www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 6 June 2015, Page No. 12757-12765

Cloud Architecture for Big Data

Pramila Joshi
Birla Institute of Technology Extension Centre Noida,
A-7, Sector – 1, Noida - 201301, Uttar Pradesh, India
pramila@bitmesra.ac.in

Abstract: ***With the advent of service computing and cloud computing, more and more services are emerging on the Internet, generating huge volume of data. The overwhelming service-generated data become too large and complex to be effectively processed by traditional approaches. How to store, manage, and create values from the service-oriented big data become an important research problem. On the other hand, with the increasingly large amount of data, a single infrastructure which provides common functionality for managing and analyzing different types of service-generated big data is urgently required.***

Nowadays, users are accessing multiple data storage platforms to accomplish their operational and analytical requirements. Efficient integration of different data sources is important. For example, an organization may purchase storage from different vendors and need to combine data with different format stored on systems from different vendors. Data integration, which plays an important role for both commercial and scientific domains, combines data from different sources and provides users with a unified view of these data. How to make efficient data integration with the 4V (volume, velocity, variety, and veracity) characteristics is a key research direction for the big data platforms.

To address this challenge, this paper describes how cloud and big data technologies are converging to offer a cost-effective delivery model for cloud-based big data analytics. It also includes how cloud computing is an enabler for advanced analytics with big data and how IT can assume leadership for cloud-based big data analytics in the enterprise by becoming a broker of cloud services.

Keywords: Big Data, Cloud, Hadoop, HDFS, MapReduce, DataNode, NameNode, closter, Racks, Block Report, Heartbeat, Replica

1. Introduction

Big data is the latest buzzword, used to describe a huge volume of both structured and unstructured data that is so large and complex that it's difficult to process them using traditional database and software techniques. In most of today's scenarios the data is too large or it moves too fast that it exceeds current processing capacity. Big data has the potential to give companies a competitive edge so that they can improve operations and make faster, more intelligent decisions. [1]

1.1 5 Vs of Big Data

Big Data is a great concept. It is here to change our world completely and is not a passing fad that will go away. To understand well, it is often described using five Vs: Volume, Velocity, Variety, Veracity and Value. [2]

Volume refers to the tremendous amounts of data being generated every second. Think of all Facebook and twitter messages, pictures, video clips, emails, sensor data etc. we are generating and sharing every second. And we are not referring to Terabytes but Zettabytes or Brontobytes. Only on Facebook alone we are sending 10 billion messages per day, pressing the "like' button 4.5 billion times and uploading 350 million new pictures each and every day. If we take into account all the data generated in the world till 2008, the same amount of data is being every minute ! This phenomenon makes data sets too large to store and analyse using traditional database technology. With the advent of big data technology we can now store and use these large data sets with the help of distributed systems, where different parts of the data are stored in different locations and integrated together by software. [3]

Velocity is the speed at which new data is being produced and then around. Just imagine how various social media messages go viral in seconds, how fast fraudulent activities

in credit card transactions can be detected and are checked for, or the ability of trading systems to analyse social media networks to pick up signals that trigger decisions to purchase or sell shares. Big data technology has enabled us now to analyse the data while it is being generated, without even storing it into databases.[3]

Variety refers to the different types of data. Earlier data used to be structured that could easily fit into tables or relational databases. In fact, more than 80% of the world's data is now unstructured take for example pictures, video clips or social media conversations and therefore can't easily be put into tables. Big data technology enables us to now exploit various types of data (structured and unstructured) including messages, social media updates, pictures, sensor data, video or voice recordings and bring them together with more traditional, structured data.[3]

Veracity refers to the fact how trustworthy the data is, despite being so disordered. With multiple forms of big data, quality and accuracy are not controllable (just think of hash tags posts with Twitter, abbreviations and colloquial speech as well as the reliability and accuracy of content) but big data and analytics technology now allows us to work with these type of data. The volumes often make up for the lack of quality or accuracy.[3]

Value: Then there is another V to take into account when looking at Big Data: Value! It is all well and good having access to big data but unless we can turn it into value it is useless which clearly implies that the most important V of Big Data is the 'value' it generates. It is important that businesses make a business case for any attempt to collect and invest into big data. It is so easy to fall into the buzz trap and embark on big data initiatives without a clear understanding of costs and benefits. [3]

2. Big Data Analysis Pipeline

As shown in the following fig, the analysis of big data typically involves multiple distinct phases.

