THE BIG DATA IN THE CLOUD AND HADOOP TECHNOLOGY

Dr Madhumita Dash[1], Mr. R N Panda[2]

[1]*Professor, Orissa Engineering College, BBSR, (India)*

[2] *Assistant Professors, DSMS, Durgapur, (India)*

ABSTRACT

This paper present a Cloud based Big Data platform for collecting and transforming distributed data into knowledge or insightful information. In recent years, data has gained more visibility and importance to organizations and governments worldwide. Innovative technologies have emerged to more efficiently collect process and disseminate data or information, bringing transformational value to these organizations. Big data is being generated by everything around us at all times like systems, sensors and mobile devices and particularly in the cloud computing. As Big data is arriving from multiple sources at an alarming velocity, volume and variety to extract meaningful value from big data, the need optimal processing power, analytics capabilities and skills. In the current information age, the transformation of data to useful information is playing a vital role in driving effectiveness and efficiency. Innovation in social computing, the proliferation of mobile devices, and the emergence of cloud computing have magnified data availability and accessibility. These changes have caused a shift from traditional data management systems and processes to an emerging paradigm of Big Data. Technologies such as Hadoop and Cloud Computing offer a lot of promise to solve some of the problems emanating from Big Data. When these technologies are applied correctly, useful information, hidden patterns and unknown correlations in the data, will be discovered.

Keywords: Big Data, Analytics, Cloud, Grid, Computing, Distributed Computing, Hadoop

I. INTRODUCTION

Big data is changing the way people within organizations work together. It is creating a culture in which business and IT leaders must join forces to realize value from all data. Insights from big data can enable all employees to make better decisions—deepening customer engagement, optimizing operations, preventing threats and fraud, and capitalizing on new sources of revenue. But escalating demand for insights requires a fundamentally new approach to architecture, tools and practices. The challenge now exists with big data for data scientists. Instead of developers writing application code, data scientists designing analytic models for extracting actionable insight from large volumes of diverse, rapidly changing data sets. The problem is that no matter how awesome those analytic models may be, they don't do anyone any good if they can't be quickly executed in the production environment. Big data refers to all the new and emerging types of data available from sensors to social media and other sources. It includes unstructured information such as emails, video, PDFs and photos. Combine that with traditional data about customers, products and transactions, and it's easy to see that

organizations are being presented with more information than they ever had before. This represents both a challenge as well as a huge opportunity. DataOps, the set of best practices that improve coordination between data science and operations, has therefore become a critical discipline for any IT organization that wants to survive and thrive in a world where real-time business intelligence is a competitive necessity. One technology that IT organizations are turning to in order to augment and modernize their IT environments is Apache Hadoop .Hadoop is an open source project and provides a platform to store vast amounts of information: what we lovingly refer to as big data. Hadoop is not meant to replace traditional data management solutions such as relational databases, but instead offers a cost effective way to deal with the growing volumes of information coming at us at great speed. However, the value does not come from storing information, but instead from being able to get information out and make sense of it by applying it to business decisions, applications and actions. In order to effectively capitalize on big data stored in Hadoop, the platform must be enterprise ready. This means it must include tools that traditional data management environments have such as integration, performance, security and administration just to name a few.

Speed counts. Business opportunities often have short shelf-lives. In many cases, they may be as fleeting as a website visit or a phone call. So, for all practical purposes, slow results can be no results. Also, business intelligence is increasingly being delivered in the form of mobile apps that salespeople, marketers and executive decision-makers are consuming in real time as they head into meetings or hop onto planes. Their need and expectation is that if Google can give them answers in a fraction of a second, their BI apps should be able to do so as well.

The Cloud Is Not a Panacea. It's awesome that lots of relatively inexpensive processing capacity is available on-demand from cloud service providers. But not every big data performance issue can be solved by spinning up a bunch of Hadoop VMs. In fact, big data performance bottlenecks are often caused by front-end data intake and transformation issues. All the cloud-based analytic processing capacity in the universe won't help you with these bottlenecks.

