International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT206613

Coming Together of Big Data and Cloud Computing : A Review

Muneeba Afzal Mukhdoomi[1], Dr. Ashish Oberoi[2], Er. Ankur Gupta[3]
[1]M.Tech Scholar, Department of Computer Science and Engineering, RIMT University, Punjab, India
[2]Professor, Department of Computer Science and Engineering, RIMT University, Gobindgarh, Punjab, India
[2]Assistant Professor, Department of Computer Science and Engineering, RIMT University, Gobindgarh, Punjab, India

Article Info
Volume 6, Issue 6
Page Number: 118-137
Publication Issue :
November-December-2020

Article History
Accepted : 15 Nov 2020
Published : 25 Nov 2020

ABSTRACT

Big data stands for sheer amount of data that is growing unceasingly at a rapid pace. Big Data demands high-powered, robust, reliable, fault-tolerant tools and techniques in order to make it convenient to process, analyse and uproot new insights from Big Data. Big data refers to huge, heterogeneous amount of details, facts and data generating at constantly rising rate. The data sets in Big Data are too bulky or extensive, as a result classical data handling application software are not competent enough to administer them. On the other hand, Cloud computing is a resourceful technology providing high computing power, scalability, computing resources as and when required for processing, storage, analytics and visualization of Big Data. Therefore, cloud computing can be regarded as a feasible and applicable technology which promises to handle Big Data challenges and also provides here and now infrastructures with all the mandatory resources. This paper will mainly review processing of big data cloud using Hadoop and spark in cloud, advantages of driving Big Data using cloud computing and applications of Big data in Cloud.

Keywords: Big Data, Hadoop, Mapreduce, Spark, Cloud Computing

I. INTRODUCTION

"Big Data is a vast amount, high rate and wide array of information asset". This is the most aged definition of big data, first originated by Doug Laney. Big Data simply means large sized data sets both structured and unstructured which is generally counted in zettabytes[1]. Formerly there was the concept of traditional data for which traditional technologies were available to handle. This data continued to grow with blistering speed, thus Big Data took over this traditional data. Since then, a profound transformation took place in management of data and now we have made headway from an age when ledgers and files were used for keeping the record of data to this digital age where tremendous amount of data is structured in a most organized and methodical way. In this new age of Big Data, the approximation of data generated per day is more or less 44 zettabytes which is eye-opening[2]. The way data from distinctive sources like Facebook, instagram, twitter, telecom companies, healthcare industries, business organizations is growing endlessly on a large scale (40TB per second), the jet engine of Boeing creates

approximately 10 TB of data every thirty minutes, the four engines on the jet can produce 640 TB of data which are quite a number and therefore it can be said that big data has almost exceeded Moore's law and hence demands cost-efficient, fault tolerant, scalable processing engine and storage infrastructures[3]. The challenge to store, process, and analyze these large chunks of datasets has compelled many organizations and individuals to take on cloud computing. Cloud computing (CC) is robust technology that carries out analytics, storage and processing of big data at high throughput and thus makes big data productive. For processing, storage and executing large sized big data sets, Cloud computing technology (CCT) is viewed as a fortune-teller[4]. The integration of big data with cloud computing is the most popular and trending technology in today's digital epoch. The elementary framework of Big Data and cloud computing is different. Big Data is a complex and voluminous data while as cloud computing is all about on demand resourceful infrastructure. Cloud computing is a powerful technology that handles big data through complex computing and deals with analysis and systematically extracts value from big data, provides effective management, scalability and reduces infrastructure cost as well. Thus, the integration of Big Data and cloud computing offers a solution which is both scalable and adaptive for Big Data and business analytics; therefore, it is a globally accepted technology. For Example Amazon Elastic MapReduce (EMR) illustrates how the elastic cloud computing (EC2) approach provides a powerful computing framework for processing, storage and analysis of data. It also includes automated resource provisioning according to the demands. Therefore, combination of Big Data and cloud computing can yield money-making results for organizations [2].

1.1 Big Data: Concept and Traits

There is no precise way to describe Big Data as it can be defined in many ways. It was believed that the data from different sources like social media, healthcare industries, and business industries had generated so greatly that the traditional systems and infrastructures were unable to process, analyse and accommodate such large data sets. So, there was a call for high-powered technologies which could operate Big Data and as a result new processing platforms such as Hadoop and MapReduce came into existence which made Big Data unchallenging and productive[5]. The 5 V's of big data are described below:

1.1.1 Volume

The amount of data being generated from different sources is termed as "volume of Big Data."This data from different sources (social media, medical studies, research studies etc.) can be of any format like audio, videos, text, image, etc. For example, the number of Facebook users is nearly 2.7 billion which gives rise to nearly 4 petabytes of data per day. Billions of images, posts, videos, tweets etc are uploaded by people every day [6].

1.1.2 Velocity

Velocity in Big Data refers to the speed of data being generated from different sources every millisecond. The pace at which data is generated is directly proportional to the amount of data that is stored. Thus, the term velocity is not confined to speed of incoming data only but also manages the flow of data being received from incoming sources at a very fast rate [6]

1.1.3 Veracity

In Big Data, veracity of data stands for exactness or accuracy of data. The data is obtained from different sources like Facebook, LinkedIn, Twitter, etc. Hence it is hard to differentiate between outdated and accurate data. We can't rely on all the data coming to us from different sources, therefore, normalization of data is conducted. It formats all data sets and lessens

data redundancy. As a result, this data become reliable enough for sales and business decisions [6].

1.1.4 Variety

Big data can be structured, unstructured and semi-structured; data can be of any format like email, PDF, image, video, audio, etc. and is generally called as "heterogeneous" data. Therefore, due to the variety in data, there occur a lot of unavoidable errors which is why combination of data through traditional tools is not possible [6].

1.1.5 Value

The value in Big Data opens up a way to extract insights from big data in order to make decision-making easy and hence enlarges the success of business companies [7]. Each set of big data contains a hidden treasure called "value" which upon proper processing and analysis is intricate from Big Data, it acts as a turning point for the market, business and society in general[6]. The better way to acknowledge the attributes of Big Data is to classify it further as follows .[7]

TYPES	NAME OF SOURCE	EXAMPLES /FRAMEWORK
DATA SOURCES	Social Data Search Data Machine Data / IoT Crowdsourced data Transactional data	Facebook ,Twitter ,Youtube Amazon ,ebay ,Wikipedia,quora Smartphones ,GPS,Sensor data Open StreetMaps,Flicker,Picasa,Instagram Fianancial credit card payments,invoices,
CONTENT FORMAT	Structure Data Unstucture Data Semi-Structured Data	WeblogStatictics,spreadsheets,barcodes Email,chats ,IoT sensor ,satellite imagery XML ,tab delimited files ,JSON
DATA STAGING	Cleaning Categorization Normalization	RapidMiner ETL
DATA PROCESSING	Batch Processing Real Time Processing	MapReduce system simple scalabe streaming system (S4),storm
DATA STORAGE	Key-Value Column-oriented Document Graph	Scalaris ,Redis ,Berkeley DB HBase, Hypertable ,BigTable MongoDB,Terrastore,RethinkDB HyperGpraphDB,Neo4j,AllegroGraph

