ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

Cloud Reference Architecture for Big data Analytics

Archna Aggarwal[1], Nisha Pandey[2]

M.Tech (CSE), Department of Computer Science, Shri Ram College of Engineering and Management, Palwal, Haryana, India[1]

Assistant Professor, Department of Computer Science, Shri Ram College of Engineering and Management, Palwal, Haryana, India [2]

ABSTRACT: In many areas, information technology has made possible the availability of enormous information for analysts and decision-makers. To capture, store, distribute, manage and analyze the larger sized information new techniques, technologies and tools are required. These large sized data sets are called big data. Data sets are not only large, but are also very huge in velocity and variety, making them difficult to handle using traditional techniques and tools are referred as Big data. Structured, semi-structured and unstructured data that has the potential to be mined for information is big data. Big data can be analyzed for better decisions and strategic business moves for insights. New algorithms, techniques, architecture, and analytics are needed to manage and extract value and hidden knowledge from big data because it is diverse and complex.

KEYWORDS: Big data, Data Analytics, Cloud Computing, Private Cloud, Public Cloud, Hybrid Cloud, Layered Architecture, Hadoop.

I. INTRODUCTION

Big data is the term used to refer such voluminous amount of data that are not managed with current methodologies or data mining software tools due to their large size and complexity. Big data is said to be a new term but, the act of gathering and storing large amounts of information for eventual analysis is ages old. The concept of big data was coined by industry analyst Doug Laney in the early 2000s. He defines big data as the three Vs: Volume, Velocity and Variety. Big data computing is an emerging data science paradigm of multidimensional information mining for scientific discovery and business analytics over large-scale infrastructure. Although the size of the data is used to know whether this dataset is Big Data is not firmly defined and continues to change over time, most analysts and researchers considers big data as a data set from 50-60 terabytes to many peta bytes (1000 terabytes per petabyte).

A survey on big data computing in cloud and future research directions in several domains of science, engineering, and business for the development of analytics and visualization tools is presented by Assuncao, Buyya and Calheiros[3]. A big data processing model is presented by Zhu, Wu and Ding[4], from the data mining perspective. Issues in big data such as storage and data transport technologies followed by methodologies for big data analytics are discussed by Espinosa and Kaisler[5]. Chen and Chen[15] presented many tools and technologies for the management challenges of big data diversity, integration, cleaning, reduction, indexing and query analysis.

New economic systems are needed to define relationships between producers, distributors and consumers of goods and services as new business domains are growing. Instead of always relying on experience or pure intuition, it is necessary to make serious decisions to use critical data sources. A survey on big data architectures and framework the industry is done by National Institute of Standards and Technology Big Data Public Working Group [1].

A. Applications of Big Data Computing

Enterprises: In many industries around the world for various enterprises big data analytics techniques are playing a very important role. Enterprises get an enormous amount of data from social media and other sources and store it for

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

meaningful and future use. Handling such a large amount of data requires specialized analytics tools and techniques. Big data provide those tools and techniques to handle such large data. Also, Decision makers and business people make faster decisions in real time and less expense which is not possible using traditional analytics tools when data doesn't have to commute to work and back.

Clustering: Clustering is one of the important techniques for data mining and popularly used in big data analytics also. Using clustering algorithm like K-means, users can divide data into groups on specific data dimensions. With clustering, it becomes simple to identify and address groups by various dimensions like purchasing behaviour, type of products purchased, customer type, products, information searched, etc.

Banking: Big data offer a number of advantages for banks. It helps in fraud detection and prevention by ensuring that no unauthorized transactions will be made, by providing a level of safety and security. Also helps in risk management by locating and presenting big data on a single large scale that makes it easier to reduce the number of risks to a manageable number. Big Data plays a crucial role in incorporating the banks prerequisites into a central, functional stage. This reduces the possibility of the bank losing data, or ignoring fraud.

