Hadoop: How Big Data Analysis is Impacting American Corporations

Written by Curren Mehta

There is a new technology being used by American corporations that has the potential to improve corporate profitability and change the way corporations interact with consumers. This article explores the new technology called Hadoop and how it is used in the emerging field of Big Data Analysis. Data of magnitudes never seen before can now be analyzed quickly and efficiently and the largest American companies are already benefitting.

Introduction to Big Data Analytics

Every time someone sends a photo, writes an email, visits a website, or buys something on the Internet, he/she is creating a trail of data. In fact, according to researchers at IBM, each day humans create 2.5 quintillion bytes of data, that's 2.5 * 10^19 bits, which is so much that in the last two years alone, 90% of the world's data has been created [1]. This is an alarming amount of data and until recently, computer scientists did not know how to effectively and efficiently process such massive amounts of data. However, armed with the developments in large computing power and analytical theory, computer scientists are now striving to use the surfeit of data to revolutionize the American Economy. Enter Big Data Analytics, the new technique relying on the Hadoop framework that engineers are using to derive insights from the deluge of data and power American corporations.

Splitting up the Data

From the perspective of a computer scientist at a major corporation, the techniques of Big Data Analytics can be divided in multiple steps. Underscoring these steps are two main principles of Big Data Analytics: Distributed File Systems and MapReduce. The Distributed File System is responsible for splitting up large amounts of data, and MapReduce is responsible for analyzing this data. In the 2000s Google created the framework of Hadoop (now developed by Apache), which takes care of both the Distributed File System and MapReduce aspects, and allows them to run complex analysis on large amounts of data. As Cloudera CEO Mike Olson stated in an interview, Hadoop is used to "solve problems where you have a lot of data…[and] you want to run analytics that are deep and computationally extensive" [2]. The first step in dealing with the massive amounts of data is to split it up into smaller pieces so that processers can run computations on it. That is the job of the Hadoop Distributed File System (HDFS). In the HDFS, the data is split into pieces called "blocks" that are generally 128MB or 256MB in size [3]. Because computers can only process a certain amount of data per unit time (due to certain hardware limitations), the only feasible way to deal with the amount of data is to place the data blocks onto multiple different servers, or computers.

The HDFS allows corporations to store and process more data than they could ever fit one machine, or server, by splitting up the data into blocks and placing each one into its own server, which are all connected into a "cluster" [2]. In the HDFS, each separate server that the data is split and placed on is called a "data node" and then a single server, known as the "NameNode", keeps track of which blocks of data reside on which node [3]. Consider, for example, that you have a massive ten thousand hour movie and that this movie has been split into smaller data blocks and put into servers by HDFS. You need some

way to know how to reconstruct the movie (i.e. to know the order in which to assemble the individual clips in order to get back the full movie in correct order). This is what the NameNode does; it keeps track of which data is stored where and what relation each block of data has to another. However, one problem is presented with this methodology. If one of the data nodes, or servers, gets destroyed or stops working, then you have lost an important part of your overall file (e.g. losing a segment of the movie in the previous example). To account for this, the HDFS duplicates each different block of data and puts it onto a different data node; the NameNode also keeps track of where the duplicates are housed [3]. Figure 1, shows the relationship between the different nodes:

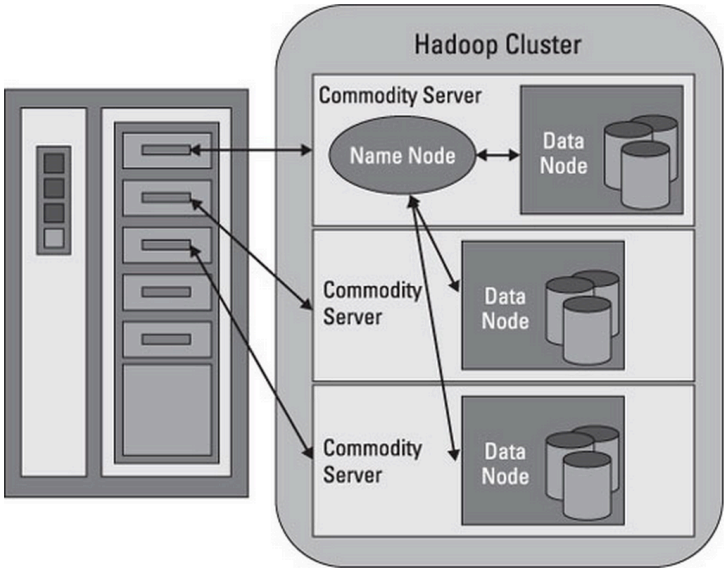

Figure 2: The relationship between the Data Nodes and the single NameNode is shown [3]