First, big data are sampled and recorded from multiple data sources (e.g., from large-scale complex service computing systems).

Second, since the collected data may or may not be in a format ready for analysis, we need to extract certain information from the data producing sources to detect and correct the inaccurate records.

Third, because of the data may be heterogeneous in nature, data integration and representation are needed.

In the end, after all above phases are complete, data analysis and modeling is conducted on the resulting integrated and cleaned big data.

Finally, data interpretation and visualization are done since big data analytics alone is of limited value if users are not able to interpret the result. [4] [5]

Figure 1 : Big Data Analysis Pipeline

3. The rise of cloud computing and cloud data stores

The rise of cloud computing and cloud data stores has played a very important role in the emergence of big data. It has remarkable advantages over conventional physical deployments. However, cloud platforms come in various forms and sometimes have to be integrated with traditional architectures. [6]

This leads to a dilemma for decision makers who are managing big data projects. How and which cloud computing platform they should chose based on their computing needs. These projects may need unpredictable, shattering, or large computing power and a vast storage. At the same time business stakeholders expect quick, less costly, and dependable results. [7]

Big Data Projects need a highly professional cloud storage which should have almost 100% availability, highly durable, and has the ability to scale from bytes to petabytes. The most prominent solution is seen in Amazon's S3 cloud storage with 99.9% monthly availability and 99.999999999% durability per year which is even less than an hour outage per month. The durability can be explained with an example. Suppose a customer stores 10,000 objects then the probability of losing one object is every 10,000,000 years on average. S3 is able to achieve this by storing data in multiple devices with error checking and self-healing processes to detect and repair errors and device failures. Process is fully transparent to the user and does not require any actions or knowledge of facts. A company can also construct and achieve a similar reliable storage but that would mean a heavy capital expenditures and big operational challenges on its part. Companies like Facebook or Google have the expertise and scale to do this economically. But big data projects and start-ups, however,

can be benefitted only by using a cloud storage service. They can balance by trading off capital expenditure for an operational one, which is excellent idea since it requires no capital expenditure or risk. It provides reliable and scalable storage solutions of a quality otherwise unachievable. [8]

The idea empowers new big data projects with a feasible option to start on a small scale with low budget. Once the product is successful these storage solutions can scale virtually abundantly. Cloud storage is a highly effective unlimited data sink. Its advantage is that many projects in order to improve their computing performances also scale horizontally. In horizontal scaling when data is copied in parallel by cluster or parallel computing processes the throughput scales linear with the number of nodes reading or writing. [75%] [8]

4. What is cloud storage

Cloud storage is a kind of service in which data is maintained, managed and backed up remotely and made available to users over a network (typically the Internet). Cloud Providers offer services that can be classified into three categories.

4.1. Software as a Service (SaaS):

In SaaS model, customers are offered an application model, as a service on demand. The application software runs on the cloud & multiple end users are serviced. It has two benefits. First, the customer doesn't need to do beforehand investments in servers or software licenses and second, the provider is also benefitted in the form of reduced costs, since it only has to maintain and host a single application. Today SaaS is offered by companies such as Google, Salesforce, Microsoft, Zoho, etc. [9] [10] [11] [12]

4.2. Platform as a Service (PaaS):

In Paas, a development platform and environment is provided as a service so that customers can build, run and manage their own applications, without worrying about building and manage complexity of infrastructure. In order to manage the scalability requirements of the applications, Paas service provides a predefined combination of Operating System and application servers, such as LAMP platform (Linux, Apache, MySql and PHP), restricted J2EE, Ruby etc. Google's App Engine, Force.com, etc are some of the popular Paas examples.

