Big Data Storage Technologies and Challenges of Big Data Storage and Management

Ms. S. K. Totade, Mr. Sanghesh B. Bele

Deptt. Of MCA, Vidya Bharati Mahavidyalaya, Amravati.

Ms. Apurva R. Raut

Deptt. Of Computer Application, VidyaBharati Mahavidyalaya, Amravati.

Abstract - Big data is a term to refer to huge data sets, have high Velocity, high Volume and high Variety and complex structure with difficulties of management, analyzing storing and processing. Due to characteristics of big data it becomes very difficult to management, analysis, storage, Transport and processing the data using the existing traditional techniques. This paper introduces Big Data Analysis and Storage Technologies, Challenges of Big Data Storage and Management and Suggestions for Big Data Storage and Management. Storage and Management are major concern in this era of big data. The ability for storage devices to scale to meet the rate of data growth, enhance access time and data transfer rate is equally challenging. These factors, to a considerable extent, determine the overall performance of data storage and management.

Keywords: *Big Data Definition, Characteristics, Data Storage Technologies, Challenges Of Big Data Storage And Management, Suggestion For A Big Data Storage And Management.*

Big Data Definition

"Big Data is at the foundation of all the megatrends that are happing."

Chris Lynch, Vertica Systems

Data is everywhere, in fact the amount of digital data that is growing at a rapid rate, and changing the way. This is the field that didn't even exist 20years back. Now data is growing faster than ever faster before and by the year 2020; about 1.7megabytes of new information will be created every second for every human being on the planet.

Data is essentially just raw bits of information science.

Big data is basically a term that describes large amount of data. Big data sets those are so big and complex that traditional data-processing application software is in adequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy and data source. There are number of concepts associated with big data: originally there were 3 concepts volume, variety, and velocity. Other concepts later attributed with big data are veracity (i.e., how much noise is in the data) and value.

The position of big data storage within the overall big data value chain can be seen in **figure 1.** Big data storage systems typically address the volume challenges by making use of distributed shared nothing architectures. This allows addressing increased storage requirements by scaling out new nodes providing computational power and storage. New machines can seamlessly be added to a storage cluster and storage system takes care of distributing the data between individual nodes transparently. Storage solutions also need to cope with the velocity and variety of data.

Velocity is important in the sense of query latencies, i.e. how long does it take to get a reply for a query? This is particularly important in the face of high rate of incoming data. For instance, random write access to a database can slow down query performance considerably if it needs to provide transactional guarantees. In contrast, variety relates to the level of effort that is required to integrate the work with the data that originates from a large number of different sources. For instance graph databases are suitable storage systems to address these challenges.

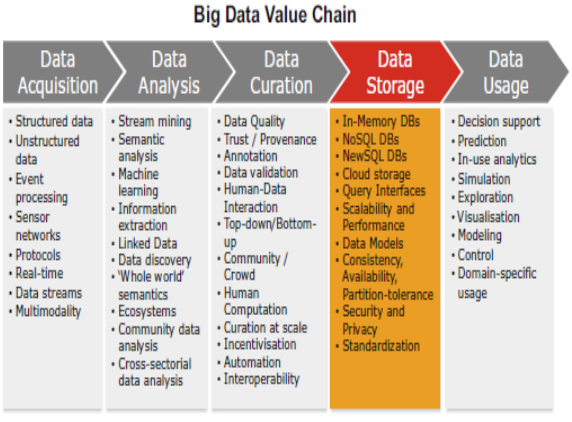

Figure 1: Data Storage in the big data value chain

Big Data Characteristics

The three Vs (volume, velocity and variety) are known as the main characteristics of big data. The characteristics are described in **figure 2** below.

Volume: refers to amount of data and there are many factors that can contribute to the volume increase in data it could amount to hundreds of terabytes or even pet bytes of information generated for everywhere. The number of sources of data for an organization is growing.

Figure 2: 3 Vs of big data

Figure 2 shows that the data volume is growing from megabytes (106) to petabytes(1015) and beyond.

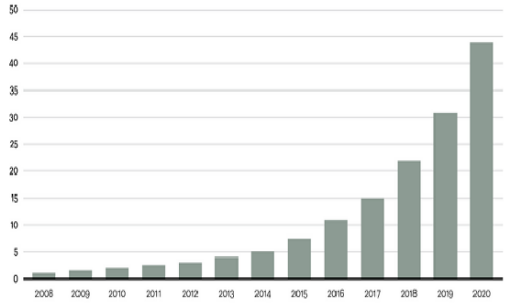

Figure 3: Data volume growth by year in zettabytes

Figure 3 indicates that the volume of data stored in the world would be more than 40 zettabytes(1021) by 2020 .

