The Cloud As An Enabler For Big Data Analytics

Prajakta Rawool[1], Swapnil Salvi[2]
[1]*Master Of Computer Application, IMCOST, Thane*
[2]*Master Of Computer Application, IMCOST, Thane*

***Abstract* --** In this data-driven society, we are collecting a massive amount of data from people, actions, sensors, algorithms and the web; For handling this large amount of data known as “Big Data” is become difficult. Extracting useful knowledge from this huge digital datasets requires smart and scalable analytics services, programming tools, and applications.Big data presents a grand challenge for database and data analytics research. Cloud computing is on-demand network access to computing resources, provided by an outside entity. Cloud computing offers the promise of big data implementation to small and medium sized businesses.This paper describes how cloud and big data technologies provides cost-effective delivery model for cloud-based big data analytics. It also includes the background and definition of cloud computing and big data , and how cloud computing is an enabler for advanced analytics with big data.

***Index Terms*-- Cloud Computing, Cloud Computing Service Model, Big Data Analytics, Types of Big Data Analytics, Emerging Technologies In Big Data.**

I. INTRODUCTION:

Nowadays, information technology opens the door through which humans step into a smart society and leads to the development of modern services such as:Internet e-commerce, modern logistics, and e-finance. It also promotes the development of emerging industries, such as Telematics, Smart Grid, Intelligent Transportation, Smart City, and High-End Equipment Manufacturing.

Two IT initiatives are currently top of mind for organizations across the globe: Big data analytics and cloud computing.
Big data analytics offers the promise of providing valuable insights that can create competitive advantage, spark new innovations, and drive increased revenues. As a delivery model for IT services, cloud computing has the potential to enhance business agility and productivity while enabling greater efficiencies and reducing costs.

Both technologies continue to evolve. Organizations are moving beyond questions of what and how to store big data to addressing how to derive meaningful analytics that respond to real business needs. As cloud computing continues to mature, a growing number of enterprises are building efficient and agile cloud environments, and cloud providers continue to expand service offerings. It makes sense, then, that IT organizations should look to cloud computing as the structure to support their big data projects.

Big data environments require clusters of servers to support the tools that process the large volumes, high velocity, and varied formats of big data. Clouds are already Deployed on pools of server, storage, and networking resources and can scale up or down as needed. Cloud computing offers a cost-effective way to support big data Technologies and the advanced analytics applications that can drive business value.

II. CLOUD COMPUTING:

Today, the most popular applications are Internet services with millions of users.
Websites like Google, Yahoo! and Facebook receive millions of clicks daily. This generates

terabytes of invaluable data which can be used to improve online advertising strategies and user satisfaction. Real time capturing, storage, and analysis of this data are common needs of all high-end online applications. To address these problems, a number of cloud computing technologies have emerged in last few years.

"Cloud computing" is the next natural step in the evolution of on-demand information technology services and products. Cloud computing is the delivery of computing services over the Internet. Cloud services allow individuals and businesses to use software and hardware that are managed by third parties at remote locations. Examples of cloud services include online file storage, social networking sites, webmail, and online business applications.

The cloud computing model allows access to information and computer resources from anywhere that a network connection is available. Cloud computing provides a shared pool of resources, including data storage space, networks, computer processing power, and specialized corporate and user applications.

The following definition of cloud computing has been developed by the U.S. National Institute of Standards and Technology (NIST):

> Cloud computing is a model for enabling 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.

III. CLOUD COMPUTING SERVICE MODELS:

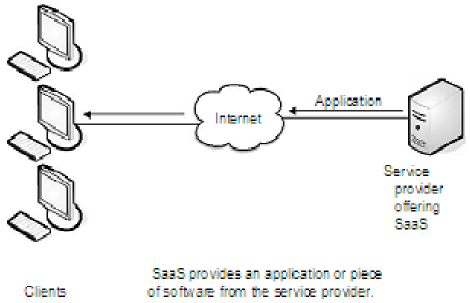
The term services in cloud computing is the concept of being able to use reusable, fine-grained components across a vendor's network. This is widely known as "as a service".

.

Cloud Providers offer services that can be grouped into three categories.