Health care: Big data [6, 7]has a great potential in health care analytics. It has the power to identify, prevent and cure various diseases and support health care industry. Now, doctors and other healthcare practitioners can make informed decisions as they have access to a wide range of clinical data that is stored systematically with the help of big data.

Governance: Surveillance system analyzing and classifying streaming acoustic signals, transportation departments using real-time traffic data to predict traffic patterns, and update public transportation schedules. Security departments analyzing images from aerial cameras, news feeds, and social networks or items of interest. Social program agencies gain a clearer understanding of beneficiaries and proper payments. Tax agencies identifying fraudsters and support investigation by analyzing complex identity information and tax returns. Sensor applications such stream air, water, and temperature data to support cleanup, fire prevention, and other programs.

Financial Services: In today's scenario, the most serious challenges facing financial institutions are to retain customers and meet consumer expectations.

Web analytics: Several websites are experiencing millions of unique visitors per day, in turn creating a large range of content. For improving response time, understanding limitations of their sites, offering more targeted ads, and so on, companies increasingly want to be able to mine the data which requires tools to perform complicated analytics on data that far exceed the memory of a single machine or even in cluster of machines.

B. Characteristics of Big Data

Big data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures, analytics, and tools in order to enable insights that unlock new sources of business value. The main features characterize big data are volume, velocity, variety, variability and complexity.

Volume: Volume refers to amount or size of data. Volume of data stored in enterprise repositories have grown from megabytes and gigabytes to petabytes.

Velocity: Velocity refers to the speed of data processing. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.

Variety: Different types of data and sources of data. Data variety exploded from structured and legacy data stored in enterprise repositories to unstructured, semi structured, audio, video, XML etc.

Variability: Data flows are extremely inconsistent with periodic peaks in addition to the velocity and variety of data. Daily, seasonal and event-triggered peak data load is difficult to manage. Unstructured data is more difficult to manage.

Complexity: In current scenario it becomes difficult to link, match, cleanse and transform data across systems data as it is complex and comes from multiple sources. However, it becomes necessary to connect and correlate relationships, hierarchies and multiple data linkages otherwise data becomes uncontrollable and unmanageable.

Data produced by different devices and applications is involved in big data. Some of the sources of Big Data are shown in the fig. 1.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

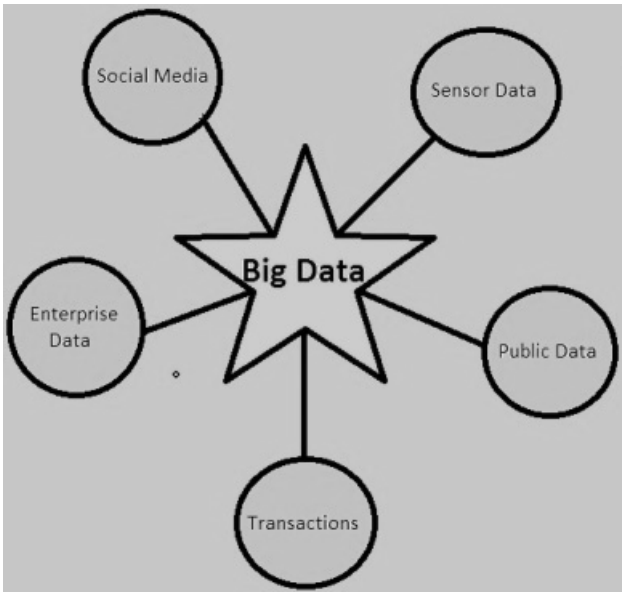

Fig.1: Sources of Big Data

C. Big Data Analytics

Analytics is the process of analyzing the data using statistical models, data-mining techniques, and computing technologies [2]. It combines the traditional analysis techniques and mathematical models to derive information. Big data analytics is the use of tools and processes to derive insights from large volumes of data. This data has either one of the three characteristics large volume, high velocity or extreme variety. Big data analytics main aim is to derive correlations and conclusions from data that were previously incomprehensible by way of traditional tools like spreadsheets. Big data analytics makes use of tools like Hadoop, SAS, R and so on which might be more powerful than formerly used rows and columns. Big data analytics makes use of those tools to derive conclusions from each organized and unorganized facts to provide insights that had been previously beyond our attain.

