# Performance Analysis of Big Data and Cloud Computing

- Abstract ***A cloud framework refers to the aggregation of components like development tools, middleware and database services, needed for cloud computing, which aids in developing, deploying and managing cloud based applications strenuously, consequently making it an efficacious paradigm for massive scaling of dynamically allocated resources and their complex computing. Big Data Analytics (BDA) delivers data management solutions in the cloud architecture for storing, analysing and processing a huge volume of data. This paper presents a survey for performance based comparative analysis of cloud-based big data frameworks from leading enterprises like Amazon, Google, IBM, and Microsoft, which will assist researcher, IT analysts, reader and business user in picking the framework best suited*** 1. ***for their work ensuring success in terms of favourable outcomes.***

## Introduction

- **Distributed Big Data Analytics for Enterprises: Needs, Challenges and the Solution**

***Online social media platform like Facebook, Whatsapp, Instagram and various other online enterprise applications are ruling the enterprise era, propagating an extensive bulk of data (in petabyte and Exabyte) on daily basis. This tremendous volume of data, generated through messaging, satellite imaging, social media, email and a lot more, can be both structured and unstructured, and is referred to as Big Data. [1, 6, 9]. Consequently, various enterprises are facing an arduous task of tracking and handling such an enormous amount of data, so much that the highly priced data warehouses, developed by these enterprises, are getting encumbered, resulting in consequential processing load [5]. In order to clear out this bottleneck, Big Data Analytics (BDA) comes into play, wherein several big data tools, techniques and methodologies are acquired by organisations for extracting the Right Data from these huge sets of unstructured data [7]. One such tool that is being opted by enterprises for bringing offloading solution in warehouse data and processing functions is Apache Hadoop, which is an open source. Hadoop, when amalgamated with various data warehouses, can turn out to very cost effective and highly business-like. Other frameworks being adopted by enterprises are Pig, Hive, Jaql, R-programming, and many more. Analysing these large sets of data with BDA is favouring IT companies in advance predictions and better decision making, thereby, boosting their revenue***

***growth and making them achieve a competitive advantage in the market [10]. However, the gained popularity and the huge acceptance of BDA in the IT industry is escalating many issues and challenges, addressing the big size and the big price of big data [9]. The cloud brings solution to this problem. Cloud Computing is a wall-to-wall, user friendly archetype for rapid provisioning of services over the internet, sanctioning on demand access to a common pool of cloud resources with minimal delivery management.***

## Bringing Together Big Data and Cloud: A Perfect Consolidation

***While cloud is all about delivering pay-as-apply, on demand, flexible and scalable services, BDA emphases on revolutionising its information assets represented by 3 V's symbolizing Volume, Velocity and Variety, into another V symbolizing Value (to organisations' business) [2]. Illustration of each of these big data dimensions is given below.***

- ***Volume: The amount of data propagated***
- **Velocity: The speed with which the data is propagated**
- **Variety: The heterogeneity of data type propagated.**

***Cloud Computing offers the possibility of accommodating a massive volume of data over the internet through hardware virtualisation, thus, adding to the availability, scalability and accessibility of Big Data [3, 9]. Moreover, cloud computing also delivers exclusive statistical tools for resourceful processing and analyses of big data through a service termed as Big Data as a Service (BDaaS) [9, 12]. Subsequently, both big data and cloud unify together to bring in value to enterprises by enhancing the agility, elasticity, accessibility and the ease of processing of cloud based big data, and, by reducing its cost of ownership and implementation complexity of big data solutions.***

