Big Data Analytics in Cloud Computing Environment

Mahammad Shabana
Research Scholar,
Dept of CSE,
CSJ MU, Kanpur.

Dr. Ravindranath
Professor,
Dept of CSE,
CSJ MU, Kanpur.

ABSTRACT:

This paper deals data processing in cloud computing environments using Big Data applications. It travels around some important areas of analytics and Big Data. One of the best qualities of cloud is sharing of resources and data into data centers on internet. At present various levels of services required to improve execution efficiency. In today's world Cloud is using big data processing technology to enhance application aggregation, data aggregation and data utilization. Cloud computing is best powerful technology for complex computing. It is used to eliminate expensive computing hardware, dedicated space, and software. Cloud computing is observed large growth in the scale of huge data. How to address big data is a great challenging and time- demanding task. It needs a large computational infrastructure for successful data processing & analysis. In this study the role of big data in cloud computing environment is reviewed. The definition, classification of big data with their characteristics and some discussions of cloud computing are expressed. The relationship between big data & cloud computing, storage systems, Hadoop technology are also elaborated.

KEY WORDS: Cloud computing, Big data.

1. INTRODUCTION:

One of the core challenge in the context of Cloud computing is the management of very huge volumes of data. This is totally independent of the resource type which is shared in the Cloud – data bases are either directly visible or accessible to clients as part of the Infrastructure, or are hidden behind service interfaces. It means that data required be partitioning and replicating across dissimilar data centers on internet. Best search engines such as Amazon, Google, have begun to establish new data centers for providing Cloud computing applications. Society is becoming increasingly more instrumented and storing vast amounts of data. Analytics solutions that structured and unstructured data are important as they can help organizations to gain insights not only from their privately acquired data, but also from huge amounts of data publicly available on the Web. This paradigm is being well liked termed as Big Data. The continuous increase in the volume of data captured by organizations, such as the increase of social media, Internet of Things (IoT), & multimedia, has produced an enormous flow of data in either structured or unstructured format. Big data are distinguished by three aspects: (a) numerous data, (b) data cannot be classified into regular relational databases, and (c) data are generated, captured, and processed speedily. The progress in data storage and mining technologies allow for the preservation of increasing large amount of data described by a change in the nature of data held by organizations. The rate at which new data is being generated is staggering. A serious challenge for researchers and practitioners is that this growth rate extends their ability to design appropriate cloud computing platforms for data analysis and update intensive workloads. Cloud computing is one of the most important shifts in modern ICT and service for enterprise applications and has become a powerful architecture to perform huge scale and complex computing. Advantages of cloud computing include virtualized resources, parallel processing, data security, & data service integration with scalable data storage. Cloud computing can not only reduce the cost and restriction for automation and computerization by individual systems and organizations but can also provide less infrastructure, maintenance cost, efficient management, and user access .

As a result of the above said advantages, a number of applications that hold various cloud platforms have been developed and resulted in a enormous increase in the scale of data generated and consumed by such applications. In cloud computing, first adopters of big data are practitioners that utilized Hadoop clusters in highly scalable and elastic computing environments provided by vendors, such as IBM, Microsoft, and Amazon. One of the base technologies applicable to the implementation of cloud computing is Virtualization. In a big data environment, the basis for many platform attributes required to access, store, analyze, & manage distributed computing components is achieved through virtualization. Virtualization is a process of sharing resources & isolation of underlying hardware to increase computer resource utilization, efficiency, and scalability. The major goal of this study is to implement a wide range investigation of the status of big data in cloud computing environments and deliver the definition with their characteristics, and classification of big data along with some more discussions on cloud computing. The relationship between big data & cloud computing, storage systems of big date, and the new concept Hadoop technology are also discussed.

Additionally, research challenges will focus on scalability, availability, data integrity and data transformation, data quality and data heterogeneity, privacy, some of legal and regulatory issues, and governance.

