International Journal of Science Engineering and Advance Technology

A Study on Big Data analysis Environment

[1] Varaprasad. Sabbithi, [2]Dr.John Mathew Rapaka
[1]Assoc.Prof, Dept of ECE,
Srinivasa Institute of Engineering and Technology,
Amalapuram, JNTUK, A.P, India
[2]Professor, Dept.of.CSE,
Srinivasa Institute of Engineering and Technology,
Amalapuram, JNTUK, A.P, India

ABSTRACT

The use of electronics to implement IT solutions has generally known about big data since recent century. IT was use full for compute a vast applications for implementing business processes, its value intention has always became the ability to handle large amounts of data more efficiently, and more accurately than human beings. The purpose of this paper is to implement the performance and measure the capacity of big data solutions, which must be taken into account for such solutions to be viable. This paper gives an overview of big data and the benefits that it offers, describes the performance and capacity aspects which helps to create big data solutions, and suggests what to be done to implement bigdata in future technology .

Key word- Big data, business process, stability.

I.INTRODUCTION

.

Does big data represent an incremental change for IT or a major transformation? After all, technologies for data warehouses, data mining, and business intelligence have been with us for years, and manufacturing applications have long used analytics to respond quickly to variances found by sorting through large volumes of rapidly arriving process data. The premise of this paper is that the proper answer is "both." Just as cloud computing enables new ways for businesses to use IT because of many years of incremental progress in the area of virtualization, big data now enables new ways of doing business by bringing advances in analytics and management of both structured and unstructured data into mainstream solutions [1].

Consider these examples:

- A major US retailer adds weather data to its distribution algorithms so that it can model delivery paths and can use disparate sources for improved logistics.
- A major Indian telecommunications firm analyzes billions of call records daily to target customers for special offers, which results in reducing churn and increasing loyalty among customers.

Big data solutions now enable us to change the way we do business, in ways that were not possible just a few years ago, by taking advantage of previously unused sources of information.

Down through the years of human history, the most successful decisions that were made in the world of business were based on the interpretation of available data. Every day, 2.5 quintillion bytes of data are created—so much that 90% of the data in the world today has been created in the last two years. Correct analysis of the data is the key success factor in being able to make better decisions that are based on the data. Given the quantity and complexity of the data that is being created, traditional database management tools and data processing applications simply cannot keep up, much less make sense of it all. The challenges for handling big data include capture, storage, search, sharing, transfer, analysis, and visualization. The trend to larger data sets is due to the additional information that can be derived from analysis of a single large set of related data, compared to separate smaller sets with the same total amount of data. Some estimates for the data growth are as high as 50 times by the year 2020[2].

II. What "big data" means

Big data is a phenomenon that is characterized by the rapid expansion of raw data. This data that is being collected and generated so quickly that it is inundating government and society. Therefore, it represents both a challenge and an opportunity. The challenge is related

to how this volume of data is harnessed, and the opportunity is related to how the effectiveness of society's institutions is enhanced by properly analyzing this information. It is now common place to distinguish big data solutions from conventional IT solutions by considering the following four dimensions:

- *Volume.* Big data solutions must manage and process larger amounts of data.
- *Velocity.* Big data solutions must process more rapidly arriving data.
- *Variety.* Big data solutions must deal with more kinds of data, both structured and unstructured.
- *Veracity.* Big data solutions must validate the correctness of the large amount of rapidly arriving data.

As a result, big data solutions are characterized by real-time complex processing and data relationships, advanced analytics, and search capabilities. These solutions emphasize the flow of data, and they move analytics from the research labs into the core processes and functions of enterprises. A typical big data lifecycle involves the capture and input of data, managing and processing of the data, and the presentation of this data to the user. The performance and capacity challenges from the perspective of the four big data dimensions, based on the lifecycle shown in Figure 1-1 volume, velocity, variety, and veracity.