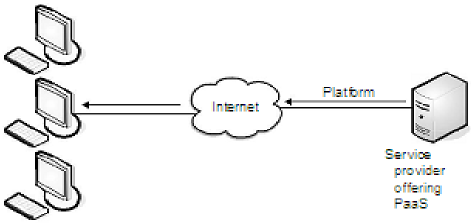
1. Software as a Service (SaaS):

Software as a Service (SaaS) is the model in which an application is hosted as a service to customers who access it via the Internet. When the software is hosted off-site, the customer doesn't have to maintain it or support it. On the other hand, it is out of the customer's hands when the hosting service decides to change it. The idea is that you use the software out of the box as is and do not need to make a lot of changes or require integration to other systems.


Application
Internet
Service provider offering SaaS
Clients
SaaS provides an application or piece of software from the service provider.

2. Platform as a Service (PaaS):

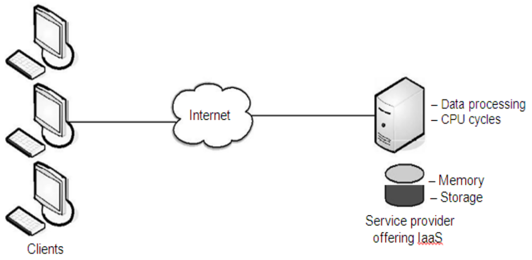
Following on the heels of SaaS,Platform as a Service (PaaS) is another application delivery model. PaaS supplies all the resources required to build applications and services completely from the Internet, without having to download or install software.PaaS generally offers some support to help the creation of user interfaces, and is normally based on HTML or JavaScript. PaaS also supports web development interfaces such as Simple Object Access Protocol (SOAP) and Representational State Transfer (REST), which allow the construction of multiple web services, sometimes called mashups. The interfaces are also able to access databases and reuse services that are within a private network.



PaaS allows clients to access a computing platform over a cloud computing solution.

3. Infrastructure as a Service (Iaas):

IaaS provides basic storage and computing capabilities as standardized services over the network. Rather than purchase servers, software, racks, and having to pay for the datacenter space for them, the service provider rents those resources.Additionally, the infrastructure can be dynamically scaled up or down, based on the application resource needs. Resources are typically billed based on a utility computing basis, so providers charge by how many resources are consumed



IaaS allows service providers to rent hardware resources.

IV. BIG DATA ANALYTICS:

Today, many organizations are collecting, storing, and analyzing massive amounts of data. This data is commonly referred to as “big data”. Big data refers to huge data sets that are orders of magnitude larger (volume); more diverse, including structured, semi-structured, and unstructured data (variety); and arriving faster (velocity) than you or your organization has had to deal with before.

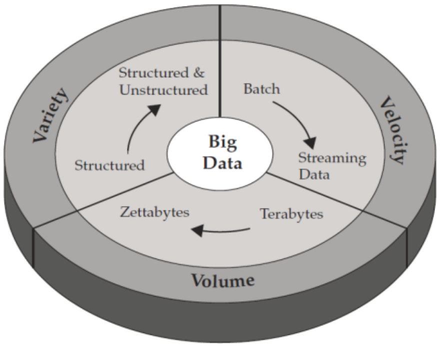
At present the industry does not have a unified definition of big data; big datahas been defined as:

“Big Data is high volume, high velocity, and/or high variety information assets thatrequire new forms of processing to enable enhanced decision making, insight discovery, and process optimization”.

—Gartner.

Big data has three main characteristics:

1) “**Volume**” which indicates a very large volume of data.
2) “**Variety**” which indicates heterogeneity in data that we have collected for processing and analysis this data variety includes structured, unstructured and semi-structured data.
3) “**Velocity**” which indicates the speed for data processing in terms of response time. This response time could be a batch, real-time or stream response-time.



Big data comes from sensors, devices, video/audio, networks, log files, transactional applications, web, and social media - much of it generated in real time and in a very large scale.Big Data requires huge amounts of storage space. While the price of storage continued to decline, the resources needed to leverage big data can still pose financial difficulties for small to medium sized businesses. A typical big data storage and analysis infrastructure will be based on clustered network-attached storage. Data storage using cloud computing is a viable option for small to medium sized businesses considering the use of Big Data analytic techniques.