Big Data Workloads are Diverse. Big data is not just one single thing. One day, it's predictive analytics. The next day, it's mobile data serving. The day after that, it's transaction processing .If your infrastructure can't handle all these different types of workloads reliably, and in real time.

II. BIG DATA TECHNOLOGY

The definition of Big Data has yet to be formalized. During a recent workshop at the National Institute of Standards and Technology (NIST), there were discussions on collaborating to define Big Data. So far, several definitions are being used with no general consensus on the definition. The most popular one defines Big Data as "a massive volume of both structured and unstructured data that is complex and of diverse data types that traditional database and software techniques cannot be used for efficient processing." Gartner defines Big Data as "high volume, velocity and/or variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision-making, and process automation." This definition uses the three Vs (Volume, Velocity, and Variety) to define Big Data (see Figure 1). To relieve the pressure that big data is placing on your IT infrastructure, you can host big data and analytics solutions on the cloud. Achieve the scalability, flexibility, expandability and economics that will provide competitive advantage into the future.

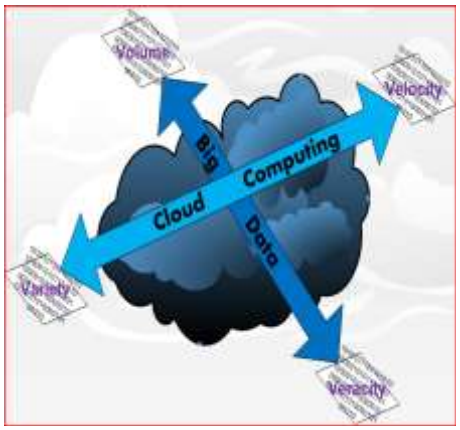

Figure 1 Volume, Velocity, and Variety) To Define Big Data

Other definitions include value or veracity to make it the forth V. Big Data sources include machine-to-machine (M2M), web and social media, transaction data, biometrics, and human-generated. Due to the increase in data sources and the corresponding exponential increase in the amount of data generated, the processing capacity of conventional database systems has become inadequate. Data set sizes have become too big, the data creation rates too high, and existing database structures do not fit the varied data types. Without the right tools, it becomes impossible to collect, store, and analyze this data to reveal practical insights [9]. Big data technology must support search, development, governance and analytics services for all data types—from transaction and application data to machine and sensor data to social, image and geospatial data, and more.

Systems: The infrastructure must capitalize on real-time information flowing through your organization. It must be optimized for analytics to respond dynamically—with automated business processes, better agility and improved economics—to the increasing demands of big data.

Privacy: To protect the reputation and brand, platform must comprise stringent policies and practices around privacy and data protection, safeguarding all of the data and insights on which your business relies.

Governance**:** The right platform instills trust, so that can act with confidence. It controls how information is created, shared, cleansed, consolidated, protected, maintained, retired and integrated within your enterprise.

Storage**:** To achieve economies and efficiencies, that must run certain analytics close to the data, while it is in motion. But for data elect to store, infrastructure must embody a defensible disposal strategy that reduces the run rate of storage, legal expense and risk.

Security**:** As an infuse analytics into the organization, data security becomes more central to competitive advantage profile. Infrastructure must have strong security measures built in to guard the organization against internal and external threats.

III. CLOUD COMPUTING

Cloud Computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models [19], [20], [22]. Cloud Computing consists of three service models, Software as a Service

(SaaS), Platform as a service (PaaS), and Infrastructure as a Service (IaaS); and four deployment models, Private cloud, Hybrid cloud, Public cloud and Community cloud.

Service Models

The three service models include Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS).

Software as a Service (SaaS)

This is a software delivery model in which the provider gives customers on-demand access to the applications hosted in a cloud infrastructure. The infrastructure is managed by the provider while the consumer has only limited user-specific application configuration settings. SaaS is increasingly becoming a common delivery model for most business applications. The consumer usually pays a subscription fee instead of a licensing fee.

Platform as a service (PaaS)

This service delivery model allows the customer to rent the cloud infrastructure (virtualized servers and associated services) to run consumer-created or acquired applications or to develop and test new ones. The infrastructure is managed and controlled by the provider; the consumer has some control over the deployed applications and possibly application hosting environment configurations.

