International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 01, January-2017

A Study of Big Data Analytics in Clouds with a Security Perspective

Dr. M. Bhanu Sridhar
Associate Professor, Dept. of CSE
GVP College of Engineering for Women
Visakhapatnam, India

A. Koushik
Assistant Professor, Dept. of CSE
Satya Institute of Technology and Management
Visakhapatnam, India

***Abstract*— Big Data is a sparkling topic with plenty of scope for research. With the advent of social media, the data has started to cross the limits of a system, server and even a data center. On the other hand, Cloud Computing is another area in the IT field where different services like Software, Infrastructure, storage etc. are offered as services online. The main advantage here is that small business firms need not accumulate data storage as a physically existing 'thing'. At this juncture, it has to be realized that Big Data in Cloud is not only for storage but also for analyzing. This paper concentrates upon the recent trends in Big Data storage and analyzing, in the clouds, and also points out the security short comings. Some light is also thrown into the future scope of this concept where Internet of Things (IoT) beckons the researchers.**

Keywords- Big Data, Cloud Computing, Security, SaaS, IaaS, IoT, Analyzing of Big Data

I. INTRODUCTION

Storage of data is a common procedure followed by anyone. The data starts to grow as years pass and is accumulated. It is up to the concerned company/individual to utilize the data so as to find out exactly what has happened and what might happen in the future.

Since the advent of Internet and much more after social media introduction, the data has grown by leaps and bounds every day. The word 'BIG DATA' was first mentioned by Michael Cox and David Ellsworth [1]. Apache Software Foundation took up the challenge of storing and processing the 'big data' and out came Apache Hadoop, Apache Pig, Apache Hive and Apache Mahout. Most recently, Apache Spark has been creating waves in the corporate IT field.

Big data [2] refers to managing, analyzing and capturing different data sets where size, complexity and rate of growth varies for each of them. The benefits that can be accomplished with big data analytics are products can be redeveloped effectively, maintenance costs can be reduced, offering deeper insight from enterprise perspective, customizing websites in real time, creating new revenue streams and analysis of risks can be performed effectively.

Cloud computing [3], [4] is also an emerging research area where different services can be provided to the users on demand. The different services that can be provided are Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). In SaaS, software is deployed over Internet is delivered as on demand service and its main characteristics are easy access to commercial software, no handling of software upgrades and patches and provides an API for integration between different pieces of software. In PaaS, platform for creation of software over the web is delivered as on demand service and its main characteristics are integration with web services and databases via the common standards, providing web based user interface (UI) tools that help to create, modify, test and deploy different UI scenarios, providing support for multiple concurrent users utilizing same development application and support for development team collaboration. In IaaS servers, network, storage and operating systems are delivered as on-demand service and its main characteristics are including multiple users on single piece of hardware, resources are distributed as a service and providing support for dynamic scaling.

The different criteria that should be considered when analyzing big data in clouds are how to incorporate scalable, efficient and low-cost data storage platform and support for application development which involves modeling, mining, exploration and analysis of parallel execution of massive amount of data sets [3].This paper deals with deeper analysis of big data in clouds, frameworks used for analysis and its pros and cons, security issues and challenges to be considered when analyzing big data in cloud and Internet of Things(IoT) and paves way for future research and development.

II. BIG DATA IN CLOUD

Due to massive amount of information available worldwide Big Data is becoming tremendous challenge for today's rapidly changing traditional markets when performing in-depth analysis. The different areas where Big Data is being available are Social media, mobile phone details, transactional data, documentations such as financial statements, insurance forms, medical records and customer correspondence, RFID tags, weather information, Internet of things (IoT), traffic patterns, communication events etc. Big data is generally defined by five V's (Variety, Velocity, Volume, Veracity, and Visibility) [2] [3].

A. Big Data Business Drivers

The big data goals for many different types of organizations fall into major categories [2][5] such as

- Revenue
  Design and execute big data analytics use cases that increase revenue, lower costs and improve efficiency in business operations
- Customer services

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 01, January-2017

Improve customer understanding, obtain behavioral insight into client transactions and attract variety of customers

- Business development

  Introduce new products and services, deciding what to outsource without affecting customer experience, gaining new competitive insight into markets.

- Business agility and governance

  Plan with greater confidence, make better decisions, Ensure regulatory compliance and lower costs.

- IT and operational optimization

  Develop a strategy that uses existing enterprise areas for optimizing the applications

B. Cloud Computing

It refers to providing different types of services on demand to various categories of users. The different types of cloud for deployment of big data in cloud are as follows [4]

- Public Cloud

  This allows different services to be accessible to the public on commercial basis by Cloud service provider. The major benefits are cost effectiveness, reliability, flexibility, scalability and its disadvantages are low security, less customizable.