Velocity: refers to data speed measures the velocity of information creation, gushing and collection, velocity is the most misunderstood big data characteristic. The data velocity is also about the rate changes, and about combining datasets that are coming with different speeds. The velocity of data also describes bursts of activities, rather than the usual steady tempo where velocity frequency equated to only real-time analytics.

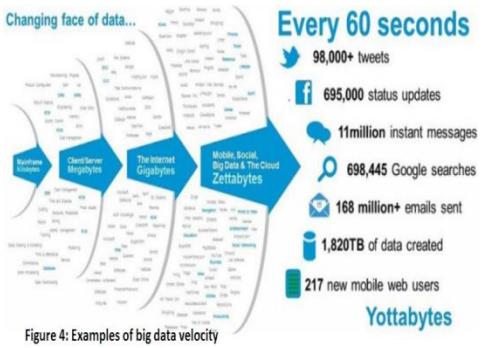

Figure 4: Examples of big data velocity

Figure 4 shows few examples of the pace of data. Data speed administration is significantly more than a bandwidth issue; Figure 2 also reflect velocity as characteristics of big data, showing how it requires near real-time and/or real-time analytics.

Variety: Other than typical structured data, big data contains text, audio, images, videos, and many more unstructured and semi-structured data, which are available in many analog and digital formats. From an analytic perspective, variety of data is the biggest challenge to effectively use it. Some researchers believe that, taming the data variety and volatility is the key of big data analytics.

Figure 5 shows the comparison between increment of unstructured, semi-structured data and structured data by years.

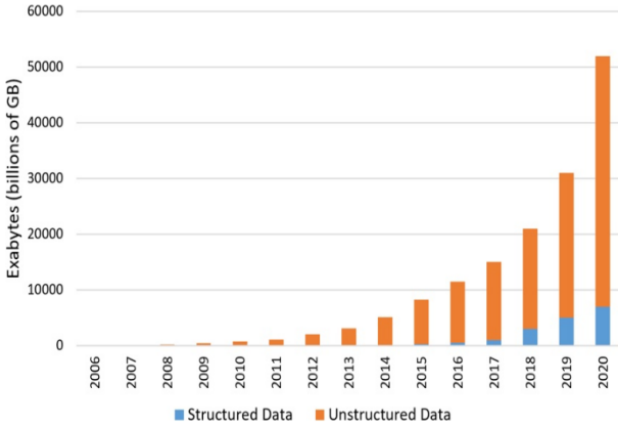

Figure 5: Growth of data variety by years

This paper investigates and analyzes big data storage technology from the following four aspects: distributed file system, NoSQL database, new-type data storage technology of MPP architecture and database all-in-one machine.

- **Distributed File Systems:** File system such as Hadoop file system(HDFS)

HDFS is an integral part of the Hadoop framework and has already reached the level of de-facto standard. It has been designed for large data files and is well suited for quickly ingesting data and bulk processing Data Storage Technologies

- **Big Data Querying Platforms:** Technologies that provide query facades in front of big data stores such as distributed file systems or NoSQL databases. The main concern is providing a high-level interface, e.g. via SQL like query languages and achieving low query latencies.

- **NoSQL Databases:** Probably the most important family of big data storage technologies are NoSQL database management systems. NoSQL databases use data models from outside the relational world that do not necessarily adhere to transactional properties of atomicity, consistency, isolation, and durability (ACID).

- **NewSQL Databases:** A modern form of relational databases that aim for comparable scalability as NoSQL

databases while maintaining the transactional guarantees made by transactional database systems.

- **Big Data Querying Platforms:** Technologies that provide query facades in front of big data stores such as distributed file systems or NoSQL databases. The main concern is providing a high level interface, e.g. via SQL like query languages and achieving low query latencies.

1. NoSQL Databases :

NoSQL databases are design for scalability, often by sacrificing consistency. Compared to relational databases they often use low-level, non-standardized query interfaces, which make them more difficult to integrate in existing applications that expect an SQL interfaces. The lack of standard interfaces makes it harder to switch vendors. NoSQL databases can be distinguished by the data models they use.

- Key-Value Stores: Key-value stores allow storage of data a schema-less way. Data objects can be completely unstructured or structured and are accessed by a single key. As no schema is used, it is not even necessary that data objects share the same structure.
- **Columnar Stores:** "A column-oriented DBMS is a database management system(DBMS) that stores data table as sections of columns of data rather than as rows of data, like most relational DBMSs". Such databases are typically sparse, distributed, and persistent multi-dimentional sorted maps in which data is indexed by a triple of a row key, column key, and a timestamp.
- **Document Databases:** In contrast to the values in a key- value store, documents are structured. However, there is no requirement of common schema that all document are must adhere to as in the case for records in a relational databases. Thus document databases are referred as a storing semi-structured data similar to key-value stores, documents can be querying their internal structure, such as requesting all documents that contain a field with a specified value.
- Graph Databases: Graph databases store in graph structured making them suitable for storing highly associative data such as social networking graphs.

