IMPACT: International Journal of Research in Engineering & Technology (IMPACT: IJRET)
ISSN(E): 2321-8843; ISSN(P): 2347-4599
Vol. 3, Issue 1, Jan 2015, 11-16

HADOOP: THE NEXT BIG THING IN INDIA! THE BIG DATA REVOLUTION

NAKUL JADHAV & TANVI DESHPANDE

Department of Electronics and Telecommunication, D. Y. Patil College of Engineering,

Savitribai Phule University of Pune, Maharashtra, India

ABSTRACT

The world is full of data. It's everywhere and increasing rapidly even as we read this line. Search engines are trying to accumulate this influx of information commonly referred as Big Data. That's where Hadoop comes in. Hadoop provides OS level abstraction. India provides both Big Data and a possible pool of software experts who are trained in Hadoop. As the amount of data generated by businesses is increasing at a vast rate, the opportunities around it are exploding in terms of both scale and variety, so it is quite evident that Hadoop and other big data technologies have lot of scope in the coming times. But the idea is still in fledgling state in India and awareness about what we call the 'Big Data revolution' is necessary. This paper focuses on the scope and future of Hadoop with context to Big Data in India. Also the areas which can benefit from such revolution have been discussed.

KEYWORDS: Big Data, Hadoop, Revolution, India, Scope, Future

INTRODUCTION

We live in the data age. It's not easy to measure the total volume of data stored electronically, but an IDC estimate put the size of the "digital universe" at 0.18 zetta bytes in 2006 and is forecasting a tenfold growth by 2011 to 1.8 zettabytes.1 A zettabyte is1021 bytes, or equivalently one thousand exa bytes, one million peta bytes, or one billion terabytes. That's roughly the same order of magnitude as one disk drive for every person in the world. It's given a name-Big Data.

Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process them using traditional data processing applications.

So the big problem is to distribute and process this data. Also the data has to be stored and that requires a lot of space as mentioned above. Search engines and computer systems have and will face problems in computing such amount of data. To simplify this problem Hadoop came into existence.

Hadoop was created by Doug Cutting, the creator of Apache Lucene, the widely usedtext search library. Hadoop has its origins in Apache Nutch, an open source web search engine, itself a part of the Lucene project. Nutch was started in 2002, and a working crawler and search system quickly emerged. However, they realized that their architecture wouldn't scale to the billions of pages on the Web. Help was at hand with the publication of a paper in 2003 that described the architecture of Google's distributed file system, called GFS, which was being used in production at Google. In 2004, they set about writing an open source implementation, the Nutch Distributed Filesystem (NDFS). In 2004, Google published the paper that introduced Map Reduce to the world. At around the same time, Doug Cutting joined Yahoo! which provided a dedicated team and the resources to turn Hadoop into a system that ran at web scale. Since then, Hadoop

has seen rapid mainstream enterprise adoption. Hadoop's role as a general-purpose storage and analysis platform for big data has been recognized by the industry, and this fact is reflected in the number of products that use or incorporate Hadoop in some way. It is now sponsored by Apache Software Foundation, which promotes collaborative development of free and open source software such as OpenOffice.

So in short-Apache Hadoop is an open-source software framework for distributed storage and distributed processing of Big Data on clusters of commodity hardware. Its Hadoop Distributed File System (HDFS) splits files into large blocks (default 64MB or 128MB) and distributes the blocks amongst the nodes in the cluster.

HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

If you remember nothing else about Hadoop, keep this in mind: It has two main parts - a data processing framework and a distributed filesystem for data storage. There's more to it than that, of course, but those two components really make things go. The distributed filesystem is that far-flung array of storage clusters noted above - i.e., the Hadoop component that holds the actual data. By default, Hadoop uses the cleverly named Hadoop Distributed File System (HDFS), although it can use other file systems as well.

HDFS is like the bucket of the Hadoop system: You dump in your data and it sits there all nice and cozy until you want to do something with it, whether that's running an analysis on it within. Hadoop or capturing and exporting a set of data to another tool and performing the analysis there. But Hadoop is not really a database: It stores data and you can pull data out of it, but there are no queries involved - SQL or otherwise. Hadoop is more of a data warehousing system - so it needs a system like MapReduce to actually process the data.

