IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.144100

Integration of Big Data and Cloud Computing

Ranjeet R. Pawar[1], Sameer V. Mulik[2]

HOD, Information Technology, Bharati Vidyapeeth Institute of Technology, Navi Mumbai, India[1]

Lecturer, Information Technology, Bharati Vidyapeeth Institute of Technology, Navi Mumbai, India[2]

Abstract: New technologies needs data to be refined and narrowed down for further processing on it. Big data concept fixes this problem of dealing with large amount of data by performing algorithms which are much easier than the traditional methods which are more complex time consuming, costly, and requires high amount of space. As we are working on data refining, we need to work with raw data which requires high amount of space physically obtaining this space is highly expensive as we require such hardware and software. Also, the refined data or the end product of such hefty data is huge so its storage also requires huge space. Cloud computing provides a platform where the problem of storage is solved. In our paper we present the correlation of both Big Data and Cloud Computing together. We can work on and store huge data with ease. Providing a user and pocket friendly platform. We will discuss on topic of analysis of Cloud based big data in Microsoft Azure, Amazon Web Service, Google Cloud using.

Keywords: Big Data, Cloud Computing, Microsoft Azure, Google Cloud, Amazon Web Service

I. INTRODUCTION

With recent technological advancements, the amount of data available is increasing day by day. For example, sensor networks and social networking sites generate overwhelming flows of data. In other words, big data are produced from multiple sources in different formats at very high speeds. At present, big data represent an important research area. Big data are rapidly produced and are thus difficult to store, process, or manage using traditional software. Big data technologies are tools that are capable of storing meaningful information in different types of formats. For the purpose of meeting users' requirements and analyzing and storing complex data, a number of analytical frameworks have been made available to aid users in analyzing complex structured and unstructured datal. Several programs, models, technologies, hardware, and software have been proposed and designed to access the information from big data. Now a days the traditional method of using data which had referring previous documentation, paper work and estimates is not used. The traditional method is time consuming as well as less accurate. So, to overcome all these problems and make a less time consuming and near accurate solution we can look forward to Big Data. While we work on big data the raw as well as processed and refined data needs to be stored over a platform. Storage devices and software are expensive. As we will require a large amount of space, the entire project will exceed the budget causing financial issues. To overcome this issue, we can take help of cloud services. Cloud servers can be used to store data on a server which will require no physical devices resulting in pocket friendly alternative of other hardware devices which could cost fortune. Together Big Data and Could Computing are making a new revolution by making the data refining as well as storing effective and near accurate. So, in this paper we will see the use of Hadoop with Microsoft Azure, Amazon Web Service and Google Cloud and compare it.

II. BIG DATA

Big Data is a collection of data that is huge in volume yet growing exponentially with time. It is data so large a size and complex that none of the traditional data management tools can store it or process it efficiently. Big data is also data but in huge sizes. Big data has become the revolution of Information Technology which is transforming industries around the world. Big data is a combination of technology and data that integrates reports and accesses all available data filtering, reporting, and correlating insights achievable with previous data technologies. Current usage of the term big data tends to refer to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from big data, and seldom to a particular size of data set. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.

For example-

- The New York Stock Exchange is an example of Big Data that generates about one terabyte of new trade data per day.
- The statistics show that 500+terabytes of new data get ingested into the databases of social media sites Facebook, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments, etc.

Big data is effective because of its 5 characteristics

1. **Volume:** It is the amount of data produced from different sources.
2. **Variety:** It represents the variety of sources from which data is produced.
3. **Velocity:** It is the measure of the speed with which the data is produced.
4. **Veracity:** It represents the quality of data that is gathered and measures its accuracy.
5. **Value:** It shows the value of data after it is analyzed.

III. HOW DOES BIG DATA WORK

Big data analytics refers to collecting, processing, cleaning, and analyzing large datasets to help operationalize their big data. organizations

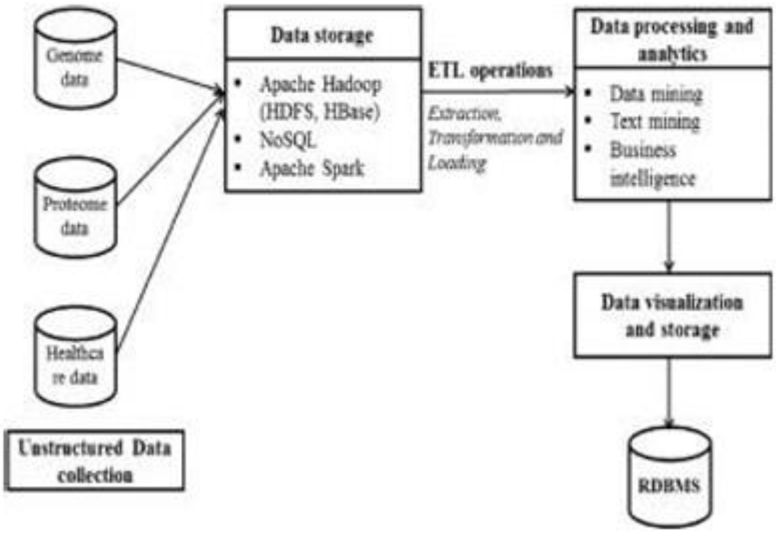
Genome data
Proteome data
Healthcare data
Unstructured Data collection
Data storage
• Apache Hadoop (HDFS, HBase)
• NoSQL
• Apache Spark
ETL operations
Extraction, Transformation and Loading
Data processing and analytics
• Data mining
• Text mining
• Business intelligence
Data visualization and storage
RDBMS