Big data analytics is a set of advanced technologies designed to work with large volumes of heterogeneous data that include different data sets such as structured/unstructured and streaming/batch, and different sizes from terabytes to zettabytes. Analyzing big data allows analysts, researchers, and business users to make better and faster decisions using data that was previously inaccessible or unusable.

With big data analytics, data scientists and others can analyze huge volumes of data that conventional analytics and business intelligence solutions can't touch. Using advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics, and natural language processing, businesses can analyze previously untapped data sources independent or together with their existing enterprise data to gain new insights resulting in significantly better and faster decisions.The technological advances in storage, processing, and analysis of Big Data include:

(a) The rapidly decreasing cost of storage and CPU power in recent years;

(b) The flexibility and cost-effectiveness of datacenters and cloud computing for elastic computation and storage;

(c) The development of new frameworks which allow users to take advantage of these distributed computing systems storing large quantities of data through flexible parallel processing.

V. TYPES OF BIG DATA ANALYTICS:

There are Four types of big data analytics that really aid business:

a) **Prescriptive analytics** is really valuable, but largely not used. According to Gartner, 13% of organizations are using predictive but only 3 percent are using prescriptive analytics. Prescriptive analytics uses predictive models, localized rules, scoring and optimization techniques to recommend one or more courses of action, and show the expected outcome ofeach.Once you have an idea of what is happening in your business and what is likely to happen next, you need to choose what action to take. With prescriptive analytics, you can evaluate different ways to proceed, and understand the consequences of each of those actions faster. You can use prescriptive analytics to make more informed business decisions in real time.

b) **Predictive analytics** use big data to identify past patterns to predict the future. It is the practice of extracting information from existing data sets in order to determine patterns and predict future outcomes and trends. Predictive analytics does not tell you what will happen in the future. It forecasts what might happen in the future with an acceptable level of reliability, and includes what-if scenarios and risk assessment.

c) **Diagnostic analytics** starts during the descriptive analytics phase and extends into predictive analytics. Diagnostic analytics are used for discovery or to determine why something happened. It's the type of analytics that gets into root cause analysis and data discovery and exploration.Understanding why things happened is essential to improving business operations and processes. You can't fix things if you don't know why they're broken. You can use diagnostic analytics to drill down into all types of data coming from many different sources. Interactive visualizations can uncover patterns and correlations that explain why revenue is down or sales are up or assets are failing—and this insight can drive predictive models.

d) **Descriptive analytics** is the most common type of analytics used by just about every organization in every industry. It serves as a foundation for more advanced analytics. When it comes to data analysis, you need to start by fully understanding **what has happened** and **what is happening now**. You can use descriptive analytics, along with diagnostic analytics, to examine key performance indicators and key metrics to understand how your company is performing and to evaluate business processes. Business intelligence and data mining can help you drill down into data to get a single view of the past and the present.

VI. EMERGING TECHNOLOGIES FOR BIG DATA:

1) Column-oriented databases

Traditional, row-oriented databases are excellent for online transaction processing with high update speeds, but they fall short on query performance as the data volumes grow and as data becomes more unstructured. Column-oriented databases store data with a focus on columns, instead of rows, allowing for huge data compression and very fast query times. The downside to these databases is that they will generally only allow batch updates, having a much slower update time than traditional models.

Advantages:

- ✓ Column-oriented organizations are more efficient when an aggregate needs to be computed over many rows but only for a notably smaller subset of all columns of data, because reading that smaller subset of data can be faster than reading all data.
- ✓ Column-oriented organizations are more efficient when new values of a column are supplied for all rows at once, because that column data can be written efficiently and replace old column data without touching any other columns for the rows.

Disadvantages:

- ✓ Transactions are to be avoided or just not supported
- ✓ Queries with table joins can reduce high performance
- ✓ Record updates and deletes reduce storage efficiency
- ✓ Effective partitioning/indexing schemes can be difficult to design.

2) Schema-less databases, or NoSQL databases

There are several database types that fit into this category, such as key-value stores and document stores, which focus on the storage and retrieval oflarge volumes of unstructured, semi-structured, or even structured data. They achieve performance gains by doing away with some (or all) of the restrictions traditionally associated withconventional databases, such as read-write consistency, in exchange for scalability and distributed processing.

