Next Generation Infrastructure for Big Data - A Challenge

Praveen Goyal Santosh[1], Kumar Bhandare[2]
[1,2]Department of Computer Science & Engineering
AISECT University, Bhopal (M.P.) India.

ABSTRACT
The Internet has engender an explosion in data growth in the form of Data Sets, called Big Data that are so large they are difficult to store, manage, and canvass using traditional RDBMS which are tuned for Online Transaction Processing (OLTP) only. Not only is this new data heavily unstructured, voluminous and stream rapidly and difficult to tackle but even more importantly. The infrastructure cost of Hardware and Software required to crackle it using traditional RDBMS to drive any analytics or business intelligence online (OLAP) from it is prohibitive. To capitalize on the Big Data trend, a new maker of Big Data technologies (such as Hadoop, Google App Engine, Microsoft Azure and others) many combines have emerged which are leveraging new parallelized processing. Commodity hardware, open source software, and tools to capture and analyse these new data sets and provide a price/performance that is 10 times better than existing Data Warehouse/Business Intelligence system. This paper presents an overview of the cloud computing scenario today. It provides the advantages and disadvantages of cloud, different examples of the cloud services, different enterprises in the field of cloud computing are being mentioned in the paper.

Keywords: Cloud computing, Big Data, OLTP, Hadoop, Map Reduce, RTAP.

I INTRODUCTION

Big data is [1] [2] [3] certainly one of the biggest buzz phrases in IT today. Combined with virtualization and cloud computing, big data is a technological capability that will force data centres to significantly transform and evolve within the next five years. Similar to virtualization, big data infrastructure is unique and can create an architectural upheaval in the way systems, storage, and software infrastructure are connected and managed. Unlike previous business analytics solutions, the real-time capability of new big data solutions can provide mission critical business intelligence that can change the shape and speed of enterprise decision making forever. Hence, the way in which IT infrastructure is connected and distributed warrants a fresh and critical analysis. Numerous technological innovations are driving the dramatic increase in data and data gathering. This is why big data has become a recent area of strategic investment for IT organizations. For example, the rise of mobile users has increased enterprise aggregation of user statistics geographic, sensor, capability, data that can, if properly synthesized and analyzed, provide extremely powerful business intelligence. In addition, the increased use of sensors for everything from traffic patterns, purchasing behaviours, and real-time inventory management is a primary example of the massive increase in data. Much of this data is gathered in real time and provides a unique and powerful opportunity if it can be analyzed and acted upon quickly. Machine-to-machine interchange is another often unrecognized source of big data. The rise of security information management (SIM) and the Security Information and Event Management (SIEM) industry is at the heart of gathering, analyzing, and proactively responding to event data from active machine log files. At the heart of this trend is the ability to capture, analyze, and respond to data and data trends in real time. Although it may be clear that new technologies and new forms of personal communication are driving the big data trend, consider that the global Internet population grew by 6.5% from 2010 to 2011 and now represents over two billion people. This may seem large, but it suggests that the vast majority of the world's population has yet to connect. While it may be that we never reach 100% of the world's population online (due to resource constraints, cost of goods, and limits to material flexibility), increasingly those that are online are more connected than ever. Just a few years ago, it was realistic to think that many had a desktop (perhaps at work) and maybe a laptop at their disposal. However, today we also may have a connected smartphones and even a tablet computing device. So, of today's two billion connected people, many are connected for the vast majority of their waking hours, every second generating data:

In 2011 alone, mankind created over 1.2 trillion GB of data.

(a) Data volumes are expected to grow 50 times by 2020.

(b) Google receives over 2,000,000 search queries every minute.

(c) 72 hours of video are added to YouTube every minute.

(d) There are 217 new mobile Internet users every minute.

(e) T witter users send over 100,000 tweets every minute (that's over 140 million per day)

(f) Companies, brands, and organizations receive 34,000 "likes" on social networks every minute.