4.3. Infrastructure as a Service (IaaS):

IaaS offers basic storage and computing capabilities as virtualized services over the network. Resources such as hard disk, RAM, CPU cores, servers, data centre space, storage systems, networking equipment etc. are tied together and made available on rent to handle workloads. The customer can build and deploy their own software on this rented infrastructure. Amazon, GoGrid, 3 Tera, etc. are some common examples.

5. Types of Clouds

Cloud computing can be categorized in three forms: public clouds, private clouds, and hybrids clouds. Different types have different levels of security and management features. [13]

5.1 Public Clouds

A public cloud is the one in which the resources are available to general public over the Internet. Although public clouds are efficient in shared resources, they are not as secure as private clouds. A public cloud benefits are [14]

- They are easy and inexpensive to set-up because hardware, application and bandwidth costs are borne by the provider.
- Scalability to meet needs.
- No resources are wasted because you pay for what you use.
- They are good for collaboration projects.

5.2 Private Clouds

In a private cloud the services and infrastructure are maintained on a private network. Private clouds provide high level of security and control over enterprise and customer data, because it is implemented within the enterprise firewall, but the company still has to bear the cost purchasing and maintaining all the software and infrastructure. A private cloud is the obvious choice when [15]

- Your business revolves around your data and applications. Therefore, control and security are of supreme importance.
- Your company is part of an industry that must adhere to the norms of strict security and data privacy issues.
- Your business is capable and large enough to run a next generation cloud data center efficiently and effectively on its own.[16]

5.3 Hybrid Clouds

A hybrid cloud is a mix and integrated service of both public and private cloud service platforms. The organization offers some resources in house and some are managed externally. A hybrid cloud manages each aspect at your business in the most efficient environment possible. Problem is that the organization has to keep track of multiple different security platforms and ensure that all aspects its business can communicate with each other. Here are a couple of situations which suggest when to go for a hybrid environment. [17]

- If your company has dynamic or highly changeable workloads highly used in Big data processing tasks.
- Hybrid clouds can be used as storage to a company's accumulated business, sales, test and other data. The company can then run analytical queries in the public cloud, which can be scaled to support demanding distributed computing tasks.
- Your company offers services that deal in different vertical markets. Here public cloud can be used to interact with the clients but private cloud keeps their data secured. [18]

6. Hadoop as Big Data Solution on Cloud

The emergence of Hadoop has changed the big data scenario. Hadoop can deal with structured, unstructured and semi-structured data sources and help you gain new improved business insights. Also, the huge volumes of data which were previously too costly to store can now be collected and analysed in one place at an affordable price. [19] Hadoop as a Service, as offered by Qubole Data Service (QDS) is a cloud computing solution for big data which makes medium and large-scale data processing accessible, easy, fast and less costly. Operational challenges of running Hadoop are reduced or eliminated, so you can focus on business growth. [19]

6.1 Advantages of Hadoop – why Hadoop in the cloud

Since cloud computing offers unlimited scaling and on-demand access to compute and storage capacity, it is the perfect match for big data processing. Qubole's Hadoop as a Service has many benefits over on-premise solutions. [19]

- **On-Demand Elastic Cluster**

One big advantage of Hadoop cluster is , unlike static, on-premise clusters, they are able to scale up or down base on data processing requirements. It is easy to add or remove extra nodes automatically from clusters depending on data size to improve performance. [19]

- **Integrated Big Data Software**

Hadoop platform consists of two main components, HDFS and MapReduce. HDFS is a reliable fully distributed file system which includes full integration with the Hadoop MapReduce, Hive, Pig, Oozie, Sqoop, Spark and Presto. Data integration and data pipelining provide a complete solution that works with your current pipeline.

