Sonali A. Tinkhede *et al*, International Journal of Computer Science and Mobile Applications,
Vol.3 Issue. 1, January- 2015, pg. 33-38 **ISSN: 2321-8363**

Big Data- The Vast Growing Technology with its Challenges and Solutions

Sonali A. Tinkhede
Student of Master of Engineering in (CS & IT)
HVPM's college of Engineering and Technology
Amravati, India
sonali.tinkhede@gmail.com

Dr. S. P. Deshpande
Associate Professor and Co-ordinator (CSE)
HVPM's College of Engineering and Technology
Amravati, India
Shrinivasdeshpande68@gmail.com

Abstract-

Big Data is the extremely large datasets that their sizes are beyond the ability of capturing, managing, and processing by most software tools and people. Big data the word, or catch-phrase, used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques. In most enterprise scenarios the data is too big or it moves too fast or it exceeds current processing capacity. The term big data is also used by vendors, may refer to the technology which includes tools and processes that an organization requires to handle the large amounts of data and storage facilities. The term big data is believed to have originated with Web search companies who had to query very large distributed aggregations of loosely-structured data. But data is expanding faster than ever and now facing big data challenges too. In this paper we are trying to address some of these challenges and their possible solutions.

Keywords: Big Data, structured and unstructured data, challenges, Lightweight Evaluation and Architecture Prototyping (LEAP), NoSQL, Fully Homomorphic Encryption (FHE).

I. Introduction

Big Data is termed extremely large datasets that their sizes are beyond the ability of capturing, managing, and processing by most software tools and people [1]. For example, Search engines, social networking, and online advertising, as well as education, health care, and medicine, etc. As the datasets are so long it have to face the challenges including capture, curation, storage, [2] search, sharing, transfer, analysis [3] and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases in medical sector, link legal citations, combat crime, and determine real-time roadway traffic conditions or other real-time applications [4] [5] [6]."

Big Data is emerging from the realms of science projects at Web companies to help companies like telecommunication giants understand exactly which customers are unhappy with service and what processes caused the dissatisfaction, and predict which customers are going to change carriers. To obtain this information, billions of loosely-structured bytes of data in different locations needs to be processed until the needle in the haystack is found. The real business impact is that big data technologies can do this in weeks or months, four-or-more-times faster than traditional data warehousing approaches.

As in the research of 2012, limits on the size of data sets that are feasible to process in a reasonable amount of time were on the order of exabytes of data [7]. The limitations also affect Internet search, finance and business informatics [8]. Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs,

cameras, microphones, radio-frequency identification readers, and wireless sensor networks [9] [10]. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s [11], as of 2012, every day 2.5 exabytes of data were created [12]. The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization [13].

Big data is difficult to manage with most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers" [14] 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. "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." [15].

Big Data is essential to run an organizations. But data is expanding faster than ever and now we are facing big data challenges too: not just large volumes, but data that changes rapidly, comes in more varieties and from more sources. Structured, internal data is increasingly being supplemented by unstructured data like audio, video and sensors, and data from external sources like the internet, social media and third parties [16]. Until recently, it wasn't possible to interrogate unstructured data, like email and conversations, in conjunction with structured data, such as spreadsheets and databases. But now it's becoming easier to combine and analyze vast, rapidly changing datasets, using new technologies and techniques that can mine it in more meaningful ways than before.

"Big Data" originally emerged as a term to describe datasets whose size is beyond the ability of traditional databases to capture, store, manage and analyze. However, the scope of the term has significantly expanded over the years. Big Data not only refers to the data itself but also a set of technologies that capture, store, manage and analyze large and variable collections of data to solve complex problems. The aim at proliferation of real time data from sources such as mobile devices, web, social media, sensors, log files and transactional applications, Big Data has found a host of vertical market applications, ranging from fraud detection to R&D.

The research report presents an in-depth assessment of the Big Data ecosystem including key market drivers, challenges, investment potential, vertical market opportunities and use cases, future roadmap, value chain, case studies on Big Data analytics, vendor market share and strategies. Despite challenges relating to privacy concerns and organizational resistance, Big Data investments continue to gain momentum throughout the globe. SNS Research estimates that Big Data investments will account for nearly $30 Billion in 2014 alone. These investments are further expected to grow at a CAGR of 17% over the next 6 years. Hence it is necessary to address the issues and challenges that Big Data and its using companies has to face as early and effectively as possible.