International Data Corporation (IDC) predicts that the market for big data technology and services will reach $16.9 billion by 2015 with 40% growth over the prediction horizon. Not only will this technology and services spend directly impact big data technology providers for related SQL database technologies, Hadoop or Map Reduce file systems, and related software and analytics software solutions, but it also will impact new server, storage, and networking infrastructure that is specifically designed to leverage and optimize the new analytical solutions.

II BIG DATA

Big data refers [4] to the collection and subsequent analysis of any significantly large collection of data that may contain hidden insights or intelligence (user data, sensor data, machine data). When analyzed properly, big data can deliver new business insights, open new markets, and create competitive advantages. Compared to the structured data in business applications, big data (according to IBM) consists of the following three major attributes:

(a) **Variety**—Extends beyond structured data and includes semi-structure or unstructured data of all varieties, such as text, audio, video, click streams, log files, and more.

(b) **Volume**—Comes in one size: large. Organizations are awash with data, easily amassing hundreds of terabytes and petabytes of information.

(c) **Velocity**—Sometimes must be analyzed in real time as it is streamed to an organization to maximize the data's business value.

III USE CASES OF BIG DATA

There are many examples [5] [6] of big data use cases in virtually every industry imaginable. Some businesses have been more receptive of the technologies and faster to integrate big data analytics into their everyday business than others. It is evident that organizations embracing this technology not only will see significant first-mover advantages but will be considerably more agile and cutting edge in the solutions and adaptability of their offerings. Use case examples of big data solutions include:

(a) Financial services providers are adopting big data analytics infrastructure to improve their analysis of customers to help determine eligibility for equity capital, insurance, mortgage, or credit.

(b) Airlines and trucking companies are using big data to track fuel consumption and traffic patterns across their fleets in real time to improve efficiencies and save costs.

(c) Healthcare providers are managing and sharing patient electronic health records from multiple sources imagery, treatments, and demographics and across multiple practitioners. In addition, pharmaceutical companies and regulatory agencies are creating big data solutions to track drug efficacy and provide more efficient and shorter drug development processes.

(d) Telecommunications and utilities are using big data solutions to analyze user behaviours and demand patterns for a better and more efficient power grid. They are also storing and analyzing environmental sensor data to provide insight into infrastructure weaknesses and provide better risk management intelligence.

(e) Media and entertainment companies are utilizing big data infrastructure to assist with decision making around customer lifecycle retention and predictive analysis of their user base, and to provide more focused marketing and customer analytics.

There are productized use cases and concrete examples of big data for every industry and company size. Therefore, whether or not your business currently is using a big data solution, your competitors likely are. The real question is how can you better optimize your environment to create a faster, more efficient solution that gives you a competitive edge? Why is this so pressing? According to research by McKinsey Global Institute (MGI) and reported by McKinsey's Business Technology, analyzing large data sets will become and has already become for a large number of businesses a key planning tool. With the caveat that the correct policies and enablers must be considered and implemented, big data will become a critical tool in developing plans for:

(i) Competitive planning and research
(ii) Future productivity and product growth
(iii) Product and services innovation
(iv) Customer satisfaction (or as delineated in the study, "Consumer Surplus")

IV EXAMPLES OF BIG DATA

This section provides with some of the real life examples of cloud computing services.

(a) Social Networking: The fig. 1 most famous example of cloud computing [7] are the social networking websites like Facebook, Twitter, My space, LinkedIn and many others which doesn't seems to be a part of cloud computing at first glance. In social networking user finds people he already knows or like to know and shares information with them. As the user shares information with people related to him, he ultimately shares the information with peoples who are running the service. Social networking can also be used by business for its promotion among its customers.