Figure 1. Big Data Classification

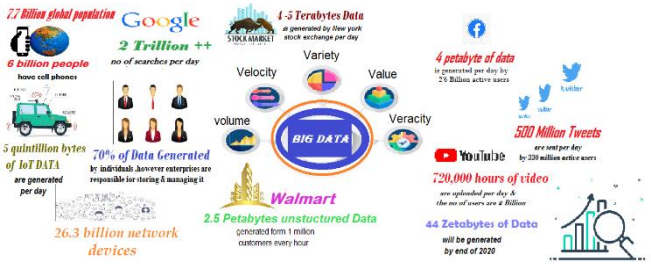

Figure 2. Big Data Info gram

1.2 Cloud Computing:

The word "cloud" in cloud computing is an analogy for "the internet", hence Cloud Computing is a computing system where storing and accessing of data is done over the internet. The key concept of cloud computing was first chalked out by John McCarthy in the 1960s. McCarthy believed that "computation may someday be organized as a public utility"[8]. According to the IEEE computer society, "cloud computing is a paradigm in which information is permanently stored in servers on the Internet and cached temporarily on clients that include desktop computers, entertainment centres, table computers, note books, wall computers, handheld devices, sensors and monitors"[9]. In disparity to the former computing like grid & cluster computing, cloud computing is a service-oriented computing and not application-oriented[10]. Cloud computing provides its users with services and resources which includes data storage, servers, databases, networking and computing power. Cloud Computing has become adored and more preferred due to its files storing system in remote databases instead of hard disc drives or local storage devices. As a result, one can access his data anywhere in the world with any device that is connected to the internet [11][12].Following are the services offered by Cloud:

Iaas: It is also known as hardware as a service (HaaS). It is used by system administrative. IaaS is one of the types of service models in cloud computing. It contributes IT infrastructure like hardware, servers, virtual machines, network resources etc. to the customers. Here "pay-as-per use" model is used in order to invoice the customers for usage of services. Examples are GoGrid, Aws (EC2)[13]

Saas: It is also known as "on-demand software" or "web based software" , used by end users. This technique of software delivery is wholly contingent on internet. In this model, software is accredited to customers based on subscription instead of buying it.

Also, it relinquishes customers from installation and running of applications on their personal computers. Examples are Google Apps, Dropbox, GoToMeeting [14]

Paas: This is particularly used by developers. Here customers are provided with a platform for developing applications like tools, programming languages, run time environment which together can be used to develop, run and manage an application on cloud. Therefore, clients need not to buy or install the software and hardware on their local computers. Examples are Microsoft Azure, Manjrasoft Aneka[14]

1.2.1 Definitions of cloud computing:

In order to comprehend the concept of cloud computing, let us have a look at some definitions of cloud computing given below:

"A Cloud is a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provision demand presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers"[15]

"A model for enabling ubiquitous convenient, on-demand network access to a shred pool of configurable computing resources (networks, servers, storage, application and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [16]

Cloud Computing Deployment Models:

Public cloud: Public cloud provide services which are open to all and are deployed globally over the public internet; it is handed by third force cloud providers. The services like virtual machines (VMs), applications or storage present in public cloud is either free or available inexpensively on demand to general public or group of industries Eg. Azure. Some advantages of public cloud are a lower cost, no maintenance, high reliability [2][17][18]

Private Cloud: Private cloud services are highly secured and its access is confined to certain authorized organizations and users only, it is deployed locally over the private network. In Private Cloud hardware and software specifically belong to an organization. Also private cloud services are hosted by single organization Eg. Hewlett Packard Enterprise (HPE). Some main advantages of private cloud are flexibility, security and high scalability[2][17][18]

Hybrid Cloud: It is a combination of both private and public cloud. It inherits the traits or advantages of both private and public cloud, therefore, maximal benefaction can be achieved from hybrid cloud model. The most common benefit of hybrid cloud is its "cloud bursting strategy", due to which data and applications are portable to public cloud as per the requirements of the organizations. Examples of some hybrid providers are Amazon, Microsoft, Google, and Cisco. [2][17][18]

The paper is further organized as follows: Sec. II take up an outline on integration of Big Data and cloud computing, Sec. III undertakes the work done in this field, Sec. IV presents the predictive analysis of Big Data. Advantages of putting Big Data in cloud are reviewed in Sec. V. Case studies in Big Data province is presented in Sec. VI, VII concludes the paper.

II. Unification of Big Data and Cloud Computing Technology

With the profuse development in technology, Big Data sets are increasing at blistering pace and therefore needs scalable & high-speed data platform. On the other hand cloud computing is such technology which provides wide-ranging space for storage of data sets. Cloud computing technology also combines all the distributed IT resources present in different locations into one virtual domain or

platform which helps in analytics of big data in order to make strategies for better business decisions. Consequently, it can be said that coming together of Big Data and Cloud Computing turns out to be complementary to each other [12]

Cloud provides Storage for Big data:

International Data Corporation (IDC) visualizes that Big Data size is approaching 44 Zettabytes by the end of 2020 which is scarcely possible to store. However, cloud storage comes into rescue which hands over unlimited distributed storage support to Big Data Eg. Google file system (GFS) and open Hadoop Distributed File System (HDFS) of GFS developed by Hadoop team [19]. A Cloud friendly NoSQL database was developed to address the Big Data storage issue. NoSQL stores massive data sets on multiple servers[12][16] .Cloud Computing technology comes up with scalable storage through which users directly accesses Cloud service providers (CSP) in order to dynamically align storage requirements of their data. Data Storage as a service (DaaS) is one of the solutions provided by cloud where users are granted with access to a data storage system. Cloud storage has ability to act as backup in case a malfunction and/or crash happen to the local system hard drives [12]

Computing of Big Data using Cloud:

In order to fetch the valuable insights from Big Data sets, data should be analysed appropriately to make proper business decisions. This problem can be skilfully dealt with, by taking on a bunch of more and more RAMs and CPUs into cloud computing technology. High Performance Computing (HPC) is momentous for data analysis. It includes many computing models like grid computing or cloud computing, favourable enough to tackle big data computing complications [12]

Big Data processing using cloud computing technology:

In order to process Big Data, data sets are extracted from multiple database servers situated across the universal locations. Maintaining all such servers stationed in different locations is a costly process for an organization but switching to CC turns it into a cost-effective process.CCT is a powerful platform & can process data through its distributed virtual servers placed everywhere across the globe. Therefore, it reduces the cost of big data processing remarkably. Besides this, CC provides many computing resources required for big data processing like faster CPU's, larger disks and RAM's[2]

