ISSN No:-2456 –2165

Big Data: Introduction and Security Issues of Big Data

Prof. Mayuri Bharat Dandwate, Prof. Rutuja Vilas Kotkar

PIRENS Institute of Computer Technology(PICT), LONI.

Abstract:-Big data, which refers to the data sets that are too big to be handled using the existing database management tools, are emerging in many important applications, such as Internet search, business informatics, social networks, social media, genomics, and meteorology. Big Data is high volume, high velocity, high variety information assets that demand cost effective forms of information processing that enable enhanced insight, decision making and process automation. Big data presents a grand challenge for database and data analytics research. Big Data poses a grand challenge on the design of highly scalable algorithms and systems to integrate the data and uncover large hidden values from datasets that are diverse, complex, and of a massive scale.

Keywords:-Big Data, Volume, Velocity, Variety, NoSQL, Storage.

I. INTRODUCTION

Big data is often described by 3V's: the extreme volume of data, the wide variety of data types and the velocity at which the data must be processed.

- Volume: Huge amounts of data , from datasets with sizes of terabytes to zettabyte.
- Velocity: Huge amounts of data from transactions with high refresh rate resulting in data streams coming at fast speed and the time to act on the basis of these data streams will often be very short. There is a variation from batch processing to real time streaming.
- Variety: Data come from various data sources. For the first, data can come from both internal and external data source. Mainly, data can come in various format such as transaction and log data from various applications, structured data as any database table, semi-structured data like as XML data, unstructured data like as text, images, video streams, audio statement, and many more. There is a deviation from sole structured data to growing more unstructured data or the mixture of the both.

Key enablers for the growth of "Big Data" are:
- Increase of storage capacities
- Increase of processing power
- Availability of data

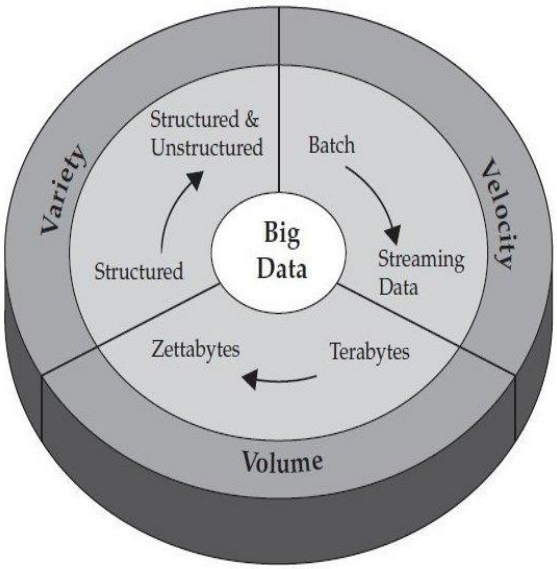

Figure-1 V's of Big Data

A. Tools used in Big Data

1. NoSQL :-Databases Mongo DB, Couch DB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper.
2. MapReduce:- Hadoop, Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum.
3. Storage:-S3, Hadoop Distributed File System.
4. Servers:- EC2, Google App Engine, Elastic, Beanstalk, Heroku.
5. Processing:- R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, BigSheets, Tinkerpop

B. Applications of Big Data

- Big data in manufacturing sector
- Big data for product distribution
- Big data in Marketing field
- Price Management using Big data

- Big data in Sales
- Big data in Human Resources
- Big data in Banking
- Big data in HealthCare
- Big data in Media and Entertainment
- Big data in Airlines

II. BIG DATA SAFETY

Every day, the security of confidential information is gaining more and more consideration. By this, we can understand that the security is the extremely highly ranked for consideration any enterprise. However, with the ease of adoption of web-based, smartphones, and cloud-based applications, the secret information has become easy to access from different platforms. Such platforms are highly vulnerable to hacking, especially if they cannot be handled properly.

Unlike earlier, companies are now collecting and using lots and lots of customer data. A lack of data security can bring some serious security issues and organization's reputation will be at stake. As far as Big Data is concerned, a company's financial and reputation things can change overnight.

