big data processing problems.

Keywords: Distributed Computing, Hadoop, Statistics, Mahout, R

1. Introduction

Today, we're surrounded by data. In one day, the information is consumed by global internet traffic to fill 168 million DVDS. In one day, people sent 294 billion emails, write 2 million blog posts, download 35 million apps. This has been demonstrated by Mayer-Schonberger and Cukier (2013). We regularly encounter limitations due to the exponential growth of data. It is difficult to go trough petabyte, exabyte even zettabyte of data to figure out which websites are popular, what kinds of ads appeal to people, and what books are in demand. When we are facing big data sets, we must face three challenges. First of all, the sets of big data are too large to use on-hand database management tools or traditional data processing applications. So, the difficulties include capture, curation, storage, search, sharing, transfer, analysis, and visualization. Besides, big data sets include different types of data. It isn't just the number.It also involves pictures, viedos, sounds, and so on. We need to use diverse methods to deal with different types of data. Last but not the least, computing speed is also a very important point.

So the traditional tools are becoming inadequate to process such big data sets. Around 2004, Google published two paper describing the MapReduce (Ghemawat et al. 2003) framework and the Google File System (Dean et al. 2004). Doug Cutting saw this opportunity and used these two technologies to Nutch, which was a subproject of Apache Lucene (http://hadoop.apache.org/). Hadoop started out as a subproject of Nutch. In 2005, hadoop was born, which was named after Doug Cutting son's toy elephant. In 2006, Doug Cutting was hired by Yahoo! and improved the Hadoop as an open source project. Today, Hadoop is a core part of the computing infrastructure for many web companies, such as Yahoo!, Facebook, Twitter, eBay, and Baidu.

2. Distributed Computing and Hadoop

In the past decades, we built bigger and bigger severs to solve the large scale data sets. But it isn't the best solution. An alternative that has gained popularity is to tie together many commodity or low-end computers as a single functional distributed computing. And the distributed computing refers to the use of distributed systems to solve computational problems. In distributed computing, a problem is divided into many tasks, each of which is solved by one or more computers. Actually, distributed computing is a wide and varied field. Hadoop is only one of the distributed computing. Which is better, bigger severs or Hadoop? Lam (2010) made a comparison. A high-end computer with four I/O channels each having a throughput of 100 MB/sec will reguire three hours to read a 4 TB data set. With hadoop, this same data set will be divided into smaller (typically 64 MB) blocks that are spread among many computers in the cluster. The cluster computers can read the data set in parallel and provide a much higher throughput. And, such a cluster of commodity computers is cheaper than one high-end computer.

Hadoop is an open source framework for writing and running distributed applications with amounts of data. Figure 1 illustrates what Hadoop is. A Hadoop cluster has many parallel machines that store and process big data sets. Generally speaking, the Hadoop cluster is in a remote location. The clients use their own desktop computers to submit computing jobs to Hadoop.

The principal advantages of Hadoop are as follows(Lu 2010):

Robust:Hadoop supports the running of applications on large clusters of commodity hardware. Sometimes the data is broken up. Computation on a piece of data takes place on the machine where that piece of data resides. This feature makes Hadoop popular in industry.

Scalable: Hadoop scales linearly to handle larger data by adding more nodes to the cluster. Hadoop is designed to be a scale-out architecture operating on a cluster of commodity PC machines. Adding more resources meaning adding more machines to the Hadoop cluster.

Accessible: Hadoop ties together many commodity or low-end computers. So even college students can quickly and cheaply create their own Hadoop cluster. Hadoop allows clients to quickly write efficient parallel code.

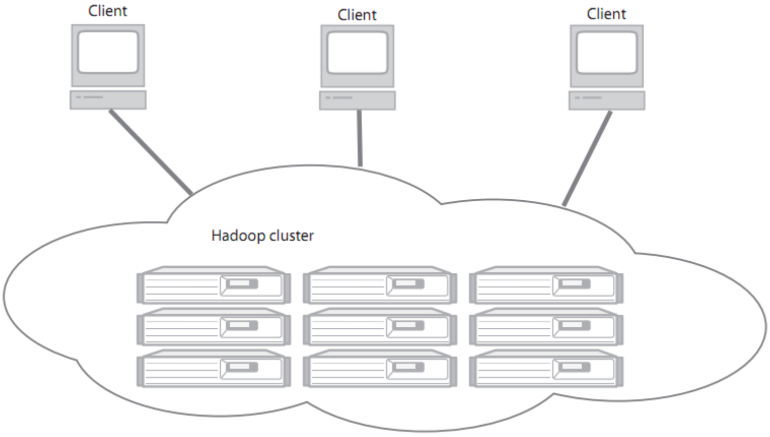

Figure1: Hadoop cluster is a set of commodity machines networked together. Clients send jobs into Hadoop cluster and obtain results.