- **Simplified Cluster Management**

One need not worry about devoting extra time and resources to manage nodes, set up clusters and infrastructure scaling because Qubole Data Service offers a fully managed Hadoop-based cluster.

- **Lower Costs**

No beforehand investment is required for on-site hardware or IT support ,in Hadoop Cloud. Costs are slashed by 90% because of spot instant pricing as compared to on-demand instances. Pay as you go model allows you pay for space only when you use it with auto-scaling clusters facility.

7. What is Hadoop

Hadoop is an open source software project for storing and processing huge volumes of structured and unstructured data. It uses multiple commodity servers as hardware which are highly fault tolerant and scalable to any extent. It is designed in a way that it can manage petabytes of data by scaling up to hundreds or thousands of physical storage servers or nodes. [20]

Hadoop was developed by Yahoo as an open source project in 2005. It was written basically in Java to handle distributed data storage and distributed processing of large data sets which is today widely considered to be components of Hadoop. A program to process large data sets is distributed across the nodes which (preferably) process the part of the data already stored with them. In this technique programs are deployed to the nodes rather than data making it faster since programs are smaller in size as compared to large volumes of data. This is a radical change from conventional data processing where data is streamed and pipelined to the processing cores. In Hadoop large data sets are split between a map and a reduce phase. Hadoop consists of a storage part (HDFS) and a processing part (Map Reduce). [19]

Hadoop, uses a concept known as MapReduce that is composed of two separate functions.

1. The Map step inputs data and breaks it down for processing across nodes within a Hadoop instance. These "worker" nodes may in turn break the data down further for processing.

2 In the Reduce step, the processed data is then collected back together and assembled into a format based on the original query being performed.

7.1. The Hadoop Distributed File System (HDFS)

HDFS is a fault tolerant and self-healing distributed file system designed to turn a cluster of industry standard servers into a massively scalable storage volume. It is designed and developed mainly to handle large-scale data processing workloads where scalability, flexibility and throughput are very crucial. HDFS can accept data in any format regardless of its schema. It is capable of optimizing for high bandwidth streaming, and can also scale up to of 100 PetaBytes and beyond. [21]

7.2. Key HDFS Features:

- Scale-Out Architecture - Add servers to increase capacity
- High Availability - Serve mission-critical workflows and applications
- Fault Tolerance - Automatically and seamlessly recover from failures
- Flexible Access – Multiple and open frameworks for serialization and file system mounts
- Load Balancing - Place data intelligently for maximum efficiency and utilization
- Tunable Replication - Multiple copies of each file provide data protection and computational performance
- Security - POSIX-based file permissions for users and groups with optional LDAP integration [22]

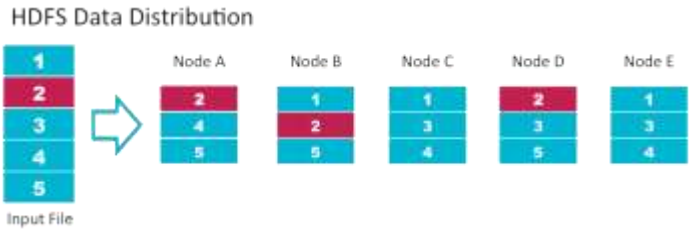

Figure 2 : HDFS Data Distribution

Data in HDFS is replicated across multiple nodes for compute performance and data protection.

