Volume 5, Issue 3, March 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Study on the Role of Big Data in Social Media

Tadepalli Sarada Kiranmayee

Abstract— In recent years there has been a rapid growth in big data and social networking sites. There are millions of active users in the social networking sites who are connected by the means of laptops, smartphones and tablets. Social Media has become the main communication network for the people around the world. This is generating huge amount of data is being generated because of the user activities like posts, likes and comments etc. This paper is the study about the importance of Big Data and its different technologies used in social media..

Keywords - Social Media, Big Data

I. INTRODUCTION

Gartner had defined the Big Data as follows: Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making [1]. Big Data helps in explaining the exponential growth and also the availability of structured and unstructured data.

Unstructured data refers to information that either does not have a pre-defined data model and/or is not organized in a predefined manner. The forms unstructured data of social media are Word Doc's, PDF's and Other Text Files - Books, letters, other written documents, audio and video transcripts; Audio Files - Customer service recordings, voicemails; Presentations - PowerPoints, SlideShares; Videos- personal video, YouTube uploads; Images - Pictures, illustrations, memes; Messaging - Instant messages, text messages.

In contrast to, the unstructured data, structured data is data that can be organized easily. It is clean, analytical and usually stored in databases. 1) Machine Generated structured data: Sensory Data - GPS data, manufacturing sensors, medical devices; Point-of-Sale Data - Credit card information, location of sale, product information; Call Detail Records - Time of call, caller and recipient information; Web Server Logs - Page requests, other server activity. 2) Human Generated: Input Data - Any data inputted into a computer: age, zip code, gender, etc.

This paper will describe how the social media takes the help of big data to deal with this structured and unstructured data.

II. CHARACTERISTICS OF BIG DATA

The characteristics of Big Data can be explained with Doug Laney 3V's.[2]

A. *Volume*: Scale is certainly a part of what makes Big Data big. The internet-mobile revolution, bringing with it a torrent of social media updates, sensor data from devices and an explosion of e-commerce, means that every industry is swamped with data- which can be incredibly valuable, if you know how to use it. 100 terabytes of data are uploaded daily to Facebook; Akamai analyses 75 million events a day to target online ads; Walmart handles 1 million customer transactions every single hour. 90% of all data ever created was generated in the past 2 years.

B. *Velocity* : In 1999, Wal-Mart's data warehouse stored 1,000 terabytes (1,000,000 gigabytes) of data. In 2012, it had access to over 2.5 petabytes (2,500,000 gigabytes) of data.Every minute of every day, we upload 100 hours of video on Youtube, send over 200 million emails and send 300,000 tweets. 'Velocity' refers to the increasing speed at which this data is created, and the increasing speed at which the data can be processed, stored and analyzed by relational databases. The possibilities of processing data in real-time is an area of particular interest, which allows companies to do things like display personalized ads on the web pages you visit, based on your recent search, viewing and purchase history.

C. *Variety* : Gone are the days when a company's data could be neatly slotted into a table and analyzed. 90% of data generated is 'unstructured', coming in all shapes and forms- from geo-spatial data, to tweets which can be analyzed for content and sentiment, to visual data such as photos and videos.

The '3 V's' certainly give us an insight into the almost unenvisionable scale of data, and the break-neck speeds at which these vast datasets grow and multiply. But only 'Variety' really begins to scratch the surface of the depth- and crucially, the challenges- of Big Data. An article from 2013 by Mark van Rijmenam proposes four more V's, to further understand the incredibly complex nature of Big Data.

D. Variability : variability refers to data whose meaning is constantly changing. This is particularly the case when gathering data relies on language processing. Brian Hopkins (a principal analyst at Forrester) cited the supercomputer Watson as a prime example of this. To participate in the gameshow Jeopardy, Watson had to "dissect an answer into its meaning and […] to figure out what the right question was". Words don't have static definitions, and their meaning can vary wildly in context.

E. *Veracity:* Although there's widespread agreement about the potential value of Big Data, the data is virtually worthless if it's not accurate. This is particularly true in programs that involve automated decision-making, or feeding the data into an unsupervised machine learning algorithm.

What's crucial to understanding Big Data is the messy, noisy nature of it, and the amount of work that goes in to producing an accurate dataset before analysis can even begin.