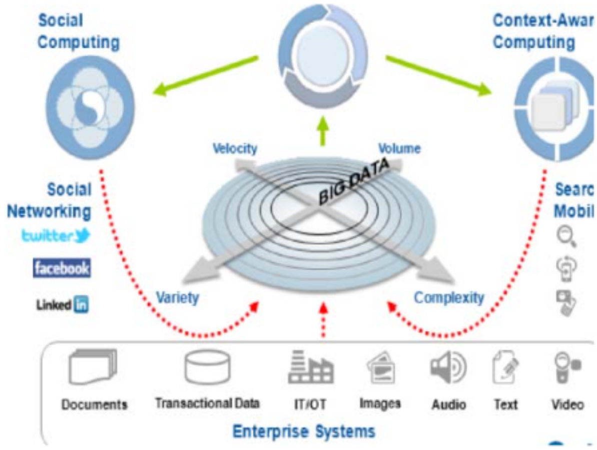


Fig. 1 Cloud Computing System

(b) Email: Some of the biggest cloud computing services are Web-based e-mail. Microsoft's Hotmail or Windows Live Mail is examples of cloud based email service. Using a cloud computing e-mail solution allows the mechanics of hosting an e-mail server and it is maintained by the people running the service. This means that we can access our e-mail from anywhere in the world.

(c) Document/Spreadsheet/Other hosting Services:

Google Docs allow users to keep and edit their documents online. Making use of Google Docs allows access and sharing of the documents from anywhere. The same document can be worked by multiple people simultaneously. Google's Picasa and Yahoo's Flickr provides hosting for the photographs that individual wants to share with other people. Comments can be placed on the photographs in the similar manner as on Facebook. But for the photographic enthusiast's some perks are provided by these photo hosting services. YouTube , a video sharing site has brought a great revolution in the field of entertainment but Although, it is not only in entertainment. DailyMotion, MetaCafe and Vimeo are some of the names in this field. In this service the users are permitted to upload their videos and the service provider take care of putting it online in a form that is easily viewed by the users.

(d) Backup Services: Services like JungleDisk, Carbonite, and Mozy allow public to automatically back up all their data to servers spread around the country or world for a surprisingly low price.

V BIG DATA CHALLENGES

In any big data project, [see fig. 1] storage capacity to accommodate the datasets must increase. However, simply adding raw capacity, without taking other infrastructure issues into account, can lead to problems and inefficient use of resources. [8] [9] The field of the life sciences provides examples of the potential impact of big data on an IT infrastructure. They include:

(a) Interdependencies of Infrastructure elements: Life sciences organizations are increasingly turning to virtualization to reduce operating costs, consolidate servers, and simplify the deployment and management of applications.

Server fig. 5 cluster nodes based on multi-core processors are now commonly used in conjunction with virtualization software to enable dozens or more applications to run as virtual machines on each physical server. Open source software, such as Hadoop and NoSQL, gives companies a way to leverage these clusters to run big data analytics.

As this architecture becomes more widely used, organizations must address several other infrastructure issues because new performance issues crop up. A single server accessing a single storage device generates predictable workloads. But matters become much more complex in a cluster running dozens of virtualized applications. The key issue becomes how to best integrate servers, storage, and network elements. The numerous applications running on the cluster all need simultaneous access to the data on storage devices. That means the storage solution will have to accommodate multiple concurrent workloads without degradation. Additionally, the network switches and adapter cards must offer the throughput and IO to sustain the required performance levels.

This places new demands on both the storage solution and the network. In particular, big data analytics requires that storage be flexible and capable of being dynamically grown to meet varying capacity and performance requirements. Because virtualized applications can be quickly and easily set up and torn down, the associated storage must support easy, dynamic provisioning. Additionally, provisioning and addition of new storage capacity must not involve taking systems offline.

From the networking perspective, server virtualization and big data analysis can change the dynamics of traffic flow within the data center network. Network links to the servers can become congested, impacting network performance and throughput. A common solution is simply to add more links. But this increases the number of switch ports needed and adds to the administrative burden on the IT staff. What is needed is a network that offers high performance scalability.