With advancement in technologies, the data available to the companies is growing at a tremendous rate. This data offers a host of opportunities to the companies in terms of strategic planning and implementation. With the help of real time big data processing, companies can use data to enhance decision making. Big Data analytics can help companies use data to impact not only on future decisions but also on present decisions.

Today, the estimated amount of data is equivalent to 1,200 exabytes, which is equal to twelve hundred billion gigabytes. That statistics is sufficient enough to fill 5 separate piles of CDs that might all attain to the moon. There is a rise in the storage capability with the rise in the amount of data. The average storage capacity of hard derive has accelerated from 10 GB in 2000 to 1TB in 2010.

With such immense rise in data come the possibilities of using it. Big data analytics allows for the use of this data to bring out relationships that are not reachable with traditional analysis techniques. The amount of possibilities and boom in analytical skills are accounting for this rise in big data analytics.

D. Benefits of Big Data Analytics

Cost Reduction: Big data helps in providing business intelligence that can reduce costs and improve the efficiency of operations. Processes like quality assurance and testing can involve many complications particularly in industries like biopharmaceuticals and nanotechnologies. Big data analytics can help industries to take better decisions by providing insights on the impact of different variables in the production process.

Improved Decision Making: Big data analytics can analyze past data to make predictions about the future. Thus businesses can not only make better present decisions but also prepare for the future. This offers them a aggressive aspect and presents a extra agile framework for selection erection or hazard handling.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

New Products and Services: Businesses can analyze past data about product launches and customer feedbacks to launch better products in futureAlong with this, real-time market analysis allows to understand business needs and changes in supply of products and changes in consumer behavior, which helps in customer oriented marketing. Analyzing consumer needs, preferences and buying behaviors can empower the increased demand for personalized services.

E. Big Data In Cloud

To build analytics and for deploying over a broadly scalable infrastructure, a new generation data-intensive platform is used, called large data in the cloud. Based on the services provided to the end users, big data in cloud are of three types: Public, Private and Hybrid.

Public big data cloud: processing and organization of large-scale data over the elastically scalable cloud infrastructure. Resources are served on the Internet as pay-to-know computing models. Examples include WindowsZure HDInsight [9], Rackspace Claudera Hadoop [10, 11], Amazon's big data computing in cloud [13] and Google Cloud Platform for big data computing [12].

Private big data cloud: Deploying big data platforms within a virtualized infrastructure on the enterprise, with more control and privacy of a single organization.

Hybrid big data cloud: Association of public and private big data structures for high availability, scalability, and disaster recovery. In this deployment, the private tasks can be migrated to the public infrastructure during peak workloads. Big data access networks and computing platform, integrate platform of data, computing, and analytics delivered as services by multiple distinct providers.

Big data computing in cloud also known as 'big data cloud' is data-intensive analytics platform of large-scale, distributed compute, and storage infrastructures. The big data cloud are large-scale distributed compute and data storages, wide range of computing facilities with seamless access to scalable storage repositories and data services, information-defined data storage, metadata-based data access instead of path and filenames, distributed virtual file system. File system could be dynamically created and mapped to the computing cluster, seamless access of computing and data, transparent access to large-scale data and compute resources, dynamic selection of data containers and compute resources, able to handle dynamic creation of virtual machines and able to access large-scale distributed data sources increasing the data location proximity, high performance data and computation, data should be high performance driven, multidimensional data handling, support for several forms of data with necessary tools for processing, analytics platform services, able to develop, deploy, and Platform for the data-intensive computing, replication mechanisms for both computing and data, and the high availability of computing and data and use analytics on the platform for data-intensive computing, support for both traditional and emerging data-intensive computing models and scalable deployment and execution of applications.