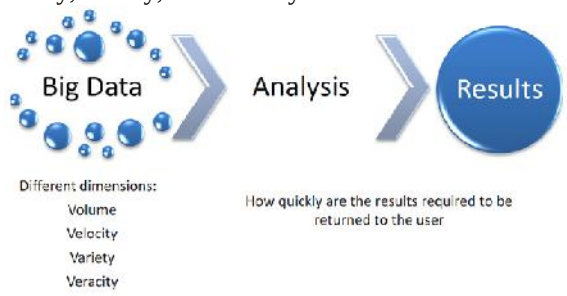

Fig.1-1 Lifecycle of big data

III. Volume: Data that is provisioned in an unprocessed state exemplifies this type of data. This data helps run the business when the data is rolled up into categories that consist of correctly aligned data. This type of data also provides a starting point for further discovery and analysis.

Scalability: The size of big data is easily recognized as an obvious challenge. Big data is pushing scalability in storage, with increases in data density on disks to match. The current Redundant Array of Independent Disks (RAID) approach that is in widespread use does not provide the level of performance and data durability that enterprises dealing with escalating volumes of data require. For example, committing data from memory to disk can increase overhead and cause processing delays if multiple disks are involved in each commit process. Moreover, as the scale of data increases, the mean time between failures (MTBF) falls. For example, a system with a billion cores has an MTBF of one hour. The failure of a particular cluster node affects the overall calculation work of the large infrastructure that is required to process big data transactions.

Furthermore, a large percentage of the data might not be of interest. It can be filtered and compressed by an order of magnitude. The challenge is to filter intelligently without discarding data samples that might be relevant to the task. For example, data that is related to time or location might be subject to wide variances yet still be valid. Data volume is increasing faster than computing resources and processor speeds that exist in the marketplace. Over the last five years, the evolution of processor technology largely stalled, and we no longer see a doubling of chip clock cycle frequency every 18 - 24 months. Now, due to power constraints, clock speeds are largely stalled and processors are being built with increasing numbers of cores. In the past, people who were building large data processing systems had to worry about parallelism across nodes in a cluster. Now, you must deal with parallelism within a single node. Unfortunately, parallel data processing techniques that were applied in the past for processing data across nodes do not directly apply for intra-node parallelism, because the architecture looks very different. For example, there are many more hardware resources, such as processor caches and processor memory channels that are shared across cores in a single node. Furthermore, the move toward packing modern processors with multiple sockets (each with tens of cores) adds another level of complexity for intra-node parallelism [3].

Finally, with predictions of "dark silicon," specifically that power considerations in the future are likely to prohibit us from using all of the hardware in the system continuously, data processing systems will probably be required to actively manage the power consumption of the processor. These unprecedented changes require us to rethink how data processing components are designed, built, and operated. To resolve these problems, there are solutions such as *database sharding* (breaking data into small pieces and running the pieces in parallel). This approach is based on *shared nothing architecture*, which means no shared components between environments. Rather than storing application data in a single database on a single server with a

shared processor, memory, and disk, the database is divided into several smaller shards, each of which can be hosted on independent servers with dedicated processor, memory, and disk. This greatly reduces resource contention. The key benefit is that smaller databases are faster. An increase in the number of nodes leads to a potential increase in the number of failures. Furthermore, there is a gradual shift from using hard disk drives (HDDs) to store persistent data. HDDs had far slower random I/O performance than sequential I/O performance. However, HDDs are increasingly being replaced by solid-state drives (SSDs), and other technologies, such as phase change memory, are around the corner. These new storage technologies require a rethinking of storage subsystem design for processing data, especially regarding high availability and fault tolerance.

In addition, the increase of unstructured data has a large impact on scalability. Data reliability relates to the limits of data density at tolerable device-level bit error rates. Traditional RAID does not provide the levels of data durability and performance for dealing with escalating volumes of data. For disks, there is the end of life, stress modes, and overheating in data centers to consider.

A shift that is underway is the transformative change of the traditional I/O subsystem. For many decades, HDDs were used to store persistent data. HDDs had far slower random I/O performance than sequential I/O performance. Data processing engines formatted their data and designed their query processing methods to work around this limitation. Countering these problems requires new methods to improve rebuild times on high-density disk drives and to reduce susceptibility to data corruption induced by disk error. The ideal system maximizes the mean time to failure and minimizes the mean time to recovery. Even more important, it provides fault tolerance for a higher number of drive failures (that is, it minimizes the potential for concurrent failure and shrinks the exposure window). Using erasure code algorithms that provide a faster protection level for bit error protection and flexibility also helps counter the problems [4].