2. NewSQL Databases :

NewSQL databases are modern form of relational databases that aim for comparable scalability with NoSQL databases while maintaining the transactional guarantees made by transactional database systems. The expectation is that NewSQL systems are about 50 times faster than traditional OLTP RDBMS.

3. Big Data Query Platforms :

Big data query platform provide query facades on top of underlying data stores. They typically offer an SQL like query interface for accessing the data, but differ in their approach and performance.

4. Cloud Storage :

As cloud computing grows in popularity, its influence on big data grows as well. While Amazon, Microsoft, and Google build on their own cloud platforms, other companies including IBM, HP, Dell, Cisco etc., build their proposal around Open Stack, an open source platform for building cloud systems, Cloud in general, and particularly cloud storage, can be used by both enterprises and end users. For end users, storing their data in the cloud enable access from everywhere and from every device in a reliable way. As cloud storage is a service, applications using this storage have less control and may experience decreased performance as a result of networking. These performance differences need to be taken into account during design and implementation stages.

Challenges of big data storage and management

With the rate of data explosion, storage system of organizations and enterprises are facing major challenges from huge quantities of data, and ever increasing of generated data. Data irrespective of its size play a vital role in the industry. Value can be created from large data set. For example, Facebook increases its ad revenue by mining its user personal preferences and creating profiles, showcasing advertisers which products they are most interested in. Google also uses data from Google search, Google hangouts, YouTube, and Gmail accounts to profile user's behavior.

In spite of numerous benefits that can be gained in large data set, big data demand for storage and processing poses a major challenge. The total size of data that will be generated by the end of 2015 is estimated at 7.9 zettabytes(ZB), and by 2020, is expected to reach 35 ZB. It is clear that big data has outgrown its current infrastructure, and pushes the limit on storage capacity and storage network. Existing traditional techniques cannot support and perform effective analysis, due to large scale of data.

Big Data Storage Management

Due to massive increase, and the heterogeneous nature of application data, one main challenge of big data is effectively, manage the petabyte (PB) of data being generated daily. Storage management encompasses technologies and process organization to improve data storage performance. Big data require more efficient technologies in processing large quantities of data within an acceptable time frame. A wide range of techniques and technologies have been developed and adopted to manipulate, analyze and visualize big data. Technologies such as massive parallel processing (MPP) database, data mining grids, distributed file system, cloud computing platforms, and scalable storage systems are highly desirable. The development of Map-Reduce, with Yahoo's Pig, alongside Facebook's Cassandra applications, has gotten the attention of the industry. Google file system, GFS, is designed to meet the increasing demands of big data, such as scalability, reliability and availability. GFS is composed of clusters, which is made up of hundreds of storage servers that support several terabytes of disk space. This meets the scalability issue of big data. Hadoop is free version of Map-reduce implementation by Apache Foundation. Hadoop distributed file system(HDFS), is a distributed file system designed to run on commodity hardware. HDFS can store data across thousands of servers. All data in HDFS is reduced into block size chunk, and distributed across different nodes and are managed by the Hadoop cluster

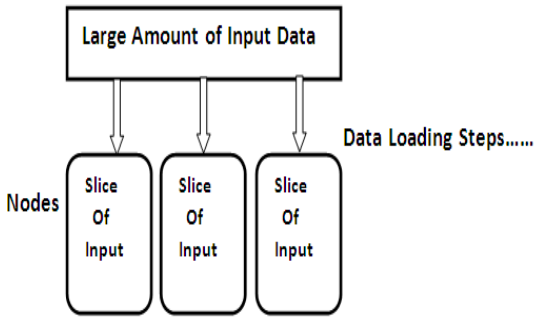

Figure 6: Distributed data across nodes at load time

Figure 6 shows the distribution of data across different data nodes to enhance performance of the entire Hadoop system. Storage vendors such as NetApp, EMC, Hitachi Data Systems, and many more are offering storage management solutions to big data inclined companies. EMC VPLEX enables manageability of storage area network, through a vital storage infrastructure that consolidates heterogeneous storage devices.

Suggestion for a Big Data Storage and Management

System Big data storage and management has been challenge in the face of the increasing volume of organizational data. This paper suggests approaches that we think in our opinion efficiently mitigate challenges of big data storage and management.