2. PURPOSE OF STUDY:

The main objective of this research paper specifies data big data management in cloud computing environment. In this we discuss about some topics like data processing based on cloud, data security, data storage technologies, data security, data privacy and trust in big data process on cloud, user friendly cloud access for big data processing, inter-cloud technology for big data.

3. CLOUD COMPUTING:

This section represents overview of cloud computing, including its definition and a comparison with related concepts.

3.1 Definitions:

The term cloud used mainly as a marketing term in a variety of contexts to represent many dissimilar ideas. A standard definition of cloud computing, in which resources (e.g., CPU and storage devices) are provided as general utilities that can be leased and released by practitioners through the Internet in an on-demand fashion. Infrastructure providers & service providers are the two traditional service providers in a cloud computing environment The prominence of cloud computing has made a large impact on the Information Technology (IT) industry over the past few years, where big search engines such as Google, Amazon and Microsoft try to provide more powerful, reliable and cost-efficient cloud platforms, and business enterprises look to reshape their business models to get benefit from this new paradigm. Indeed, cloud computing provides a variety compelling features that make it attractive to business organizers, as shown below.

No up-front investment:

Cloud computing make use of a pay-as you-go pricing model. To start obtaining benefit from cloud computing, a service provider does not need to invest in the infrastructure. Simply hire resources from the cloud according to its personal needs and pay for the usage.

Low operating cost:

Cloud environment resources can be rapidly allocated and de-allocated on demand; a service provider no longer needs to provision capacities according to the max load. This provides large savings since resources can be released to save on operating costs when service demand is low.

Greatly scalable:

Infrastructure providers pool large amount of resources from data center on internet and create them simply accessible. A service provider can expand its service to large scales in order to handle at a great rare increase in service demands (e.g., flash-crowd effect). This model is called surge computing.

3.2. Related Technologies:

Cloud computing is frequently compared to the following technologies:

Grid Computing:

An example of distributed computing is Grid computing that coordinates worked resources to attain a common computational objective. The growth of Grid computing was originally handled by scientific applications which are commonly computation intensive. Cloud computing is close to Grid computing in that it also hires distributed resources to obtain application-level objectives. Yet, cloud computing takes one step further by holding virtualization technologies at multiple levels (hardware and application platform) to perceive resource sharing and dynamic resource provisioning.

Utility Computing:

Utility computing represents the replica of providing resources on-demand &charging customers based on utilization rather than a flat rate. Cloud computing can be perceived as a awareness of utility computing. It acquires a utility-based pricing scheme completely for economic reasons. On-demand resource provisioning & utility based pricing, service providers can frankly maximize resource utilization and lower their operating costs.

Virtualization:

Virtualization is a technology that conceptual away the details of physical hardware and supply virtualized resources for high-level applications. A virtualized server is usually called a virtual machine. Virtualization forms the basis of cloud computing, as it produces the capability of pooling computing resources from cluster of servers & dynamically assigning or reassigning virtual resources to applications on-demand.

Autonomic Computing:

Initially coined by IBM in 2001, autonomic computing points at building computing systems have the ability of self-management, i.e. responding to internal and external observations without human intervention. The main objective of autonomic computing is to overwhelm the management complexity of today's computer systems. Yet cloud computing displays certain autonomic features such as automatic resource provisioning, its objective is to reduce the resource cost rather than to reduce system complexity. In summary, cloud computing holds virtualization technology to obtain the idea of providing computing resources as a utility. It shares clear aspects with grid computing and autonomic computing yet differs from them in other features. Therefore, it offers good benefits and force unique challenges to meet its requirements.

3.3 Types of clouds:

There are many cases to consider when working an enterprise application to the cloud environment. For example, some service providers are mainly interested in reducing operation cost, while others may like better high reliability and security. There are different types of clouds, each with its own profits and pitfalls.

