ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Managing Big-Data Over Social Media: A Survey

[I]Priya Sharma, [II]Keshav Gupta
[I,II]School of CSE, Galgotias University, Greater Noida, U.P, India

Abstract
The increase in data storage over the server side has led to the introduction of the big data. This era marks the vital and plenty use of the social networking which has made the global interaction just on click of the applications. The increase in data collection over the socialnetwork has increase from a tiny KB to PB. This datacollection has no definite mass for memory demand forstorage .The current graphs from various sites shows agreat variations for data collection. So we can't stick to one particular technique to resolve the data storage issue. We need to compress on various level. In this paper, I am explaining various ongoing trend for Big-data handlingover the Social networks.

Keywords
Bigdata, volume, velocity, variety, veracity, social media, Hadoop and Map Reduce.

I. Introduction

The big data is collection of huge amount of data which usually exceeds from Terabytes to Petabytes .The Big data is useful for storage of large amount of data like RBI and US Investment banks has huge amount of data for judiciary and business purposes. Also this century has marked the need of big data for Social networking . Facebook ,twitter, etc are the companies which needs to handle the Big-Data and are using various techniques to handle this issue. There is a lot of data collect by Client-Side every day.

1. Google processes information from many sources in Petabyte (PB), Facebook produces log information of around 10 PB per month, many companies processes information of tens of PB or TB for on-line commercialism per day.[1,2]
2. Newly advancement in technology make it very simple in bringing up the data or information. For example nearly on an average , seventy two hours of videos are uploaded to You Tube in each and every minute. Hence, the main challenge is of collecting and getting right information from widely distributed data.[1,2]
3. Twitter-a large social networking website tuned towards quick communication. over one hundred forty million active users publish over four hundred million 140- character "Tweets" daily [3,4].

Big-data is characterized by following four factors:-

- Volume:- Big-data implies to the abundance amount of data. Every day on Social Media, massive data is created which can be seen as one of the example of the volume of Big-data [5].
- Variety:- It refers to the Structured, Unstructured and Semi-Structured data. While interacting online data comes in various forms like e-mail, photos, videos, PDFs, audio, etc. The storage of this type of data is of great concern [5].
- Velocity:- It ensures the flow of data from various sources like business processes, networks, human interaction in social media etc. The data flow is the continuous process [5].
- Veracity:- Veracity of Big-data is the noise present in the data. The data which is chosen for mining, storage and analysis should be free from noise or we can say that it should be meaningful in nature [5].

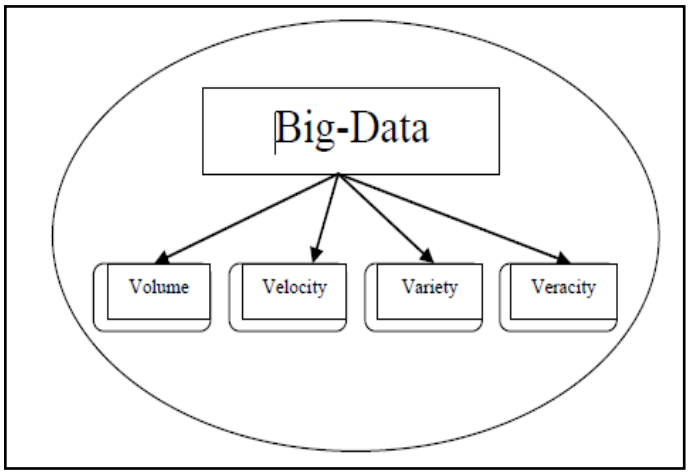

Fig. 1 : Big Data with its Vs

II. Big Data History

Considering the evolution and complexity of big data systems, previous descriptions are based on a one-sided viewpoint, such as chronology [8] or milepost technologies [9]. In this survey, the history of big data is presented in terms of the data size of interest. Under this framework, the history of big data is tied tightly to the capability of efficiently storing and managing larger and larger datasets, with size limitations expanding by orders of magnitude. Specifically, for each developed . Thus, history of big data (huge information) will be roughly split into the subsequent stages:

Megabyte to Gigabyte

In the 1970s and 1980s, historical business data introduced the earliest "big data" challenge in moving from megabyte to gigabyte sizes. The urgent need at that time was to house that data and run relational queries for business analyses and reporting. Research efforts were made to give birth to the "database machine" that featured integrated hardware and software to solve problems. The underlying philosophy was that such integration would provide better performance at lower cost. After a period of time, it became clear that hardware-specialized database machines could not keep pace with the progress of general purpose computers. Thus, the descendant database systems are soft- ware systems that impose few constraints on hardware and can run on general-purpose computers.

Gigabyte to Terabyte

In the late 1980s, the popularization of digital technology caused data volumes to expand to several gigabytes or even a terabyte, which is beyond the storage and processing capacity of an individual large computer system. Data parallelization was proposed to extend storage capabilities and to improve performance by distributing data and related tasks, such as building indexes and evaluating queries, into disparate hardware. Based on this idea, several types of parallel databases were built, including shared memory databases, shared-disk databases, and shared- nothing databases, all as induced by the underlying hardware architecture. Of the three types of databases, the sharednothing architecture, built on a complex cluster of single machines - each with its own processor, memory and disk [10] has witnessed great success. Even in the past few years, we have witnessed the blooming of commercialized products of this type, such as Teradata [11], Netezza [12], Aster Data [13], Greenplum [14], and Vertica [15]. These systems exploit a relational data model and declarative relational query languages, and they explored the use of divide-and- conquer parallelism to partition data for storage.