7.3. MapReduce

MapReduce is heart of Hadoop which is highly scalable parallel processing framework across hundreds and thousands of servers in Hadoop cluster that works hand in hand with HDFS. Instead of moving data to the computing/processing location , processing is done at the location itself. Data storage and computation both coexist on the same physical nodes in the cluster. MapReduce processes exceedingly large amounts of data without being affected by traditional bottlenecks like network bandwidth by taking advantage of this data proximity. [22]

7.4. Key MapReduce Features:

- Scale-out Architecture - Add servers to increase processing power
- Security & Authentication - Works with HDFS and HBase security to make sure that only approved users can operate against the data in the system
- Resource Manager - Employs data locality and server resources to determine optimal computing operations
- Optimized Scheduling - Completes jobs according to prioritization
- Flexibility – Procedures can be written in virtually any programming language
- Resiliency & High Availability - Multiple job and task trackers ensure that jobs fail independently and restart automatically

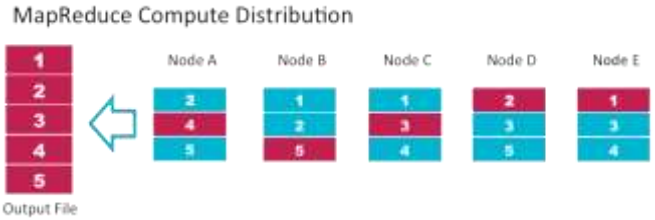

Figure 3 : Map Reduce Compute Distribution

MapReduce divides workloads into multiple tasks executed in parallel for fast access.

7.5. Challenges of employing Hadoop

The challenge of employing Hadoop is two-fold.

Firstly, setting up and managing a Hadoop cluster and doing it efficiently and in a cost effective manner is largely unachievable for small and medium sized organizations. But big enterprises like Yahoo or Facebook who do this in house for their large scale data, is more economical. [23]

Secondly, writing programs for utilizing Hadoop is a complex process. Nowadays Various tools like Hive and Pig

are available in the ecosystem around Hadoop which make Big Data processing accessible focusing on what to do with the data and avoid complexity of programming. [23]

8. How it is done

Typical big data projects focus on scaling or adopting Hadoop for data processing. For large scale data processing MapReduce has proved highly effective. It has tools like Hive and Pig on top of Hadoop so that it is feasible to process huge data sets easily. For example what Hive does is transform SQL like queries to MapReduce jobs. [24]

Lets us understand how it is done with the help of clouds. Massively large data sets of log files are collected in a cloud. Amazon's S3 is one such example of cloud data sink. Hadoop, Hive, and Pig are also used to access data from database directly with Hadoop. Qubole is a leading provider of cloud based services in this space. They come with unique database adapters that can unlock data instantly, which otherwise would be inaccessible or require significant development resource. One great example is their mongoDB adapter. It gives Hive table like access to mongoDB collections. Qubole scales Hadoop jobs to extract data as quickly as possible without overpowering the mongoDB instance.

A cloud service provider offers Hadoop clusters which scale automatically with the demand of the client. This helps in attaining performance to the maximum for massive jobs and optimal savings when little and no processing is going on. A good example is Amazon Web Services Elastic MapReduce, which allows scaling of Hadoop clusters. [25] However, the scaling is not done automatically with the demand but requires user actions.[24] The scaling itself is not optimal since it does not utilize HDFS well and loses data locality feature of Hadoop. This means that an Elastic MapReduce cluster wastes resources when scaling. Moreover Amazon's Elastic MapReduce needs a customer to explicitly request a cluster every time when it is required and removes it when it is no longer needed. There is also no user friendly interface for interaction with or exploration of the data which subsequently results in extra operational burden and excludes all but the most proficient users.

9. How Qubole handles Hadoop cluster

As far as Qubole is concerned it scales and handles Hadoop clusters differently. Here the clusters management is transparent to the user and no action is needed from the client. It stops the clusters when no activity is taking place clusters are stopped thus ending further expenses. The system can automatically detect demand for new clusters and starts a new cluster if needed e.g. when a user queries Hive. It does this even faster than Amazon raises its clusters on explicit user requests. The clusters here have a user defined minimum and maximum size and they can scale as needed so that they offer optimal performance at minimal expense.