Public clouds:

A cloud in which service providers provide their resources as services to the common public. Public clouds propose several key benefits to service providers, which include no initial capital investment on infrastructure & shifting of risks to infrastructure providers. Nevertheless, public clouds need fine-grained control over data, network and security settings, which hampers their successfulness in many business scenarios.

Private clouds:

Private Cloud's also recognized as internal clouds, private clouds are designed for incompatible use by a single organization. A private cloud may be built and managed by the firm or by external providers. A private cloud provides the highest degree of control over performance, reliability & security. Although, they are frequently criticized for being similar to traditional proprietary server farms and do not supply benefits such as no up-front capital costs.

Hybrid clouds:

A hybrid cloud is a mixture of public and private cloud representations that tries to address the restrictions of each approach.

In a hybrid cloud, a part of the service infrastructure moves in private clouds and the remaining part moves in public clouds. Hybrid clouds provide more flexibility than both public and private clouds. Significantly, they provide safe control and security over application data when compared to public clouds. On the down side, scheming a hybrid cloud needs carefully determining the greatest split between public and private cloud components.

Virtual Private Cloud:

Another solution to addressing the shortcomings of both public and private clouds is called Virtual Private Cloud (VPC). A VPC is essentially a platform moving on top of public clouds. The main difference is that a VPC influences virtual private network (VPN) technology that permits service providers to design their own topology & security settings such as rules of firewall. Essentially, VPC is a more holistic design since it not only virtualizes servers and applications, but also concealed communication network as well. In advance, for most companies, VPC offers seamless transition from a proprietary service infrastructure to a cloud-based infrastructure, payable to the virtualized network layer. For most service providers, choosing the best cloud model is dependent on the business scenario. For example, computation-intensive scientific applications are great deployed on public clouds for cost-effectiveness. Conceivably, certain types of clouds will be more popular than other clouds.

3.4Cloudcomputing characteristics:

Cloud computing provides several major features that are different from conventional service computing, which we outline below:

Multi-tenancy:

In a cloud environment, services possessed by multiple providers are co-located in a single data center. The performance & management issues of these services are shared among infrastructure provider and the service providers. The layered architecture of cloud computing offers a natural division of duties: the possessor of each layer only needs to focus on the certain objectives associated with this layer. Nevertheless, multi-tenancy also introduces troubles in understanding and managing the interactions surrounded by various stakeholders.

Shared resource pooling:

The infrastructure provider offers a pool of computing resources that can be dynamically assigned to large number of resource consumers.

Such dynamic resource assignment capability provides much flexibility to infrastructure providers for engaging their own resource usage and operating costs. For instance, an IaaS provider can hold VM migration technology to attain a high degree of server consolidation, while minimizing cost, maximizing resource utilization such as power consumption and cooling.

Geo-distribution and ubiquitous network access:

Clouds are commonly accessible through the Internet & use the Internet as a service delivery network. So any device with Internet connectivity, be it a mobile phone, a (personal digital assistant) PDA or a laptop, is allowed to access cloud services. Furthermore, to obtain high network performance and localization, more number of today's clouds consists of data centers located at many sites around the globe. A service provider can easily hold geo-diversity to obtain maximum service utility.

Service oriented:

cloud computing acquires a service-driven operating model. Hence it places a powerful emphasis on service management. In a cloud computing environment, each IaaS, PaaS and SaaS provider gives its service according to the Service Level Agreement (SLA) hold talks with its customers. Therefore SLA assurance is a critical objective of every provider.

Dynamic resource provisioning:

One of the salient key features of cloud computing is that computing resources can be acquired and released on the fly. Compared to the conventional model that provisions resources as stated by peak demand, dynamic resource provisioning permits service providers to obtain resources based on the current demand, which can greatly lower the operating cost.

Self-organizing:

Since on-demand resources can be allocated or de allocated, service providers are empowered to lead their resource consumption according to their personal needs. Additionally, the automated resource management feature produces high agility that enables service providers to reply quickly to speedy changes in service demand.