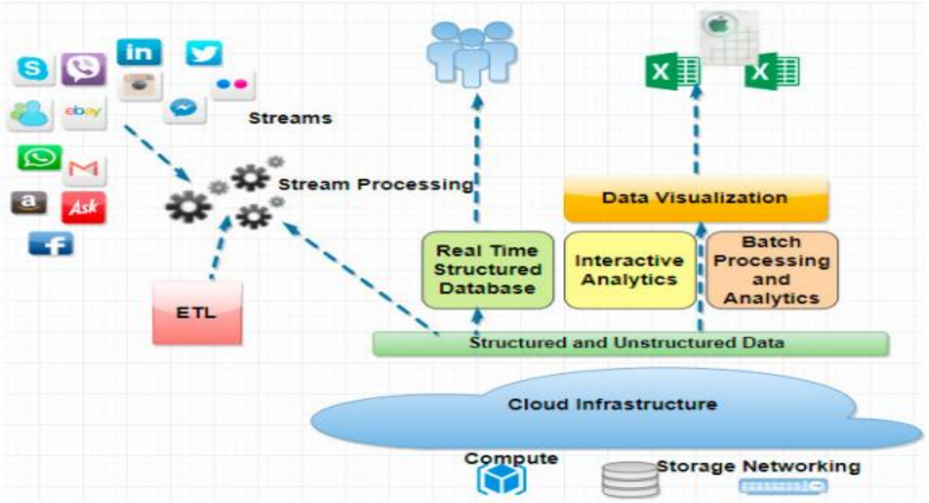

Fig 1. High Level Architecture Diagram of Cloud Based Big Data Framework [31]

## Big Companies Offering a Bigger Solution: Cloud Supported Big Data

***Due to the overall reduced cost of ownership [9] and other previously discussed benefits provided by cloud based big data, many leading organisations are offering cloud based big data frameworks, to be ahead of the game***

***This paper presents a comparison survey for performance analysis of cloud based big data enterprise solution frameworks for Big Data Analytics, Big Data Storage [22] and Big Data Warehouse, provided by Amazon, Google, IBM and Microsoft cloud computing platforms, on the basis of various parameters like mode of software, type of data, data sources, operating system, and many more. The main purpose of this comparative analysis is not to denounce any of the mentioned cloud based big data tools. However, the objective is to demonstrate significant usage of each in various fields.***

***Succeeding this introductory section (1), Section 2 demonstrates frameworks for various cloud based big data tools launched by the leading enterprises like the Amazon, Google, IBM and Microsoft. Section 3 presents the summarisation of these tools demonstrating their comparative features in tabular format. Section 4 concludes this related work.***

## Related Work

### Varied Collection of Cloud Based Big Data Tools

***In this section, three tools each for Big Data Analytics, Big Data Storage and Big Data Warehouse are selected from the cited leading vendors, with their concise features and benefits.***

## Amazon Web Services (AWS)

***A wide spectrum of BDA frameworks are delivered by AWS which assist in building and deploying BDA applications in an easy and quick manner.***

### Amazon ElasticSearch Service (Big Data Analytics Framework)

***Amazon ES is an open source search and a managed service for creating domain, and, for deploying and operating scalable AWS cloud ElasticSearch clusters [15], which delivers capabilities like, real time application monitoring, log analytics, and clickstream analytics. Some of its benefits are integration with other AWS services [14], easy usage, scalable cluster, high accessibility, and support for Open-Source APIs and Tools. [16, 17, 19, 31].***

### Amazon S3 (Big Data Storage Framework)

***Amazon S3 is a vastly scalable, secure and robust big data storage with a broad spectrum of engines, with a wide variety of use cases, one of which is Big Data Analytics. Some of the advantages of using this framework are elastic management with the most flexible set of administration and storage management solutions, incomparable robustness, consistence and scalability, in-place query, integration with maximum vendor solutions, most wide ranging security and compliance competences, simplified and accelerated data transfer. [15-17, 19, 22].***

### Amazon Redshift (Data warehousing Framework)

***Amazon Redshift is one of the well-recognised, fully managed AWS data warehouse platforms, offering quick, simplified and cost-friendly analysis of data with standard SQL and prevailing Business Intelligence Tools. Some of the features of Amazon Redshift are speedy query performance, cost-effectiveness, simplified and highly protected framework, elastic and scalable cluster, compatibility with numerous SQL clients and vast expandability.***

## Google Cloud Platform (GCP)

***Google Cloud Platform delivers an array of some powerful tools for diverse purposes ranging from big data analytics, data warehouse, to database and storage, and a lot more. Some of these standard tools, each for big data analytics, big data storage, and big data warehouse framework, are summarised below.***

### Google Cloud Dataproc (Big data Analytics Framework)

***Google Cloud Dataproc, as indicated by its name, is a fully-organised and automated cloud based Apache Hadoop and Spark service for speedy, simplified and economical cluster management operations. Some of its features and benefits are fast cluster scaling, cost-effectiveness, and open-source framework. [19, 20, 31].***

### Google Cloud Storage (Big Data Storage Framework)

***Designed specifically for enterprises and developers, Google cloud storage is an incorporated object storage which performs numerous tasks, ranging from real-time data processing to data archiving (with Coldline and Nearline storage solutions) to data analytics. Some of the benefits of its usage are high availability and low pricing, refined archiving and storage, seamless and effortless data transition, cost-effectiveness, enhanced security for enterprisecritical resources, and partnership with leading vendor solutions. [19, 20, 22].***