Advantages:

- ✓ NoSQL databases are more scalable and provide superior performance.
- ✓ Both structured and unstructured data can be stored as there is no fixed data model. This flexibility gives organisations access to much larger quantities of data.
- ✓ It is easy to change how data is stored using refactoring or batch processing.

Disadvantages:

- ✓ Schema-less have poor integrity.
- ✓ It faces Ambiguity Problem.
- ✓ Performance Suffers.

3) MapReduce

This is a programming paradigm that allows for massive job execution scalability against thousands of servers or clusters of servers. Any MapReduce implementation consists of two tasks:

i. The "Map" task, where an input dataset is converted into a different set of key/value pairs, or tuples;

ii. The "Reduce" task, where several of the outputs of the "Map" task are combined to form a reduced set of tuples (hence the name).

Advantages:

- ✓ It resolves or avoids several complications of distributed computing.
- ✓ It allows unlimited computations on an unlimited amount of data.
- ✓ It actually simplifies the developer's life.

Disadvantages:

- ✓ When your processing requires lot of data to be shuffled over the network.
- ✓ MapReduce is not suitable for a large number of short on-line transactions.
- ✓ Here are some usecases where MapReduce does not work very well.:
  1. When you need a response fast. e.g. say < few seconds (Use stream processing, CEP etc instead)
  2. Processing graphs
  3. Iterations - when you need to process data again and again.
  4. Joining two large data sets with complex conditions

4) Hadoop

Hadoop is by far the most popular implementation of MapReduce, being an entirelyopen source platform for handling Big Data. It is flexible enough to be able to work with multiple data sources, either aggregating multiple sources of data in order to do large scale processing, or even reading data from a database in order to run processor-intensive machine learning jobs. It has several different applications, but one of the top use cases is for large volumes of constantly changing data, such as location-based data from weather or traffic sensors, web-based or social media data, or machine-to-machine transactional data.

Advantages:

- ✓ Hadoop is a highly scalable storage platform, because it can store and distribute very large data sets.
- ✓ Hadoop also offers a cost effective storage solution for businesses' exploding data sets.
- ✓ Hadoop is so fast that it can able to efficiently process terabytes of data in just minutes, and petabytes in hours.
- ✓ It is simple and robust coherency model.
- ✓ Ability to rapidly process large amounts of data in parallel.

Disadvantages:

- ✓ Rough manner: Hadoop Map-reduce are rough in manner. Because the software under active development.
- ✓ Programming model is very restrictive: Lack of central data can be preventive.
- ✓ Joins of multiple datasets are tricky and slow: No indices! Often entire dataset gets copied in the process.Still single master which requires care and may limit scaling.
- ✓ Cluster management is hard: In the cluster, operations like debugging, distributing software, collection logs etc are too hard.

5) Hive:

Hive is a "SQL-like" bridge that allows conventional BI applications to run queries against a Hadoop cluster. It was developed originally by Facebook, but has been made open source for some time now, and it's a higher-level abstraction of the Hadoop framework that allows anyone to make queries against data stored in a Hadoop cluster just as if they were manipulating a conventional data store. It amplifies the reach of Hadoop, making it more familiar for BI users.

Advantages:

- ✓ Fits the low level interface requirement of Hadoop perfectly.
- ✓ Supports external tables which make it possible to process data without actually storing in HDFS.
- ✓ It has a rule based optimizer for optimizing logical plans.
- ✓ Supports partitioning of data at the level of tables to improve performance.

Disadvantages:

- ✓ Hive perform poorly when you need low-latency execution for simple queries.
- ✓ No support for update and delete.
- ✓ No access control implementation.
- ✓ Correlated sub queries are not supported.

6) PIG

PIG is another bridge that tries to bring Hadoop closer to the realities of developers and business users, similar to Hive. Unlike Hive, however, PIG consists of a "Perl-like" language that allows for query execution over data stored on a Hadoop cluster, instead of a "SQL-like" language. PIG was developed by Yahoo!, and, just like Hive, has also been made fully open source.