Map Reduce runs as a series of jobs, with each job essentially a separate Java application that goes out into the data and starts pulling out information as needed. Using Map Reduce instead of a query gives data seekers a lot of power and flexibility, but also adds a lot of complexity.

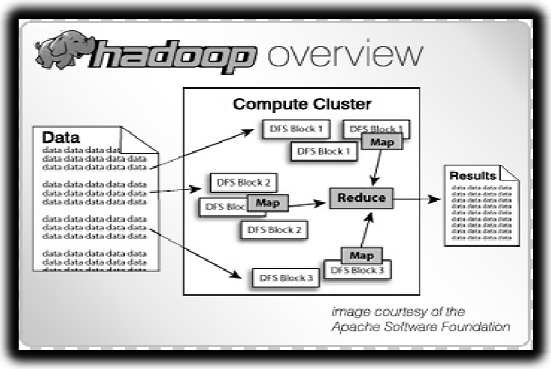

Figure 1: Basic Block Diagram of Hadoop

HADOOP FOR INDIA

Even though this framework and it's operation has been operational in the world for some time now, it is relatively new to Indian markets. Most of the Indian undergrads don't even know the term Big Data or what Hadoop exactly each. Like all western technology making its way into Indian mainstream markets, Hadoop to finds itself in the relative unknown. The potential is huge. A country of 2.252 billion people and all the data that comes with it. Plus as an

emerging global superpower, the need to create its own mark on the technological world is essential. As Facebook made its way into the hearts of Indian masses, the scope for new technology is emmense. And Hadoop is the next big thing! Following graph shows companies where Big Data is implemented.

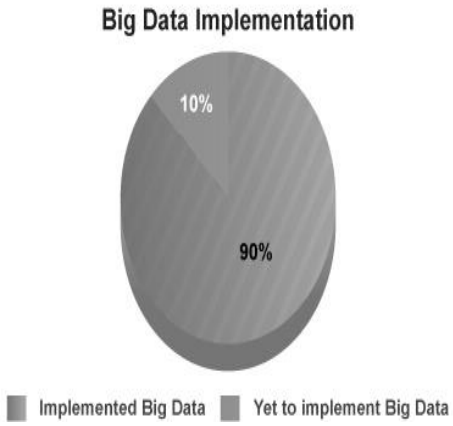

Figure 2: Implementation of Big Data in Markets

Figure 3: Overview of the Current Indian Market

Big Data and Hadoop skill could mean the difference between having your dream career and getting left behind. Dicehas quoted, "Technology professionals should be volunteering for Big Data projects, which makes them more valuable to their current employer and more marketable to other employers."

According to 90 executives who participated in the 'The Big Data Executive Survey 2013' conducted by NewVantage Partners LLC, supported by the Fortune 1000 senior Business & Technology executives, 90% of the organizations surveyed are already doing something with Big Data.

Hadoop skills are in demand – this is an undeniable fact! Hence, there is an urgent need for IT professional to keep themselves in trend with Hadoop and Big Data technologies. The above info graph shows how many organizations are influenced by Big Data and looking to implement them, if not already.

The Indian Big Data industry is predicted to grow five-fold from the current level of $200 mn to $1 bn by 2015 which is 4% of the expected global share. At the same time Gartner has predicted that there is going to be significant gap in job openings and candidates with Big Data skills**.** This is the right time to take advantage of this opportunity. This skill gap

in Big Data can be bridged through comprehensive learning of Apache Hadoop that enables professionals and freshers alike, to add the valuable Big Data skills to their profile. Technology professionals should be volunteering for Big Data projects, which make them more valuable to their current employer and more marketable to other employers**.**