Fig. big data architecture

1. **Collect Data collection looks different for every organization.** With today's technology, organizations can gather both structured and unstructured data from a variety of sources — from cloud storage to mobile applications to in- store IoT sensors and beyond. Some data will be stored in data warehouses where business intelligence tools and solutions can access it easily. Raw or unstructured data that is too diverse or complex for a warehouse may be assigned metadata and stored in a data lake.

2. **Process Data Once data is collected and stored**, it must be organized properly to get accurate results on analytical queries, especially when it's large and unstructured. Available data is growing exponentially, making data processing a challenge for organizations. One processing option is batch processing, which looks at large data blocks over time. Batch processing is useful when there is a longer turnaround time between collecting and analyzing data. Stream processing looks at small batches of data at once, shortening the delay time between collection and analysis for quicker decision-making. Stream processing is more complex and often more expensive.

3. **Clean Data big or small** requires scrubbing to improve data quality and get stronger results; all data must be formatted correctly, and any duplicative or irrelevant data must be eliminated or accounted for. Dirty data can obscure and mislead, creating flawed insights.

4. **Analyze Data** Getting big data into a usable state takes time. Once it's ready, advanced analytics processes can turn big data into big insights.

Some of these big data analysis methods include:

- • Data mining sorts through large datasets to identify patterns and relationships by identifying anomalies and creating data clusters.
- • Predictive analytics uses an organization's historical data to make predictions about the future, identifying upcoming risks and opportunities.
- • Deep learning imitates human learning patterns by using artificial intelligence and machine learning to layer algorithms and find patterns in the most complex and abstract data.

DOI: 10.17148/IJARCCE.2025.144100

IV. CLOUD COMPUTING

Cloud computing refers to storing data over network of servers over internet to perform manipulation, manage and process data. No personal system, local server or storage device is used.
Cloud computing can be classified in following categories: -

1. **IaaS:** Infrastructure as a service provides various infrastructure support to the users, with the only difference being that the infrastructure doesn't exist in physical form. In the IaaS model, all the services like hardware, software, servers, storage, and other infrastructure are provided by a third party through virtualization. Examples of IaaS include Linode, Amazon Web Services, Microsoft Azure.
2. **PaaS:** Platform as a service delivers a framework for developers which they can use to build and create applications. All the platform's work is managed by third-party service providers, while developers can manage their applications. Examples of PaaS include AWS Elastic Beanstalk, Heroku and Force.com.
3. **SaaS:** In Software as a service various application software are provided to the users, which they can use without installing on their systems. The user is not responsible for the software's working, and they have to modify settings according to their usage.

V. HOW DOES CLOUD COMPUTING WORK

Cloud computing stores data on remote servers which can be accessed over internet. It is divided into front-end and back-end. There is connection on each side using internet.

1. **Front-end**: Front-end is the client or the user side. It has the client computers or computer networks. It also requires the application software required for the connection with the cloud server. Every cloud system has different accessibility scheme
2. **Back-end**: Backend is the cloud section of the. At the back end there are various computer servers and data storage system. The cloud can include any program from data processing to video games. Each application has its dedicated server

Cloud computing and Big data Cloud computing has changed the perspective of handling data. Big data on other hand helps to minimize the huge data and compress it to easy and refined data. We require a platform to store this refined data. There are some options for the following:

1. **Big Data on AWS**: Cloud platforms need a cost-effective way to process vast amounts of data they are storing. At the centre of Amazon's analytics offerings is AWS Elastic MapReduce (EMR), a managed Hadoop, Spark and Presto solution. EMR takes care of setting up an underlying EC2 cluster and provides integration with a number of AWS services including S3 and DynamoDB. Clusters can be created and deleted on demand to process specific jobs or kept running for extended periods of time. Clusters typically take around 15 minutes to provision before job execution begins. Data Pipeline can read and write data from most AWS storage services and supports a range of data processing activities including EMR, Hive, Pig, and can execute Unix/Linux shell commands. AWS makes it very easy to create predictive models without the need to learn complex algorithms. To create a model user are guided through the process of selecting data, preparing data, training and evaluating models through a simple wizard-based UI. It is also possible to create models via the AWS SDK. Once trained the model can be used to create predictions via online API (request / response) or a batch API for processing multiple input records. To making sense of data through dashboards and data visualizations AWS offers Quick Sight (currently in preview). Dashboards can be built from data stored across most AWS data storage services and supports a number of third-party solutions.