10. HDFS Architecture

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It is similar with existing distributed file systems in many ways.[26] However it different from other distributed file systems in the way that it is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS is highly suitable for applications that have large data sets. Its throughput is very high when it accesses to application data. HDFS also enables streaming access to file system data by relaxing a few POSIX requirements. Though HDFS is now an Apache Hadoop subproject but it was originally developed as infrastructure for the Apache Nutch web search engine project.

10.1. NameNode and DataNodes

HDFS follows master/slave architecture. An HDFS cluster has a single namenode and multiple datanodes. NameNode is a master server who is responsible for managing the file system namespace and controls file access by clients. Each cluster has multiple DataNodes, which control the storage attached to the nodes that they run on. HDFS has a file system namespace which allows user data to be stored in files. Internally, a file is partitioned into one or more blocks and these blocks are stored in a set of DataNodes. The NameNode controls file system namespace operations for example opening, closing, and renaming files and directories. It also governs the mapping of blocks to DataNodes. The DataNodes handle read and write requests from the file system's clients. The DataNodes are also responsible for block creation, deletion, and replication upon instruction from the NameNode.[25][27]

The NameNode and DataNode are software which are designed to run on commodity machines. These machines run a GNU/Linux based operating system (OS). HDFS is developed using Java which means any machine that supports Java can run the NameNode or the DataNode software. Since Java language is highly portable it means that HDFS can be deployed on a variety of machines. A typical deployment example can be dedicated machine that runs only the NameNode software. And rest of the other machines in the cluster runs one instance of the DataNode software. The architecture does not allow multiple

DataNodes to be run on the same machine but in a real deployment that is rarely the case.

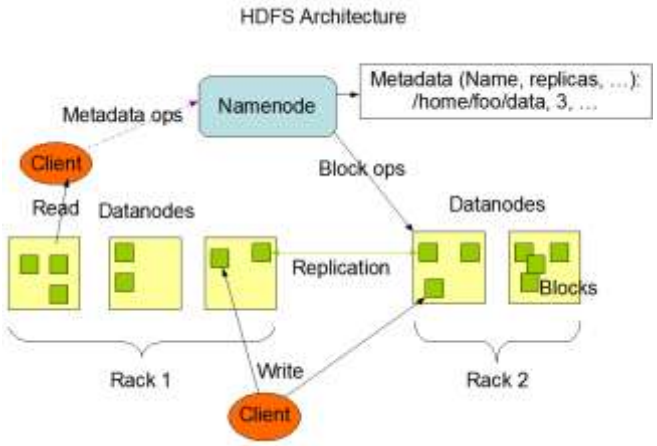

Figure 4 : HDFS Architecture

The architecture of the system is greatly simplified by the existence of a single NameNode. The NameNode is the controller and repository for all HDFS metadata. The design of the system is such that user data never flows through the NameNode.

10.2 The File System Namespace

HDFS supports a traditional hierarchical file organization. One can create directories and store files inside these directories. The file system namespace hierarchy is similar to most other existing file systems; one can create and remove files, move a file from one directory to another, or rename a file. HDFS does not yet implement user quotas. HDFS does not support hard links or soft links. However, the HDFS architecture does not preclude implementing these features. [25]

The NameNode follows the file system namespace. Name node records any change made to the file system namespace or its properties. An application can specify the replication factor of a file. A replication factor is number of copies/replicas of a file that should be maintained by HDFS. Namenode stores this information. [25]

10.3. Data Replication

HDFS is designed such that it can store very large files reliably in a large cluster across machines. Each file is stored as a sequence of same sized blocks except the last block. This promotes fault tolerance. The replication factor and block size are configurable per file. As discussed earlier an application can specify the number of replicas of a file and This replication factor can be specified at file creation time and can be changed later. Files in HDFS are write-once and strictly follow the rule of one writer at any time. [25]

The NameNode is responsible for all decisions regarding replication of blocks. Every Datanode in the cluster periodically sends a Heartbeat and a Blockreport to namenode. Presence of a heartbeat in form of receipt implies that the DataNode is functioning properly. List of all blocks on a DataNode is contained in a Blockreport.