Terabyte to Petabyte

During the late 1990s, when the database community was cherishing its finished task on the parallel database, the rapid development of Web 1.0 led the whole world into the Internet era, along with massive semistructuredor unstructured web pages holding terabytes or petabytes (PBs) of data. The resulting need for search companies was to index and query the mushrooming content of the web. Unfortunately, although parallel databases handle structured data well, they provide little support for unstructured data. Additionally, systems capabilities were limited to less than several terabytes. To address the challenge of web-scale data management and analysis, Google created Google File System (GFS) [16] and Map Reduce [17] programming model. GFS and Map Reduce enable automatic data parallelization and the distribution of large-scale computation applications to large clusters of commodity servers. A system running GFS and Map Reduce can scale up and out and is there- fore able to process unlimited data. In the mid-2000s, user-generated content, various sensors, and other ubiquitous data sources produced an inspiring flow of mixedstructuredata, which called for a model shift in computing architecture and large-scale data processing mechanisms. NoSQL databases, which are scheme-free, fast, highly scalable, and reliable, began to emerge to handle these data. In Jan. 2007, Jim Gray, a database software pioneer, referred the shift as the fourth prototype [18]. He conjointly argued that only development of new computing tools is able to carry out the various operations such as manage, visualize and analyze the immensely created data.

Petabyte to Exabyte

Under current development trends, data stored and analyzed by big companies will undoubtedly reach the PB to Exabyte magnitude soon. However, current technology still handles terabyte to PB data; there has been no revolutionary technology developed to cope with larger datasets. In Jun. 2011, EMC published a report entitled Extracting Value from Chaos [19]. The concept of big data and its potential were discussed throughout the report. This report ignited the enthusiasm for big data in industry and academia. In the years that chased, most the dominating trade firms, as well as, EMC, Oracle, Microsoft, Google, Facebook, started to evolve big data projects. In Mar. 2012, the Obama executives declare that the US would invest 200 million dollars to initiate a big data study plan. The effort can involve a variety of federal agencies, together with federal agency, the National Institutes of Health, and therefore the National Science Foundation [8]. This endeavor aims to foster the development of advanced data management and analysis methods.

III. Literature Review

The data collected over social network shows no particular graph for everyday data collection. So there is a need to classify the data collected over the social network on the basis of size.

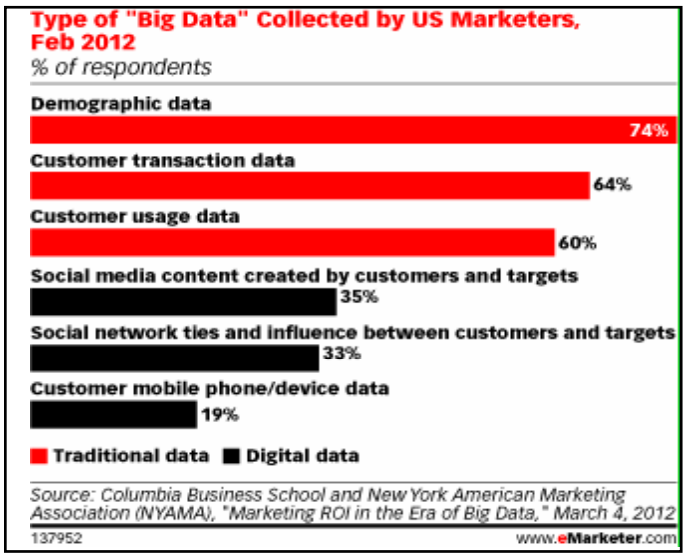

Fig 2: Big- data collected by US marketers, Feb 2012.

In 2010,Google's Eric Schmidt famously stated that: "every two days we now create as much information as we did from the dawn of civilization up till 2003"[20].Big-data is the massive amount of data which is created every minute by a single click of the user. This data is used to extract valuable information for the better insights of public sectors, private sectors and customers. As stated by IBM, 2.5 quintillion bytes of data is created by the user every single day , which refers to that most of data which is readily available is created in the last 2 years. This amount of data is very big in nature. All the pictures, smileys, posts, videos, likes, comments etc comes under 2.5 quintillion bytes of data each and every day [6].

Table : Data collected over social media in every minute, by Intel,2012.

Sources	Data Visuazation (approx.)
Facebook	6 million views
Google	2+ million queries searched
YouTube	1.3 million video seen
Twitter	320+ new account created

On the importance of Compressing data:

1. File compression technique

File system compression takes a reasonably simple approach to reduce the storage space by transparently applying the compression to each and every file as it is written to the disk. that's why in

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

this concept we use small and simple compression tools like DiskDoubler and SuperStor Pro [7].