Advantages:

- ✓ Decrease in development time. This is the biggest advantage especially considering vanilla map-reduce jobs' complexity, time-spent and maintenance of the programs.
- ✓ User can control the execution of every step.
- ✓ Speaking of UDFs, you could write your UDFs (User Defined Function) in Python.
- ✓ Pig is one of the best tool to make the large unstructured data to structured.
- ✓ Pig as a base pipeline where it does the hard work and you just apply your UDF in the step that you want.

Disadvantages:

- ✓ Not mature. Even if it has been around for quite some time, it is still in the development.
- ✓ Data Schema is not enforced explicitly but implicitly.

7) WibiData

WibiData is yet another technology which attempts to bridge the disparity between big data and more traditional analytics (I'm noticing a trend here). Built on top of HBase, WibiData combines web analytics with the power of Hadoop. It's easy to see why such a technology has caught on, as it's fairly rare these days to see a web master who hasn't dealt with unstructured data at least in passing. To them, WibiData is probably a breath of fresh air; a solution which allows them to more easily work with their user data and respond to user behavior in real-time with better content, better decisions, and better communication.

Advantages:

- ✓ It allows web sites to better explore& work with their user data.
- ✓ Enabling runtime responses to user behavior.
- ✓ Application Integration: a framework for integrating with other applications

8) PLATFORA

Perhaps the greatest limitation of Hadoop is that it is a very low-level implementation of

MapReduce, requiring extensive developer knowledge to operate. Between preparing, testing and running jobs, a full cycle can take hours, eliminating the interactivity that users enjoyed with conventional databases. PLATFORA is a platform that turns user's queries into Hadoop jobs automatically, thus creating an abstraction layer that anyone can exploit to simplify and organize datasets stored in Hadoop.

Advantages:

- ✓ One big benefit of Platfora is that it frees enterprises from having to build large data warehouses for the information**.**
- ✓ It provide users with graphical images of the data.
- ✓ It helps users visualize what they are exploring, giving them deep illumination into a wide variety of business data that they already have.

9) Storage Technologies

As the data volumes grow, so does the need for efficient and effective storage techniques. The main evolutions in this space are related to data compression and storage virtualization.

Advantages:

- ✓ Stored files can be accessed from anywhere via Internet connection.
- ✓ Businesses and organizations can often reduce annual operating costs by using cloud storage.

Disadvantages:

- ✓ Many IT administrators are slow to implement the technology because the transition from a non-virtualized to a virtualized environment isn't always a smooth process.

10) SkyTree

SkyTree is a high-performance machine learning and data analytics platform focused specifically on handling Big Data. Machine learning, in turn, is an essential part of Big Data, since the massive data volumes make manual exploration, or even conventional automated exploration methods unfeasible or too expensive.

Advantages:

- ✓ It provides the ability to seamlessly migrate analytics projects to latest advance Big Data platforms.
- ✓ Skytree can install on each Hadoop node for greater scalability.
- ✓ Skytree's technology scales to handle more data, faster and more accurately than any other approach available.
- ✓ Skytree makes machine learning available to the enterprise, without the need for advanced data scientist expertise

As we can see, all of these technologies are closely associated with the cloud. Most cloud vendors are already offering hosted Hadoop clusters that can be scaled on demand according to their user's needs. Also, many of the products and platforms mentioned are either entirely cloud-based or have cloud versions themselves.

The scope of big data analytics continues to expand:

Early interest in big data analytics focused primarily on business and social data sources, such as e-mail, videos, tweets, Facebook* posts, reviews, and Web behavior. The scope of interest in big data analytics is growing to include data from intelligent systems, such as in-vehicle infotainment, kiosks, smart meters, and many others, and device sensors at the edge of networks—some of the largest-volume, fastest-streaming, and most complex big data. Ubiquitous connectivity and the growth of sensors and intelligent systems have opened up a whole new storehouse of valuable information. Interest in applying big data analytics to data from sensors and intelligent systems continues to increase as businesses seek to gain faster, richer insight more cost effectively than in the past, enhance machine-based decision making, and personalize customer experiences.