2. **Azure**: Azure Synapse Analytics is a complete data analytics platform service. It brings together Azure Data Warehouse, Azure Data Lake, Azure Data Factory, Spark and Power BI under one unified development experience. It allows you to pick and choose the languages and frameworks that best suit your skills and needs. If you prefer SQL then you have the choice of using the pre-provisioned massively parallel processing architecture, a new serverless options for querying data in a Data Lake Store or using Spark SQL. A managed Apache Spark environment including a rich interactive notebook experience is available which also comes with support for C# through .NET for Apache Spark. HDInsight comes with a local HDFS and can also connect to blob storage or Data Lake Store. Data stored on the local HDFS is lost when the cluster is shutdown. Clusters can be automatically created and deleted on a schedule using PowerShell and Automation, alternatively on-demand HDInsight clusters can be created for specific jobs invoked through Data Factory. Azure Data Factory still exists as its own standalone service used to build data processing pipelines. Data factory can read data from a range of Azure and third-party data sources, and through Data Management Gateway, can connect and

DOI: 10.17148/IJARCCE.2025.144100

consume on-premise data. Data Factory comes with a range of activities that can run compute tasks in HDInsight, Azure Machine Learning, stored procedures, Data Lake and custom code running on Batch. A SQL-like language is used to perform times series-based queries and can call into Azure Machine Learning to score data in real- time. Azure Machine Learning is a fully managed data science platform that is used to build and deploy powerful predictive and statistical models. Trained models can be published as web services for consumption either as a Realtime request/response API or for batch execution. Azure Machine Learning also comes with interactive Jupyter notebooks for recording documenting lab notes.

3. **Google Cloud Platform Cloud:** Dataproc is Google's fully managed Hadoop and Spark offering. Google boasts an impressive 90 second lead time to start or scale Cloud Dataproc clusters, by far the quickest of the three providers. An HDFS compliant connector is available for Cloud Storage that can be used to store data that needs to survive after the cluster has been shut down. There is no built-in support for on-demand clusters, however full control over the cluster is available through the Gcloud cli, REST API or SDK so this can be automated if required. Data processing pipelines can be built using Cloud Dataflow. Google has taken a different approach to AWS and Azure; both have gone with a declarative model that delegates processing work to other services such as Hadoop. Cloud Dataflow on the other hand provides a fully programmable framework, available for Java and Python, and a distributed compute platform. The programming model and SDK was recently submitted to the Apache Foundation and have become Apache Beam, which can use both Cloud Dataflow as well as Spark for pipeline execution. Cloud Dataflow supports both batch and streaming workers. Google offers Machine Learning as a fully managed platform for training and hosting TensorFlow models. It relies on Cloud Dataflow for data and feature processing and Cloud Storage for data storage. There is also Cloud Datalab, a lab notebook environment based on Jupyter. A set of pre-trained models are also available.

VI. CONCLUSION

In this article we covered the overview of big data cloud computing and the co- relation between them. We have also seen the working of Azure, AWS, and Google Cloud for utilization of big data. Cloud computing has made it easy and cheap to store data. With this decentralized storage, we are seeing a revolution in how data is managed and stored. Without data being accessible to use anywhere at any time, it has made it easier for remote teams to work on projects that involve collaborations. Big data and cloud computing can make any data related task easy when it goes hand in hand.

REFERENCES

[1]. J. Panneerselvam, L. Liu, and R. Hill, “An introduction to big data,” in Application of Big Data for National Security. Elsevier, 2015, pp.3-13.

[2]. T. Mahmood and U. Afzal, “Security analytics: Big Data Analytics for Cybersecurity: A review of trends, techniques and tools,” in 2013 2nd National Conference on Information Assurance (NCIA), Dec 2013, pp.129- 134.

[3]. A. Vailaya, "What’s All the Buzz Around “Big Data?”", IEEE Women in Engineering Magazine, December 2012.

[4]. S. Singh and N. Singh, "Big Data Analytics", 2012 Conference on International Communication, Information & Computing Technology Mumbai India, IEEE, October 2011

[5]. Amanpreet Kaur Sandhu, “Big Data with Cloud Computing: Discussions and Challenges”, March 2022.

[6]. Dimpal Tomar and Pradeep Tomar, “Integration of cloud computing and big data technology for smart generation”