Survey on Current Issues Big Data Storage

Betsy Chacko
Department of Computer Applications,
St. Mary's College,
Thrissur, Kerala-680301

Abstract— **Big data is defined as large amount of data which requires new technologies and architectures so that it becomes possible to extract value from it by capturing and analysis process. Due to such large size of data it becomes very difficult to perform effective analysis using the existing traditional techniques. Big data due to its various properties like volume, velocity, variety, variability, value and complexity put forward many challenges. Since Big data is a recent upcoming technology in the market which can bring huge benefits to the business organizations, it becomes necessary that various challenges and issues associated in bringing and adapting to this technology are brought into light. This paper introduces the different data organizations and its challenges in the era of big data. The paper also proposes a new method to data storage so as to flow easily with moving data and hence achieve an efficient data retrieval from the ocean of data.**

Keywords— Big data, structured data, unstructured data, semi-structured data, hay-stack

I. INTRODUCTION

Data is growing at a huge speed making it difficult to handle with present technology. The main difficulty in handling such data is that the volume is increasing rapidly in comparison to the computing resources. Retrieval of the moving data need a moving technology. Many organizations are struggling to deal with increasing data volumes, and big data simply makes the problem worse. How to preserve more information, without overloading the systems is the question frequently asked. Effective retrieval with optimum result is another issue. Whether the data is stored on premise or in the cloud, structured data is a critical part of an organization's managed information that needs to be considered in responding to any legal discovery. Big Data comes with almost 80% of unstructured data and hence we can think of a storage structure which accommodates both structured and unstructured data in a single information stack.

II. ORGANIZATION OF DATA AND CHALLENGES

A. Structured Data

Structured data refers to information with a high degree of organization, such that inclusion in a relational database is easy and readily searchable by simple, straightforward search algorithms or other search operations. Structured data is both queried and archived as business entities and can be retrieved as a complete response to legal discovery. It has unique limitations and requirements for data collection that require a different approach. As we know it has the advantage of being easily entered, stored, queried and analysed. At one time, because of the high cost and performance limitations of storage, memory and processing, relational databases and spreadsheets using structured data were the only way to effectively manage the data. The keys to a successful collection are to identify the best process based on the needs and limitations of the structured data system, verify the data by some means and then validate the process by the way of queries used for extraction or backup logs. Archiving structured data can speed up e-discovery, ensure compliance, reduce application overhead, and improve production system performance.

Challenge : As the volume and velocity of data become too large, the conventional approach of a structured data processing starts declining. And in this era of Big data Tsunami it does matter. Semi structured data can be an alternative in this context. Fig.1 shows the failure in scalability of Relational Database.

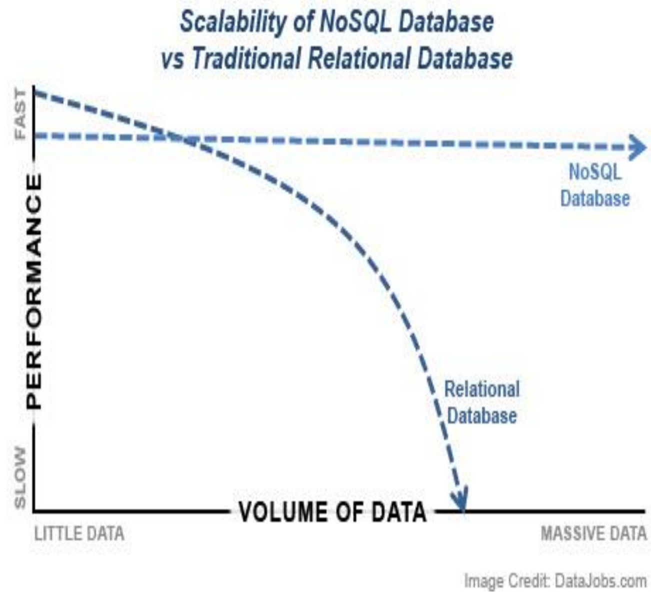

Image Credit: DataJobs.com

Figure 1: Scalability of Not Only SQL (NoSQL) Database vs. Traditional Relational Database

B .Unstructured Data

Unstructured data refers to information that does not have a pre-defined organization. It can't be so readily classified and fit into a neat box: photos and graphic images, videos, streaming instrument data, webpages, pdf files, PowerPoint presentations, emails, blog entries, wikis and word processing documents.