Processing of Big Data using HADOOP-MapReduce in Cloud environment

"Hadoop in cloud" means resource pool provided by Cloud system. It is used to run Apache Hadoop clusters. Following are the reasons for using Hadoop in cloud: Organizations are in need of Hadoop clusters but have no space for physical servers therefore CC comes into picture. Using Hadoop in cloud due to the flexibility provided by cloud in order to make changes according to need. The time for transmission of data over network gets saved if analysed data is stored in cloud. In order to utilize functions (computing, networking and storage) properly, Hadoop cluster must run in cloud. Running Hadoop in Cloud means little time, money and efforts to take care of it [20]. Transmission of analysed data (already stored in cloud) into Hadoop is exempted & hence saves time when Running Hadoop in the same cloud takes place. Running Hadoop in Cloud includes "pay only for time you use" while one needs to pay for maintenance of Hadoop servers irrespective of the usage hours. Running Hadoop in cloud means End users does not require to purchase hardware and still CC furnishes them with whatever resources they want to be served with [21]

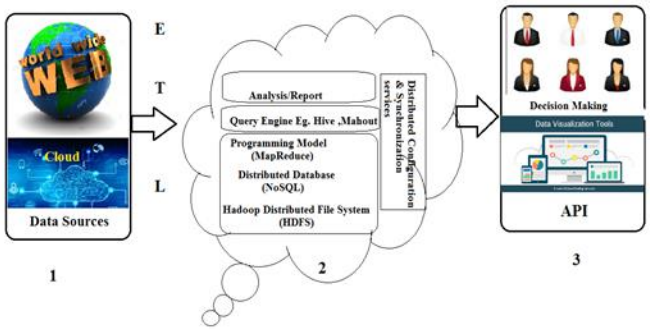

Figure 3. Use of Cloud in Big Data processing using Hadoop

The use of cloud computing in big data is showed in Fig. 3. The data which originate from distinctive sources like social networks, internet of things (IOT), business organizations is stored in cloud computing [22]. Hadoop is a java based open source software used for processing and storing of Big Data. In this module, the Data already stored in cloud undergoes Extract, transform, load (ETL) process where this data are extracted from cloud storage, transformed into a particular format and finally loaded into a fault-tolerant database like NoSQL [4]. HDFS is a distributed file system that stores massive data sets over number of nodes in a set. HDFS includes two types of nodes namely Name Node (Master) and Data Node (slave). Name Node remains updated about the directory (metadata) and location of all files in the cluster. Data node stores actual data in HDFS. A large input data set is segregated into many small data sets called "blocks" of size 64 MB. Name node performs mapping of blocks to the data nodes. Data nodes including blocks mark their presence to name node while they set into operations. Also, name node performs replication factor property in case a Data node crashes. The processing of this data occurs through MapReduce [4]. MapReduce is also called as heart of Hadoop. The processing of data is done by MapReduce. MapReduce is a programming module. The working principle of MapReduce is based on the game plan of "split-apply-combine". It works in two stages that is Map and Reduce. MapReduce includes one master Job Tracker and one slave Task for each cluster. Job Tracker handles job resource management and also schedules jobs for Task Tracker. Task tracker traces the tasks across the cluster and keeps the Job tracker updated about the progress. When job Tracker stops receiving signal or heartbeat from Task Tracker, it considers the same as task failure. Therefore Job tracker is responsible to re-schedule the job. [23].The output data from previous module undergo data visualization to represent this data in the form of graphs, charts or maps. Data visualization method facilitates it to comprehend the trends, fashions and pattern in data for complex decision making which directs the probable success in business organizations [22]

Processing of Big Data using Spark:

Researchers felt need of the moment for running Apache Spark on top of the Hadoop, due to the downside of Hadoop Apache. Spark is a freely accessible cluster computing platform or a data processing engine developed in order to provide faster processing and analytics than Hadoop. Spark working principle is based on MapReduce Algorithm. The most prominent high-point of the spark is "in-memory computation" that is processing of large data sets are being done within the RAM which makes it 100 times faster than Hadoop. Spark has been developed to handle wide range of workloads individually like batch processing, machine learning, graph data, structure data, streaming data with its own components. In order to process structured data or machine learning in Hadoop, one needs to install Hive or Mahout. Spark can be merged with other big data tools like Hadoop, Hive and Cassandra, etc. Spark is fault-tolerant which indicates that in case of any failure/bug /error in the system, spark has the power and potential to recover itself and continue work on its operations [24] The six components of the Spark are Cluster Manager (base), Apache core, The six components of the Spark are Cluster Manager (base), Apache core, Resilient distributed datasets,

Spark SQL, Spark streaming, Machine learning library (MLIB), GraphX [25]

Apache Core: This is the main engine of Spark. It performs services of the Spark like scheduling, distribution of tasks, and monitoring of application. BDAS, the Berkeley Data Analytics Stack, open source software which combines all the software components of the spark together and magnifies the power of the Apache Spark [25]

Resilient Distributed Datasets (RDD): Apache Spark is all about RDD. RDD for Spark is what MapReduce is to Hadoop. RDD defines the basic organization of data or how information is being processed in Spark. Resilient in RDD indicates that this data structure is fault resistant, distributed means data sets collected are divided and distributed among different servers. RDD can only be recreated but never be amended (immutable) [26]

RDD Operations:

Transformation: This function is executed on demand. The function creates new data sets from already existing RDD datasets, the output of transformation function is RDD which varies from the parent RDD Eg. Dataset group by key, map, pipe, flatmap and reducebykey etc are some transformation functions in Spark [26]

Action: This operation is responsible for execution in Cluster. It processes RDD and the given non-RDD values as output are stored in external storage. Some action functions are count, first, reduce, countbykey, collect, count, first, take, and for each etc[27]

Directed Acyclic Graph (DAG): This component demonstrates RDD operations in visual form. Here, vertices symbolizes RDDs and edges stands for RDDS operations [27]

Spark SQL: This segment of Spark helps in processing of Structured Data. It processes SQL queries using Data Frames programming framework [27]

Spark Streaming: With the help of Spark Streaming component, one can have fault-tolerant and high-yielding system for processing a continue delivering live stream digital data [27]

Mlib: This is a virtual library of Apache Spark which contains many machine learning algorithms like “classification”, “regression”, and “clustering” [27]
GraphX: Spark GraphX is used for faster processing of data that is organized using graphical representation called “Graph Data” [27]

Benefits of Apache Spark in cloud:

Azure Databricks: Databricks was developed by one and the same founder team of Apache Spark in the year 2013. The term Databricks can be simply referred as “Apache Spark dipped into a cloud service.” Databricks is the most figured out and gushing platform for operating Apache Spark, it includes Apache Spark basic infrastructure in addition to its new updated architecture in the most desirable way possible. With Databricks on, users need not to worry about the writing software and technical chaos like configuration files, framework and building blocks instead Databricks provide a clickable platform where user just need to set the cursor on a perimeter using mouse to get started and Databricks work wonders under the hood. Databricks is the cloud boosted platform which makes Apace Spark more productive and useful, it includes features like “pay-as-you-go model” which means users pay for only consumed services and if they need the same services again next time, they can pay then only, “role based access control” that is Databricks sets a margin between users like who can access Azure’s resources and areas. Databricks is extremely handy for organizations relying on Spark and helps them to dive into the Spark quickly, it assembles all the big

datasets into "notebooks" which makes it easy to analyse the big data. Also, Databricks plays a pivotal role in order to extract meaningful insights from Big Data [27]

III. Literature Review

This section defines the work done by Researchers in the related field.