Running Calculations on Large Data Sets

With the data split in to manageable segments (the data blocks) the first part of the Hadoop process, and by extension Big Data Analysis, is complete. But the whole goal is to be able to derive insights and find conclusions from this data, so computations must be run. That is where MapReduce comes into play. MapReduce is comprised of two separate

functions, namely Map() and Reduce(). Imagine you have a large set of data, specifically an extremely large online book, which has been split into blocks by the HDFS. Now imagine that you want to run a calculation on this data, such as counting the number of times that each word in the online book occurs. The job of Map() function will be to take the input data from a each specific HDFS data node and to parse it into the format of <key, value> pairs that match the calculation that we are trying to run [4]. In this example, the <key,value> pairs that the Map() function would generate would be in the form of <word, count> [4]. In our example this means that for each unique word in the book you would see data in the form of <word, 1> where "word" is the specific word, and 1 refers to the fact that the specific word exists in the online book. For example you may see <school, 1> and <pencil, 1> referring to the fact that each word occurs in the data set. Note that <pencil, 1> will occur as many times as the word exists in the book, because no aggregation is done at this stage. The Map() function will run *in parallel* on the data nodes. What this means is that the same Map() function runs and processes on each of the data nodes (servers) at the same time [4]. Rather than counting all the words and their frequencies in data block one, and then data block two (and so forth), it will instead do the calculations across all data nodes (servers) at the same time [4]. Going back to the example, rather than counting all the word frequencies in your book page-by-page, imagine that you had ten thousand employees and they could each count the word frequencies on one page only, and they all did this at the same time. They would be counting *in parallel* and this task would be orders of magnitude faster than if you did it yourself, page by page.

Immediately following the Map() function is the Reduce() function, which performs the final calculation and aggregation by analyzing the results of the many Map() scripts. For

example, the Map() script may have found that in data block one (part 1 of the book in our example), there were 100 occurrences of <pencil, 1> and in data block two there were 50 occurrences of <pencil, 1>. Then Reduce() would combine these to say that the word "pencil" occurred 150 times total and it would then export this final output. The image below (Figure 2), courtesy of Stanford University, illustrates this process:

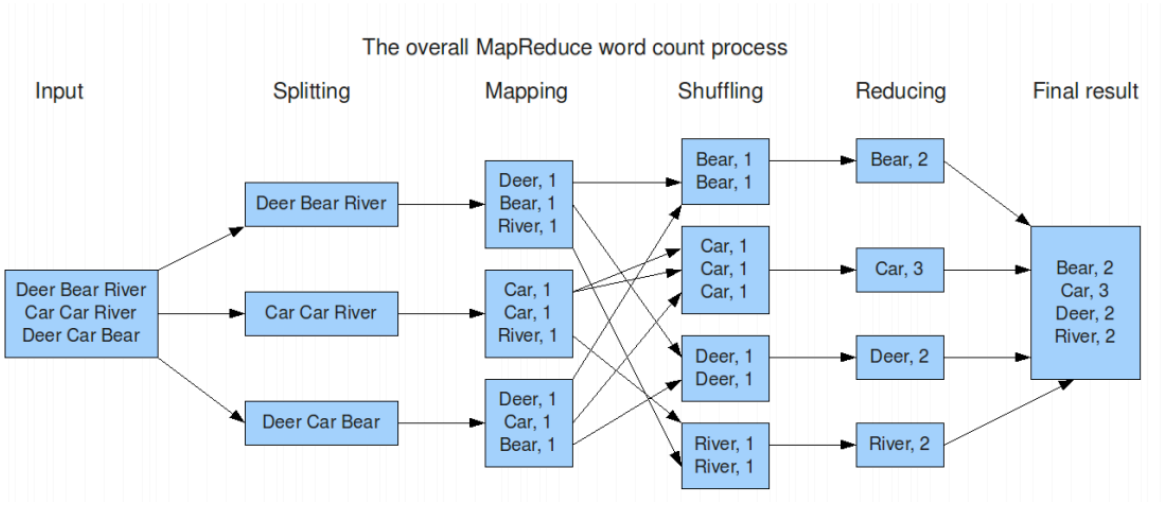

Figure 2: Shows how words can be counted using MapReduce [4]

While this example showed how MapReduce can be used to see how frequently words occur in a book, MapReduce can be used for many things and it is the individual programmers who write the Map() and Reduce() functions, most often in Java or a scripting language, based on what they are trying to accomplish. As I'll describe in more detail later, one common application of Hadoop and MapReduce is analyzing the actions of a user on a website and using that to create a better experience for the user in the future. As shown in the example, the benefit of MapReduce is that it is able to split a large task into much smaller tasks and perform these tasks simultaneously so that the computation is much quicker and more efficient. Overall, it is the HDFS that works to distribute the

massive amounts of data onto different data nodes so that it can be stored efficiently and then the MapReduce that takes advantage of the fact that the data is separated in the cluster to run the same functions on all the small sets of data at the same time and the combine to arrive at the final answer much, much faster.