- Private Cloud

  It allows different services to be accessible within an organization. Its main benefits are cost efficiency, more control, high security and privacy and its disadvantages are inflexible pricing, limited scalability and additional skills are required to maintain cloud deployment.

- Hybrid Cloud

  It is combination of public and private clouds which allows data and applications to move from one cloud to another. Its main benefits are scalability, flexibility and security and its disadvantages are network issues, security compliance and infrastructural dependency.

As it can be viewed in Figure 1 [16], a profound analysis is essential for the classification of clouds and the capabilities of each type. In comparison, each style will have its own pros and cons but after taking a specific view, it is for the customer to choose which type can be relied upon. Public clouds are generally offered by cloud computing firms for enterprise organizations and can provide more services. Here, the need of the service and the security provided by the firm play a key role in the final decision of the user. Private clouds are used for small or middle-level business companies or even for personal usage. It becomes transparent for a customer that depending on the confidentiality of the project, and the level of resources or services provided in public and private clouds, he might use the ones that are best suited for his firm. Thus comes out the hybrid cloud where some resources from the public cloud and some others from private cloud can be used to provide utmost satisfaction to the customers.

The services provided by the Public, Private and Hybrid clouds are depicted in an able manner in Figure 1 and a succinct look on the diagram provides the all needed information about the taxonomy of cloud computing.

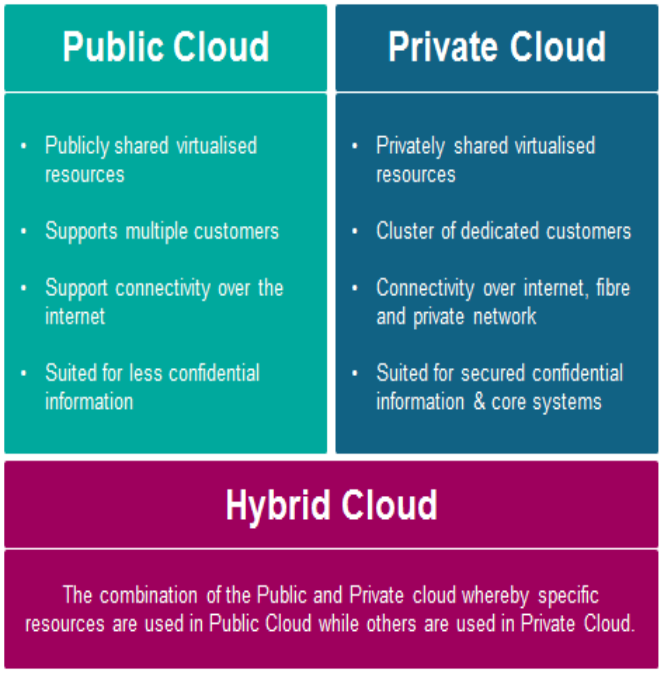

Fig 1: Classification of Clouds

C. Architectural Decisions for deploying Big Data in cloud

Big Data requires large amounts of storage and processing but traditional platforms for analysis such as data warehouses can't handle these gigantic data. So cloud computing is a means for accommodating these databases by using divide and conquer approach [3] [5]. Several architectural decisions are to be finalized before handling Big Data:

- Performance
- Scalability
- Reliability
- Availability
- Location and placement
- Sensitive data
- Disaster recovery

The key considerations to be taken for successful justification, management and deployment of Big Data in cloud are build business case, assess Big Data application workloads for deployment of Big Data into cloud, develop technical approach for deploying and managing Big Data and operationalize cloud based Big Data infrastructure.

III. CLOUD BASED BIG DATA ANALYTICS FRAMEWORKS

There are several frameworks available for storing and processing of data like Hadoop, Spark, Twister etc. Several databases like Hbase, HadoopDB etc have been used for storing data of any structure where Apache Pig, Apache Hive and so on has been used for processing data. Figure 2 elaborates the usage of Cloud Computing in Big Data Analytics [6]:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 01, January-2017

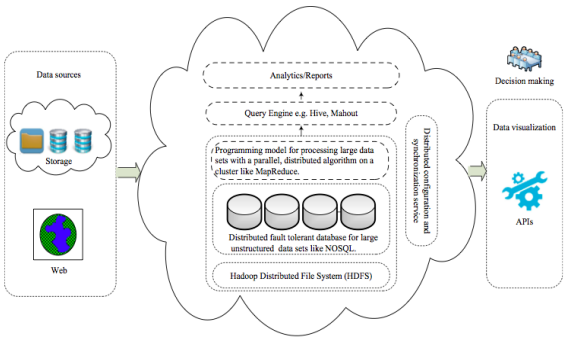

Fig 2: Usage of Cloud Computing in Big Data Analytics

Big Data Analytics require high performance processors to produce effective results for computation of data mining algorithms. There are several data mining techniques and tools that are available for extracting useful knowledge from large data sets but effective response in shorter time is the major criterion. A combination of Big Data Analytics and knowledge discovery with cloud computing systems offers an effective solution producing useful insights in shorter time. The several models for implementing Big Data analytics services are as follows [17]:

- Data Analytics Software as a Service
  It provides well defined data mining algorithms as service to end users.
- Data Analytics Platform as a Service
  It provides suitable platform for developers to build their own applications without concern about underlying infrastructure.
- Data Analytics Infrastructure as a Service
  It provides set of resources to run data mining applications

A. Hadoop

It is framework for processing large data sets across different clusters of nodes. It is open source software written in Java which implements HDFS (Hadoop Distributed File System) [7][8]. The major components of Hadoop are as follows:

- HDFS
  It holds large amount of data which provides efficient access where redundant data is stored across multiple machines which is highly fault tolerant and it is designed using low cost hardware [9]. Its main features are
  a) It is suitable for processing large amounts of distributed data.
  b) Hadoop provides command line interface to interact with HDFS.
  c) It provides an efficient approach for authentication of different nodes [10].
- Map Reduce
  It is efficient processing programming model for distributed computing using Java. Map Reduce Algorithm consists of two major tasks
  a) Map
  b) Reduce

Map takes a set of data and converts it into key/value pairs(tuples) and reduce task takes the output from a map as an input and combines those data tuples into a smaller set of tuples till desired result is obtained[8][10].

Advantages of Hadoop

- No license software is required.
- Used to design for cheap commodity hardware
- Simple programming model
- Scalability
- Robust and Fault-tolerant

Disadvantages of Hadoop

- Restrictive programming model
- Difficult to manage clusters
- Limited security
- Not suitable to handle small sets of data[8][9]

B. Spark

It is fastest cluster computing technology which extends Hadoop Map reduce model to efficiently perform more types of computations that includes interactive queries and processing of streams. Its main feature is in-memory cluster computing that increases speed of application [11][12].Its main features are as speed, support for multiple languages and provision for advanced analytics framework[11]. The major components of Spark are as follows:

- Spark Core
  It is execution engine where diverse applications are built on spark platform which provides in-memory computing
- Spark SQL
  It is component which is built on top of Spark core which provides support for structured and semi-structured data
- Spark Streaming
  It provides efficient streaming of data sets by performing RDD(Resilient distributed datasets) transformations on these data sets.
- Machine Learning Library
  It is distributed machine learning framework which runs as fast as Hadoop disk based version of Apache Mahout
- GraphX

It is distributed graph processing framework which provides an API for modeling user defined graphs and also provides an efficient optimized results.

Advantages of Spark

- Support for in-memory cluster computing platform by executing batch jobs faster than map reduce
- Support for sophisticated analytics
- Flexible and powerful
- Providing support for multiple languages
- Supports machine learning algorithms for future predictions

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 01, January-2017

Disadvantages of Spark

- Consumes a lot of memory
- It would take larger resources[11][12]

IV. SECURITY ISSUES AND CHALLENGES FOR BIG DATA ANALYTICS IN CLOUD

Security is becoming major issue for data storage in cloud based networks. Cloud computing technology comes with security issues which include networks, databases, operating systems, virtualization, resource scheduling and allocation, transaction management, load balancing and memory management [15]. The security issues associated with cloud computing environment can be categorized into several levels such as [13][14]:

- Network level

The issues and challenges associated with the network level includes network protocols and security in networks such as distributed nodes, distributed data etc

- User Authentication level

The issues and challenges associated with this level includes encryption/decryption techniques, authentication methods which includes authentication of distributed applications, access rights for nodes, logging etc

- Data level

The issues and challenges associated with this level include integrity of data and availability issues with data such as protection of data and distributed data.

- Generic level

The issues and challenges associated with this level includes different usage of security tools and usage of different technologies

Cloud security alliance in 2013 identified top ten challenges for Big Data security such as

- Secure computation in distributed programming frameworks
- Security best practices for non-relational data bases
- Secure data storage and transactions logs
- End-point input validation/filtering
- Real-time security monitoring
- Scalable and composable privacy-preserving data mining and analytics
- Cryptographically enforced data centric security
- Granular access control
- Granular audits
- Data Provenance

V. INTERNET OF THINGS (IOT)

It is becoming the next technological revolution where the revenue and data that IoT products generate will force the entire organizations to upgrade their tools and processes, technologies to accommodate this new data volume. The different ways in which IoT will have great impact on Big Data are data storage, data security, Big Data tools and technologies.

CONCLUSION

Big Data is becoming tremendous challenge for today's rapidly changing traditional markets when performing in-depth analysis. The different areas where Big Data is being available are Social Media, Mobile phone details, transactional data, documentation such as financial statements, insurance forms, medical records and customer correspondence, RFID tags, Weather information, Internet of Things (IoT), traffic patterns, communication events etc. Cloud Computing is another area in the IT field where different services like Software, Infrastructure, storage etc. are offered as services online.