(b) Unpredictable workloads: Another infrastructure issue to consider relates to the change in the way data is accessed in a big data workflow. Efforts to derive decision-making information from big data sources typically use a number of analysis tools,

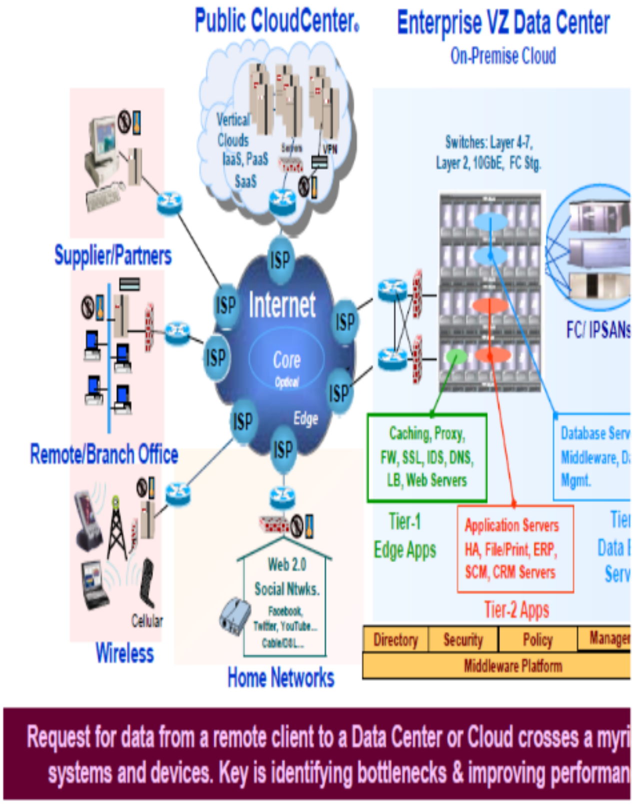


Fig. 2 System & Devices in Cloud Computing System

applied at different stages of a computational workflow. For example, in the life sciences, newer lab equipment, such as next-generation sequencers, produces a much richer set of data. This makes the raw output from today's lab equipment of interest to a more diverse group of researchers.

Each group of researchers subjects the data to a wide variety of analysis tools, all with different IO and throughput requirements. Depending on the type of analysis being performed, the workflows and applications will likely have diverse performance requirements. All of this makes big data workflows highly unpredictable.

What is required is an infrastructure that can support diverse workflows, offering high sustained throughputs. Specifically, the infrastructure must be able to handle large sorts, which are quite common in big data workflows. With large sorts, the files are typically larger than system memory and therefore cannot be retained in local cache. As a result, large sorting workloads require a file system and storage solution that can deliver high throughput and IO. The storage system must also be able to provide low latency access to file system metadata.

(c) **Data management:** Life sciences research has become more multi-disciplinary and more collaborative. This complicates data management and makes computational workflows more complex. As noted above, the richness of data from newer lab equipment makes it of interest to more types of researchers. Some groups might need instant analysis of the data in early stage research to determine which new drug candidates to move along and evaluate further. Other groups might need to re-examine the original data months or years later when a candidate moves into clinical trials or is being studied for potential adverse effects.

VI INFRASTRUTURE NEEDED

Operators could use their data access to [10] [11] [12] enhance internal processes, such as knowing customers' value, what type of content they prefer, and the type of device they carry. Similarly, decisions on the rolling out of networks and sales channels should consider the location and demographic data of potential customers. Customer care departments should use data to predict when a customer is at risk of churn and act on it. Customer data will also allow operators to reduce their losses from customer or dealer commission fraud.

Mobile operators are in a sweetspot for data generation and analysis. Data is generated every time a customer makes a call, navigates the Web, interacts with social media, or buys a product using their phone. Simply by having the phone connected to the operator's network, data is being generated, such as location, speed of movement, and even biometric data.