Khaled Alwasel, et.al presented a simulator tool called "BigDataSDNSim." It helps in modelling and simulation of Big Data Management systems like Yet another Resource Negotiator (YARN), programming models (MapReduce) and software defined networks (SDN) within cloud environment. This device imitates three technologies like SDN cloud, stream-BDMS, MapReduce into a single simulator. BigDataSDNSim is an extension of cloudSim. CloudsimSDN and IOTSim, that comes up with basic framework for survey and evaluation of Big Data applications in SDN based cloud data center.In this study, IOTSim initially designed for simulate MapReduce model is altered in terms of its logic and code to make it up to mark for Big Data management. This study demonstrates that network transmission in case of SDN based cloud data centre is enhanced by 41% in comparison to traditional network, the average job accomplishment time is upgraded by 24 % and power consumption is reduced by 22% [28].

Qingchen Zhang, et.al describes privacy preserving high-order possibilistic c-means (PPHOPCM) algorithm for clustering of big data with cloud computing. Big Data is a large voluminous heterogeneous data which makes it a herculean task to cluster Big Data. Due to increase in run time for big sized data sets, many algorithms (graph-based co-clustering, non-negative tri-factorization algorithms) designed could not achieve clustering of big data productively. In order to solve this problem a privacy preserving high-order PCM scheme (PPHOPCM) executed in BGV homomorphic encryption technique for big data clustering is designed in this paper which also reduces the threat of divulgence of private data that occurs in Distributed high-order possibilistic c-means (DHOPCM) algorithm. It is concluded in this study that PPHOPCM scheme is appropriate for clustering of heterogeneous large scale big data keeping in view the privacy concerns as well [29].

Hergen, et.al proposed a system for auditing and operating the wind farm using digital twin (DT) through augmented reality (AR). The system landscape based on cloud helps the users in complete analysis of wind farm. Wireless Experimental Centres (WEC) consists of many sensors. In order to collect all data from sensors through Message Queuing Telemetry Transport (MQTT), Raspberry Pi computers are exerted into WEC protocol. This is transferred to secure copy protocol (SCP) in a single stream. IOT micro service located in SCP receives this data sent from WEC, Load it into SAP HANA (SAP High-Performance Analytic Appliance) database, JAVA micro services located in the same SCP extracts the data from SAP HANA and supply it. Gateway server which acts as a gate between two nodes, it transfers this data into enterprise central component (ECC) backend which is located in the same cloud. Now visualisation of data occurs in order to analyse this data for wind farm through AR. Microsoft HOLOlens (smartgalsses) connected to IT Landscape via LAN is required for user to see-through the visualization of data. This technique has been enforced to many wind farms of different countries and it was feasible to keep track of all wind farms uniformly [30].

Dimple Tomar, et.al represented an ideal Model through which they distinctly explained how cloud is used in Big Data. There are many origins which give rise to large unorganized data sets, all this Data is being stored in cloud .In Hadoop, Hadoop Distributed File System (HDFS) is in charge of baseline data and record for metadata framework which are to be used

further for extracting valuable statistics and figures. MapReduce is the heart of Hadoop, which is responsible for splitting of whole input data into small segments and applies programming model to all the small data sets coequally at a time. Hive is a Hadoop based tool used for data processing and analysis, it comes up with SQL based user interface for requesting data from various database tables for data processing and analysis. Mahout is a framework that uses Hadoop library to fulfil all the growing demands in cloud. In conclusion data is visualized using tools in order to serve the examined data in the form of charts and reports for decision making purpose [22].

Vidushi Vashishth, et.al presented predictive approach for scheduling tasks in Cloud computing to overcome the flaws that occur when big data is processed in cloud environment. While processing of big data in cloud, Map phase gives rise to a number of keys that MapReduce cannot withstand and also fails to respond in a timely manner. However, in this study, a well organized approach has been presented for task scheduling in cloud environment in order to reach the best results. This paper includes working of practical swarm optimization(PSO) algorithm with certain classifier algorithms Naive Bayes classifier, random forest classifier and K nearest neighbour (kNN) algorithm to predict the best suited virtual machine (VM) for Big Data sets, PSO algorithm do not need to iterate the computing of the components in the swarm therefore makes it much faster. The outcome of experiments in this study concludes that this approach performs the job of task scheduling 100 times swiftly as compared to the conventional algorithms [31].

Jun Hu, et.al outlined the role of big data and Cloud computing for development of Internet of Energy (IOE).Internet of energy means enhancing and automating the manufacturing process of energy framework for producers. The energy internet operations are based on “controllability” and “observability” that in turn needs fast analysis of data. In this study, Lambda architecture is used for high throughput processing of big data, it includes three layers namely batch layer, speed layer and serving layer. Batch layer stores query results ahead of time. So that when query actually needs to be executed, results can be acquired directly from batch view. Speed layer handles real-time streaming data; also it reads the data in patch wise fashion and updates its speed status. Energy data coming from distinctive sources are of different formats like structured, unstructured and the unprocessed data. MySQL and postgreSQL databases are used for storing structured data; unstructured data is stored in NoSQL databases (Hbase, MongoDB, and Cassandra).The unprocessed data coming from energy internet cloud computing framework is stored directly in MongoDB. Once this data is processed using Spark or MapReduce engine, results are transferred back to MongoDB from HDFS via Kafka. Kerberos authentication protocol is used for security of data. Energy internet also begun cloud engine services. It takes care of backend applications. Users are able to modify “business logic” due to cloud functions. Background tasks helps in processing of data. [32]

ZHANG Yaoxue, et.al surveyed the models of cloud computing that are used to address the issues of big data. They shed light on solutions promised due to integration of big data and cloud technologies like Google File System (GFS) and Hadoop Distributive File System (HDFS) for file storage system. NoSQL based database engines (BigTable, Dynamo, Cassandra, Hbase, Hypertable and MongoDB) for big database management, MapReduce and Spark for Big Data Processing. This study also discussed two cloud computing models like Fog computing, Transparent computing that resolves the challenges arising in cloud due to Internet of Things running into Big Data [33].

Chtaintureena Thingom, et.al depicted the significances of combining Big Data and CC. The union of Big Data and CC supports scalability, rapid processing of immense data sets, speedy query processing and decreases the expanses for substantial infrastructure. Big Data and CCT are the fastest emerging technologies in IT these days which make it easy to bring out the insights from big data and also impacts the business firms in a huge way [34].