10.4. Replica Placement: The First Baby Steps

HDFS reliability and performance is critically dependent on the placement of replicas. HDFS distinguishes itself from most other distributed file systems because of its optimization of replica placement. This feature requires lots of tuning and experience. Rack-aware replica placement policy ensures improved data reliability, availability, and network bandwidth utilization.

Instances of a large HDFS run on a cluster of computers that are widely spread across several racks. Communication between two nodes in different racks takes place switches. Mostly network bandwidth between machines in the same rack is greater than network bandwidth between machines in different racks.

Each DataNode belongs to some rack. The NameNode determines this rack id with the help of a process outlined in Hadoop Rack Awareness. Replicas are placed on unique racks according to a simple but non-optimal policy. This ensures minimum or no data loss when an entire rack collapses and permits use of bandwidth while reading data from multiple racks. On a component failure , evenly distributes replicas in the cluster make it easy to balance load. However, cost of writes goes up due to this policy because a write needs to transfer blocks to multiple racks.

10.5. Replica Selection

In order to minimize global bandwidth consumption and read latency, HDFS responds to a read request from a replica that is closest to the reader. If a replica exists on the same rack as the reader node, then that replica used for the read request. In case of HDFS cluster spanning across multiple data centers, a replica which is resident in the local data center is chosed.

10.6. Safemode

Safemode is a special state possessed by Namenode on startup. In the safemode state replication of data blocks does not occur. DataNodes send Heartbeat and Blockreport messages to NameNode. DataNode has a blockreport that

contains the list of data blocks being hosted. Each block can have a specified minimum number of replicas. When the minimum number of replicas of that data block has checked in with the NameNode, that is when a block is considered safely replicated. NameNode comes out of the Safemode state after a configurable percentage of safely replicated data blocks checks in with the NameNode (plus an additional 30 seconds). It then determines the list of data blocks (if any) that still have fewer than the specified number of replicas. The NameNode then replicates these blocks to other DataNodes.

11. Conclusions and future directions

Organizations facing architecture decisions should evaluate their security concerns or legacy systems clearly before going for a potentially complex private or hybrid cloud deployment. A public cloud solution is often easily achievable. The questions that need to be asked are which new processes can be deployed in the cloud and which legacy process are feasible to transfer to the cloud in order to adopt a public cloud solution. It makes sense to retain a core data set or process privately but most big data projects are doing well in the public cloud due to the flexibility it provides.[24]

In a growing number of companies, business users already consume IT as a service. IT can continue to extend this role to brokering cloud-based big data analytics services. As a cloud services broker, your role is to weigh user needs against the available delivery options for your organization. This means developing a strategy for private, public, and hybrid services; driving discipline into the selection of cloud service providers; and negotiating and establishing contracts with potential cloud service providers, among other similar tasks. Organizationally, this can reduce risk and better utilize existing investments in private cloud technologies. Individual users benefit by getting the right solution to meet their needs. IT can quickly demonstrate value to the business by partnering with users to:

• Select the right private or public cloud implementation for their needs by defining technology requirements, assessing risk, and specifying deployment requirements based on corporate governance policies and regulatory compliance requirements. For example, certain workloads can be managed only in a private cloud in a specific location.

• Build or work effectively with a technology partner to develop services as required.

• Evaluate outside services for design, delivery, customization, pricing, privacy, integration, security, and support.

• Provision services from internal and external sources so that they appear seamless to users.

• Develop terms with vetted cloud service providers.

• Manage existing services, including service level agreements (SLAs) and service life cycle.

As a service broker, IT collaborates with the business on the best way to use technology for competitive advantage. With cloud-based big data analytics, the objective must be to provide the right solution for users' needs balanced against corporate governance policies, existing IT resources, performance requirements, and overall business goals. In most IT departments today, providing this consultative approach to service will require IT to reorganize to remove silos, hire or develop team members with new skills, and encourage a strong partnership with the business. The payoff will be significant, especially for big data analytics projects, which require collaboration between IT technology experts, business users, data scientists, and others who can help develop the appropriate analytics plan and algorithms to extract meaningful insights from the data.