Unstructured information is typically text-heavy and it results in irregularities and ambiguities that make it difficult to understand using traditional computer programs as compared to data stored in fielded form in databases. Nowadays it is seen that around 80-90% of all potentially usable business information may originate in unstructured form. In this context the non-relational database gives a better approach towards data processing.

If it was possible or feasible to instantly transform unstructured data to structured data, then creating intelligence from unstructured data would be easy. However, structured data is akin to machine-language, in that it makes information much easier to deal with using computers; whereas unstructured data is usually for humans, who don’t easily interact with information in strict, database format. Email is an example of unstructured data; because while the busy inbox of a corporate human resources manager might be arranged by date, time or size; if it were truly fully structured, it would also be arranged by exact subject and content, with no deviation or spread – which is impractical, because people don’t generally speak about precisely one subject even in focused emails. Spreadsheets, on the other hand, would be considered structured data, which can be quickly scanned for information because it is properly arranged in a relational database system. The problem that unstructured data presents is one of volume. Because the pool of information is so large, current data mining techniques often miss a substantial amount of the information even if efficient analytics are used. Unstructured data is a generic label for describing any data that is not in a database or other type of data structure.

Challenge : Effective information retrieval of moving data (with a maximum velocity) for an optimum result is challenging because the data is completely unstructured and finding the key for search from the unstructured heap is a difficult task.

C. Semi structured Data

The semi-structured model is a database model where there is no separation between the data and the schema, and the amount of structure used depends on the purpose. It is a cross between the two. It is a type of structured data, but lacks the strict data model structure. With semi-structured data, tags or other types of markers are used to identify certain elements within the data, but the data doesn’t have a rigid structure. For example, word processing software now can include metadata showing the author's name and the date created, with the bulk of the document just being unstructured text. Emails have the sender, recipient, date, time and other fixed fields added to the unstructured data of the email message content and any attachments. Photos or other graphics can be tagged with keywords such as the creator, date, location and keywords, making it possible to organize and locate graphics. XML and other mark-up languages are often used to manage semi-structured data. Semi-structured data is increasingly occurring since the advent of the Internet where full-text documents and databases are not the only forms of data anymore and different applications need a medium for exchanging information.

The advantages of this model are the following:

- It can represent the information of some data sources that cannot be constrained by schema.
- It provides a flexible format for data exchange between different types of databases.
- It can be helpful to view structured data as semi-structured (for browsing purposes).
- The schema can easily be changed.
- The data transfer format may be portable.

Challenge: The traditional relational data model has a popular and ready-made query language.Queries cannot be made as efficient as in a more constrained structure, such as in the relational model. Typically the records in a semi-structured database are stored with unique IDs that are referenced with pointers to their location on disk. This makes navigational or path-based queries quite efficient, but for doing searches over many records (as is typical in SQL), it is not as efficient because it has to seek around the disk following pointers.

III. BIG DATA: A MOVING TARGET

Big data refers to technologies and initiatives that involve data that is too diverse, fast-changing or massive for conventional technologies, skills and infra-structure to address efficiently. Said differently, the volume, velocity or variety of data is too great. What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain. Big data is a moving target; what is considered to be "Big" today will not be so years ahead. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."The following graph represents the data growth in every year.

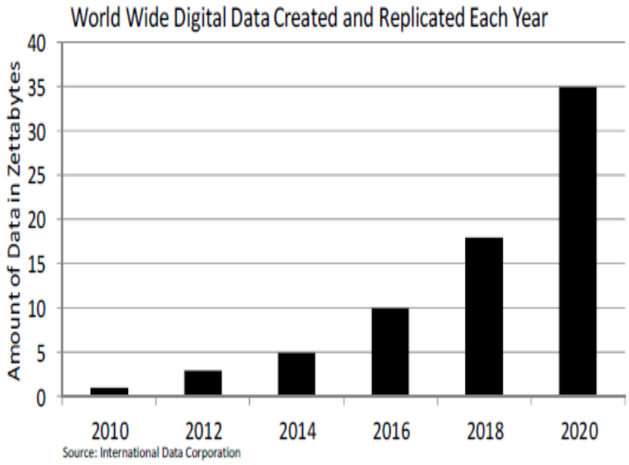

Figure 2: Data Growth in Zettabytes

Big data can be defined with the following properties associated with it:

