A Review on Big Data

Monika Yadav [1], Naveen Prakash[2],
[1,2]Ganga Institute of Technology and Management,
MDU Rohtak

Mr. Joni Birla[3]
[3]Assistant Professor
Department of Computer Science and Engineering,
GITAM, MDU Rohtak

We are living in theInformation Era – the Age of BIG data. It is clearly visible that organizations need to employ data-driven decision making to gain competitive advantage. Processing, integrating and interacting with huge data *should* make it better data which gives both more panoramic and more granular views to aid strategic decision making. This is made possible via Big Data exploiting affordable and usable Computational and Storage Resources. It remains unclear what Big Data actually is; current offerings appear as isolated silos that are difficult to integrate and/or make it difficult to better utilize existing data and systems. In this paper we are studying about Big data, its 3 V's, Architecture of Big Data, Analysis of Big data, Big Data for enterprise.

Index Terms**—Big Data, Data mining on large scale, V's architecture.**

I INTRODUCTION

Big data is a phase that shows to data sets or combinations of data sets whose size (volume), complexity (variability) and rate of growth (velocity) make them difficult to be caught, set, processed or analyzed by conventional technologies and tools such as relational databases and desktop visualization features within the time necessary to make them useful. While the size used to determine whether a particular data set is considered, big data is not firmly defined and continues to change over time, most analysts and practitioners currently refer to data sets from 30-50 terabytes(1012 or 1000 gigabytes per terabyte) to multiple Petabytes (1015 or 1000 terabytes per Petabyte) as big data.[1]

II 3 V'S OF BIG DATA

The 3Vs that define Big Data are Variety, Velocity and Volume which are explained as follows:

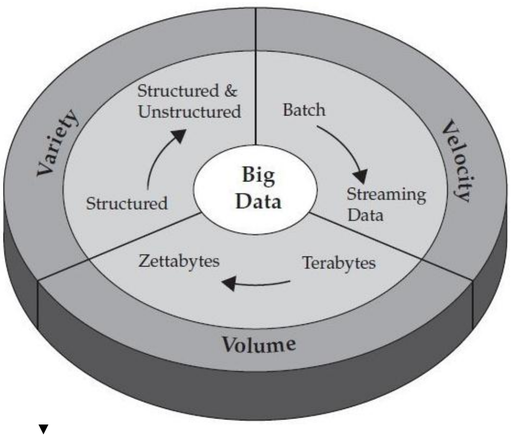

Fig.1–3 V's of Big Data

Volume: We currently see the exponential growth in the data storage as the data is now more than text data. We can find data in the format of videos, music and large images on our multimedia channels. It is very common to have Terabytes and Petabytes of the storage system for enterprises. As the database grows the applications and architecture built to support the data needs to be reevaluated quite often. Sometimes the same data is re-checked with multiple angles and even though the useful data is the same the new found intelligence creates explosion of the data. The big volume represents Big Data.

Velocity: The data growth and social media explosion have changed how we look at the data. There was a time when we believed that previous data is recent. The matter of the fact newspapers and other sources are still following that logic. However, news channels and radios have changed how fast we get the news. Today, people answered on social media to update them with the latest happening. On social media sometimes old messages (a tweet, status updates etc.) is not something interests users. They often reject old messages and pay attention to recent updates. The data movement is now almost real time and the update window has reduced to fractions of the seconds. This high velocity data represent Big Data.

Variety:Information can be stored in multiple multimedia formats. For example database, excel, access or for the matter of the fact, it can be stored in a simple text file. Sometimes the data is not even in the traditional format as we assume, it may be in the form of video, SMS, pdf (Portable document format) or something we might have not thought about it. It is the need of the organization to set it and make it meaningful. It will be easy to do so if we have data in the same format, however it is not the case most of the time. The real world has data in many different formats and that is the challenge we need to overcome with the *Big Data*. This variety of the data represents Big Data.[2]

III ARCHITECTURE OF BIG DATA

In 2000, Seisint Inc. (now LexisNexis Group) made a C++-based file-sharing framework for data storage and query. The whole system stores and distributes structured, semi-structured, and information across multiple servers. Users can build queries in a C++ dialect called ECL. ECL uses an "apply schema on read" method to infer the structure of stored data when it is queried, instead of when it is stored. In 2004, LexisNexis acquired Seisint Inc. and in 2008 acquired Choice Point, Inc. and their high-speed parallel processing platform. The two platforms were merged into HPCC (or High-Performance Computing Cluster) Systems and in 2011, HPCC was open-sourced under the Apache v2.0 License. Quant cast File System was available about the same time.