Infrastructure as a Service (IaaS)

IaaS is a delivery capability in which the consumer provisions processing, storage, networks, and other fundamental computing resources. The consumer can deploy and run arbitrary software (operating systems and applications) but does not manage or control the underlying cloud infrastructure.

Deployment Models

The deployment models include Private cloud, Hybrid cloud, Public cloud and Community cloud.

Private Cloud

With this model, the internal or corporate cloud infrastructure (systems and services) is operated solely for an organization. This gives the organization better management and control over their data and systems. It is also considered a proprietary network or a data center that supplies hosted services to a limited number of people.

Hybrid Cloud

A Hybrid Cloud is made up of at least one private cloud and at least one public cloud. An example is when a vendor has a private cloud and forms a partnership with a public cloud provider, or a public cloud provider forms a partnership with a vendor that provides private cloud platforms. In other instances, the organization owns and manages some of the cloud resources internally while others are made available externally. A hybrid cloud provides the consumer the best of both worlds.

Public Cloud

A public cloud is a cloud model in which the cloud provider makes the cloud infrastructure available to the general public; and is owned by the cloud provider. This model is also considered as external cloud. It has several advantages to include: lower cost of deployment, scalability and efficient use of resources (since you only pay for what you use).

Community Cloud

A Community Cloud allows the cloud infrastructure to be shared by several organizations and supports a specific community that has shared concerns. This model can be managed by the organizations involved or a third party, and may exist on premise or off premise.

Big Data Platform as a Service (PaaS)

The Big Data PaaS is a Cloud platform that leverages the Cloud provider's distributed capabilities, compute power, analytical tools, storage capacity, elasticity and others.

Our experiment demonstrates how conventional systems are insufficient to solve the Big Data problem. It further shows some of the benefits of adopting a Cloud solution.

IV. HADOOP OVERVIEW AND CHARACTERIZATION

Hadoop is a framework used to process large data sets in a distributed computing environment. The underlying architecture of Hadoop is HDFS (Hadoop Distributed File System). It provides fault-tolerance by replicating data blocks. Hadoop is based on Google's MapReduce in which an application can break into small parts or blocks that can be run on any node so that applications can run on systems with thousands on nodes. Hadoop framework includes several benchmarks such as Sort, Word Count, Terasort, Kmeans iterations, and NutchIndexing. These benchmarks are based on distributed computing and storage. Apache Hadoop has an architecture that is similar to the MapReduce runtime used by Google. Apache Hadoop runs on the Linux operating system. Hadoop accesses data via HDFS (Hadoop Distributed File System), which maps all the local disks of the computing nodes to a single file-system hierarchy, allowing the data to be dispersed across all the data/computing nodes. HDFS also replicates the data on multiple nodes so that failures of nodes containing a portion of the data will not affect the computations that use that data.

Hadoop runs best on physical servers. A Hadoop cluster comprises a master node called the name node and multiple child nodes called data nodes. These data nodes are separate physical servers with dedicated storage (much like your PC hard drive), instead of a common shared storage.

Hadoop is "Rack Aware" – Hadoop data nodes (servers) are installed in racks. Each rack typically contains multiple data node servers with a top of rack switch for network communication. "Rack awareness" means that the name node knows where each data node server is and in which rack. This ensures that Hadoop can write data to 3 (default) different data nodes that are not on the same physical rack, which helps prevent data loss due to data node and rack failure. When a MapReduce job needs access to data blocks, the name node ensures that the job is assigned to the closest data node that contains the data, thereby reducing the network traffic. The Hadoop system admin manually maintains this rack awareness information for the cluster.