Fig.2 depicts integrated cloud and big data access networks on cloud infrastructure for analytics development. The content from several sources like social media, web logs, scientific studies, sensor networks, business transactions, and so on are growing rapidly. Fusing the information from several sources and Deriving useful information for decision-making from such large data would be a challenging task. The various elements of big data access network are: data services, big data computing platform, data scientist, and computing cloud.

Data and platform services: There are several providers who provide services for accessing both data and platform services for computing on the data. Google data APIs (GData)[15], is an example that provide protocols for reading and writing data on the web for several services such as Google analytics, content API for shopping, spreadsheets, and YouTube.

Big data computing platform: A platform for managing various data sources, including data access, programming models, management, schedulers, security etc. Platforms include various tools to access other data platforms using web services, streaming, and APIs. Some Other data platforms are Google data, social networking, data services from relational data stores and etc.

Data scientist: analytics developers, those who have access to the computing platform.

Computing cloud: An infrastructure for computing from private/public/hybrid cloud.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

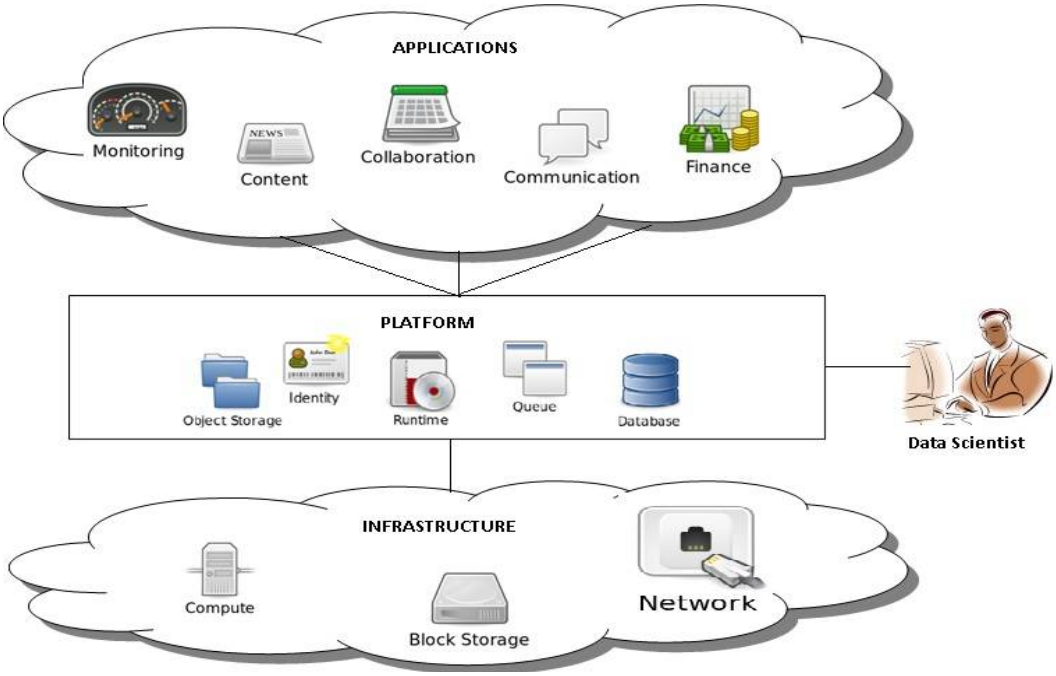

Fig.2: Big data cloud computer network

F. Big data cloud for the Enterprise

Big data cloud help enterprises to grow revenue, save money and achieve many other business objectives in any vertical by rapidly constructing their big data databases and writing analytics for mining the information. Enterprises have a number of benefits of big data cloud. They are:

Big data cloud would allow enterprises to collect billions of real-time data points on its products, resources, or customers and then repackage that instantaneously to optimize customer experience or resource utilization.