The impact of big data on networking

In 2010, Eric Schmidt, then CEO of Google, was reported as saying: "Every two days, as much information is created as has been in existence since the dawn of civilization up until 2003." Big data invariably means that enterprises must handle larger amounts of data on existing network infrastructures. This presents a huge performance and capacity challenge, particularly for the use of Apache Hadoop as a building block for big data.

A Cisco white paper explains the Hadoop data hierarchy this way:

"The Hadoop Distributed File System (HDFS) is the first building block of a Hadoop cluster. To efficiently process massive amounts of data, it was important to move computing to where the data is, using a distributed file system rather than a central system for the data. A single large file is split into blocks, and the blocks are distributed among the nodes of the Hadoop cluster." An efficient and resilient network is a crucial part of a good Hadoop cluster. The nodes in a Hadoop cluster are interconnected through the network workload processing. A network is also crucial for writing data, reading data, signaling, and for operations of HDFS and the MapReduce infrastructure.

Therefore, the failure of a networking device affects multiple Hadoop data nodes. This means that a job might need to be restarted or more loads must be pushed to the available nodes, which makes jobs take a lot longer to finish. As a result, networks must be designed to provide redundancy with multiple paths between computing nodes and, furthermore, must be able to scale. Factors such as workload patterns, rack configurations within clusters, storage area network (SAN) access, and separation of data access networks from general networks need to be considered[5].

In addition, the network must be able to handle bursts effectively without dropping packets. For this reason, it is important to choose switches and routers with queuing and buffering strategies.4

There are a few other approaches that are commonly used to address the performance and capacity challenges:

- Use proprietary networks
- Design and segregate networks that are based on traffic (for example, separate a big data infrastructure management network from a data traffic network path)
- Apply data locality by taking an extract, *load*, and transform (ELT) approach (to process and analyze data where it is stored rather than using extract, *transform*, and load (ETL), which involves moving data twice)

Cloud services

Big data and cloud services are two initiatives that are at the top of the agenda for many organizations. There is a view that cloud computing can provide the opportunity to enhance organizations' agility, enable efficiencies, and reduce costs. In many cases, cloud

computing provides a flexible model for organizations to scale their big data capabilities, as evidenced by the inclusion of MapReduce in the offerings of Amazon Web Services. However, this needs to be done with careful planning, especially estimating the amount of data to analyze by using the big data capability in the cloud, because not all public or private cloud offerings are built to accommodate big data solutions. In some cases, cloud computing environments can pose the following performance and capacity difficulties for big data:

- The movement of large data sets into and out of the cloud can be affected by the degradation in WAN transfer speeds that occurs over long distances when you are using traditional transfer protocols. It can also be affected by the "last foot" bottleneck inside of the cloud data center, which is caused by the HTTP interfaces with the underlying object-based cloud storage.
- Where there is a need to analyze low-latency, real-time data, you might need to use different approaches. If you do not have the performance necessary to process real-time data without introducing latency, it might make the results too stale to be useful.
- In other cases, the use of cloud technologies might not be appropriate for use with big data analysis, because it is more suitable for variable and random use among users. Big data analysis requires a dedicated infrastructure that is used at full capacity for hours at a time. That is, the analysis is normally performed by batch jobs. This is not to say that the cloud cannot be used for storage; however, it requires careful design to cater to big data analysis.

IV. Velocity: *In motion* data from automated sources represents velocity. Although this data is useful in signaling- or event-driven systems, this data can be difficult to relate to other parts of a business. This section describes the performance and capacity challenges that result from the increased speed of the flow of data through organizations. Some of the challenges that result from the increased velocity of big data include access latencies, the need for rapid use of the data, the need for faster response times, and the impact on the organization's security mechanisms.