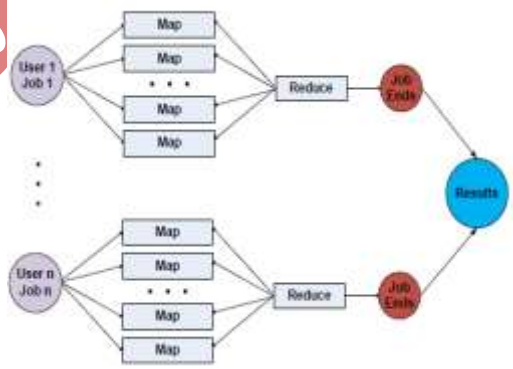

Figure 2. Hadoop MapReduce Model

V. HADOOP IN THE CLOUD

Hadoop as a Service in the Public Cloud – Hadoop distributions (Cloudera CDH, IBM BigInsights, MapReduce, Hortonworks) can be launched and run on the public clouds like AWS, Rackspace, MS Azure, IBM SmartCloud, etc., which offer Infrastructure as a Service (IAAS). In a public cloud, you are sharing the infrastructure with other customers. As a result, you have very limited control over which server the VM is being spun up and what other VMs (yours or other customers) are running on the same physical server. There is no "rack awareness" that you have access to and can configure in the name node. The performance and availability of the cluster may be affected as you are running on VM. Enterprises can use and pay for these Hadoop clusters on demand. There are options for creating your own private network using VLAN, but Hadoop cluster performance recommendation is to have a separate isolated network because of high network traffic between nodes. In all the cases with the exception of the AWS EMR, you have to install and configure the Hadoop cluster on the cloud.

MapReduce as a Service – Amazon's EMR (Elastic MapReduce) provides a quick and easy way to run MapReduce jobs without having to install a Hadoop cluster on its cloud. This can be a good way to develop Hadoop programming expertise internally within your organization or if you only want to run MapReduce jobs in your workloads.

Hadoop on S3 – You can run Hadoop using Amazon's S3 instead of HDFS to store data. Performance of S3 is slower than HDFS, but it provides other features like bucket versioning and elasticity as well as its own data loss protection schemes. This may be an option if your data is already being stored in S3 for your business (e.g. Netflix uses a Hadoop cluster using S3).

Hadoop in private Cloud – We have the same set of considerations for a private cloud deployment for Hadoop as well. However, in case of a private cloud, you may have more control over your infrastructure that will enable you to provision bare-metal servers or create a separate isolated network for your Hadoop clusters. Some of these private cloud solutions also provide a Paas layer that offers pre-build patterns for deploying Hadoop clusters easily (e.g. IBM offers patterns for deploying InfoSphere BigInsights on their SmartCloud Enterprise). In addition, you also have an option of deploying a "Cloud in a Box" like the IBM PureData System, which offers Hadoop ready in your own data center. The big reason for private cloud deployment would be around data security and access control for your data as well better visibility and control of your Hadoop infrastructure.

VI. DEPLOYING HADOOP CLUSTER IN THE CLOUD

Your enterprise should evaluate the security criteria for deploying workloads in public cloud before moving any data into the Hadoop cluster. Hadoop cluster security is very limited. There is no native security for data that will satisfy enterprise data security requirements around SOX, PII, HIPPA, etc.Evaluate Hadoop distributions that you would want to use and the operating system standards of your enterprise. Preferably go with distributions that are close to the open source Apache distributions. Hadoop distributions typically run on Linux. Hortonworks provides a Hadoop distribution for Windows that is currently available on MS Azure cloud.When using AWS, be aware that using Hadoop with S3 would tie you to Amazon's cloud. For open standards, look at OpenStack-based cloud providers like Rackspace, IBM SmartCloud, HP, etc.Look at the entire Hadoop ecosystem and not just the basic Hadoop cluster. The value from Hadoop is the analytics and data visualization

that can be applied on large data sets. Ensure that the tools you want to use for analytics (e.g. Tableau, R, SPSS etc) are available for use on the cloud provider.Get an understanding on where the data to be loaded into Hadoop comes from. Are you going to load data from your internal systems that are not on the cloud or if the data is already in the cloud. Most public clouds charge for data transmission fees if you are moving data back and forth. Hadoop clusters on VM will be slow. You may be able to use these for development and test clusters. VMware's project Serengeti is trying to address the deployment of Hadoop clusters on virtual machines without taking a performance hit. However, with this approach you will be tied to VMware's Hypervisor which should be a criterion to consider when selecting a cloud provider.