a) Variety

Data being produced is not of single category as it not only includes the traditional data but also the semi structured data from various resources like web Pages, Web Log Files, social media sites, e-mail, documents, sensor devices data both from active passive devices. All this data is totally different consisting of raw, structured, semi structured and even unstructured data which is difficult to be handled by the existing traditional analytic systems.

b) Volume

The Big word in Big data itself defines the volume. At present the data existing is in zettabytes and is supposed to increase to yottabytes in nearby future. The social networking sites existing are themselves producing data in order of terabytes everyday and this amount of data is definitely difficult to be handled using the existing traditional systems.

The Big Data Stack Data growth curve:
Megabytes - Gigabytes - erabytes - Petabytes - Exabytes - Zettabytes - Yottabytes - Brontobytes - Geopbytes

c) Velocity

Velocity in Big data is a concept which deals with the speed of the data coming from various sources. This characteristic is not being limited to the speed of incoming data but also speed at which the data flows. For example the data from the sensor devices would be constantly moving to the database store and this amount won't be small enough. Thus our traditional systems are not capable enough on performing the analytics on the data which is constantly in motion.

d) Variability

Variability considers the inconsistencies of the data flow. Data loads become challenging to be maintained especially with the increase in usage of the social media which generally causes peak in data loads with certain events occurring.

e) Complexity

It is quite an undertaking to link, match, cleanse and transform data across systems coming from various sources. It is also necessary to connect and correlate relationships, hierarchies and multiple data linkages or data can quickly spiral out of control.

f) Value

User can run certain queries against the data stored and thus can deduct important results from the filtered data obtained and can also rank it according to the dimensions they require. These reports help these people to take smart decisions and to find the business.

There is no question that organizations are swimming in an expanding sea of data that is either too voluminous or too unstructured to be managed and analysed through traditional means. Among its burgeoning sources are the clickstream data from the Web, social media content (tweets, blogs, Facebook wall postings, etc.) and video data from retail and other settings and from video entertainment. But big data also encompasses everything from call center voice data to genomic and proteomic data from biological research and medicine. Every day, Google alone processes about 24 petabytes (or 24,000 terabytes) of data. Yet very little of the information is formatted in the traditional rows and columns of conventional databases.

A single database model or technology cannot satisfy the needs of every organization or workload. Despite its success and universal adoption, this is also true for RDBMS technology. This is especially true when processing large amounts of multi-structured data we have to develope non-relational systems to deal with extreme data volumes. Web-focused companies such as Google and Yahoo that have significant volumes of web information to index and analyze. Non-relational systems are useful for processing big data where most of the data is multi-structured. They are particularly popular with developers who prefer to use a procedural programming language, rather than a declarative language such as Structured Query Language (SQL),to process data. These systems support several different types of data structures including document data, graphical information, and key-value pairs.

Two important big data trends for supporting the store and analyze approach are relational Database Management System (DBMS) products optimized for analytical workloads (often called analytic Relational DBMSs, or Advanced DBMSs) and non-relational systems (sometimes called NoSQL systems) for processing multi-structured data. A non-relational system can be used to produce analytics from big data, or to preprocess big data before it is consolidated into a data warehouse.

A. Storage and Processing Issues

1. Deploy optimized hardware and software solutions for processing different types of big data workloads, and then combine these solutions with the existing enterprise data warehouse to create an integrated information system.

2. The different varieties of raw source data exist in organizations is to be analyzed and deliver the analytical results to business users. Processing of multiple variety of data brings difficulty in analysis.
3. Supporting extreme workloads is not a new challenge for the computing industry. But since the data growth is too faster the business transaction processing systems are incapable of processing such a huge work load.
4. The storage available is not enough for storing the large amount of data which is being produced by almost everything.
5. Uploading the large amount of data in cloud will take large amount of time to get uploaded and moreover this data is changing so rapidly which will make this data hard to be uploaded in real time. At the same time, the cloud's distributed nature is also problematic for big data analysis.
6. Processing of large amount of data takes large amount of time.