Utility-based pricing:

Cloud computing engages a payer-use pricing model. The exact pricing scheme may change from one service to another service.

For example, a SaaS provider may hire a virtual machine from an IaaS provider for an hour basis. A SaaS provider that produces on-demand customer relationship management (CRM) may demand its customers based on the number of users it serves (e.g., Sales force). Utility-based pricing lowers service operating cost as it demands customers on a per-use basis. Yet, it also introduces complexities in controlling the operating cost. In this prospect, companies like VKernel provide software to help cloud clients understand, study and cut down the not necessary cost on resource consumption.

4. BIG DATA:

Right now we are living in data world. Everywhere we are seeing only data. So, important thing is how to store date & how to process data? Exactly to say, what is big data? The data which is beyond to the storage capacity and the data which is beyond to the processing power, that data is called Big data. By using traditional database technologies we cannot store, process, and analyze large volume of data.The nature of big data is out of focus and involves on severable processes to identify the data into new insights. The term "big data" is relatively new in IT industry & business firm. Nevertheless several researcher and users have utilized the term big data in previous literature. Several definitions of big data currently are alive. Meanwhile and defined big data as distinguished by three Vs: volume, variety, and velocity. The terms volume, variety and velocity were initially introduced by Gartner to narrate the elements of big data challenges. Big data is not only distinguished by the three Vs mentioned above but may also enlarge to four Vs, namely, volume, variety, velocity, and value This 4V definition is far apart recognized because it focuses the meaning and requirement of big data. Big data is a set of techniques and technologies that acquire new forms of integration to uncover great hidden values from huge data sets that are diverse, complex and of a massive scale.

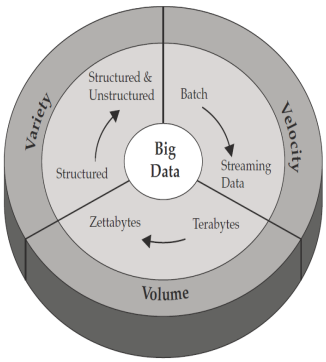

Figure: Characteristics of Big data

Volume:

The term Big Data itself is related to a size which is enormous / huge. Size of data plays very important role in determining value out of data.

Also, whether a particular data be considered as Big Data or not is dependent on the Volume of data. Hence, Volume is one most important characteristic which needs to be considered while dealing with Big Data.

Variety:

Variety refers to heterogeneous sources & the nature of data both structured and un-structured. This variety of big data poses clear issues for storage, mining & analyzing data. Relational Data Base Management System is used to store structured data. Social Networks are the best example for unstructured data. Data may be either structured, unstructured or semi structured data.

Velocity:

The term 'Velocity" represents the Processing speed of data. How fast the data is generated and processed as per client demands, determines real potential in the data. For ex social network sites, mobile devices etc.

Value:

is the most important feature of big data it refers to the process of discovering enormous hidden values from large datasets with different types and rapid generation.

Advantages of Big Data:

While taking decisions Businesses can utilize outside intelligence.
Customer service improved.
Early risks identification if any
Better operational efficiency

Classification of big data:

To better understand their characteristics, Big data are classified into different categories. The classification is more important because of large-scale of data in the cloud. The classification is based on five features: (i)big data sources, (ii)content format, (iii)big data stores,(iv) big data staging (v)big data processing. Each of these categories has its own characteristics and complexities as outlined. Data sources include internet data centre, sensing and all stores of transnational information, ranges from unstructured to highly structured are stored in different formats. More familiar is the relational database that comes in a large number of varieties. As a result of the wide variety of data sources, the captured data which differ in size with respect to redundancy, consistency and noise, etc.