1. **Big Data Storage Mediums**: All storage management solutions have at their end, storage devices. The qualities of these storage devices can have significant impact on the entire storage effectiveness; can be critical in the big data environment. We suggest the use of a hybrid storage device, which is an aggregate of hard disk drive (HDDs), and solid state drives (SSDs). HDDs provide huge storage capacity, at a relatively cheap price. This characteristic of the HDDs allows storage system to scale to meet the rate of growth of data. The disadvantage here is that, HDDs has a slow data transfer rate becoming a bottleneck for performance. On the other hand, SSDs provides avenue for high performance and reliability. They have low latency, thus providing a much faster, random access. SSDs are very expensive for the storage capacity they provide. The combination of these storage devices into a logical unit in an array should solve the storage demands posed by large datasets.

2. **Backup Strategies:** Recovery is the main objective for backup. The ability of production system to recover, and in a timely manner is very crucial in the era of big data. From this prospective, this paper recommend full back as a favorable choice over the others. Full backup ensures speedy recovery, through it takes a considerable amount of time to back up a large dataset. Applying full backup to large datasets may increase the data block repetition. Data reduplication technology, significantly reduce the volume of stored data block for every single full backup, and allow users to backup, and recover data within relatively short period of time. Rather than directly from the production system. Replication keeps copy of production data in real time.

3. **Business Continuity and Disaster Recovery:** An optimal business continuity solution, takes into account, two parameters, to a negligible level – Recovery Point Objective (RPO), which is the point in time that a production system, and data must be recovered after a disaster. Recovery Time Objective (RTO) is the time frame within which production system, and data must be recovered after a disaster. In a large dataset, the complexity of business continuity increases, with the influx of a variety of data, which must be maintained in their formats. Business continuity planning requires, saving a copy, or multiple copies of production data, through backups, local or remote replication. The use of enterprise software such as EMC Power path can be beneficial. EMC Power path provides features such as cluster support, dynamic load balancing, configuration and management, automatic path failover.

Conclusion

In the recent years, a academia pays more attention to cloud computing. Big data focuses on “data”, like data service, data acquisition, analysis and data mining which pays more attention on ability on data storage. Cloud computing focuses on computing architecture and practices. Big data and cloud computing are two sides of same issue.

Big data era has brought about an explosive growth of data. The increase of mobile applications, social media, and big data analytic initiatives has cause big data storage challenges to become even greater. Choosing the right storage devices, management tool, and efficient techniques is relevant and determines the rate of growth.

The approaches to big data storage and management can significantly, affect the entire organization. This paper examines and summarizes existing current storage technologies for big data applications. Variables such as capacity, scalability, data transfer rate, access time, and cost of storage devices, are also highlighted. Finally some suggestions are made to curb the problems posed by big data storage.

References

[1] Wikipedia, Big Data. http://en.wikipedia.org/wiki/Big data

[2] Singhal , R., Bokare, S., & Pawar, P.(2010). Enterprise storage architecture for optimal business continuity.

[3] White, T. (2012). Hadoop: The definitive guide: ”O’Reilly Media, Inc.”

[4] International Conference on Operations Excellence and Service Engineering Orlando, Florida, USA, September 10-11,2015 “Big Data Analysis and Storage”

[5] AT Kearney, “Big Data and the Creative Destruction of Today’s Business Model” 2013.

[6]Avita Katal, Mohammad Wazid and R.H. Goudar, “Big Data: Issues, Challenges, Tools and Good Practices”,978-14799-0192-0/13.

[7] Amrit Pal et al,”A Performance Analysis of MapReduce Task with large Number of Files Dataset in Big Data Using Hadoop”

International Conference on Communication Systems and Network Technologies IEEE,978-1-4799-3070-8/14,2014.

[8]Glavic, B. (2014). Big Data provenance: Challenges and implications for benchmarking. In T. Rabl, M. Poess, C.Baru, & H. A. Jacobsen (Eds.), specifying big data benchmarks (pp. 72-80). Berlin: Spriger.

[9] "Challenges of big data storage and management" Rajeev Agrawal * Global Journal of Information Technology Volume 06,Issue 1,(2016)01-10

[10]"Big Data Storage and Challenges" M.H.Padgavankar1, Dr. S.R.Gupta, M.H.Padgavankar et al/(IJCSIT) International Journal of Computer Science and Information Technologies, Vol.5(2), 2014,2218-2223

[11] http://www.nature.com/news/specials/bigdata/index.html

[12] "Big Data and Cloud Computing Issues", Awodele. O, International Journal of Computer Applications (0975-8887) Volume 133-No.12, January 2016

[13] "The Research and Application of a Big Data Storage Model", Na Liu and Jianfei Zhou, International Journal of Database Theory and Application Vol. 8, No (2015), pp.319-330 http://dx.doi.org/10.14257/ijdta.2015.8.4.32.

[14] Usha Albuquerque & Nidhi Prasad,"Career in Data Science", Employment News. Vol XLIII, No. 20 pp1, New Delhi, DOI: 18 to 24-August-2018.