Access latencies

Access latencies create bottlenecks in systems in general, but especially with big data. The speed at which data can be accessed while in memory, network latency, and the access time for hard disks all have performance and capacity implications. For big data, data movement is usually not feasible, because it puts an unbearable load on the network. For example, moving petabytes of data across a network in a one-to-one or one-to-many fashion requires an extremely high-bandwidth, low-latency network infrastructure for efficient communication between computer nodes. Consider the performance and capacity implications of Hadoop processing on disk systems that are not designed for the speed of data movement that big data requires. Hadoop is built on top of the distributed file system called Hadoop Distributed File System (HDFS). Usually, an HDFS writes the data three times for a triple mirroring scheme (replication). This is to ensure that no data is lost, because the system was originally designed for "just a bunch of disks" (JBOD), not for enterprise-class disk systems. This ensures that the Hadoop cluster can be scaled at a low cost compared to enterprise disk systems. Enterprise disk systems, such as SANs, are typically avoided because of the high level of data traffic that needs to be sustained as the cluster scales upward. For example, given a 5-node cluster with a 50 MB/s throughput on each node, the total amount of throughput that is required is 250 MB/s. If the cluster is scaled to 10 nodes, the total throughput increases. Because there is a triple redundancy feature, there are capacity implications for disk capacity. Furthermore, the replication can create loads on the servers (processor, memory, and I/O), because these are not off-loaded data replication processes. This can cause an immense load on the network while the systems try to handle the traffic. There is another impact to consider with potential disk failures. If they occur, there can be periods of high I/O, lasting hours to days, that result when the system must rebalance itself to ensure that each node is replicated the correct number of times. This is the performance and capacity load of setting up the systems [5].

Big data uses different types of analytics, such as "adaptive predictive models, automated decision making, network analytics, analytics on data-in-motion, and new visualization." Previously, data was pre-cleaned and stored in a data mart. Now, most or even all source data is retained. Furthermore, new types of feeds, such as video or social media feeds are available (Twitter, for instance). Addressing all of these feeds has a computational cost that pushes query use throughout hardware components for server, I/O, memory, network, and SAN to higher-than-expected levels. Traditional performance and capacity techniques can be used in combination with newer big data-specific

analytics techniques to ensure optimal processing time of analytic queries in a big data system.

Data interpretation requirements

First, there is the challenge of whether there is enough I/O and network bandwidth when you are pushing the data to storage. Second, there is the challenge of check pointing this data with regard to in-memory analytics systems. These systems perform computations in memory at a certain point to ensure that the results of a particular step of a complex calculation are not lost. That data needs to be written to a local disk. While data is being written to disk for safekeeping (that is, check pointing), the node is not performing computations. As the amount of data that needs to have a checkpoint increases, the amount of time that is required for the process to complete increases. To reduce the time that check pointing takes and make more time available for computations, you need more disks, or a faster, lower-latency disk (such as solid-state drives), or flash memory. The size of big data invariably means that it takes longer to identify potential bottlenecks that might affect the performance of the system. Therefore, the system must be designed to respond quickly. Cluster architectures with fast, proprietary, low-latency networks and large memory (such as RAM) are typical approaches to ensure rapid response from big data systems. To facilitate speed and real-time results, a new approach is emerging in which subsets of the big data are held and processed within a server's fast local memory rather than having to access the same information from slow disk storage or from another server. Such *in-memory computing* enables dramatically faster computation and analysis. In the world of finance, where speed is everything, the use of in-memory computing can help increase a firm's competitive advantage. Another point to consider is that for typical cluster architectures, the entire infrastructure (for example, computing nodes) needs to be of the same specification in terms of memory, processors, I/O bandwidth, and operating system version. The cluster runs at the speed of the node with the lowest specification. It is important to remember this as the infrastructure scales to meet performance needs.

Response time

Response times for results are still critical, despite the increase of data size. To ensure speed and real-time feedback from big data, a new approach is emerging where data sets are processed entirely within a server's memory. Several vendor solutions offer in-memory analytics to address the requirements for real-time analysis results (for example, IBM InfoSphere Streams and BigInsights, SAS Visual Analytics, and SAP HANA). These apply in scenarios where there is a high velocity of data or real-time return of results. The main limitation with traditional business intelligence technologies is the time that it takes to read from and store to disk (disk I/O). By storing the data in memory (RAM), this limitation is removed. However, this needs to be balanced with the higher cost of RAM, compared to disk storage, to determine whether business requirements justify the additional expense. Typically, in-memory technologies have Java virtual machines (JVMs) running on the application tier. JVMs use garbage collection to reclaim heap space from objects that are no longer being used or are out of reach. They "pause" current JVM processing to perform this task. When the size of the JVM memory heap is large (64-bit JVM), the garbage collection behaves in unpredictable ways and can cause large pauses. That is not suitable for high-velocity data analysis or real-time return of results. However, there are tuning techniques that can minimize this problem, such as the use of compressed references [5].