5. RELATIONSHIP BETWEEN CLOUD COMPUTING AND BIG DATA:

Cloud computing and big data are combined. Big data provides practitioners the ability to use commodity computing to process distributed queries across large data sets and return result in a timely manner. Cloud computing supplies the underlying engine throughout the use of Hadoop, a class of distributed data-processing platforms. Distributed fault-tolerant database and processed through a programming paradigm for huge datasets with a parallel distributed algorithm in a cluster. The main motive of data visualization is to outlook analytical results presented visually through different graphs for decision making. Based on cloud computing big data utilizes distributed storage technology rather than local storage attached to electronic device.

Big data evaluation is driven by quick-growing cloud-based applications developed using virtualized technologies. Consequently, cloud computing not only provides facilities for the computation and processing of big data but also works as a service model. The author narrated that cloud computing infra- structure can perform duties as an effective platform to address the data storage desired to perform big data analysis. Cloud computing is correlated with a new pattern for the allocation of computing infrastructure and big data processing method.

Various cloud-based technologies have to manage the new environment because dealing with big data for concurrent processing has become growingly complicated. In a cloud environment a good example of big data processing is Map Reduce; it allows for the processing of huge datasets stored in parallel in the cluster. Cluster computing displays high quality performance in distributed system environments, such as computing power, data storage, and network communications. Similarly, Bollier and Firestone emphasized the cluster capability computing to provide a hospitable context for data growth.

Nevertheless, Miller argued that the absence of data availability is very high because users offload more decisions to analytical methods; incorrect use of the methods or intrinsic weaknesses in the methods may deliver wrong and costly decisions. DBMSs are considered a part of the contemporary cloud computing architecture & act as a principal role to ensure the easy transition of applications from conventional enterprise infrastructures to new cloud infrastructure architectures. The pressure for organizations to rapidly adopt and implement technologies, such as cloud computing, challenges of big data storage and processing demands entails unforeseen risks and consequences.

6. BIG DATA MANAGEMENT SYSTEM:

Many researchers have proposed that commercial DBMSs are not suitable for processing enormously large scale data. Typical architecture's likely bottleneck is the database server while faced with heavy workloads. One database server has reduction of scalability and cost, which are two principal goals of big data processing. In order to adapt various enormous data processing models, D. Kossmann et al. exhibited four distinct architectures based on classic multi-tier database application architecture which are splitting, distributed control and caching architecture. It is understandable that the alternative providers have different business paradigms and target different kinds of applications: Google seems to be more interested in small user applications with less workloads whereas Azure is currently the most affordable service for medium to great services. Most of recent cloud service providers are utilizing hybrid architecture which is capable of fulfilling their actual service requirements.

7. CONCLUSION:

In today's world, the size of data is huge and continues to increase every day. The variety of data being generated is become larger. The velocity of data generation and growth is increasing because of the escalation of mobile devices and other device sensors connected to the Internet.. In this study, we presented a review on the growth of big data in cloud computing. We suggested a classification for big data, a conceptual view of big data, & a cloud services model. This model was compared with several big data cloud platforms. We talk over the background of Hadoop technology and its major components, namely, Map Reduce and HDFS. We represented current Map Reduce projects and related software Technologies. We also evaluated some of the challenges in big data processing.

REFERENCES:

[1] R. L. Villars, C. W. Olofson, M. Eastwood, Big data: what it is and why you should care, White Paper, IDC, 2011, MA, USA.

[2] D.E. O'Leary, Artificial intelligence and big data, IEEE Intell.Syst.28 (2013)96–99.

[3] M. Chen, S. Mao, Y. Liu, Big data: a survey, Mob. Network. Appl. 19(2)(2014)1–39.

[4] B. P. Rao, P. Saluia, N. Sharma, A. Mittal, S. V. Sharma, Cloud computing.

[5] "Big data: science in the petabyte era," Nature 455 (7209):1, 2008.

[6] S. Ghemawat, H. Gobioff, and S. Leung, "The Google file system," in ACM SIGOPS Operating Systems Review, vol. 37, no. 5. ACM, 2003, pp. 29–43.

[7] Douglas and Laney, "The importance of 'big data': A definition,"2008.