Hadoop consists of MapReduce and Hadoop Distributed File System (HDFS), as well as a number of related projects, including Hive, HBase, and others. We first cover MapReduce, where the application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. Next we explain the HDFS in more detail.

2.1 MapReduce

MapReduce is a programming model for processing large data sets with a parallel, distributed algorithm on a cluster. MapReduce program comprises a map procedure and a reduce procedure. Each phase is defined by a data processing function, and these functions are respective (Rajaraman and Ullman 2012, Wang 2012).

Map procedure performs filtering and sorting. The master node takes the input, divides it into smaller sub-problems, and distributes them to worker nodes. The worker node processes the smaller problem, and passes the answers back to its master node. Reduce procedure performs a summary operation. The master node then collects the answers to all the sub-problems and combines them in some way to form the output. Figure 2 gives an example of MapReduce for word count problem.

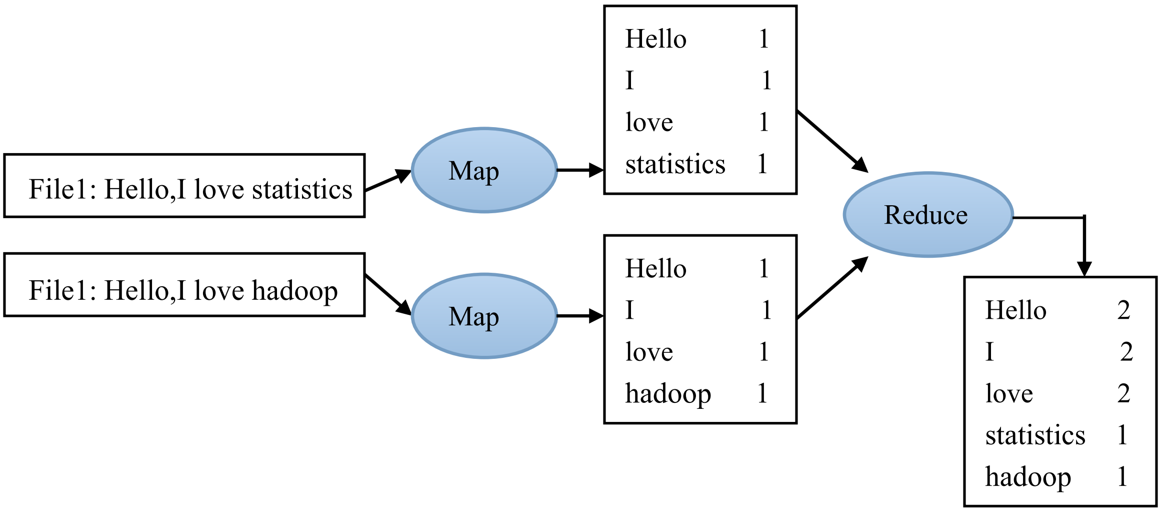

Figure2: In the Map procedure, each file is broken up. In the Reduce procedure, the counts will be aggregated.

2.2 HDFS

HDFS is a distributed, scalable, and portable file system written in Java for the Hadoop framework. We can store a big data set of 100 TB as a single file in HDFS, something that would overwhelm most filesystem. HDFS manage the file system through some important roles, NameNode, DataNode,Scendary DataNode, JobTracker, TaskTracker. Figure3 illustrate their interaction.