VII. HADOOP AS A PLATFORM FOR BIG DATA

Hadoop is a framework comprising of a parallel computing system and a parallel database. It is highly scalable and well suited for the distributed processing of large data sets across clusters of computers. Hadoop is based on a simple data model, in which any data (structured or unstructured) will fit. This contrasts the relational data model in which there must exist a schema for the data before any input . The Hadoop framework comprises of two key components or systems: MapReduce implementation and Hadoop Distributed File System (HDFS). User 1 Job 1 Map Map Map Map Reduce Job Ends User n Job n Map Map Map Map Reduce Job Ends Results .During the execution of a Hadoop service each task is either a map or reduce job.

MapReduce MapReduce is a framework introduced by Google for processing parallelizable problems across large datasets in a distributed computing environment. MapReduce is considered the execution engine.

Hadoop Distributed File System

HDFS is a scalable and distributed file system designed to store large amounts of data. It is a block-structured file system in which a file is broken into blocks of a fixed size and stored across a cluster of one or more data storage systems. A HDFS cluster primarily consists of a NameNode that manages the file system metadata and at a Nodes that store the actual data. HDFS is a master-slave architecture with the master being considered the name node while the slaves are considered data nodes.

NoSQL (Not Only SQL)

A NoSQL database provides a simple, lightweight and scalable mechanism for storing and retrieving data. In the Big Data arena, it is preferred over traditional relational databases especially when working with a huge quantity of data when the data's nature does not require a relational model. NoSQL systems allow the use of a SQL-like query language. Hbase, for example is a NoSQL database and part of the Hadoop ecosystem modeled from Google's BigTable system [6].

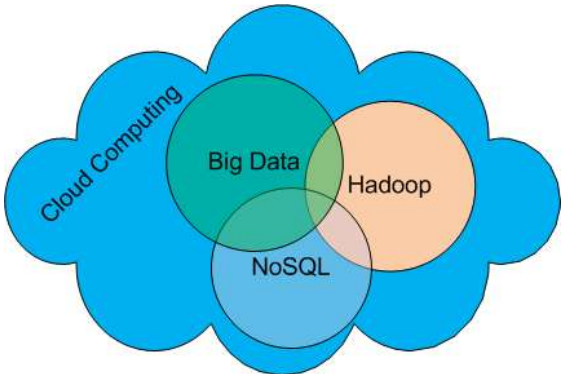

Figure 3. The Collage Of Big Data, Hadoop, Cloud Computing And Nosql

Cloud Computing has come a long way and has now attained a reasonable level of maturity that makes it a viable platform for multiple applications. A Cloud based platform as the underlying technology on which Big Data tools reside. Fig. 4, depict the collage of Big Data, Hadoop, Cloud Computing and NoSQL. This platform provides a very solid infrastructure for collecting, storing, analyzing and producing insightful information.

VII. CONCLUSIONS

This paper discuss Big Data and propose a platform which integrates the Cloud, Big Data, NoSQL, Hadoop and analytic tools to efficiently capture, store and analyze complex datasets. In future works, we will perform more experiments especially with analytics tools to examine other capabilities of the platform. Data analytics and security features require more research to highlight issues affecting the platform. Further work will include hosting a service-oriented decision support system on our Cloud platform and testing for fault tolerance, scalability, data availability and quality of service [8], [5], [8]

REFERENCES

[1] Abbadi, A. El. (2011). Big Data and Cloud Computing : Current State and Future Opportunities, 530–533.

[2] Chaudhuri, S. (2012). What Next ? A Half-Dozen Data Management Research Goals for Big Data and the Cloud. Proceedings of the 31st symposium on Principles of Database Systems, 1–4. doi:10.1145/2213556.2213558

[3] Cohen, J., Dolan, B., & Dunlap, M. (2009). MAD skills: new analysis practices for big data. Proceedings of the …. Retrieved from http://dl.acm.org/citation.cfm?id=1687576

[4] Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters. Operating Systems Design and Implementation (OSDI ', 04.