### Google BigQuery (Big Data Warehousing Framework)

***This is a highly scalable, swift, and low-priced fully-organised data warehouse for data analytics in production. Some of the advantages of using this product are quick infrastructure set-up, seamless scaling, effective analysis and quick insights, and business data and investments protection. [19, 20].***

### IBM Cloud

***The IBM cloud offers diversified cloud based Big Data services, and delivers the accurate tool for the accurate job, ranging from Big Data Analytics Framework, networking, monitoring and so on. Like in previously discussed cloud providers' selection of frameworks, one framework each Big Data Analytics, Big Data Storage, and Big Data Warehousing are picked up, and summarised here as well.***

### Analytics Engine (Big Data Analytics Framework)

***This is one of the solutions offered by the IBM cloud, which benefits enterprises in easy data analytics and resolving various existing big data issues. It is a managed cluster platform integrates with the Apache Hadoop and Apache Spark services for building and deploying analytics applications in a simplified fashion. Some of its features are open-source, scalable clusters, configurable environment. [24, 25, 31]***

### IBM Cloud Object Storage (Big Data Storage Framework)

***This cloud based storage tool is developed by IBM for the purpose of storing, managing and accessing data through REST based APIs and IBM's self-service portal. This framework can have direct connection with applications and interconnection with other IBM cloud services. Its major features are durability and reliability, regional resilience, storage classes flexibility, rest based APIs and SDKs, and IBM cloud management console. [23-25]***

### IBM Db2 Warehouse on Cloud (Big Data Warehouse Framework)

***This is a highly organised, IBM CLU Acceleration driven business-class cloud based data warehouse service for incomparable query performance. This tools is highly-organised and fully secured, compatible with Oracle and Nerezza, built for the hybrid cloud, a data store for data sciences, and unburdens analytics workload. [24, 25] 2.1.4.***

### Microsoft Azure

***Microsoft Azure offers a wide-ranging collection of cloud services, beneficial for both developers and IT professionals for constructing, deploying and managing applications ranging from mobile***

***applications to ISC (Internet-Scale Computing) solutions, via its world-wide grid of datacentres, with the support of DevOps and integrated tools. Here also, solutions for BDA, Big Data storage, and Big Data warehouse are selected and their features and benefits are summarised below.***

### Azure HDInsight (Big Data Analytics Solution)

***This solution offers 99.99% SLA for a single instance of virtual machine, unlike other industrial services which offer SLA on the critical virtual machines only. For instance, it offers provision for optimised clusters creation for Spark, Hadoop, Kafka, HBase Storm, and Microsoft R Servers supported by 99.9% SLA. Some of its significant features are global availability, high security and compliance, highly-productive platform for research and development, cost-effectiveness, as well as great extensibility. [27-31]***

### Azure Blob Storage (Big Data Storage Solution)

***Blob Storage is an easy and cost-effective storage for exabytes of varied unstructured data like audio, videos, images and a lot more, in tiers classified as hot, cool and archive based on the frequency of data access. Its features are strong data integrity confirming availability of its latest version everywhere, flexibility to perform modifications for application enhancement and bandwidth usage reduction, various blob types like page, block and append blob providing flexibility for storage optimisation, and automatic geo-replication for easy empowerment of improved.***

### Azure SQL Data Warehouse (Big Data Warehouse Solution)

***SQL Data Warehouse is a framework for massively-parallel processing with SQL analytics enabling elastic and independent scaling of compute and storage, with the capability of its effortless integration with big data stores for building a hub for cubes and data marts. Some of its features and benefits of usage are infinite scaling, elasticity and extensibility, high security and compliance, and compatibility with Microsoft and other leading vendors leveraging other valued technologies. [27-30] 3***.