The different frameworks like Hadoop, Spark, and Twister are used for analyzing and processing Big Data in cloud computing. In spite of its benefits, there are many security issues and challenges to be faced for storing of data in cloud based networks at different levels such as Network level, User Authentication level, Data level and Generic level. Lastly Internet of Things (IoT) is also starting to have an enormous impact on Big Data analytics. IoT and cloud computing are inter-related and the security of sensor messages surely presents a topic for research. Big Data Analytics, which is being loaded and utilized from the Clouds these day's poses a serious challenge to the corporate business companies in the area of security.

The paper, after discussing Big Data, Analytics and Cloud Computing, proposes the budding idea of using the Big Data Analytics from the Cloud – like MS Azure Solutions, SAP HANA and other services. The paper proposes that security of these services must be more beefed up to avoid irreparable damages to the mined Big Data in the Clouds. The area of Big Data Analytics in Cloud Computing ultimately is the tower to climb for the researchers to bring out satisfactory solutions to the ever-emerging problems in the field.

REFERENCES

[1] Cox, Michael, and David Ellsworth. "Application-controlled demand paging for out-of-core visualization." *Proceedings of the 8th conference on Visualization'97*. IEEE Computer Society Press, 1997.

[2] Rohit Chandrashekar, Maya Kala, Dashrath Mane, "Integration of Big Data in Cloud computing environments for enhanced data processing capabilities". International Journal of Engineering Research and General Science Volume 3, Issue 3, Part-2, May-June, 2015.

[3] Charlotte Castelino, Dhaval Gandhi, Harish G. Narula, Nirav H. Chokshi, "Integration of Big Data and Cloud Computing", International Journal of Engineering Trends and Technology (IJETT) – Volume 16 Number 2 – Oct 2014.

[4] Ali Gholami, ErwinLaure."Security and Privacy of Sensitive data in Cloud Computing: A Survey of Recent Developments", NETCOM, NCS, WiMoNe, CSEIT, SPM – 2015,pp. 131–150, 2015.

[5] Zhigao Zheng, Ping Wang, Jing Liu, Shengli Sun, "Real-Time Big Data Processing Framework: Challenges and Solutions", Appl. Math. Inf. Sci. 9, No. 6, 3169-3190 (2015).

[6] Samiya Khan, Kashish AraShakil, MansafAlam, "Cloud Based Big Data Analytics: A Survey of Current Research and Future Directions", Journal of Contemporary Psychotherapy,2015.

[7] Ritu Aggrawal, Gautam Kumar, "Business Intelligence by Cloud Computing Using Hadoop Single Node Cluster", International Journal of Electrical Electronics & Computer Science Engineering, Special Issue - TeLMISR 2015, ISSN: 2348-2273.

[8] SanthoshVoruganti, "Map Reduce a Programming Model for CloudComputing Based On Hadoop Ecosystem", International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 3794-3799.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 01, January-2017

[9] Nikhil Gupta, Komal Saxena, "Cloud Computing Techniques for Big Data and Hadoop Implementation", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 3 Issue 4, April – 2014.

[10] Samira Daneshyar, Majid Razmjoo, "Large-scale data processing using MapReduce in Cloud computing Environment", International Journal on Web Service Computing (IJWSC), Vol.3, No.4, December 2012.

[11] J. Boehm, K. Liu, C. Alis, "Sideloading - Ingestion of Large Point Clouds Into the Apache Spark Big Data Engine", ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLI-B2, 2016, pp.343-348.

[12] Damien Graux, Louis Jachiet, Pierre Genevs,NabilLaya, "SPARQLGX: Efficient Distributed Evaluation of SPARQL with Apache Spark", International Semantic Web Conference,2016.

[13] R.Saranya, V.P.MuthuKumar, "Security issues associated with Big Data in cloud computing", International Journal of Multidisciplinary Research and Development,Volume 2, Issue 4, 580-585,April 2015.

[14] Elmustafa Sayed Ali Ahmed, Rashid A. Saeed, "A Survey of Big Data Cloud Computing Security", International Journal of Computer Science and Software Engineering (IJCSSE), Volume 3, Issue 1, December 2014,ISSN (Online): 2409-4285.

[15] Shantanu Kalbhor, Hiteshkumar Jain, Kaushiki Upadhyay, "Providing classification and security of Big Data in Cloud computing", International Journal of Technical Research and Applications e-ISSN: 2320-8163, Volume 4,Issue 2(March-April, 2016), PP. 302-304.

[16] http://cloudacademy.com/blog/cloud-migration-benefits-risks/

[17] Domenico Talia ,"Clouds for Scalable Big Data Anaytics", IEEE Computer Society, Volume 46, Issue 5, May 2013, Pages 98-101