Sreekant Rallpallia, et.al published regarding integration and analysis of healthcare big data in cloud environment using Hadoop framework. Healthcare industry is one of the biggest sources of big data which should be analysed for keeping documentation details of patients, their complete medical records, and diagnosis for treatment. Security and confidentiality is another issue related to healthcare data, the data should be accessible to only certain amount of people concerned to it. In this study it is forecasted, the integration of big data and Cloud computing security model containing "identity-based proxy re-encryption" (IB-PRE) should be used by healthcare providers. The main aim of IB-PRE security method is to acquire the identity credentials like E-mail or IP address from the users; as a result data remains secure and confidential. Therefore, it is concluded that combination of big data & cloud computing framework proves to be beneficial for healthcare industry [35].

Liwei Kuang, et.al proposed a model called "T3R5" which includes three tensor and five processes. T3R5 is initiated to solve the challenges related to incorporation of fast growing volume and variety of Cyber-Physical-Social System (CPSS) data like extraction of highly valuable information from small sized data, setting up relations between heterogeneous data objects by estimating analogy between them and to update ranking result after measuring the priority of CPSS data etc. Data representation (R1), the first process of T3R5 is used to convert heterogeneous CPSS data into low-rank tensors that are shaped as per their previous format into components in a tensor space. Dimensionality Reduction(R2) make use of Higher-Order Singular Value Decomposition (HOSVD) to root out the superior data generated in tensors which is small scale but highly informative. Relation Establishment (R3) as the name suggests is used to map relations between CPSS data items. Data Rank (R4) includes two algorithms; the Multi-linear algorithm is used to determine the rank between CPSS data items and an incremental algorithm is used to renovate the ranks using build up data. Data retrieval (R5) helps in extraction of informative data items used for powerful application. Additionally, three tensor types (Td, Tf, Ts) and a rank vector (Vr) are used to represent the data characteristics, distinguished features, multi-aspect similarities and data ranks. In this study, investigational results convey that the proposed model is workable to integrate the CPSS big data on cloud [36].

K.Nithiya, et.al discussed combination of cloud computing and Big Data for keeping a check on black money holders. In order to deal with the complexity of Big Data, Fast Range Aggregate Query (FastRAQ) is proposed that provides high speed estimated results for Range-aggregate queries in Big Data world. The whole Big Data is broken down into separate data sets using balanced partitioned algorithm, for each data set a local estimation sketch is generated. When a range aggregate query request takes place, FastRAQ acquires result directly by summing up local estimates from all portions. The balanced partitioning algorithm uses stratified sampling model which involves division of whole data into groups based on their attributes and further partitions groups into subgroups keeping in view the number of servers available and data distribution strategy. In this study, particulars and transactions of user's bank account maintained in several different banks are tracked with the aim of keeping an eye on black money

holders so that administration can trace them effortlessly [37].

Venkata Narasimha Inukollu, et.al focused mainly on security issues of cloud computing correlated with Big Data. In this study various security measures which would improve the security of cloud computing have been discussed. The proposed solution motivates the use of many technologies to diminish the security problem. File encryption, since machines contain data in cluster form which is more liable to be hacking. As a result, stored data should be encrypted. Network encryption, the remote procedure calls (RPC) should happen over secure socket layer (SSL) with the purpose that securing useful information in data. Logging, all the users in charge of MapReduce jobs should be logged and also these logs are required to be audited everyday in order to keep an eye on malicious user, if any. Software Format Node Maintenance, all those nodes responsible for running software is to be formatted systematically to keep the system secure and updated. Nodes Authentication, all new nodes joining the cluster should be authenticated using Kerberos technique to limit the entry of any evil node. Rigorous System Testing of Map Reduce jobs, all map reduce jobs written by developers should be tested in a distributed environment to ensure its robustness. Honeypot nodes, which are actually a Trap node, should be implanted in the cluster, in order to trap the hacker with ease. Therefore, it is concluded, since Cloud is a widely used technology in industry and research fields. Using proposed approaches, cloud environments can be a secured platform for complex business performance. Thus, security is of paramount importance for organizations running on these cloud environments [38].

Ibrahim, et.al discussed uplift of Big Data through cloud. There are many organizations where Big Data is used along with cloud computing technology such as A.Swiftkey needs to handle multiple terabytes of data and therefore uses Apache Hadoop which runs on Amazon S3 & Amazon EC2. The biggest online travel company named RedBus uses Google's BigQuery for processing of large sized data sets. For Twitter Amazon cloud infrastructure is implemented for data acquisition and data analysis [4]

Linquan Zhang, et.al propounded and advocated two algorithms called "online lazy migration (OLM)" and "randomized fixed horizon control (RFHC)". These algorithms are used for upgrading the routes and options for data centres in order to pass on the data to cloud environment. Competitive analysis is being conducted for comparing performance of online algorithm to offline algorithm. Experimental and graphical results shows that the OLM algorithm achieves a worst-case competitive ratio lower than 2.55 while as RFHC provides an increased size of lookahead unit with the decreased competitive ratio [39].

Rainer Schmidt, et.al proposed a framework for strategic alignment of cloud base architecture for big data. This framework is an outcome for integration of two models that is service model (Iaas, Paas, Saas, Daas) and deployment model (hybrid, private, public) used for implementing big data pipelines. In DaaS pipeline, high degree of measurability and quickness is attained. On the other hand, downsides of DaaS leads to public or private cloud service big data pipeline that builds up the efficiency of a venture. Therefore, it is concluded the use of cloud computing services and models to initiate Big Data in organizations is profit-making particularly for small medium enterprises (SME). This proposal lowers the cost for executing Big Data in SME and also upgrades the status of IT companies [40]

Table 2: Table of Comparison

Author	Year	Description	Technique	Results
Khaled Alwasel , Rodrigo N.Calheiros , Saurabh Garg, RajkumarBuyya, Rajiv Ranjan.	2019	An approach is proposed for simulation and modelling of Big Data applications in SND based cloud data centres	"BigDataSDNSim" platform is used to carry out modelling and simulation of big data applications in cloud computing environment	In comparison to legacy network, average network transmission time improved by 41%, job completion by 24% and power consumption reduced by 22%
Qingchen Zhang, Laurence T. Yang, Zhikui Chen, PengLi .	2018	An Algorithm is presented for heterogeneous clustering of Big Data with BGV security technique	"PPHOPCM" algorithm with BGV technique is applied for clustering large amount of data securely on cloud	Experimental results shows that PPHOPCM can conveniently cluster a large amount of heterogeneous data sets in cloud computing without the threat of disclosing private data
Hergen Pargmann DörtheEuhausen, Robin Faber.	2018	A scheme is suggested for auditing and operating the wind farm	Auditing and operating of wind farm using digital twin(DT) through Augmented reality(AR)	This technique has been enforced to many wind farms of different countries and it was feasible to keep track of all wind farms in parallel