[8] J. Dean and S. Ghemawat, "Map reduce: simplified data processing on large clusters," Communications of the ACM,vol. 51, no. 1, pp. 107–113, 2008.

[9] D. Borthakur, "The hadoop distributed file system: Architecture and design," Hadoop Project Website, vol. 11, 2007.

[10] D. Kossmann, T. Kraska, and S. Loesing, "An evaluation of alternative architectures for transaction processing in the cloud," in Proceedings of the 2010 international conference

[11] A. Rabkin and R. Katz, "Chukwa: A system for reliable large-scale log collection," in USENIX Conference on Large Installation System Administration, 2010, pp. 1–15.

[12] S. Sakr, A. Liu, D. Batista, and M. Alomari, "A survey of large scale data management approaches in cloud environments, "Communications Surveys & Tutorials, IEEE, vol. 13

[13] R. Cumbley, P.Church, Is Big Data creepy? Comput.LawSecur.Rev. 29 (2013)601–609.

[14] Y. Cao, C. Chen, F. Guo, D. Jiang, Y. Lin, B. Ooi, H. Vo, S. Wu, and Q. Xu, "Es2: A cloud data storage system for supporting both oltp and olap," in Data Engineering (ICDE),2011 IEEE 27th International Conference on. IEEE, 2011,pp. 291–302.

[15] F. Chang, J. Dean, S. Ghemawat, W. Hsieh, D. Wallach,M. Burrows, T. Chandra, A. Fikes, and R. Gruber, "Big table: A distributed structured data storage system," in 7th OSDI,2006, pp. 305–314.

[16]. Aguilera, M. K., A. Merchant, M. A. Shah, A. C. Veitch, and C. T. Karamanolis, \Sinfonia: A New Paradigm for Building Scalable Distributed Systems," SOSP 2007.

[17]. Aguilera, M., W. Golab, and M. Shah, \A Practical Scalable Distributed B-Tree, "VLDB 2008.

[18]. Amazon Elastic Compute Cloud: http://aws.amazon.com/ec2/, Retrieved date: Sep. 27, 2009.

[19]. Apache Hadoop, http://hadoop.apache.org/, Retrieved date: Sep. 27, 2009.

[20]. Apache HBase, http://hadoop.apache.org/hbase/, Retrieved date: Sep. 27, 2009.

[21] Sang Woo Han, Jong Won. A multi-agent-based management system for pervasive collaborative computing environment[C]. IEEE international conference on computing and communications.USA: Institute of electrical and electronics engineers, 2009:1-6.

[22] Han Xu, CaoYongcun. The application of computer-supported collaborative technologies

[23] Han J, Kamber M. Data mining concepts and techniques [M]. San Francisco: Morgan kaufmann, 2006. Massachusetts: MIT press, 2004:191-211.

[24]S.Kaisler,F.Armour,J.A.Espinosa,W.Money,BigData:IssuesandChallengesMovingForward,SystemSciences (HICSS),2013,in: Proceedings ofthe46th Hawaii International Conference on, IEEE, 2013,pp.995–1004.

[25] T. Hey, S. Tansly, and K. Tolle, editors. The Fourth Paradigm: Data- Intensive Scientific Discovery. October 2009.

[26] V. Uhlig, J. LeVasseur, E. Skoglund, and U. Dannowski. Towards scalable multiprocessor virtual machines. In Proceedings of the 3rd Virtual Machine Research & Technology Symposium, San Jose, CA, 2004.

[27] J. Hu, L. Fang, Y. Cao, H. J. Zeng, H. Li, Q. Yang and Z. Chen, “Enhancing Text Clustering by Leveraging Wikipedia Semantics”, Proceeding of SIGIR’08, Singapore, (2008) July, pp. 179-186.

[28] S. Chakrabarti, “Mining the web: Discovering Knowledge from Hypertext Data”, Morgan Kaufmann Publishers, (2003).