## Comparison of Cloud Based Big Data Enterprise Solutions Frameworks

***The comparative analysis of above summarised products is demonstrated below in Table II, wherein comparison is done among respective tools of a field, like the selected Big Data Analytics tool of AWS is compared with Big Data Analytics tools of the rest three organisations. The main purpose of this comparative analysis is not to denounce any of the mentioned cloud based big data tools. However, the objective is to demonstrate significant usage of each in various fields.***

| 1. | Big Data Analytics | Amazon ElasticSearch Service | Google Cloud Dataproc | IBM Analytics Engine | Azure HDInsight |
|---|---|---|---|---|---|
| | *Mode of Software* | Open-Source | Open-Source | Open-Source | Open-Source |
| | *Types of Data* | Structured, semi-structured and unstructured | Structured, semi-structured and unstructured | Unstructured | Unstructured |
| | *Data Sources* | Amazon S3, Amazon Kinesis Firehose, and Amazon DynamoDB | Google Bigtable, Google Cloud Storage, and Google BigQuery | IBM Cloud Object Storage | Blob Storage |
| | *Supported Operating System* | CentOS, Ubuntu, and Amazon Linux | Debian 8 | CentOS 7 | Ubuntu 14, Ubuntu 16, and Windows Server 2012 R2 |
| | *Applications* | Logs analytics, real-time applications monitoring, and clickstream analytics | Batch processing, querying, streaming, and machine learning | Data analytics, enterprise solution for various Big data problems, and analytics applications development and deployment | Stream and Batch data analytics |
| | *Service Integration* | Yes | Yes | Yes | Yes |
| | *Deployment* | Zonal | Zonal | Regional | Regional |

| | | | | |
|---|---|---|---|---|
| *of Compute Nodes* | | Managed | node, 59 data nodes) | 1 control node for MPP Engine) |
| *Deployment Locality* | Availability zone, Region | Region | Region | Region |
| *Supported Data Format* | CSV, Avro, Parquet, SequenceFile, TSV, Grok, RCFile, TSV, ORC and RegexSerDe | CSV, Avro, JSON (Newline delimited only) | Delimited format such as CSV | Parquet, Orc, flat delimited text, RC |
| *Storage Format* | Columnar | Columnar | Columnar | Columnar |
| *Data Sources* | Amazon S3, Amazon Dynamo DB, Amazon EMR, AWS Data Pipeline | Google Cloud Storage, Google Cloud Dataflow, readable data sources | Data file on network, data stores like Amazon S3 or IBM Cloud Object Storage, and Db2® server | Azure Blob Storage |
| *Data Loading Methods* | COPY from S3, Streams from Amazon Kinesis Firhose | Streaming Upload, Bulk Upload, Google Analytics Premium | Load from Cloud, Load from File, Load Geospatial Data, Load Twitter Data, Load Public Data | Data Load via PolyBase, SQLBulkCopy API, Bulk-Load data with SSIS (SQL Server Integration Service) and BCP Command, Azure Data Factory |
| *Query Language* | PostgreSQL | Standard SQL (Beta), Legacy BigQuery SQL | SQL Reference, CLPPlus, SQL PL, PL/SQL | PolyBase T-SQL |
| *Integration with ETL and BI Tools* | Yes | Yes | Yes | Yes |
| *Backup Retention Policy* | 1-35 days | 7 days | 2 days | 7 days |

## Conclusion and Future Work

***This work is based upon comparative analysis of the three selected big data cloud based solutions i.e. Big data Analytics, Big Data Storage and Data Warehouse delivered by the cited leading enterprises. The Big Data provisioning of operative processes for data sets collection, being massively huge and way too complex, outlines the need of large scale and real time development of numerous tools in Big Data research.***

***Opting for a suitable cloud platform provider is subjected to the business outlook, internal constrictions, workrelated requirements and the behaviour of applications being migrated. Where AWS provisions a well-integrated platform promising extensive geographical delivery of its services, GCP deals with pay-per-second billing model confirming pay-exactly-for-what-you-use, as well as a rich set of services like tensorflow offering varied machine.***

***learning applications. MS Azure embraces the possibility of running Microsoft RServer based models with big data, supporting 99.99% SLA for a single instance of virtual machine. IBM cloud provides a one-stop solution with strong technical approach towards solving customers' issues.***

***This survey covers the in-depth understanding of these enterprise solution frameworks, their services and features, as well as their applications and use cases, with the formulated comparison among them, thereby benefiting researchers, IT analytics, business users as well as readers in quicker and enhanced decision support, promoting innovative work, enhancement and implementation of such upcoming valuable frameworks in near future.***

***This concept of cloud based big data framework foresees its application in offering reasonable and powerful computational solutions to problems against the computationally unapproachable conventional machine learning algorithms, empowering better predictive modelling and decision making, as well as speedy and precise real-time analysis with enhanced occupancy out of the same core big data framework merged with cloud embedding, profiting in gaining better business insights.***