II. Big-data Characteristics and some facts towards Need

A. Big data has the following characteristics:

- Very large, distributed aggregations of loosely structured data – often incomplete and inaccessible [17]:
  - Petabytes/exabytes of data,
  - Millions/billions of people,
  - Billions/trillions of records,
  - Loosely-structured and often distributed data,
  - Flat schemas with few complex interrelationships,
  - Often involving time-stamped events,
  - Often made up of incomplete data,
  - Often including connections between data elements that must be probabilistically inferred,
- Applications that involved Big-data can be:
  - Transactional (e.g., Facebook, PhotoBox), or,
  - Analytic (e.g., ClickFox, Merced Applications).

B. Some Unbelievable Facts:

We surely see a lot of hype surrounding big data**.** 'Big Data' is coming as a phenomenon that is changing the world as we know it. How much processing demand of data is growing in a day and in future are mention in some following facts [18].

- Over 90% of all the data in the world was created in the past 2 years.
- The total amount of data being captured and stored by industry doubles every 1.2 years.
- Every minute we send 204 million emails, generate 1,8 million Facebook likes, send 278 thousand Tweets, and up-load 200 thousand photos to Facebook.
- Google alone processes on average over 40 thousand search queries per second, making it over 3.5 billion in a single day.
- Around 100 hours of video are uploaded to YouTube every minute and it would take you around 15 years to watch every video uploaded by users in one day.
- Facebook users share 30 billion pieces of content between them every day.
- If you burned all of the data created in just one day onto DVDs, you could stack them on top of each other and reach the moon – twice.
- 1.9 million IT jobs will be created in the US by 2015 to carry out big data projects. Each of those will be supported by 3 new jobs created outside of IT – meaning a total of 6 million new jobs thanks to big data.
- It is expected that by 2020 the amount of digital information in existence will have grown from 3.2 zettabytes today to 40 zettabytes.

III. Some Big Data Challenges and their Solution

New data sources, ranging from diverse business transactions to social media, high-resolution sensors, Healthcare and the Internet of Things, are creating a digital tidal wave of big data that must be captured, processed, integrated, analyzed, and archived**.** But for that it is necessary to overcome some pretty big obstacles. These big challenges are Scalability, Availability, Manageability, Performance, Cost and security are just on the top to concern. Some of these challenges and their possible solution are address in this section as follows:

A. Challenge of Scalability:

Big data systems storing and analyzing petabytes of data are becoming increasingly common in many application areas. These systems represent major, long-term investments requiring massive scale software and system deployments. With analysts estimating data storage growth at 30 to 60 percent per year, organizations must develop a long-term strategy to address the challenge of managing projects that analyze exponentially growing data sets with predictable, linear costs. With big data you want to be able to scale very rapidly and elastically. Whenever and wherever you want. Across multiple data centers and the cloud if need be.

Big data systems are inherently distributed systems. Hence, software architects must explicitly deal with issues of partial failures, unpredictable communications latencies, concurrency, consistency, and replication in the system design. These issues are exacerbated as systems scales to utilize thousands of processing nodes and disks, geographically distributed across the data centers. For example, the probability of failure of a hardware component increases with scale. To mitigate the risks associated with scale and technology, a systematic, iterative approach is needed to ensure that initial design models and database selections can support the long-term scalability and analysis needs of a big data application.

The solution-approach developed at the SEI is based on principles drawn from proven architecture and technology analysis and evaluation techniques to help the Department of Defense (DoD) and other enterprises develop and evolve systems to manage big data having following characteristics.

- Replicate data across clusters and data centers to ensure availability in the case of disk failure or network partitions. Replicas must be kept consistent using either master-slave or multi-master protocols. The latter requires mechanisms to handle inconsistencies due to concurrent writes, typically based on Lamport clocks [21].

- Design components to be stateless and replicated and to tolerate failures by dependent services, for example, by using the Circuit Breaker pattern and returning cached or default results whenever failures are detected. This pattern ensures that failures do not rapidly propagate across components and allow applications an opportunity for recovery.
- At the SEI, have developed a lightweight risk reduction approach that we have initially named Lightweight Evaluation and Architecture Prototyping (for Big Data), or LEAP(4BD). The approach is based on principles drawn from proven architecture and technology analysis and evaluation techniques such as the Quality Attribute Workshop and the Architecture Tradeoff Analysis Method. LEAP(4BD) leverages our extensive experience with architecture-based design and assessment and customizes these methods with deep knowledge and experience of the architectural and database technology issues most pertinent to big data systems

You can scale up to the heavens as data is increasing like endless but not with our traditional relational database systems. Most NoSQL solutions have their own scaling limitations, the data stores running on horizontally-scaled commodity hardware [20]. These NoSQL databases achieve high scalability and performance using simpler data models, clusters of low-cost hardware, and mechanisms for relaxed data consistency that enhance performance and availability.