In 2004, Google published a paper on a process called Map Reduce that uses a similar architecture. The Map Reduce concept provides a parallel processing model, and an associated implementation was released to process huge amounts of data. With Map Reduce, queries are split and distributed across parallel nodes and processed in parallel (the Map step). The results are then gathered and delivered (the Reduce step). The framework was very successful, so others wanted to replicate the algorithm. Therefore, an implementation of the Map Reduce framework was adopted by an Apache open-source project named Hadoop.

MIKE2.0 is an open approach to information management that acknowledges the need for revisions due to big data implications identified in an article titled "Big Data Solution Offering". The methodology addresses handling big data in terms of useful permutations of data sources, complexity in interrelationships, and difficulty in deleting (or modifying) individual records.

2012 studies showed that multiple-layer architecture is one option to address the issues that big data presents. A distributed parallel architecture distributes data across multiple servers; these parallel execution environments can dramatically improve data processing speeds. This type of architecture inserts data into a parallel DBMS, which implements the use of Map Reduce and Hadoop frameworks. This type of framework looks to make the processing power transparent to the end user by using a front-end application server.

Figure 2 gives Layered Architecture of Big Data System. It can be decomposed into three layers, including Infrastructure Layer, Computing Layer, and Application Layer from top to bottom. [3]

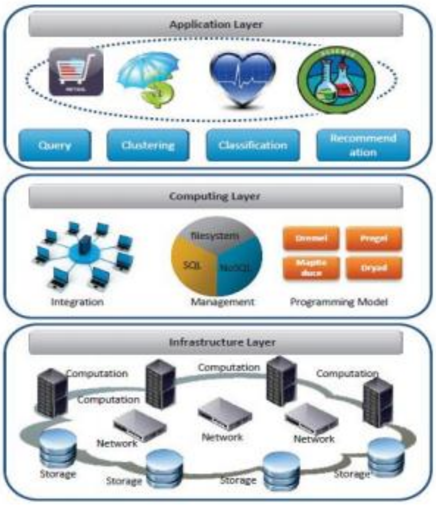

Fig. 2-Architecture of Big Data

IV ANALYSIS OF BIG DATA

Big data has to be collected, massaged, linked together and interpreted for it to be of any use to anyone. Companies and other entities need to filter the vast amount of available data to get to what's most relevant to them. Fortunately, hardware and software that can process, store and analyze huge amounts of information are becoming cheaper and faster, so the work no longer requires massive and prohibitively expensive supercomputers. Some of the software is becoming more user friendly so that it doesn't necessarily take a team of programmers and data scientists to wrangle the data (although it never hurts to have knowledgeable people who can understand your requirements).

Companies take advantage of cloud computing services so that they don't even have to buy their own computers to do all that data crunching. Data centers, also called server farms, can distribute batches of data for processing over multiple servers, and the number of servers can be scaled up or down quickly as needed. This scalable distributed computing is accomplished using innovative tools like Apache Hadoop, MapReduce and Massively Parallel Processing (MPP). NoSQL databases have been developed as more easily scalable alternatives to traditional SQL-based database systems.

Much of this big data processing and analysis is aimed at finding patterns and correlations that provide insights that can be exploited or used to make decisions. Businesses can now mine massive amounts of data for information about consumer habits, their products' popularity or more efficient ways to do business. Big data analytics can be used to target relevant ads, products and services at the customers they believe are most likely to buy them, or to create ads that are more likely to appeal to the public at large. Companies are now even starting to do things like send real-time ads and coupons to people via their smartphones for places that are near locations where they have recently used their credit cards.

It's not just for making us buy stuff, however. Businesses can use the information to improve efficiency and practices, such as finding the most cost-effective delivery routes or stocking merchandise more appropriately. Government agencies can analyze traffic patterns, crime, utility usage and other statistics to improve policy decisions and public service. Intelligence agencies can use it to, well, spy, and hopefully foil criminal and terrorist plots. News outfits can use it to find trends and develop stories, and, of course, write more articles about big data.