VII. CLOUD AND BIG DATA

Big Data and cloud computing go hand-in-hand. Both Cloud and Big Data is about delivering value to enterprise by lowering the cost of ownership. The cloud enables big data processing for enterprises of all sizes by relieving a number of problems, but there is still complexity in extracting the business value from a sea of data. Forrester has defined big data as

"*Technologies and techniques that make capturing value from data at an extreme scale economical*."

Cloud computing enables companies of all sizes to get more value from their data than ever before, by enabling blazing-fast analytics at a fraction of previous costs. This, in turn drives companies to acquire and store even more data, creating more need for processing power and driving a virtuous circle.

Cloud has glorified the "As-a-Service" Model by hiding the complexity and challenges involving in building a scalable elastic self-service application. The same is the requirement for Big Data Processing. Hadoop in a similar way hides the complexity of the large scale distributed processing from the end user perspective. The user write "Map-Reduce" programs or familiar known constructs with "Hive" or "Pig" and are able to seamless do the big data crunching without worrying about the complexity of node failures, linear scalability, replication, fault-tolerance elasticity etc., where Hadoop silently provides the large scale distributed capabilities behind the scene. Thus the simplification provided by Cloud and Big data is the prime reason for the mass adoption of Big Data and Cloud.

Cloud delivery models offer exceptional flexibility, enabling IT to evaluate the best approach to each business user's request. For example, organizations that already support an internal private cloud environment can add big data analytics to their in-house offerings, use a cloud services provider, or build a hybrid cloud that protects certain sensitive data in a private cloud, but takes advantage of valuable external data sources and applications provided in public clouds.

- ✓ Investments in big data analysis can be significant and drive a need for efficient, cost-effective infrastructure.
- ✓ Data services are needed to extract value from big data.

VIII. CONCLUSION

Big data needs a lot of space – so moving it to the cloud makes sense .Many organizations utilises the cloud for their big data analytics.

Benefits of implementing big data technology through cloud computing are cost savings in hardware and processing, as well as the ability to experiment with big data technology before making a substantial commitment of company resources. Several models of cloud computing services are available to the businesses to consider, with each model having trade-offs between the benefit of cost savings and the concerns data security and loss of control.

In production, the cloud is used for analytics (17%), and real-time and batch processing (43%). According to the survey, organizations are using the cloud as a platform for more production, with 56% interested in putting more than 10 TB of data in the cloud, and 20% willing to put more than 100 TB of data in the cloud.Customer demand and mobility were among others. Interestingly, the pollsters noted that business and IT agreed on the cloud's positive traits.

All of these results point us to the conclusion that there should be more information and education about the cloud and its big data analytics attributes before more enterprises will embrace the cloud's technology.

REFERENCES:

[1] Big Data Processing in Cloud Computing Environments- ChangqingJi, Yu Li, WenmingQiu, UchechukwuAwada, Keqiu Li

[2] IBM Analytics-Tools- http://www.ibm.com/analytics/us/en/analytics-technology/

[3] Ericsson, 2013. Ericsson Mobility Report. [online] Ericsson. Available at: http://www.ericsson.com/res/docs/2013/ericsson-mobility-report-june-2013.pdf

[4] Chapter1.Cloud Computing Basics.pdf

[5] Gartner IT Glossary (n.d.). Retrieved from http://www.gartner.com/it-glossary/big-data/

[6] Online Learning for Big Data Analytics Irwin King, Michael R. Lyu and Haiqin Yang

[7] The Basic Of *Cloud Computing* at http://csrc.nist.gov/publications/drafts/800-145/Draft-SP-800-145_cloud-definition.pdf.

[8] Intel and IBM, Combat Credit Card Fraud with Big Data, 2013, Intel Corporation.

[9] BIG DATA ANALYTICS: Profiling The Use Of Analytical Platforms In User Organizations, By Wayne Eckerson, *director of Research, Business Applications and Architecture Group, Tech Target, September 2011*

[10] BigData: https://en.wikipedia.org/wiki/Big_data

PrajaktaRawoolis currently studying in Mumbai University at IMCOST, Thane since 2012, currently pursuing MCA with excellent academics. She is having interest in Studying New Technologies.

SwapnilSalviis currently studying in Mumbai University at IMCOST, Thane since 2012, currently pursuing MCA with excellent academics . He is having interest in Studying New Technologies.