B. Challenges related to Cost

Big data applications employ many thousands of compute-and-storage resources. Regardless of whether these resources are capital purchases or resources hosted by a commercial cloud provider, they remain a major cost and hence it is necessarily be reduce. If we are trying to meet even one of the challenges with RDBMS or even most NoSQL solutions can cost a pretty penny. Doing big data the right way it should not be costly to manage our budget.

Straightforward resource reduction approaches such as data compression are common ways to reduce storage costs [21]. Elasticity is another way that big data applications optimize resource usage, by dynamically deploying new servers to handle increases in load and releasing them as load decreases. Decreasing resource utilization. For example, Facebook built HipHop, a PHP-to-C++ transformation engine that reduced its CPU load for serving web pages by 50 percent. At the scale of Facebook's deployment, this represents a very significant resource reduction and cost savings.

C. Challenge of speed and continuous availability

Traditional WAN-based transport methods cannot move terabytes of data at the speed dictated by businesses; they use a fraction of available bandwidth and achieve transfer speeds that are unsuitable for such volumes, introducing unacceptable delays in moving data into, out of, and within the cloud. When you rely on big data to feed your essential, revenue-generating 24/7 business applications, even high availability needs high speed processing and continuous availability [23].

One Solution is Aspera On Demand delivers scale-out transfer capacity Built on our patented fasp™ transfer technology. It enables efficient, large-scale workflows, with enterprise-grade security, a variety of client options like web, mobile, embedded, and applications for ingest, sharing, collaboration and exchange of big data, available on demand as a subscription service. By Delivering on the Promise Aspera has solved the big data challenge [22].

D. Challenges of Security

The biggest challenge for big data from a security point of view is the protection of user's privacy. Big data frequently contains huge amounts of personal identifiable information, personal account information and therefore privacy of users is a huge concern. Because of this a big data security breach will potentially affect a much larger number of people. This in itself can be a security challenge as removing unique identifiers might not be enough to guarantee that the data will remain anonymous. When storing the data organizations will face the problem of encryption [24]. Data cannot be sent encrypted by the users if the cloud needs to perform operations over the data.

While using big data a significant challenge is how to establish ownership of information. If the data is stored in the cloud a trust boundary should be establish between the data owners and the data storage owners. An additional problem is that software commonly used to store big data, such as Hadoop, doesn't always come with user authentication by default. This makes the problem of access control worse, as a default installation would leave the information open to unauthenticated users.

There are a number of general security recommendations that can be applied to big data:

- If you are storing your big data in the cloud, you must ensure that your provider has adequate protection mechanisms in place. Make sure that the provider carries out periodic security audits and agree penalties in case that adequate security standards are not met.
- Create an adequate access control policy that allow access to authorized users only.
- Protect both the raw data and the outcome from analytics should be adequately protected. Encryption should be used accordingly to ensure no sensitive data is leaked.
- Protect data in transit should be adequately protected to ensure its confidentiality and integrity.
- Use real-time security monitoring to access the data. Threat intelligence should be used to prevent unauthorised access to the data.
- When producing information for big data, it should be adequately anonymised, removing any unique identifier for a user. A solution to the problem of encryption is to use "Fully Homomorphic Encryption" (FHE**),** which allows data stored in the cloud to perform operations over the encrypted data so that new encrypted data will be created.
- Adequate access control mechanisms will be key in protecting the data. The approach is to protect the information using encryption that only allows decryption if the entity trying to access the information is authorised by an access control policy.
- Big data solutions often rely on traditional firewalls or implementations at the application layer to restrict access to the information.

E. Other challenges

In an online world where nanosecond delays can cost you sales so in this, challenges like Performance, availability, manageability, etc have to face at any time to any organization. Big data must move at extremely high velocities no matter how much you scale or what workloads your database must perform. The data handling hoops of RDBMS and most NoSQL solutions put a serious drag on some performance point of view. When you rely on big data to feed your essential, revenue-generating 24/7 business applications, even high availability is not really high enough. Your data can never go down to be available. Manageability is Staying ahead of big data using RDBMS technology is a costly, time-consuming and often futile endeavor. And most NoSQL solutions are plagued by operational complexity and arcane configurations. So, it's seems hard to manage. Thus if we want to use big data to make fruitful for any applications at any time it must be able to meet its challenges in Performance, availability, manageability in reduce cost.