In essence, big data allows entities to use nearly real-time data to inform decisions, rather than relying mostly on old information as in the past. But this ability to see what's going on with us in the present, and even sometimes to predict our future behavior, can be a bit creepy.[4]

V BIG DATA FOR THE ENTERPRISE

With Big Data databases, enterprises can save money, grow revenue, and achieve many other business objectives, in any vertical.

Build new applications: Big data might allow a company to collect billions of real-time data points on its products, resources, or customers – and then repackage that data instantaneously to optimize customer experience or resource utilization. For example, a major US city is using MongoDB to cut crime and improve municipal services by collecting and analyzing geospatial data in real-time from over 30 different departments.

Improve the effectiveness and lower the cost of existing applications: Big data technologies can replace highly-customized, expensive legacy systems with a standard solution that runs on commodity hardware. And because many big data technologies are open source, they can be implemented far more cheaply than proprietary technologies. For example, by migrating its reference data management application to MongoDB, a Tier 1 bank dramatically reduced the license and hardware costs associated with the proprietary relational database it previously ran, while also bringing its application into better compliance with regulatory requirements.

Realize new sources of competitive advantage: Big data can help businesses act more nimbly, allowing them to adapt to changes faster than their competitors. For example, MongoDB allowed one of the largest Human Capital Management (HCM) solution providers to rapidly build mobile applications that integrated data from a wide variety of disparate sources.[5]

VI CLOUD COMPUTING

Cloud computing refers to a broad set of computing and software products that are sold as a service, managed by a 3rd-party provider and delivered over a network. Infrastructure-as-a-Service (IaaS) is a flavor of cloud computing in which on-demand processing, storage or network resources are provided to the customer. Sold on-demand with limited or no upfront investment for the end-user, consumption is readily scalable to accommodate spikes in usage. Customers pay only for the capacity that is actually used (like a utility), as opposed to self-hosting, where the user pays for system capacity it is used or not.

As compared to self-hosting, IaaS is:

Inexpensive: To self-host an application, one has to pay for enough resources to handle peak load on an application, at all times. Amazon discovered that before launching its cloud offering it was using only about 10% of its server capacity the vast majority of the time.

Tailored: Small applications can be run for very little cost by taking advantage of spare capacity. Bandwidth, processing and storage capability can be added in relatively small increments.

Elastic: Computing resources can easily be added and released as needed, making it much easier to deal with unexpected traffic spikes.

Reliable: With the cloud, it’s easy and inexpensive to have servers in multiple geographic locations, allowing content to be served locally to users, and also allowing for better disaster recovery and business continuity.

Overall, cloud computing provides better agility and scalability, together with lower costs and faster time to market. However, it does require that applications be engineered to take advantage of this new infrastructure; applications built for the cloud need to be able to scale by adding more servers, for example, instead of adding capacity to existing servers.

On the storage layer, traditional relational databases were not designed to take advantage of horizontal scaling. A class of new database architectures, dubbed NoSQL databases, is designed to take advantage of the cloud computing environment. NoSQL databases are natively able to handle load by spreading data among many servers, making them a natural fit for the cloud computing environment. Part of the reason NoSQL databases can do this is that related data is always stored together, instead of in separate tables.

In fact, MongoDB is built for the cloud. Its native scale-out architecture, enabled by 'sharding,' aligns well with the horizontal scaling and agility afforded by cloud computing. Sharding automatically distributes data evenly across multi-node clusters and balances queries across them. In addition, MongoDB automatically manages sets of redundant servers, called 'replica sets,' to maintain availability and data integrity even if individual cloud instances are taken offline. To ensure high availability, for instance, users can spin up multiple members of a replica set as individual cloud instances across different availability zones and/or data centers. MongoDB has also partnered with a number of leading cloud computing providers, including Amazon Web Services, Microsoft and Soft Layer.[6]

REFERENCE

[1] http://www.ijsrp.org/research-paper-1014/ijsrp-p34125.pdf
[2] http://blog.sqlauthority.com/2013/10/02/big-data-what-is-big-data-3-vs-of-big-data-volume-velocity-and-variety-day-2-of-21/
[3] file:///C:/Users/Khushboo/Downloads/5-ciima-2013-13-2-chan-1-14.pdf
[4] http://computer.howstuffworks.com/internet/basics/what-is-big-data-1.htm
[5] https://www.mongodb.com/big-data-explained
[6] https://en.wikipedia.org/wiki/Big_data