Dimple Tomar , MandeepTomar	2018	A model is proposed to demonstrate the combination of Big Data and cloud computing for smart generation	A three sectional module containing all necessary components for processing ,storing and analysis of big data in cloud environment	It is concluded that integration of big data and cloud computing technology renders the opportunities for business organizations by providing them with critical business opportunities ,als o offers many computing resources for driving big data in cloud computing
Vidushi Vashishth, Anshuman Chhabra , Apoorvi Sood.	2017	This paper presents predictive approach for scheduling tasks in Cloud computing and to overcome the flaws that occurs when big data is processed in cloud environment	practical swarm optimization(PSO) algorithm with certain classifier algorithms Naive Bayes classifier, random forest classifier and K nearest neighbour (kNN) algorithm is used to predict the VM best suited to schedule tasks for datasets	It is concluded that this approach performs the job of task scheduling 100 times swiftly as compared to the conventional algorithms.
Rui Fu, FengGao, RongZeng , Jun Hu , Yi luo, Lu Qu	2017	This paper outlined the role of big data and Cloud computing for development of Internet of Energy(IOE)	A combined platform of Big Data (Hadoop, lambda architecture, Hbase, MongoDB, Kerberos)and cloud computing services for Energy Internet	A framework containing both big data and cloud computing platform satisfactory for characteristics of energy data

ZHANG Yaoxue, REN Ju, LIU Jiagang, XU Chugui, GUO Hui, LIU Yaping	2017	A survey of Solutions to the Challenges arising due to 5v's of big data	A framework for big data in cloud computing which includes file management systems(HDFS,GFS),database management (Hbase,BTable,MongoDB,Cassandra etc),processing tools (MapReduce ,Spark) ,Query systems(Jaql,Pig,Hive). Also, this study surveys fog computing and transparent computing, to support the big data services of IoT	It is proved ,cloud computing and its related computing technologies can productively minimize the issues arisen By big data.
Chintureena Thingom, Guydeuk Yeon	2016	In This paperwork is done on Significance for combination of big data and cloud computing	Here, Integration of Big data (BD) and cloud computing technology (CCT) has been the focus of much debate, also various services offered by cloud computing to Big Data .	It is concluded that big data and cloud computing are complementary to each other.
Sreekanth Rallapallia, Gondkar RRb, Uma Pavan Kumar Ketavarapuc .	2016	A cloud based framework has been propound for processing of healthcare Big Data	A framework based on cloud computing technology is used for processing and analysing of healthcare Big Data .In order to keep this data secure and private, ID-based proxy Re-encryption(IB-PRE) schemes and Identity based on signature encryptions.	It proves to be remarkable for patients by enhancing the standard of healthcare, by providing proper diagnosis ,treatment and care .
LiweiKuang, Laurence T. Yang, Yang Liao	2015	This paper presents a model for methodical integration of heterogeneous data generated from cyber, social and physical space.	This paper proposed a model "T3R5" that contains five processes and three tensor types to solve the challenges related to incorporation of fast growing volume and variety of Cyber-Physical-Social System (CPSS) big data.	Observational results shows that the proposed model is feasible to integrate the CPSS big data on cloud.
K.Nithiya, S.Balaji	2014	A method is discussed in order to keep check on black	"FastRAQ" is proposed that provides high speed estimated results for Range-aggregate queries in Big data world	This scheme can seize black money holders bank accounts ,so

		money holders bank accounts		that government can trace them easily
Venkata Narasimha Inukollu , Sailaja Arsi Srinivasa Rao Ravuri	2014	This paper includes a study of Security issues of cloud computing related to Big data	Besides discussing many security solutions by different researchers ,a new approach containing many security solutions for big data in cloud .	Cloud environments can be secured for complex business operations.
IbrahimAbakerTargioHash, IbrarYaqoob , NorBadrulAnuar , SalimahMokhtar , AbdullahGani , SameeUllahKhan	2014	In this paper ,rise of big data in cloud computing is discussed	This study demonstrates processing of big data using Hadoop- MapReduce cluster ,various services of cloud used for big data processing and analysis .This study also sheds light on applications of big data and cloud as a pair in industries (Redbus) and online portals (twitter)	Many challenges of big data can be addressed by running big data sets over cloud computing infrastructure .CC reduces the need for maintaining expensive infrastructure and performs large scale computing
Linquan Zhang, Chuan Wu, Zongpeng Li, ChuanxiongGuo, Minghua Chen, Francis C.M. Lau	2013	This work studies effective minimum cost approach for loading Big Data into cloud	This paper proposes two online algorithms called an online lazy migration (OLM) and randomized fixed horizon control (RFHC) for upgrading the routes and options for data centres in order to pass on the data to cloud environment.	Observational results shows that the OLM algorithm achieves a worst-case competitive ratio lower than 2.55 while as RFHC provides an increased size of lookahead unit with decreased competitive ratio
Rainer Schmidt Michael Möhring	2013	This study discusses the methods for improvement of strategic alignment of cloud -centric architecture for Big Data	This framework is an outcome for integration of two models that is service models(Iaas,Paas,Saas,Daas) and deployment models (hybrid ,private ,public) used for implementing big data pipelines	This proposal lowers the cost for executing Big Data in SME and also upgrades the status of IT companies.

IV. Big Data Predictive Analytics

- By 2025, more than 50 Billion devices capable of carrying out autonomous computing will be unfolded in the world in order to gather, examine and share Big Data.
- By 2027, the profit(s) for software and services are intended to increase globally from $42 Billion to $103 billion.
- Most of the big data has been generated from the past two years which is estimated nearly to be 90% of the whole data present on planet today.
- As per International Data Corporation (IDC), the treasure trove of Data is going to outstretch to 175 zettabytes by 2025.
- The further enhancing of Data and information exponentially will lead to many challenges like analyses and understanding of this massive data and will migrate to the cloud for its better solutions.
- International Data Corporation (IDC) predicts that nearly 30% present globally will be instantaneous by 2025.
- In forthcoming years Big Data will be in highly inflated demand for business, market, finance, healthcare, manufacturing and other industries. The exploration for assets within big data is in progress.
- Big Data will emerge as a revolution for companies. Due to Big Data insights the companies generating data on peak will become highly data powered.[41]

V. Case Studies

Red Bus: Redbus, one of the biggest online ticketing companies established by phanindra Sama, charan padmaraju and sudhakar pasupunuri in 2006 Redbus contains more than 2300 bus operators providing services to nearly eighty 80000 routes. Each Process for booking produces huge amount of data stored in traditional database analysed using Hadoop framework that could not bring ease to the company's requirements for fast analysis of data and low investment plan. Therefore GoogleBigQuery is need of the hour. It scans large volume of data in (TB) seconds and (PB) minutes and provides quick services to the Redbus Customers. Also, cloud service models used for Redbus is IaaS, PaaS [42]