If one believe to be able to do anything you want with big data. Some directly created DataStax Enterprise solution are available in the market. It's the big data platform that answers all of your big data challenges. It combines the massive scalability, high performance, continuous availability and tunable consistency of real-time with analytics powered by Apache Hadoop [25] and enterprise search with Apache Solr, creating a smartly integrated secure platform for the most demanding big data enterprise workloads.

IV. Conclusion

Big data is a relative term describing a situation where the volume, velocity and variety of data exceed an organization's storage or compute capacity for accurate and timely decision making. Big data is not just about helping an organization be more successful – to market more effectively or improve business operations. It reaches to far more socially significant issues as well. Big Data is do much large that is not handle by any organization normally using traditional DBMS approaches. If you've ever tried to wrangle big data capabilities out of your traditional relational database solution, you already have a pretty good idea of the hurdles you face. For the DoD, Healthcare industries, the challenges of big data are daunting. Military operations, intelligence analysis, logistics, and health care all represent big data applications with data growing at exponential rates and

the need for scalable software solutions to sustain future operations. As we seen in this paper, Big Data is ever expanding and it's the growing needs of the today's world. The challenges we see in this paper should be address properly as fast as possible. Organization should be interested in working to ensure appropriate technology selection and software architecture design for their big data systems. So, it will work to increase the benefits to grow the business opportunities towards success and profit for all.

References

[1] D. Krishna. "Big Data". http://www.irmac.ca/1011/Big%20Data%20v2.1.pdf.
[2] Kusnetzky, Dan. "What is "Big Data?"". ZDNet.
[3] Vance, Ashley (22 April 2010). "Start-Up Goes After Big Data With Hadoop Helper". *New York Times Blog*.
[4] "Data, data everywhere". *The Economist*. 25 February 2010. Retrieved 9 December 2012.
[5] "E-Discovery Special Report: The Rising Tide of Nonlinear Review". Hudson Global. Retrieved 1 July 2012. by Cat Casey and Alejandra Perez
[6] "What Technology-Assisted Electronic Discovery Teaches Us About The Role Of Humans In Technology — Re-Humanizing Technology-Assisted Review". Forbes. Retrieved 1 July 2012.
[7] Francis, Matthew (2012-04-02). "Future telescope array drives development of exabyte processing". Retrieved 2012-10-24.
[8] "Data Crush by Christopher Surdak". Retrieved 14 February 2014.
[9] Hellerstein, Joe (9 November 2008). "Parallel Programming in the Age of Big Data". *Gigaom Blog*.
[10] Segaran, Toby; Hammerbacher, Jeff (2009). *Beautiful Data: The Stories Behind Elegant Data Solutions*. O'Reilly Media. p. 257. ISBN 978-0-596-15711-1.
[11] Hilbert & López, 2011 .
[12] "IBM What is big data? — Bringing big data to the enterprise". www.ibm.com. Retrieved 2013-08-26.
[13] Oracle and FSN, "Mastering Big Data: CFO Strategies to Transform Insight into Opportunity", December 2012
[14] Jacobs, A. (6 July 2009). "The Pathologies of Big Data". *ACMQueue*.
[15] Magoulas, Roger; Lorica, Ben (February 2009). "Introduction to Big Data". *Release 2.0* (Sebastopol CA: O'Reilly Media) (11).
[16] http://www.reportsnreports.com/reports/288019-the-big-data-market-2014-2020-opportunities-challenges-strategies-industry-verticals-and-forecasts.html.
[17] "Enterprise Big-data" by David Floyer**,** on Nov 21, 2014.
[18] http://www.linkedIn.com/The Eye-Opening Facts Everyone Should Know _ Bernard Marr _ LinkedIn.html.
[19] http://www.csc.com/business_drivers/offerings/82042-big_data_storage_solutions.
[20] http://www.datastax.com/big-data-challenges.
[21] http://www.sei.cmu.edu/uls /Addressing the Software Engineering Challenges of Big Data » SEI Blog.html.
[22] http://cloud.asperasoft.com/big-data-cloud/
[23] http://www.xenolytix.com/verticals.html
[24] https://www.mwrinfosecurity.com/articles/big-data-security---challenges-solutions/
[25] Apache Hadoop. Available at http://hadoop.apache.org.
[26] http://en.wikipedia.org/wiki/Big data.