Big data cloud provide pay-as-go consumption models and other services similar to cloud services.This pricing model would effectively reduce both the cost of the applications development by minimizing the cost of development tools.

Realize new sources of information and build applications to gain competitive advantage. The information can be quickly mingled from various big data databases, and applications can be built rapidly for many platforms such as hand-held and mobile devices.

The speed at which it has been upgraded and growth in the amount of data sharing within the organization allows businesses and other organizations to respond to customers' demand more quickly and accurately.

II. RELATED WORK

A survey on big data architectures and framework from the industry is done by the National Institute of Standards and Technology Big Data Public Working Group[4]. According to this survey big data is a term used to describe the large amount of data in the networked, digitized, sensor-laden, information-driven world. Data may be deprived of traditional technical perspectives and data growth can eliminate scientific and technological advancements in data analytics, due to the availability of opportunities with big data. Many comparisons were identified, through compilation, review and comparison of Big Data Architecture implementation. These similarities between the architecture supported in the development of NBDRA Even though each Big Data system is tailored to the needs of the particular implementation, certain key components appear in most of the implementations. Three general components were observed in the surveyed architectures. They are Big Data Management and Storage, Big Data Analytics and Application Interfaces, and Big Data Infrastructure [1].

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

Bernice Purcell et.al. Stated that Big Data is comprised of large data sets that can't be handle by traditional systems. Big data includes structured data, semi-structured and unstructured data. The data storage technique used for big data includes multiple clustered network attached storage (NAS) and object based storage. Unstructured and semi-structured data is processed by using Hadoop architecture and to locate all relevant data uses map reduce which select only the data directly answering the query. The advent of Big Data has posed opportunities as well challenges to business [8].

S. Vikram Phaneendra & E. Madhusudhan Reddy et.al. emphasized that data was less in olden days and was easily handled by RDBMS but now it becomes difficult to handle huge data through RDBMS tools, which is preferred as "big data". They specified that big data differs from other data in 5 dimensions such as volume, velocity, variety, value and complexity. They depicted hadoop architecture to handle big data systems. Hadoop architecture handle large data sets, scalable algorithm does log management application of big data can be found out in financial, retail industry, health-care, mobility, insurance. They also focused on the challenges like data privacy, search analysis, etc, that need to be faced by enterprises while handling big data[17].

According to Kiran kumara Reddi & Dnvsl Indira et.al. big data is combination of structured, semi-structured ,unstructured homogenous and heterogeneous data .The author suggested to use nice model to handle transfer of huge amount of data over the network .Under this model, these transfers are relegated to low demand periods where there is ample ,idle bandwidth available . This bandwidth can then be repurposed for big data transmission without impacting other users in system. The Nice model uses a store and forward approach by utilizing staging servers. The model is able to accommodate differences in time zones and variations in bandwidth. They suggested that new algorithms are required to transfer big data and to solve issues like security, compression, routing algorithms [18].

Jimmy Lin et.al. used Hadoop which is currently the large –scale data analysis " hammer" of choice, but there exists classes of algorithms that aren't " nails" in the sense that they are not particularly amenable to the MapReduce programming model . He focuses on the simple solution to find alternative non-iterative algorithms that solves the same problem. The standard MapReduce is well known and described in many places .Each iteration of the pagerank corresponds to the MapReduce job. The author suggested iterative graph, gradient descent & EM iteration which is typically implemented as Hadoop job with driven set up iteration &Check for convergences. The author suggests that if all you have is a hammer, throw away everything that's not a nail [19].

Albert Bifet et.al. states that the fastest and most powerful way to obtain useful knowledge in real time is streaming data analysis which allows organizations to react quickly when problem appear or detect to improve performance. "Big data" is the term referred to huge amount of data created every day. Apache hadoop, apache big, cascading, scribe, storm, apache hbase, apache mahout, MOA, R, etc. are tools used for mining big data. He advised that our ability to handle multiple exabytes of data depends primarily on the existence of a diverse variety of datasets, techniques, and software frameworks [21].