Mining Twitter in Cloud: The analysis of large volume of datasets from Twitter is conducted using cloud computing technology. PageRank algorithm is applied to whole user base of the Twitter in order to obtain the user rankings. The computation process of the Twitter is done in two phases that is crawling phase and the processing phase. Crawling Phase includes mining of all data sets from the twitter while as in processing phase PageRank algorithm was employed to determine all the received data. A graph accommodating nearly fifty billion nodes (users) and two billion edges (replies, retweets etc.) were web crawled in a crawling phase. Thus, it is concluded that Amazon cloud infrastructure provides an inexpensive solution for data analysis and collection [43]

Nokia: Nokia is a well known multi-national telecommunication company. Thousands of users use Nokia handsets to communicate, share data and capture photos .Hence, Nokia accumulates a lot of Big Data. In order to storage, analyse and extract value from this Data Nokia uses Hadoop petabytes-scale cluster which is interconnected to teradata enterprise data warehouse (EDW), oracles and MySQL data marts[44]

VI. Advantages of Cloud Computing for Big Data

The combination of Big Data with cloud computing technology is known as "Big Data clouds" There are

various worthy causes for computing big data using Cloud computing. The benefits of Big Data Clouds are:
a) Resilience: Cloud Computing technology has the potential to withstand with the challenges of Storing Big Data as per the demand which can be either increasing or decreasing storage space after the insights are taken from the big data sets
b) Processing of Big Data Sets: Due to large size of big data sets it is complex to process such massive data faster, Cloud computing technology makes it possible to analyse and process the big data sets swiftly
c) Instant Infrastructure: Cloud computing technology comes up with a fast infrastructure for Big Data analytics with extensible domain which on-the-other-hand companies had to set up themselves
d) Inexpensive Big Data Analytics: For Big Data Analytics companies needs to maintain a data centre which swallow company's Big Budget, shifting Big Data to Cloud Computing results in low-price data Analytics.
e) Reduced Complexity: In order to execute solutions or technique (Hadoop, Spark etc.) for Big Data analytics, processing and visualization many components are needed while as cloud works outs all these components simply and hence reduces the complexity of computing Big Data sets [2]
f) Cloud as Storage: Storing Big data in cloud refines the efficiency of Big Data.
g) Remote Servers: Cloud includes multiple remote servers which makes it possible to deal with mighty sized big data sets at the same time [45]

VII. Conclusion

This is the age of Big Data, where massively large, unorganised data is continuously generated at an increasingly faster pace. According to the figures, almost 1 petabytes of data is generated in a single day. This massive data demand a stable and efficient infrastructure to handle and interpret this data which is being offered by Hadoop and Spark programming models in the hands of CCT. Another crux of this review is integration of Big Data and CC, in order to discover the powerful platform which is of utmost importance to present-day enterprises. Their combination inside the same infrastructure can reduce expenditures on a large scale; nevertheless it is an overwhelming technical venture. This review includes characteristics of big data, brief study of Hadoop, Spark architecture, advantages of CC for Big Data; related case studies dealing with combined platform of big data and CC.

VIII. REFERENCES

[1]. S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big data: Issues and challenges moving forward," Proc. Annu. Hawaii Int. Conf. Syst. Sci., pp. 995–1004, 2013, doi: 10.1109/HICSS.2013.645.

[2]. M. Islam and S. Reza, "The Rise of Big Data and Cloud Computing," Internet Things Cloud Comput., vol. 7, no. 2, p. 45,2019,doi:10.11648/j.iotcc.20190702.12.

[3]. R. L. Villars, C. W. Olofson, and M. Eastwood, "Big Data: What It is and Why You Should Care," IDC White Pap., pp. 7–8, 2011.

[4]. I. A. T. Hashem, I. Yaqoob, N.B. Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, "The rise of 'big data' on cloud computing: Review and open research issues," Inf. Syst., vol. 47, pp. 98–115, 2015, doi: 10.1016/j.is.2014.07.006.

[5]. Mayer-Schonberger,V.,& Cukier, K. Big data.

[6]. M. A. U. D. Khan, M. F. Uddin, and N. Gupta, "Seven V's of Big Data understanding Big Data to extract value," Proc. 2014 Zo. 1 Conf. Am. Soc. Eng. Educ. - "Engineering Educ. Ind. Involv. Interdiscip. Trends", ASEE Zo. 1 2014, 2014, doi: 10.1109/ASEEZone1.2014.6820689

[7]. J. Hurwitz, A. Nugent, F. Halper, and M. Kaufman, Big Data for Dummies, no. 2013.

[8]. Retrieved from https://en.wikipedia.orglwikiCloud_computing.asp

[9]. V. V. Arutyunov, "Cloud computing: Its history of development, modern state, and future considerations," Sci. Tech. Inf. Process., vol. 39, no. 3, pp. 173–178, 2012, doi: 10.3103/S0147688212030082.

[10]. V. Kale and V. Kale, "CloudComputing Basics," Creat. Smart Enterp., no. August 2013, pp. 141–171, 2017, doi: 10.1201/9781315152455-6.

[11]. Retrieved from https://www.investopedia.com/terms/c/cloud-computing.asp

[12]. Guo, H., Goodchild, M., & Annoni, A. Manual of Digital Earth.

[13]. Shawish A., Salama M. (2014)Cloud Computing: Paradigms and Technologies. In: Xhafa F., Bessis N. (eds) Inter-cooperative Collective Intelligence: Techniques and Applications. Studies in Computational Intelligence, vol 495.Springer,Berlin,Heidelbe https://doi.org/10.107/978-3642-35016-0_2

[14]. B. P. Rimal, E. Choi, and I. Lumb, "A taxonomy and survey of cloud computing systems," NCM 2009 - 5th Int. Jt. Conf. INC, IMS, IDC, pp. 4451,2009,doi:10.1109/NCM.2009.218.

[15]. S. Namasudra, P. Roy, and B. Balusamy, "Cloud computing: Fundamentals and research issues," Proc. - 2017 2nd Int. Conf. Recent Trends Challenges Comput. Model. ICRTCCM 2017, pp. 7–12, 2017,doi:10.1109/ICRTCCM.2017.49.

[16]. Mell, P., & Grance, T. (2020). The NIST definition of cloud computing.16

[17]. L. Savu, "Cloud computing: Deployment models, delivery models, risks and research challanges," 2011 Int. Conf. Comput. Manag. CAMAN 2011, 2011, doi: 10.1109/CAMAN.2011.5778816.

[18]. Y. Jadeja and K. Modi, "Cloud computing - Concepts, architecture and challenges," 2012 Int. Conf. Comput. Electron. Electr. Technol. ICCEET 2012, pp. 877–880, 2012, doi: 10.1109/ICCEET.2012.6203873.

[19]. Z. Tang, "On Study of Application of Big Data and Cloud Computing Technology in Smart Campus On Study of Application of Big Data and Cloud Computing Technology in Smart Campus," 2017, doi: 10.1088/1755-1315/.