As more companies move toward big data and test the limits of applications such as the Hadoop cluster, the JVM issues are becoming more apparent. Different solutions are being proposed or made available to address them. For example, a new version of Hadoop implements CRC32C by using hardware support to improve performance. People who have the skill set required for tuning large JVMs are rare. Therefore, it is advisable to engage the primary vendors who are deploying the in-memory analytic technology to configure the JVM. This helps to ensure that it runs in an optimal manner, rather than taking the risk of doing it in-house and then encountering difficulty in resolving performance and capacity problems.

V. Variety: This infinitely variable source of knowledge can lead to new business insight or nothing at all. Finding the relevant "needles" in this giant "haystack" is the ongoing challenge with this constantly evolving data source. Big data encompasses various data types: structured, unstructured, and semi-structured. All of these need to be harnessed for an organization to get the true value of big data. Integrating and analyzing various data sources enable organizations to gain great insight that is derived from the data that is available for decision making. Businesses that broaden the range of data sources also achieve better value. This involves combining a wide range of data formats that are traditionally not stored on the same platform and are constantly changing. These unique circumstances result in several performance and

capacity challenges. Most big data projects face more challenges from variety and fewer from data volumes.

Tuning

The rise of information from a variety of sources, such as social media, sensors, mobile devices, videos, and chats, results in an explosion of the volume of data. Previously, companies often discarded the data because of the cost of storing it. However, with frameworks such as Hadoop and relatively inexpensive commodity servers, it is now feasible for companies to store the data. Along with inexpensive servers and storage comes the potential complexity of deployment and management of a very large infrastructure. Similarly, tuning this infrastructure for the expected big data workloads is still a challenge for which proven, reusable patterns are still in the early stages of development.

Hadoop is a well-known open source framework for big data distributed applications. Tuning a Hadoop cluster requires understanding the Hadoop framework and all of the components of the stack, which include Hadoop MapReduce, the JVM, the network, OS, hardware, background activities, and, possibly, the BIOS, as shown in Figure 1-2.

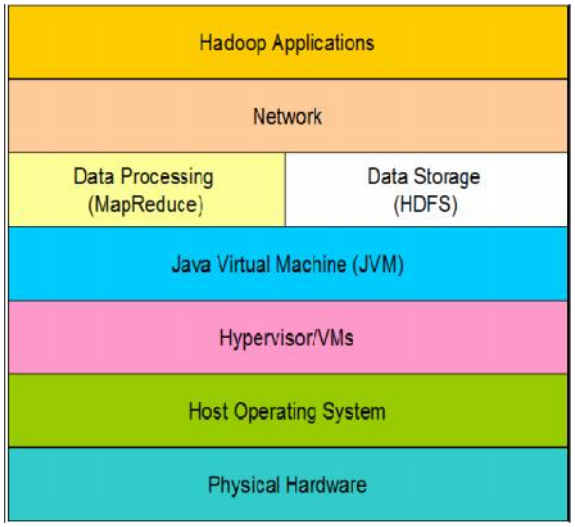

Figure 1-2 Technology stack of a Hadoop environment

Hadoop is a well-adopted, standards-based, open-source software framework built on the foundation of Google's MapReduce and Google File System papers. It's meant to leverage the power of massive parallel processing to take advantage of Big Data, generally by using lots of inexpensive commodity servers.

Accessible—Hadoop runs on large clusters of commodity machines or on cloud computing services such as Amazon's Elastic Compute Cloud (EC2).

Robust—Because it is intended to run on commodity hardware, Hadoop is architected with the assumption of frequent hardware malfunctions. It can gracefully handle most such failures.

Scalable—Hadoop scales linearly to handle larger data by adding more nodes to the cluster.

Simple—Hadoop allows users to quickly write efficient parallel code.