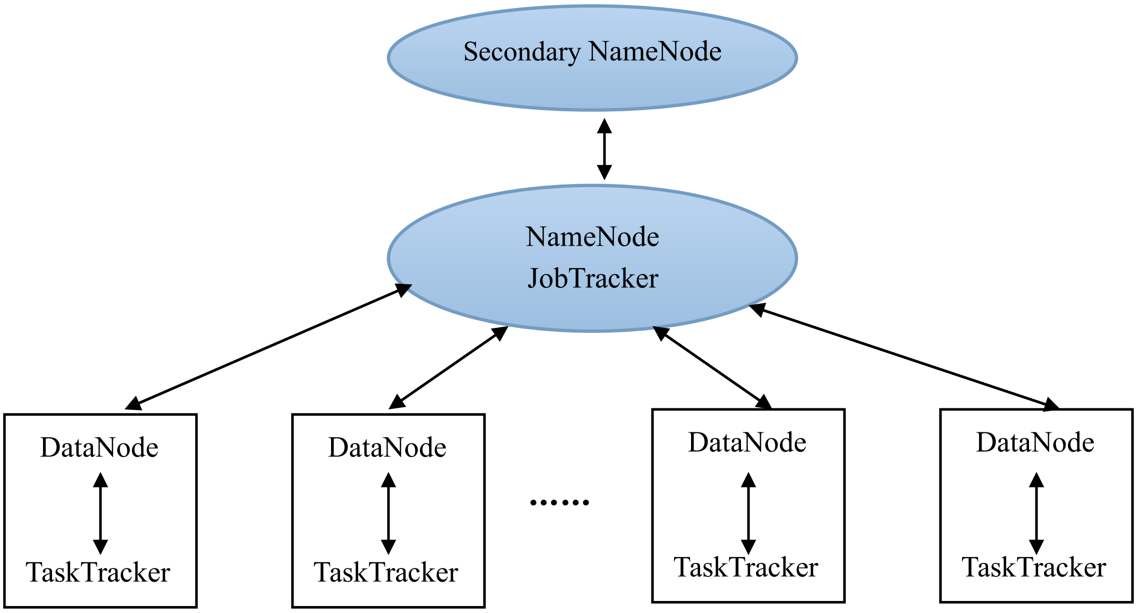

Figure3: Topology of a typical Hadoop cluster . NameNode and JobTracker are masters, the DataNodes and TaskTrackers are slaves.

The NameNode is the master of HDFS that directs the slave DataNode to perform the low-level I/O tasks. The function of the NameNode is the bookkeeper of HDFS. It keeps track of how the files are broken down into the file blocks, which nodes store these blocks, and the overall health of the distributed filesystem.

The DataNode is the actual files on the local filesystem. When we read or write a HDFS, the file is broken down into the file blocks and the NameNode will tell the client which DataNode each block resides in.

The Secondary NameNode is an assistant for monitoring the state of the HDFS. Like the NameNode, each cluster has one Secondary NameNode, and it typically resides on its own machine as well. The Secondary NameNode differs from the NameNode in that this process doesn't receive or record any real-time changes to HDFS.

The JobTracker is the liaison between application and Hadoop. The JobTracker determines the execution plan by determining which files to process, assigns nodes to different tasks, and monitors all tasks as they're running. There is only one JobTracker daemon per Hadoop cluster.

TaskTracker is responsible for executing the individual tasks that the JobTracker assigns. The responsibility of the TaskTracker is to constantly communicate with the JobTracker. If the JobTracker fails to receive a heartbeat from a TaskTracker within a specified amount of time, it will assume the TaskTracker has crashed and will resubmit the corresponding tasks to other nodes in the cluster.

3. Hadoop in statistics

As been well known, statistics is the study of the collection, organization, analysis, interpretation and presentation of data. Now, the big data exceeds the ability of commonly used statistical software to capture, manage, and process the data within a tolerable elapsed time. So, Hadoop is a very versatile tool that allows statistican to access the power of distributed computing.

3.1 Mahout

Mahout is a new open source project by the Apache Software Foundation. The goal of Mahout is to creat scalable machine-learning algorithms that are free to use under the Apache license(http://mahout.apache.org/). The field is closely related to statistics. Maybe the machine-learning algorithms aren't a new field. But Mahout is different. Mahout prodeces machine-learning algorithms on the Hadoop platform in order to deal with the big data.

The core algorithms of Mahout are batch based collaborative clustering, and classification. Although there are still various algorithms missing, Mahout is young and definitely growing.

3.2 Hadoop and R

As a data analyst or statistician, we are familiar with the R programming language. R is very widely used, has a very effective core development group, and has a vast number of user contributed packages that add up to. However, R can't deal with the complex big data. And more and more people use Hadoop to run MapReduce jobs or access HBase tables. The idea is interacting R programmer and Hadoop. Statistician can run R on a massively distributed system without having to understand the underlying infrastructure. So statistician can keep the mind on analysis and not the implementation details.