Applications: Big Data Analytics in Industry

Big Data Analytics is still a very new field of computer science and engineering and many corporations are beginning to see the impact of hiring quantitative employees whose job it is to analyze the large swaths of data that the companies are inadvertently collecting. As a reference, in 2013 IBM predicted that by 2015 (just 2 years later) there would be 4.4 million new jobs in Big Data Analytics [5]. These newly hired computer scientists are tasked with using high-level programming languages and mathematics to write the MapReduce functions that take these large data sets and turn them into valuable information. Furthermore, companies are now hiring statisticians and data scientists to consider new trends that data can provide, which the computer scientists can then use Hadoop to look for. Nearly all companies collect data, whether they intend to or not. Internet companies may intentionally collect data such as which products people are buying or inadvertently and unknowingly collect data about how long visitors are staying on their site. However, now companies can make use of the data they didn't even know they were collecting, by using Big Data. In the past few years, some companies, the "early adopters", or those that adopt new practices and technologies before the rest of the industry does, began to use Big Data Analytics to find out important things about customers and use those to drive their success. For example a company can use Big Data Analytics to analyze the billions of pieces of data of user interaction on a website and then

figure out how long the average user is on their website. As more companies strive to keep up with their competitors, the skills of computer scientists versed in Big Data Analysis will be increasingly in demand and the fabric of the all industries around us will grow more sophisticated and responsive.

Consider this first real-life example: Target is one of America's largest chain retail stores, with yearly revenues of over $72 billion [6]. Target creates a unique guest ID for each Target customer and this "tracks purchase history, credit card use, survey responses, customer support incidents, email click-throughs, web site visits and more" [7]. Collecting this much data for each customer results in an amazingly large amount of stored data and information, and Target leaned on mathematician and programmer Andrew Pole to discover useful insights from this information [8]. So how did target use all of this data and what did they use Big Data Analysis for? Well it turns out, Target was able to analyze this massive set of data and predict when or if each customer was going to get pregnant [7]. Target was then able to email and send pregnant women coupons and advertisements for pregnancy-related clothing and items and drive up their sales [7]. This example shows how by using Big Data Analysis to analyze seemingly unrelated or unimportant raw data, Target was able to create a prediction model that helped drive up its revenues. Target's rival, Walmart, is also getting into Big Data Analysis. According to a 2013 article, Walmart now possess information and data on 60% of Americans – roughly 145 million people [9]. Even more, some people claim that Walmart "collects information on what shoppers buy, where they live and what they like via in-store Wi-Fi" [9]. All of this results in a lot of data. Using Big Data Analysis, Walmart is able to analyze over one hundred million web-search keywords daily and analyze terabytes of data on customer actions each day [10]. One way

that Walmart is able to use all of this data to help customers have a better experience when they are shopping online by predicting what a specific user wants to see and curating the shopping experience to that user [11]. In fact, in the first year after Walmart created WalmartLabs, their hub for Big Data, Hadoop and computer science, the number of customers completing a purchase at Walmart's website after making a search jumped by 20% [11]. As Gibu Thomas, the worldwide head of Walmart's mobile division stated, "[by] leveraging big data, we are also developing predictive capabilities to automatically generate a shopping list for our customers based on what they and others purchase each week" [12]. Now certain customers with frequent Walmart shopping histories are able to discover new products on Walmart.com via the predictive shopping list. These sentiments clearly show that Walmart not only finds investing in Big Data Analysis technologies to be a profitable endeavor, but there is also direct evidence that they benefitted from it.

Conclusion

The key theme in these examples (e.g. the Walmart and Target examples) is that corporations are able to benefit from data they were *already* inadvertently collecting by now using Big Data Analysis and Hadoop to analyze it. The fact that Walmart and Target both already used big data technologies and have used them for a couple years now most likely stems from the fact that these companies are large enough and wealthy enough to afford to invest in new technology as an early adopter. However, as the technology for Hadoop becomes more commonplace and more scientists and engineers learn how to use it and learn of its potential, smaller companies and startups too will begin readily using Big Data Analysis. If the results from Target and Walmart are a predictor for the future, then Big Data Analysis and Hadoop stand to revolutionize corporate America for years to come.

References

[1] http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

[2] http://radar.oreilly.com/2011/01/what-is-hadoop.html

[3] http://www.dummies.com/how-to/content/hadoop-distributed-file-system-hdfs-for-big-data-p.html

[4] http://hci.stanford.edu/courses/cs448g/a2/files/map_reduce_tutorial.pdf

[5] https://www-03.ibm.com/press/us/en/pressrelease/41733.wss

[6] http://www.forbes.com/sites/maggiemcgrath/2014/02/26/target-profit-falls-46-on-credit-card-breach-and-says-the-hits-could-keep-on-coming/

[7] http://www.crmsearch.com/retail-big-data.php

[8] http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

[9] http://www.huffingtonpost.com/2013/11/26/walmart-data_n_4344879.html

[10] https://datafloq.com/read/walmart-making-big-data-part-dna/509

[11] http://www.bloomberg.com/news/articles/2014-06-03/retailers-use-big-data-to-turn-you-into-a-big-spender

[12] http://www.cnbc.com/id/100759264#