VI. Veracity: Veracity represents the valuable information that is trusted throughout an enterprise. It is both relevant and flexible. Its lineage can be traced to source data. Groups of items roll up into summaries as needed. Attributes such as cardinality and uniqueness are maintained to assist with enterprise integration. Every day, 2.5 quintillion bytes of data are created. This data comes from digital pictures, videos, posts to social media sites, intelligent sensors, purchase transaction records, and cell phone GPS signals, to name a few. In short, big data is messy data. Veracity deals with uncertain or imprecise data. If the data is error-prone, the information that is derived from it is unreliable, and users lose confidence in the output. Cleaning the existing data and putting processes in place to reduce the accumulation of dirty data is crucial. It is very important that companies get data quality right. To achieve improved efficiency, they need better-quality data. To address the performance and capacity challenges that arise from lack of veracity, it is important to have data quality strategies and tools as part of a big data infrastructure. The aim of the data quality strategies is to ascertain "fit for purpose." This involves evaluating the intended use of big data within the organization and determining how accurate the data needs to be to meet the business goal of the particular use case. The data quality approaches that the organization adopts need to include several strategies:

- Definition of data quality benchmarks and criteria
- Identification of key data quality attributes (such as timeliness and completeness)
- Data lifecycle management and compliance
- Metadata requirements and management
- Data element classification

Furthermore, big data governance is critical. The approach needs to be collaborative and to be clear about what needs to be 100% precise and what is "good enough," in percentage terms. The collaborative process must identify the strategic data stores and the specific data items that are critical from a governance perspective. Security also plays a key role by ensuring that fake, fraudulent, or unauthorized data is not introduced into big data. Data must be properly protected and not distributed to unauthorized recipients. The final piece of the Big Data puzzle is the low-cost hardware and software environments that have recently become so popular. These innovations have transformed technology, particularly in the last five years. Capturing and exploiting Big Data would be much more difficult and costly without the contributions of these cost-effective advances.

VII. Conclusion

Big Data and cloud computing have already begun to change the research landscape. Researchers have begun to embrace both in an effort to continue to produce cutting edge research. Big facilities like the Pathfinder projects for the Square Kilometre Array and the Large Hadron Collider produce Big Data, but Big Data can also come from sensor networks and crowd-sourced repositories. The volume of data being captured often provides a resource well beyond the original purpose, and it heralds a new way of thinking for many researchers. New skills are needed and this is where communities and the associated platforms are critical to success. Over the next few years, cloud computing services will prove key to the development of research data communities. Virtual Laboratories from numerous disciplines will exist, with dozens of communities forming around these resources. Communities will develop platforms that will be able to cross disciplines, and make the use of Big Data a natural extension of research activity. The next few years will provide an opportunity to observe and understand how cloud computing and Big Data changes how researchers work. The combination of community and research platforms will enable far greater collaboration and in turn, better research outcomes. The reuse of platforms and Big Data datasets will be made possible by the ability of cloud computing proliferate customized VMs throughout a research community.

This future will not be without challenges of its own. It is imperative the due diligence be paid to issues such as security and skills development, as well as improving the stability of the underpinning technology. As more research finds its way into the cloud, frailties of the system will be exposed, and will need to be addressed decisively. While these risks exist and need to be attended, the potential benefits are enormous. The simple fact that Big Data offers such a rich opportunity for research, and is reusable in ways beyond the original purpose, justifies the effort to capture and retain this scale of information. Research communities that form in precincts, around disciplines or even around Big Data, can create collaborative platforms that are shareable and repeatable. The future of cloud computing is all but assured, growing with the same inexorability as the Internet itself has over the last decade. Provided we understand this growth and the opportunities it presents, it can only serve to enrich research as we know it.

References:

[1] Lane, J. (2010) "Let's make science metrics more scientific", Nature 464, 488–489.

[2] Lane, J. & Black, D. (2012) "Overview of the Science of Science Policy Symposium", J. Pol. Anal. Manage. 31, 598–600.

[3] Largent, M. A. & Lane, J. I. (2012) "Star Metrics and the Science of Science Policy", Review of Policy Research 29, 431–438.

[4] Brenda L. Dietrich, Emily C. Plachy, Maureen F. Norton (2014) "Analytics Across the Enterprise: How IBM Realizes Business Value from Big Data and Analytics"

[5] Martin Oberhofer, Eberhard Hechler Ivan Milman (2014) "Beyond Big Data: Using Social MDM to Drive Deep Customer Insight, IBM Publishers.