The diversity of data availability allows mobile operators to achieve a depth of customer profiling beyond that of other industries. Operators have the potential to know customers' whereabouts, their network of contacts, content preferences, wealth, and product preferences. It is entirely foreseeable that, in the not-so-distant future, mobile operators will also be able to generate revenues from the packaging and selling of this data. Traditional revenues such as voice and SMS are under pressure from Web-based services and over-the-top (OTT) providers, and with mobile broadband now reaching its peak of profitability in developed economies, telecoms will need to pursue new ventures. But in order to move on to this stage, operators need to set up the basic processes, frameworks, and technical infrastructures needed to capture and manipulate Big Data.

Contact centre text mining and telecom bandwidth throttling. Monitoring real-time contact centre and social media for surges in keyword frequency could be used as a lead indicator to infrastructure bottlenecks and be used as an input for throttling network traffic.

Co-location analysis from cell phone towers. Can we ask "needle in a haystack" queries to isolate collocation events from the massive call detail record (CDR) data ocean using Hadoop and columnar architectures?

Multi device event stream analysis correlating firewall, IDS, and switch activities in real time. How can we get a 360-degree view of an intrusion from the patterns of event data across devices?

VII HADOOP

This section provides some of the managing Big Data.

(a) Hadoop MapReduce and Hadoop Distributed File System (HDFS)

Hadoop is a framework that provides open source libraries for distributed computing using MapReduce software and its own distributed file system, simply known as the Hadoop Distributed File System (HDFS). It is designed to scale out from a few computing nodes to thousands of machines, each offering local computation and storage. One of Hadoop's main value propositions is that it is designed to run on commodity hardware such as commodity servers or personal computers, and has high tolerance for hardware failure. In Hadoop, hardware failure is treated as a rule rather than an exception. [6] [10]

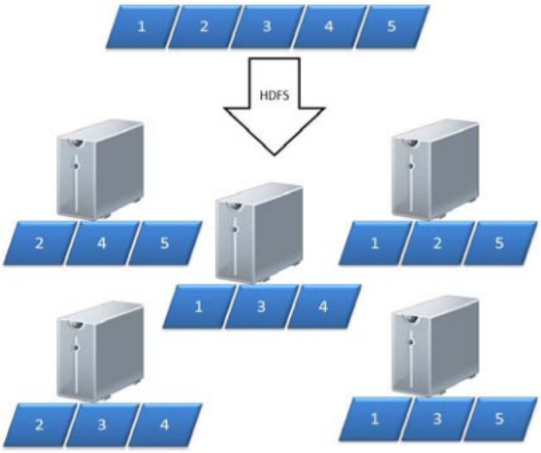


Fig. 4 Illustration of distributed file storage using HDFS

(b) HDFS

The HDFS is a fault-tolerant storage system that can store huge amounts of information, scale up incrementally and survive storage failure without losing data. Hadoop clusters are built with inexpensive computers. If one computer (or node) fails, the cluster can continue to operate without losing data or interrupting work by simply re-distributing the work to the remaining machines in the cluster. HDFS manages storage on the cluster by breaking files into small blocks and storing duplicated copies of them across the pool of nodes. The figure 2 illustrates how a data set is typically stored across a cluster of five nodes. In this example, the entire data set will still be available even if two of the servers have failed. Compared to other redundancy techniques, including the strategies employed by Redundant Array of Independent Disks (RAID) machines, HDFS offers two key advantages. Firstly, HDFS requires no special hardware as it can be built from common hardware. Secondly, it enables an efficient technique of data processing in the form of Map Reduce.