F. *Visualization*: Once it's been processed, you need a way of presenting the data in a manner that's readable and accessible- this is where visualization comes in. Visualizations can contain dozens of variables and parameters- a far cry from the x and y variables of your standard bar chart- and finding a way to present this information that makes the findings clear is one of the challenges of Big Data. It's a problem that's spurned a burgeoning market- new visualization packages are appearing all of the time, with AT&T announcing their offering, Nanocubes, just this week.

G. *Value*: The potential value of Big Data is huge. Speaking about new Big Data initiatives in the US healthcare system last year, McKinsey estimated if these initiatives were rolled out system-wide, they "could account for 300 billion to 450 billion in reduced health-care spending, or 12 to 17 percent of the $2.6 trillion baseline in US health-care costs". However, the cost of poor data is also huge- it's estimated to cost US businesses $3.1 trillion a year. In essence, data on its own is virtually worthless. The value lies in rigorous analysis of accurate data, and the information and insights this provides.

In essence, when the media talk about Big Data, they're not just talking about vast amounts of data that are potential treasure troves of information. They're also talking about the business of analyzing this data- the way we pick the lock to the treasure trove. In the world of Big Data, data and analysis are totally interdependent- one without the other is virtually useless, but the power of them combined is virtually limitless.

IV. DIFFERENT ARCHITECTURES IN BIG DATA

These are the architectures in Big Data which are described below:

A. Map Reduce [3] : MapReduce is the heart of Hadoop. It is this programming paradigm that allows for massive scalability across hundreds or thousands of servers in a Hadoop cluster. The Hadoop concept is fairly simple to understand for those who are familiar with clustered scale-out data processing solutions.

The term MapReduce actually refers to two separate and distinct tasks that Hadoop programs perform. The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce job is always performed after the map job.

B. *Apache Hadoop*: The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

The project includes these modules: Hadoop Common: The common utilities that support the other Hadoop modules, Hadoop Distributed File System (HDFS™): A distributed file system that provides high-throughput access to application data, Hadoop YARN: A framework for job scheduling and cluster resource management, Hadoop MapReduce: A YARN-based system for parallel processing of large data sets. [4]

C. *HBase* [5] : is an open source, non-relational, distributed database modeled after Google's BigTable and written in Java. It is developed as part of Apache Software Foundation's Apache Hadoop project and runs on top of HDFS (Hadoop Distributed Filesystem), providing BigTable-like capabilities for Hadoop. That is, it provides a fault-tolerant way of storing large quantities of sparse data (small amounts of information caught within a large collection of empty or unimportant data, such as finding the 50 largest items in a group of 2 billion records, or finding the non-zero items representing less than 0.1% of a huge collection).

HBase features compression, in-memory operation, and Bloom filters on a per-column basis as outlined in the original BigTable paper. Tables in HBase can serve as the input and output for MapReduce jobs run in Hadoop, and may be accessed through the Java API but also through REST, Avro or Thrift gateway APIs.

HBase is not a direct replacement for a classic SQL database, although recently its performance has improved, and it is now serving several data-driven websites,including Facebook's Messaging Platform.

D. *Apache Hive*: is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis.[2]While initially developed by Facebook, Apache Hive is now used and developed by other companies such as Netflix. Amazon maintains a software fork of Apache Hive that is included in Amazon Elastic MapReduce on Amazon Web Services. Apache Hive supports analysis of large datasets stored in Hadoop's HDFS and compatible file systems such as Amazon S3 filesystem. It provides an SQL-like language called HiveQL with schema on read and transparently converts queries to map/reduce, Apache Tez and in the future Spark jobs. All three execution engines can run in Hadoop YARN. To accelerate queries, it provides indexes, including bitmap indexes.

By default, Hive stores metadata in an embedded Apache Derby database, and other client/server databases like MySQL can optionally be used. [5]

D. *Apache Pig :* [6] is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

At the present time, Pig's infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs, for which large-scale parallel implementations already exist (e.g., the Hadoop subproject). Pig's language layer currently consists of a textual language called Pig Latin, which has the following key properties:

1. Ease of programming. It is trivial to achieve parallel execution of simple, "embarrassingly parallel" data analysis tasks. Complex tasks comprised of multiple interrelated data transformations are explicitly encoded as data flow sequences, making them easy to write, understand, and maintain.
2. Optimization opportunities. The way in which tasks are encoded permits the system to optimize their execution automatically, allowing the user to focus on semantics rather than efficiency.
3. Extensibility. Users can create their own functions to do special-purpose processing.