[5] Demirkan, H., & Delen, D. (2012). Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. Decision Support Systems, in press(0), 1–10. doi:10.1016/j.dss.2012.05.048

[6] Ghemawat, S., Gobioff, H., & Leung, S.-T. (2003). The Google file system. ACM SIGOPS Operating Systems Review, 37(5), 29. doi:10.1145/1165389.945450

[7] Greenberg, A., Hamilton, J., Maltz, D. A., & Patel, P. (2009). The Cost of a Cloud : Research Problems in Data Center Networks, 39(1), 68–73.

[8] Herodotou, H., Lim, H., Luo, G., Borisov, N., & Dong, L. (2011). Starfish : A Self-tuning System for Big Data Analytics. Systems Research, (1862), 261–272. Retrieved from http://www.cs.duke.edu/~hero/files/cidr11_starfish.pdf

[9] LaValle, S., Lesser, E., & Shockley, R. (2011). Big data, analytics and the path from insights to value. MIT sloan management …, 52(2), 21–31. Retrieved from http://sloanreview.mit.edu/the-magazine/2011-winter/52205/big-data-analytics-and-the-path-from-insights-to-value/

[10] Long, P., & Siemens, G. (2011). Penetrating the Fog: Analytics in Learning and Education.

[11] Manyika, J., Chui, M., Brown, B., & Bughin, J. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, (June). Retrieved from

http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Big+data+:+The+next+frontier+for+innovation+,+competition+,+and+productivity#0

[12] McGuire, T., Chui, M., & Manyika, J. (2012). Why Big Data Is The New Competitive Advantage. Ivey Business Journal, 76(4), 1–4. Retrieved from http://web.ebscohost.com/ehost/pdfviewer/pdfviewer?sid=370d66cc-740c-4531-a268-28955083c818@sessionmgr11&vid=2&hid=14

[13] Padhy, R. P. (2013). Big Data Processing with Hadoop-MapReduce in Cloud Systems, 2(1), 16–27.

[14] Prekopcsk, Z., Makrai, G., & Henk, T. (2011). Radoop : Analyzing Big Data with RapidMiner and Hadoop. Technology. Retrieved from http://www.prekopcsak.hu/papers/preko-2011-rcomm.pdf

[15] Rajaraman, A., & Ullman, J. D. (2010). Mining of Massive Datasets.

[16] Russom, P. (2011). Big data analytics. October, 19(September) 2011, 40. Retrieved from http://faculty.ucmerced.edu/frusu/Papers/Conference/2012-sigmod-glade-demo.pdf

[17] Tsuchiya, S., & Lee, V. (n.d.). Big Data Processing in Cloud Environments, 159–168.

[18] Begoli, E., & Horey, J. (2012). Design Principles for Effective Knowledge Discovery from Big Data. 2012 Joint Working IEEE/IFIP Conference on Software Architecture and European Conference on Software Architecture, 215–218. doi:10.1109/WICSA-ECSA.212.32

[19] Peter Mell, Tim Grance. “Effectively and Securely Using the Cloud Computing Paradigm”, NIST, Information Technology Laboratory.

[20] NIST Cloud Computing Standards Roadmap. “NIST CCSRWG – 070 Eleventh Working Draft”. May 2, 2011

[21] Madoka Yuriyama, Takayuki Kushida, “Sensor-Cloud Infrastructure - Physical Sensor Management with Virtualized Sensors on Cloud Computing”, IBM Research - Tokyo, March 17, 2010

[22] Peter Mell and Tim Grance. “The NIST Definition of Cloud Computing”. 10-7-09

[23] Cloud Computing, 227-234. Ieee. doi:10.1109/CLOUD.2011.20

[24] Benson, K., Dowsley, R., & Shacham, H. (2011). Do you know where your cloud files are? Proceedings of the 3rd ACM workshop on Cloud computing security workshop - CCSW ’11, 73. New York, New York, USA: ACM Press. doi:10.1145/2046660.2046677

[25] Hauswirth, M., & Decker, S. (2007). Semantic Reality – Connecting the Real and the Virtual World Position Paper, 1-4.