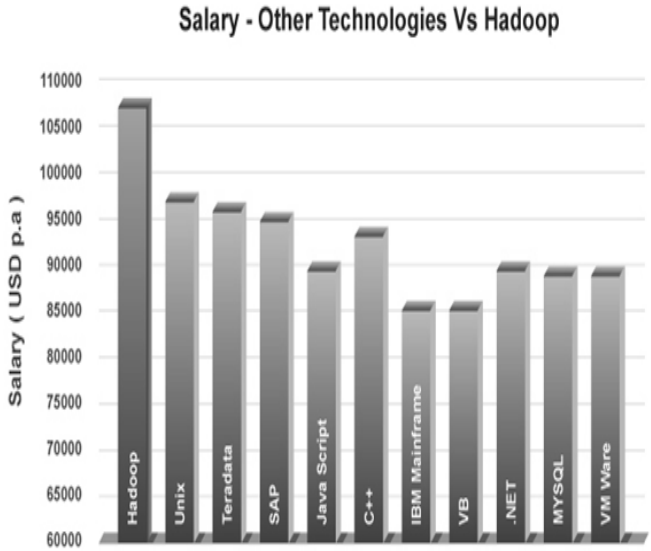

Figure 4: Salaries for Hadoop Trained Professionals

PROPOSED AREA OF GROWTH

Another government plan includes introducing soil-cards for farmers across the country. Imagine the data and the amount of information that will be needed to be processed. Indian India has gone through a big political upheaval in the recent general elections.

The country which had a lull in terms of economic strategy and development plans is now back on the road of targeted development of business. The Prime Minister of India has launched many big and influential projects. One of them being the ‘Make in India’ campaign. This is where Hadoop experts will reap rewards for their skills. Right now most of the companies are operating their Hadoop services outside India. Companies like Amazon, Alibaba, Google, Facebook, Twitter, etc. are now targeting the Indian markets under the ‘Make in India’ campaign. Soon they will be on the prowl for Big Data and Hadoop developers to cut their cost of employing developers outside India. As our country is a never ending market for e-commerce, the demand for Big Data developers is going to keep growing. The undergrads need to sharpen their skills in data management and upgrade to the alien-but-awesome companies like TCS, Infosys which are already collaborating with the govt. for various projects will be on lookout for Hadoop developers for such projects.

Also, ”The Digital India” initiative launched by Indian PM Narendra Modi aims at ensuring that government services be made available to citizens electronically by reducing paperwork. When paperwork has to be eliminated and a lot of digitalized data has to be handled, Hadoop is just the right tool you need to do it efficiently.

CONCLUSIONS

The Hadoop Big Data market is growing everyday. Technology giants are investing on Hadoop. The lack of well

trained human force is the problem everyone is facing right now. Hadoop have a well supported community and they are introducing new technologies on top of Hadoop stack. Certainly Big Data and Hadoop stack will have a huge scope in technology world and obviously in Indian IT industry. The Indian IT hub can provide manpower and skill plus an already available market. So developers need to strike the iron when it's still hot. A head start in the world of Hadoop development will reap huge dividends once Big Data becomes a common household name amongst the Indian IT community. As the Times of India put it aptly in their article on Hadoop-' As the world moves to a digital age, there is literally an explosion of data and Hadoop makes it possible to stay on top of it. If a few years ago, megabytes and gigabytes used to be the extent of data, experts are now talking of several petabytes.

Say, if 1MB represents a spoonful of sand the napetabyte is equivalent to about as much and on a mile long beach. Every major internet company -- be it Google, Twitter, LinkedIn or Facebook -- uses some form of Hadoop to sort and search the vast amounts of data that their hundreds of millions of users are creating every second.' So mastering Hadoop can be said to be the next frontier for the Indian developer.

REFERENCES

1. Tom White. Hadoop:The Definitive Guide. May 2012, (3rd Edition), OReilly pp 1-10.
2. Edureka blog (2012,April 12).Hadoop statistics[Online].Available:http://www.edureka.co/blog/5-reasons-to-learn-hadoop/
3. Brian Proffit. (2013, May 23)Hadoop:What it is and how does it works, http://www.readwrite.comHadoop-what-it-is-and-how-it-works
4. The Times of India. (April 04, 2013) 5 things about hottest IT skill Hadoop.
5. Definitions-Big Data and Hadoop, Wikipedia.
6. Opinions- http://www.quora.com/What-is-the-scope-to-begin-a-Career-in-Big-Data-using-Hadoop-in-India-as-a-fresher