A. Big Data Security Problems

Usually, for any data, the security issue raises if there are no proper security measures taken either by firewalls or antivirus softwares or both. These measures effectively work if the data is within the system. But what if the data goes beyond just your system – it can be Cloud?

In many cases, the distributed systems' computations have a less protection, say one or two levels. This kind of protection level is not at all suggested. At some or the other point, connections security and the encryption of access control will be not effective and inaccessible to dept. of IT who solely depend on it. Automated data transformation needs extra security norms, which are frequently not available. Suggested detailed audits are not done periodically on Big Data because of the huge amount of data being involved. Often non-relational databases keep evolving and thus makes it hard for security solutions to maintain the needed. Performing unethical surveys like IT specialists practicing data mining could give them a lot of personal information too, without the knowledge of users. Whenever a system gets really massive amounts of information, it definitely must be validated to stay reliable and exact. Because of the Big Data size, its origin is not always tracked or checked.

1. *Distributed Frameworks*: Most big data implementations actually distribute huge processing jobs across many systems for faster analysis. Hadoop is a well-known instance of open source tech involved in this, and originally had no security of any sort. Distributed processing may mean less data processed by any one system, but it means a lot more systems where security issues can crop up.
2. *Non-Relational Data Stores:* Think NoSQL databases, which by themselves usually lack security (which is instead provided, sort of, via middleware).
3. *Storage:* In big data architecture, the data is usually stored on multiple tiers, depending on business needs for performance vs. cost. For instance, high-priority "hot" data will usually be stored on flash media. So locking down storage will mean creating a tier-conscious strategy.
4. *Endpoints:* Security solutions that draw logs from endpoints will need to validate the authenticity of those endpoints, or the analysis isn't going to do much good.
5. *Real-Time Security/Compliance Tools:* These generate a tremendous amount of information; the key is finding a way to ignore the false positives, so human talent can be focused on the true breaches.
6. *Data Mining Solutions:* These are the heart of many big data environments; they find the patterns that suggest business strategies. For that very reason, it's particularly important to ensure they're secured against not just external threats, but insiders who abuse network privileges to obtain sensitive information – adding yet another layer of big data security issues.
7. *Access Controls:* Just as with enterprise IT as a whole, it's critically important to provide a system in which encrypted authentication/validation verifies that users are who they say they are, and determine who can see what.
8. *Granular Auditing:* can help determine when missed attacks have occurred, what the consequences were, and what should be done to improve matters in the future. This in itself is a lot of data, and must be enabled and protected to be useful in addressing big data security issues.
9. *Data Provenance:* primarily concerns metadata (data about data), which can be extremely helpful in determining where data came from, who accessed it, or what was done with it. Usually, this kind of data should be analyzed with exceptional speed to minimize the time in which a breach is active. Privileged users engaged in this type of activity must be thoroughly vetted and closely monitored to ensure they don't become their own big data security issues.

III. HOW CAN BIG DATA SECURITY BE IMPROVED?

Here are some ways to strengthen Big Data security:

- Focus on application security, rather than device security.
- Isolate devices and servers containing critical data.
- Introduce real-time security information and event management.
- Provide reactive and proactive protection.

IV. CONCLUSION

The availability of Big Data, low-cost commodity hardware, and new information management and analytic software have produced a unique moment in the history of data analysis. Big Data clearly improves business efficiency. As compute power and storage capacities continue to rise, and costs continue to decline, Big Data and analytics are playing an increasingly important role. As organizations make greater use of big data, it is likely that there will be increased concerns about individual privacy issues. It has been need of the day in various sectors along with security of data.

REFERENCE

[1]. Viktor Mayer-Schoenberger & Kenneth Cukier, "Big Data: A Revolution That Will Transform How We Live, Work, and Think"
[2]. Dio L Herben, "Big Data, Big Analytics: Emerging Business Intelligence"
[3]. EMC Education Services, "Data Science and Big Data Analytics"
[4]. https://www.oracle.com/in/big-data/index.html
[5]. https://en.wikipedia.org/wiki/Big_data
[6]. http://bigdata.teradata.com/US/Big-Data-Quick-Start/How-Big-Data-Works