2. Storage array compression

When the storage comes beyond the simple file system and there is a need felt to store data in blocks, there comes idea of array. this array storage system is implemented using the concepts that came from Sun Microsystem called sun fire *4500 [7].

3. NAS compression

If the client side in a particular client/server architecture do not support the processing of DiskDoubler and SuperStor Pro then there comes a alternate for doing this termed as Storwize STN-6000p [7].

4. Backup Storage Compression

There is a concept of tape hardware compression. It offers these advantages: It does not slow things down. It is available on most of the modern hardware system and it provides compression ratio more than 2:1. But ,it is not mainly used as hardware implementation rather than it is used as a virtual software concept [7].

5. Data -deduplication

The searching of various data on any search engine like Google comes with a problem of data replication. Once a data is replicated ,the storage database is uploaded with higher number of repeatable data and hence results in redundancy problem. there are two approaches for resolving these issues, the one is on the client side and other on server side. The client side deals this using Source-Based Data- deduplicatin (Puredisk and Avamar).The server side handles it using Target-Based Data-deduplication (Sepaton S2100 and Quantum DXi) [7].

IV. Technology Used

In the present scenario, Big-data plays a very important role in digitized world. Big-data is nothing but extensive amount of data which is growing each and every second rapidly. This enormous data cannot be stored using traditional approaches, therefore it demands another way for its storage. Hadoop is a tool through which we store Big-data. It is an open source framework(freely available) with additional feature of scalability and fault tolerance for data storage and processing. It uses HDFS(Hadoop Distributed File System) for storage of data from distributed sources across cluster of computers. These cluster of computers are commonly referred to as Data Nodes. MapReduce is the programming interface model for Hadoop. It is responsible for scheduling, monitoring and allocating jobs to data nodes. The workflow management of Hadoop is carried out by Apache Oozie.

V. Conclusion

The discussed techniques are par excellence for the current scenario. The ideas implemented here are being practically used in the companies in direct or indirect way. But in the era of online lifestyle, there is always a possibility of increase data storage requirement . So, there will be always a new idea for new necessities for Big-Data storage.

References

[1] *Min Chen·Shiwen Mao·Yunhao Liu, Big Data: A Survey© SpringerScience+Business Media New York. online: 22 January 2014*
[2] *Mayer-Sch¨onberger V, Cukier K, Big data: a revolution that will transform how we live, work, and think. Eamon Dolan/ Houghton Mifflin Harcourt,2013*
[3] *Shamanth Kumar Fred Morstatter Huan Liu,Twitter Data Analytics, August 19,2013 Springer.*
[4] *https://blog.twitter.com/2012/twitter-turns-six.*
[5] *Kevin Normandeau, Beyond volume variety and velocity is the issue of Big-data veracity, September 12, 2013.*
[6] *Bikram K.Singh,Big-data annd its use in social marketing, Nov 27, 2014.*
[7] *Brian Peterson, Top five data storage compression methods, July 2008.*
[8] *V. R. Borkar, M. J. Carey, and C. Li, Big data platforms: What"s next?" XRDS, Crossroads, ACM Mag. Students, vol. 19, no. 1, pp. 44–49, 2012.*
[9] *V. Borkar, M. J. Carey, and C. Li, Inside big data management: Ogres, onions, or parfaits?" in Proc. 15th Int. Conf. Extending Database Technol., 2012, pp. 3–14.*
[10] *D. Dewitt and J. Gray, ,,,,Parallel database systems: The future of high performance database systems," Commun. ACM, vol. 35, no. 6, pp. 85–98, 1992.*
[11] *Teradata. Teradata, Dayton, OH, USA [Online]. Available: http://www.teradata.com/,2014*
[12] *Netezza. Netezza, Marlborough, MA, USA [Online]. Available: http://www-01.ibm.com/software/data/netezza,2013*
[13] *Aster Data. ADATA, Beijing, China [Online]. Available: http://www.asterdata.com/,2013*
[14] *Greenplum. Greenplum, San Mateo, CA, USA [Online]. Available: http://www.greenplum.com/,2013*
[15] *Vertica [Online]. Available: http://www.vertica.com/,2013*
[16] *S. Ghemawat, H. Gobioff, and S.-T. Leung, The Google file system, in Proc. 19th ACM Symp. Operating Syst. Principles, 2003, pp. 29–43.*
[17] *T. Hey, S. Tansley, and K. Tolle, The Fourth Paradigm: Data-Intensive Scientific Discovery. Cambridge, MA, USA: Microsoft Res., 2009.*
[18] *J. Dean and S. Ghemawat, Mapreduce: Simplified data processing on large clusters, Commun. ACM, vol. 51, no. 1, pp. 107–113, 2008*
[19] *J. Gantz and D. Reinsel, Extracting value from chaos, in Proc. IDC iView, 2011, pp. 1–12*
[20] *Tamsin Oxford, social mining part 1:how Big-data is transforming customer insights, Aug 13,2014.*
[21] *Shankar Ganesh Manikandan, Siddarth Ravi, Big Data Analysis using Apache Hadoop, ©2014 IEEE*