[20]. Retrieved from https://www.oreilly.com/libabry/view/moving-hadoop-to-/9781491959626/ch01.html

[21]. Retrieved from https://blog.syncsort.com/2017/06/bigdata/5-reasons-hadoop –in-the-cloud

[22]. D. Tomar and P. Tomar, "Integration of Cloud Computing and Big Data Technology for Smart Generation," 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, 2018, pp. 1-6, doi: 10.1109/CONFLUENCE.2018.8443052.

[23]. J. Hurwitz, A. Nugent, F. Halper, and M. Kaufman, Big Data for Dummies, no. 1. 2013.

[24]. Z. Han and Y. Zhang, "Spark: A Big Data Processing Platform Based on Memory Computing," Proc. - Int. Symp. Parallel Archit. Algorithms Program. PAAP, vol. 2016-Janua, pp. 172–176, 2016, doi: 10.1109/PAAP.2015.41

[25]. M. Franklin, "The Berkeley Data Analytics Stack: Present and future," pp. 2–3, 2014, doi: 10.1109/bigdata.2013.6691545.

[26]. M. Mittal and V. E. Balas, Big Data Processing Using Spark in Cloud. 2018.

[27]. R. Ilijason, Beginning Apache Spark Using Azure Databricks. 2020.

[28]. K. Alwasel, R. N. Calheiros, S. Garg, R. Buyya, and R. Ranjan, "BigDataSDNSim: A Simulator for Analyzing Big Data Applications in Software-Defined Cloud Data Centers," 2019, Online. Available: http://arxiv.org/abs/1910.04517.

[29]. Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "PPHOPCM: Privacy-preserving High-order Possibilistic c-Means Algorithm for Big Data Clustering with Cloud Computing," IEEE

Trans. Big Data, vol. 7790, no. c, pp. 1–1, 2017, doi: 10.1109/tbdata.2017.2701816.

[30]. 30 H. Pargmann, D. Euhausen, and R. Faber, "Intelligent big data processing for wind farm monitoring and analysis based on cloud-Technologies and digital twins: A quantitative approach," 2018 3rd IEEE Int. Conf. Cloud Comput. Big Data Anal. ICCCBDA 2018, pp. 233–237, 2018, doi: 10.1109/ICCCBDA.2018.8386518.

[31]. V. Vashishth, A. Chhabra, and A. Sood, "A predictive approach to task scheduling for Big Data in cloud environments using classification algorithms," Proc. 7th Int. Conf. Conflu. 2017 Cloud Comput. Data Sci. Eng., pp. 188–192, 2017, doi: 10.1109/CONFLUENCE.2017.7943147.

[32]. Lu Qu, R., 2017. big data and cloud computing for energy internet. (china) international Electrical and energy confrence,.

[33]. Y. Zhang, J. Ren, J. Liu, C. Xu, H. Guo, and Y. Liu, "A survey on emerging computing paradigms for big data," Chinese J. Electron., vol. 26, no. 1, pp. 1–12, 2017, doi: 10.1049/cje.2016.11.016.

[34]. S. Nepal and M. V Ramakrishna, "Proceedings of the International Conference on Data Engineering," pp. 22–31, 1999, doi: 10.1007/978-981-10-1678-3.

[35]. S. Rallapalli, G. Rr, U. Pavan, and K. Ketavarapu, "Impact of Processing and Analyzing Healthcare Big Data on Cloud Computing Environment by Implementing Hadoop Cluster," Procedia - Procedia Comput. Sci., vol. 85, pp. 16–22, 2016, doi: 10.1016/j.procs.2016.05.171.

[36]. L. Kuang, L. T. Yang, and Y. Liao, "An Integration Framework on Cloud for Cyber-Physical-Social Systems Big Data," IEEE Trans. Cloud Comput., vol. 8, no. 2, pp. 363–374, 2020, doi: 10.1109/TCC.2015.2511766.

[37]. K. Kedharewsari, V. Maria Anu, and V. Rajalakshmi, "Integration of big data & cloud computing to detect black money rotation with range - aggregate queries," Int. J. Eng. Technol., vol. 8, no. 2, pp. 768–773, 2016, doi: 10.18535/ijecs/v5i6.23.

[38]. V. N. Inukollu, S. Arsi, and S. Rao Ravuri, "Security Issues Associated with Big Data in Cloud Computing," Int. J. Netw. Secur. Its Appl., vol. 6, no. 3, pp. 45–56, 2014, doi: 10.5121/ijnsa.2014.6304.

[39]. L. Zhang, C. Wu, Z. Li, C. Guo, M. Chen, and F. C. M. Lau, "Moving big data to the cloud: An online cost-minimizing approach," IEEE J. Sel. Areas Commun., vol. 31, no. 12, pp. 2710–2721, 2013, doi: 10.1109/JSAC.2013.131211.

[40]. R. Schmidt and M. Möhring, "Strategic alignment of cloud-based architectures for big data," Proc. - IEEE Int. Enterp. Distrib. Object Comput. Work. EDOC, pp. 136–143, 2013, doi: 10.1109/EDOCW.2013.22.

[41]. The future of big data: 5 predictions from experts for 2020-2025. (2020). ,retrieved from https://www.itransition.com/blog/the-future-of-big-data

[42]. Google,Casestudy:HowredBususesBigQueryto MasterBigData. ⟨https://developers.google.com/bigquery/case-studies/⟩, (accessed 22.07.14).

[43]. A casestudy, CloudComputing (CLOUD), 2010, in: Proceedings of IEEE 3rd International Conferenceon,IEEE,Miami,FL,2010, pp. 107–114.

[44]. Cloudera, Nokia : Using Big Datato Bridge the Virtual & Physical Worlds. ⟨http://www.cloudera.com/content/dam/cloudera/documents/ClouderaNokia-case-study-final.pdf⟩, (accessed24.07.14)

[45]. S. A. El-seoud, H. F. El-sofany, M. Ashraf, and F. Abdelfattah, "Big Data and Cloud Computing : Trends and Challenges Big Data and Cloud Computing : Trends and

Challenges," no. April, 2017, doi: 10.3991/ijim.v11i2.6561.

Cite this article as :

Muneeba Afzal Mukhdoomi, Dr. Ashish Oberoi, Er. Ankur Gupta, "Coming Together of Big Data and Cloud Computing : A Review", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 6, pp. 118-137, November-December 2020. Available at doi : https://doi.org/10.32628/CSEIT206613
Journal URL : http://ijsrcseit.com/CSEIT206613

AUTHOR'S PROFILE

MS. Muneeba Afzal Mukhdoomi completed her Bachelors of Technology (B.Tech) Degree in Computer Science and Engineering from School of Technology, University of Kashmir. She is currently pursuing Masters of Technology (M.Tech) in Computer Science and Engineering from RIMT University. Her areas of focus include Big Data, Artificial intelligence, Internet of things (IOT) and Natural Language Processing. She has participated in several hands-on workshops, Webinars and completed short term courses on Data Science, AI, JAVA , C , C++ programming and Web Designing. She is interested in pursuing a career in Research Field in future.