Sameer Agarwal et.al. Presents a BlinkDB, a approximate query engine for running interactive SQL queries on large volume of data which is massively parallel. BlinkDB uses two key ideas: an adaptive optimization framework that builds and maintains a set of multi-dimensional stratified samples from original data over time, and a dynamic sample selection strategy that selects an appropriately sized sample based on a query's accuracy or response time requirements [23].

III. LAYERED ARCHITECTURE - BIG DATA CLOUD

The architecture[16] of big data computing in cloud represented as four-layered model is shown in Fig.3. The cloud infrastructure layer handles the elastic scalable computing, storage, and networking infrastructure. The big data fabric layer addresses the several tools for data management, access, and aggregation. The tools and technologies for data access and processing, programming environments for designing the analytics and scheduling models for execution, and so on are provided in the third layer i.e. the platform layer. The top layer is the big data analytics, focused on analytics usage, and publishing standards to offer them as services. Functional description of each layer is given as follows:

Cloud infrastructure: large-scale management of dynamic and elastic scalable large infrastructure of compute and storage resources as services. Virtualization technologies are used for on-demand provisioning of the resources based on SLAs and QoS parameters. The services in this layer are as follows: large-scale elastic infrastructure to set up big

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

data platform on demand, dynamic creation of virtual machines, large-scale data management for file/block/objected-based storages on demand, ability to move the data in seamless across the storage repositories, and able to create the virtual machines and auto mount the file system with the compute node.

Big data fabric: To connect multiple cloud infrastructures standards, interoperable protocol APIs are offered by this layer. This layer provides tools and APIs through which storage, compute, and application services can be accessed.

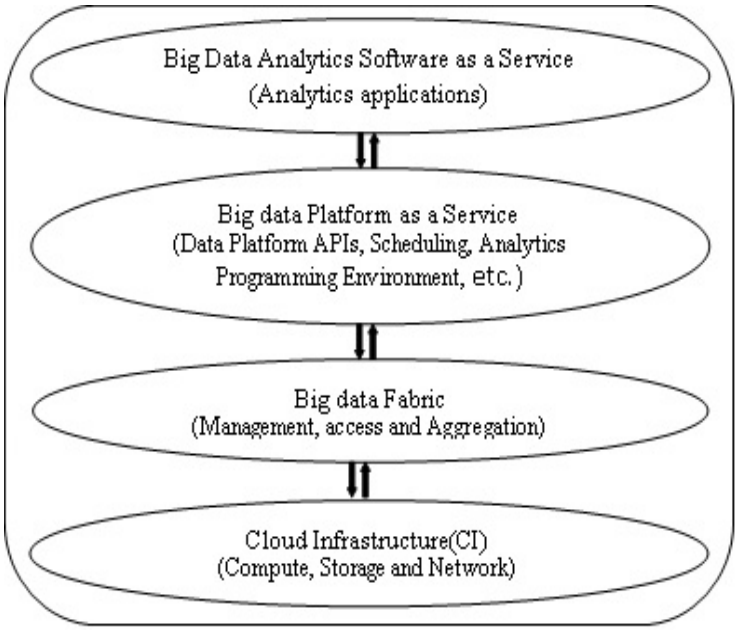

Fig.3: Big Data Cloud Reference Architecture

Big data platform as a service: This basic layer offers several platform services to work with storage/data, and computing services based on SLAs and QoS. This layer consists of middleware management tools such as schedulers, data management tools such as NoSQL tools, and data-intensive programming models for data processing. This layer would mainly focus on development of tools and software development kits (SDKs) that are essential for the design of analytics.

Big data analytics: Big data analytics offered as services, where users could quickly work on analytics without investing on infrastructure and pay only for the resources consumed. This layer organizes the repository of software appliances and quickly deploys on the infrastructure and delivers the end results to the users; the pricing would be computed based on the usage, QoS provided, and so on.