(c) Map Reduce

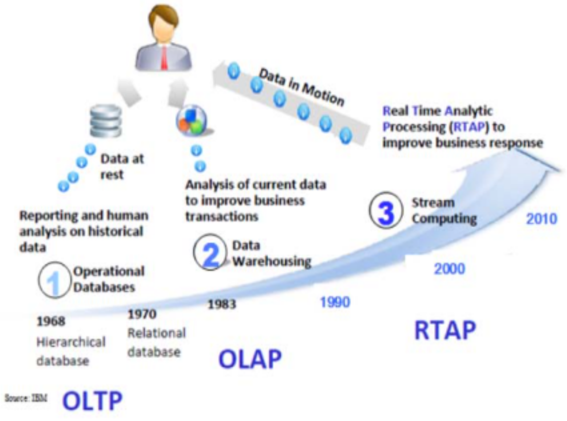
Most [11] enterprise data management tools (database management systems) are designed to make simple queries run quickly. Typically, the data is indexed so that only small portions of the data need to be examined in order to answer a query. This solution, however, does not work for data that cannot be indexed, namely in semi-structured form (text files) or unstructured form (media files). To answer a query in this case, all the data has to be examined. Hadoop uses the MapReduce technique to carry out this exhaustive analysis quickly.

Map Reduce is a data processing algorithm that uses a parallel programming implementation. In simple terms, MapReduce is a programming paradigm that involves distributing a task across multiple nodes running a "map" function. The map function takes the problem, splits it into sub-parts and sends them to different machines so that all the sub-parts can run concurrently. The results from the parallel map functions are collected and distributed to a set of servers running "reduce" functions, which then takes the results from the sub-parts and re-combines them to get the single answer.

VIII CONCLUSION

Big Data management is a recent technology which is being used at large level by the infrastructure and services industries focusing to capture potential opportunities. This paper provides the overview of the technology. how it is related with cloud is also discussed.

This paper explores some of the future needs of Big Data which may lead to the improvement and advancement of the technology in near future required large IT infrastructure.


Data in Motion
Real Time Analytic Processing (RTAP) to improve business response
Data at rest
Analysis of current data to improve business transactions
Reporting and human analysis on historical data
3 Stream Computing
2010
2 Data Warehousing
2000
1 Operational Databases
1983
1990
1968 Hierarchical database
1970 Relational database
RTAP
OLAP
OLTP

Source: IBM

Figure-5

REFERENCES

[1] Anil Vasudeva, President & Chief Analyst, IMEX Research "NextGen Infrastructure for Big Data", 2011.

[2] J. Dean and S. Ghemawat, "Mapreduce: Simplified Data Processing on Large Clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107-113, 2008.

[3] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica and M. Zaharia, "A View of Cloud Computing," Commun. ACM, vol.53, no. 4, pp. 50-58, 2010.

[4] L. Wang, J. Zhan, W. Shi and Y. Liang, "In Cloud, Can Scientific Communities Benefit from the Economies of Scale?," IEEE Trans. Parallel Distrib. Syst., vol. 23, no. 2, pp.296-303, 2012.

[5] S. Chaudhuri, "What Next?: A Half-Dozen Data Management Research Goals for Big Data and the Cloud," in Proc. 31st Symp. Principles of Database Systems (PODS'12), pp. 1-4, 2012.

[6] Jeffrey Dean, Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. http://static.usenix.org/event/osdi04/tech/dean.htmal [Accessed 30th Dec 2013].

[7] White Paper - Introduction to Big Data: Infrastructure and Networking Considerations Leveraging Hadoop-Based Big Data Architectures for a Scalable, High-Performance Analytics Platform.2012.

[8] Accunet "Building a Scalable Big Data Infrastructure for Dynamic Workflows" http://www.accunetsolutions.com [Accessed 30th Dec 2013]

[9] Prodan R, Sperk M, "Scientific computing with Google App Engine", Future generation computer systems, Elsevier, 2013

[10] Shvachko K, Hairong K, Radia S, Chansler R, "The Hadoop distributed file system", Mass Storage Systems and Technologies (MSST), pp. 1-10, IEEE symposium, 2010.

[11] Bhandarkar M, "MapReduce programming with Apache hadoop", Parallel & Distributed Processing (IPDPS), page 1, ISSN. 1530-2075, Atlanta, IEEE symposium, 2010.

[12] Bibi S, Katsaros D, Bozanis P,"Business Application Acqusition:On-premise or SaaS based solutions, Published in software IEEE, Vol.29, Issue3, pp.86-93, IEEE computer society 2012.