RHadoop is one of the open source project spearheaded by Revolution Analytics to grant data scientists access to Hadoop's scalability from their favorite language, R. RHadoop is a collection of R packages that let you run MapReduce jobs entirely from within R as well as giving you access to Hadoop files and HBase tables. RHadoop is comprised of three packages, rmr, rhdfs, rhbase. Rmr allows to write MapReduce programs in R. Rhdfs provides file level manipulation for HDFS, the Hadoop file system.Rhbase provides access to HBASE, the hadoop database.

In addition, RHive is an R extension facilitating distributed computing via HIVE query. Hive is a data warehouse system for Hadoop that facilitates easy data summarization, ad-hoc queries, and the analysis of large datasets stored in Hadoop compatible file systems. Hive provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL. RHive allows easy usage of HQL in R, and allows easy usage of R objects and R functions in Hive.

RHipe is another merger of R and Hadoop. RHipe has functions that access the

HDFS from R, that are used inside MapReduce jobs and functions for managing MapReduce jobs.It was first developed by Saptarshi Guha as part of his PhD thesis in the Purdue Statistics Department. Now there is a core development group and a very active Google discussion group. This is a very good trend for many bioinformatics people.

3.3 Case Study

Hadoop has great advantage in the field of statistical analysis. So it is popular in the real world application. We provide examples of how enterprises have used Hadoop as part of the solutions to their data processing problems.

China Mobile Limited was demonstrated by Chuck (2010).China Mobile Limited is a Chinese state-owned telecommunication company that provides mobile voice and multimedia services. The company is one of the largest mobile telecommunications companies by market capitalization today. China Mobile generates large amounts of data in the normal course of running its communication network. For example, each call generates a call data record (CDR), which includes information such as the caller's phone number, the callee's phone number, the start time of the call, the call's duration, information about the call's routing, and so forth. But traditional tools have a single server and cann't deal with the current data. So China Mobile Limited initiated an experimental project to develop a parallel data mining tool set on Hadoop and evaluated it against its current system. They named the project Big Cloud–based Parallel Data Mining (BC-PDM). By using Hadoop, the BC-PDM has a massive scalability and low cost. China Mobile Limited analyzes the big data to extract insights for improving marketing operations, network optimization, and service optimization.

Baidu incorporated on January 18, 2000, is a Chinese web services company headquartered in the Baidu Campus in Haidian District in Beijing. Baidu offers many services, including a Chinese language-search engine for websites, audio files, and images etc. To collect and analyze this stumbling data, Baidu requires its highly available back-end platform to collect, analyze, and transform millions of ratings per day. Baidu began to build a distributed platform. At the end of 2012, Baidu has three Hadoop cluster and consist of 700 machines. Everyday Hadoop processs 120 TB data, nearly 3000 assignments(Lu, 2010).

4. Discussion

Now we are facing tremendous amount of data piling up at TB data daily. Any queries against such growing data will be daunting jobs. Distributed computing and Hadoop seem good solutions for such tasks. On the other end, R is also a star in statistics community, in particular in academia. Now we have the integration of both R and Hadoop, Rhadoop, Rhive, Rhipe. We are entering into the ubiquitous computing age now. We believe that Hadoop and statistics have a better future.

Actually, Mahout is closely related to statistics. Mahout is so young that various algorithms are still missing. We statisticians can do some research on machine-learning algorithm as well.

Reference

Dean, Jeffrey, Sanjay Ghemawat (2004), "MapReduce: Simplified Data Processing on Large Clusters", http://labs.google.com/papers/mapreduce.htm.

Ghemawat, Sanjay, Howard Gobioff, Shun-Tak Leung (2003), "The Google File System", http://labs.google.com/papers/gfs.html.

Lam, Chuck (2010), *Hadoop in ation*, Manning Publications.

Lu, Jiaheng(2010), *Hadoop in ation*, China Machine Press (in chinese).

Mayer-Schonberger, Viktor, Kenneth Niel Cukier (2013), *Big Data: A Revolution*

That Will Transform How We Live, Work, and Think, Houghton Mifflin Harcourt.

Rajaraman, Anand, Jeffrey-David Ullman (2012), *Mining of Massive Datasets*, Cambridge University Press.

Wang, Bin (2012), *Mining of Massive Datasets,*Posts&Telecom Press (in chinese).