IV. CONCLUSION

With the increase in data in our daily life, it becomes difficult to handle such vast amount of data. To handle such vast amount we have an emerging concept of big data. In this paper, we studied various aspects of big data, its characteristics, applications, how it is used with cloud. Big data analytics with cloud computing helps enterprises of small to large size to reduced commitment of company resources. It provides insights to the business pertaining to improve performance, decision making support, and innovation in business models, products, and services.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

REFERENCES

[1] NIST big data public working group, "Survey of big data architectures and framework from the industry", 2014, http://jtc1bigdatasg.nist.gov/ _workshop2/07_NBD-WD_Big_Data_Architectures_Survey.pdf, 2014 [last accessed 30 April 2014].
[2] https://www.newgenapps.com/blog/what-is-big-data-analytics-benefits-challenges
[3] M. D. Assuncao, R. N. Calheiros, S. Bianchi, M. Netto, R. Buyya, "Big data computing and clouds: trends and future directions", Journal of Parallel and Distributed Computing (JPDC) 2015; 79(5): 3–15.
[4] X. Wu, X. Zhu, G. Q. Wu, W. Ding, "Data mining with big data", IEEE Transactions on Knowledge and Data Engineering 2014; 26(1): 97–107.
[5] S. Kaisler, F. Armour, J. A. Espinosa, W. Money, "Big data: issues and challenges moving forward", Proceedings of the 46th IEEE Annual Hawaii international Conference on System Sciences (HICC 2013), Grand Wailea, Maui, Hawaii, January 2013, pp. 995–1004.
[6] V. V. Mayer, K. Cukier, " Big Data: A Revolution That Will Transform How We Live, Work and Think", John Murray Press: UK, 2013.
[7] J. Ginsberg, " Detecting influenza epidemics using search engine query data", Nature 2009; 457: 1012–1014.
[8] B. Purcell, "The emergence of "big data" technology and analytics",Journal of Technology Research 2013.
[9] A. Chauhan, V. Fontama, M. Hart, W. Hyong, B. Woody, "Introducing Microsoft Azure HDInsight, Technical Overview", Microsoft press: One Microsoft Way, Redmond, Washington, 2014.
[10] Rack space, www.rackspace.com [last accessed 22 August 2014].
[11] Cloudera Hadoop, http://www.cloudera.com [last accessed 03 September 2014].
[12] Google big query, https://cloud.google.com/bigquery-tour [last accessed 15 January 2015].
[13] Amazon elastic MapReduce, developer guide, 2015, http://docs.aws.amazon.com/ElasticMapReduce/ latest/DeveloperGuide/emr-dg.pdf [last accessed 1 November 2014].
[14] J. S. Ward, A. Barker, "Undefined By Data: A Survey of Big Data Definitions", Stamford, CT: Gartner, 2012.
[15] Google-gdata, .NET library for the Google data API, http://code.google.com/p/google-gdata [last accessed 20 February 2015].
[16] R. Kune, P. K. Konugurthi, A. Agarwal, R. R. Chillarige and R. Buyya, "The anatomy of big data computing", Softw. Pract. Exper. 2016; 46:79–105 Published online 9 October 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/spe.2374
[17] S.V. Phaneendra, E. M. Reddy, "Big Data- solutions for RDBMS problems- A survey", 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013).
[18] K. K. Reddi & D. Indira, "Different Technique to Transfer Big Data : survey", IEEE Transactions on 52(8) (Aug.2013) 2348 { 2355}
[19] J. Lin, "MapReduce Is Good Enough?", The control project. IEEE Computer 32 (2013).
[20] A. Bifet, "Mining Big Data In Real Time", Informatica 37 (2013) 15–20 DEC 2012
[21] S. Agarwal, B. MozafariX, A. Panda, H. Milner, S. MaddenX, I. Stoica, "BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data", Copyright © 2013ì ACM 978-1-4503-1994 2/13/04