III. CONTRIBUTION

In the present situation there is no time to think about data storage, by that time it has come. In this paper I propose the hay-stack storage method which can act as a better storage platform for big data.

The structure follows the visualization of a hay-stack seen near paddy fields. It is built upon a strong foundation and there is a bar in the middle which strengthens the stack though the heap is wide or tall. Fig.2 shows the hay-stack structure for big data storage.

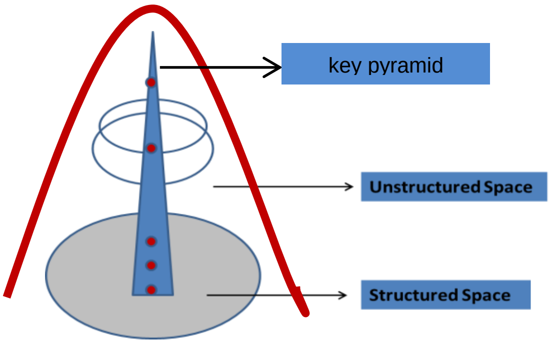

Figure 2: hay-stack storage

The method proposes an augmentation storage system. The existing relational, non-relational databases need not be restructured completely. As in Fig.2 the structure is based on a key pyramid which is the back bone of the structure. The pyramid contains the keys (red dots) of different variety of data in an ordered sequence and the circle represents the area of structured data of the corresponding key. Each circle can be considered as an independent stack of same/different design and linked through the pyramid. Now traversing through the pyramid is nothing but searching through variety of data. Building up indexes right in the beginning while collecting and storing the data will reduce the processing time considerably. Key data can be searched using any of the available search algorithms. As the *volume* of available key data increases the stack starts growing horizontally neglecting its structured nature. The search always starts with structured data followed by unstructured data and hence attain an optimal result. The pyramid gets taller and taller as *variety* increases and hence scalability can be resolved. The processing can be much faster since the data is completely built upon a single key pyramid.

The logical representation of 'hay-stack' can be a better solution for big data storage and the physical representation and its challenges shall be considered for my future work.

IV. CONCLUSION

Big data is being generated by everything around us at all times. Every digital process and social media exchange produces it. Systems, sensors and mobile devices transmit it Big data is changing the way people within organizations work together. It is creating a culture in which business and IT leaders must join forces to realize value from all data. But escalating demand for insights requires a fundamentally new approach to architecture, tools and practices. Big data is arriving from multiple sources at an alarming velocity, volume and variety. To extract meaningful value from big data, we need optimal processing power, analytics capabilities and skills. This paper described the concept of different organizations of data, its challenges in the era of big data. To accept and adapt the new technology many challenges and issues exist which need to be brought up as a future work. The challenges, issues and the proposed method will help the business organizations which are moving towards this technology for increasing the value of the business. Several new and enhanced data management and data analysis approaches help the management of big data and the creation of analytics from that data. The actual approach used will depend on the volume of data, the variety of data, the complexity of the analytical processing workloads involved, and the responsiveness required by the business.

REFERENCES

[1] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money, "Big Data: Issues and Challenges Moving Forward", *IEEE, 46th Hawaii International Conference on System Sciences,* 2013
[2] Philip Russom,Fourth Quarter 2013, TDWI best practi ces Report tdwi.orgTDWIresearch
[3] Colin White, BI Research, Using Big Data for Smarter Decision Making July 2011,Sponsored by IBM
[4] Sam Madden, "From Databases to Big Data", IEEE, Internet Computing, May-June 2012.
[5] International Journal of Application or Innovation in Engineering &Management (IJAIEM), Volume 3, Issue 3, March 2014 ISSN 2319 – 4847Understanding the big data problems and their solutions using hadoop and map-reduce
[6] http://practicalanalytics.wordpress.com/2011/05/15/new-tools-for-new-times-a-primer-on-big-data