E.*HDFS* : [7] The Hadoop Distributed File System (HDFS) is a sub-project of the Apache Hadoop project. This Apache Software Foundation project is designed to provide a fault-tolerant file system designed to run on commodity hardware.

According to The Apache Software Foundation, the primary objective of HDFS is to store data reliably even in the presence of failures including NameNode failures, DataNode failures and network partitions. The NameNode is a single point of failure for the HDFS cluster and a DataNode stores data in the Hadoop file management system.

HDFS uses a master/slave architecture in which one device (the master) controls one or more other devices (the slaves). The HDFS cluster consists of a single NameNode and a master server manages the file system namespace and regulates access to files.

V. BIG DATA IN SOCIAL MEDIA

A. *AMAZON***:** Amazon uses big data also to offer a superb service to its customers. This could be the effect of the purchase of Zapos in 2009, but it clearly helps that it ensures that customer representatives have all the information they need the moment a customer needs support. They can do this because they use all the data they have collected from their customers to build and constantly improve the relationship with its customers. But Amazon is expanding its usage of Big Data since it notices that the competition is nearing closer. As such, Amazon added a remote computing services, via Amazon Web Services (AWS), to their already massive product and service offering. AWS was launched in 2002, but only recently they added Big Data services and they now offer tools to support data collection, data storage, data computation along with data collaboration and data sharing. All are available in the cloud. The Amazon Elastic MapReduce provides a managed, easy to use analytics platform built around the powerful Hadoop framework that is used by large companies, including Dropbox, Netflix and Yelp.

However, there is more. Amazon also uses Big Data to monitor, track and secure its 1.5 billion items in its retail store that are laying around it 200 fulfilment centres around the world. Amazon stores the product catalogue data in S3. This is a simple web service interface that can be used to store any amount of data, at any time, from anywhere on the web. It can write, read and delete objects up to 5 TB of data each. The catalogue stored in S3 receives more than 50 million updates a week and every 30 minutes all data received is crunched and reported back to the different warehouses and the website.

At AWS, Amazon also hosts public big data sets at no cost. All available big data sets can be used and seamlessly integrated in AWS cloud-based solutions. Everyone can now use this public data, such as the data from mapping the Human Genome Project.[8]

B. *FACEBOOK* **:** Hive, the data warehousing infrastructure Facebook helped develop to run on top of Hadoop, is central to meeting the company's reporting needs. Facebook must balance the need for rapid results in features such as its graph tools with simplicity and ease of reporting, so it is working on another contribution to Hive that will improve the speed of queries. Improving Hive's speed is important, as the scalability that makes the tool central to the social network's needs can come at the expense of low latency.

"Hive is still a workhorse and it will remain the workhorse for a long time because it's easy and it scales," Parikh said. "Easy is the key thing when you want lots of people to be able to engage with a tool. Hive is very simple to use, so we've been focused on performance to make it even more effective."

For companies just embarking on a big data initiative, striking a balance between handling technology challenges with Hadoop and deriving insight from data will be difficult but important, Parikh said. Businesses will need to experiment and maintain a constant focus on long-term goals to ensure they build out technology in the right way. However, with constant innovations in the open source Apache Hadoop community, businesses have more resources than ever to make data central to their operations in the same way as social media giants. [9]

VI. CONCLUSIONS

Big Data is the key to social media interactions between people is that it leaves knowledge behind for others to find and reuse. This can be the original content that started the conversation or the subsequent comments, discussion, ratings, ranking, tweets, etc. These conversations will remain on the network afterwards, usually for a digital eternity. Also, finding what one is looking for in the vast sea of a million or billion human conversations is a difficult task. Thus, separating Big data and the social media is not possible but they go hand=in-hand.

REFERENCES

[1] http://www.gartner.com/it-glossary/big-data
[2] http://dataconomy.com/seven-vs-big-daapreduce/
[3] http://www01.ibm.com/software/data/infosphere/hadoop/mapreduce/
[4] https://hadoop.apache.org/
[5] http://en.wikipedia.org/wiki/Apache_HBase
[6] https://pig.apache.org/
[7] http://www.webopedia.com/TERM/H/hadoop_distributed_file_system_hdfs.html
[8] https://datafloq.com/read/amazon-leveraging-big-data/517
[9] https://datafloq.com/read/amazon-leveraging-big-data/517