References

1. http://www.webopedia.com/TERM/B/big_data.html
2. http://www.jeffince.co.uk/big-data--analytics.html
3. https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know
4. "big data science : Myth and reality", Jan 2015
5. Zibin Zheng, Jieming Zhu, and Michael R. Lyu , "Service-generated Big Data and Big Data-as-a-Service: An Overview", June 2013
6. Dona Sarkar, Dr. Asoke Nath, "Big Data – A Pilot Study on Scope and Challenges", www.ijarcsms.com Volume 2, Issue 12, December 2014
7. http://www.qubole.com/resources/articles/big-data-cloud-database-computing
8. Elmustafa Sayed Ali Ahmed and Rashid A.Saeed , "A Survey of Big Data Cloud Computing Security", "Academia.edu", , Dec 2014
9. http://www.thbs.com/knowledge-zone/cloud-computing-overview
10. Maria Evans, Tam Huynh, aria Evans, Tam Huynh, Kieu Le, Mark Singh, "cloud storage", 2011

11. Blerim Rexha, Blerta Likaj and Haxhi Lajqi ,"Assuring security in private clouds using ownCloud", ijacit.com, 2012
12. Meenaskhi, Anju Chhibber , "An overview on cloud computing technology", "international journal of advances in computing and information technology" , 2012
13. Judith Hurwitz, Robin Bloor, Marcia Kaufman, and Fern Halper, "Comparing Public, Private, and Hybrid Cloud Computing Options" , "Cloud Computing For Dummies", Oct 2009
14. Vineetha V, "Performance Monitoring in cloud", http://www.infosys.com/engineering-services/features-opinions/Documents/cloud-performance-monitoring.pdf, Jan 2012
15. http://www.brainypro.com/cloudComputing.html, Jan 2013
16. Shivi G, T Narayanan, "A Review on Matching Public, Private, and Hybrid Cloud Computing Options", International Journal of Computer Science and Information Technology Research Vol. 2, Issue 2, April-June 2014
17. S M Hemlatha, S ganesh, "A Brief Survey on encryption schemes on Cloud Environments", International Journal of Computer and organization trends", Vol 3, Issue 9, Oct 2013
18. http://searchcloudcomputing.techtarget.com/definition/hybrid-cloud, Mar 2015
19. http://www.qubole.com/hadoop-as-a-service, Jan 2015
20. http://www.computerweekly.com/feature/Big-data-storage-Hadoop-storage-basics
21. www.cloudera.com/content/cloudera/en/.../hdfs-and-mapreduce.html, march 2013
22. Atul Patil, T I Bagban, "Improved Utilization of Infrastructure of clouds by using Upgraded Functionalities", "International Journal of Innovative Research in Advanced Engineering", Vol 1, Issue 7, Aug 2014
23. www.qubole.com/resources/articles/what-is-hadoop
24. www.qubole.com/resources/articles/big-data-cloud-database-computing
25. Trapti Sharma , "Modelling Cloud Services for Big Data using Hadoop", International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015
26. hadoop.apache.org › Hadoop › Apache Hadoop Project Dist POM
27. Xianglong Ye, Mengxing Huang, Donghai Zhu, Peng Xu, "A Novel Blocks Placement Strategy For Hadoop", Conference IEEE/ACIS 11th International Conference on Computer and Information Science, 2012
28. Elmustafa Sayed Ali Ahmed and Rashid A.Saeed, "A Survey of Big Data Cloud Computing Security", "Academia.edu", Dec 2014
29. R. Sharir, "Cloud database service: The difference between dbaas, daas and cloud storage - what's the difference" , http://xeround.com/blog/2011/02/dbaas-vs-daas-vs-cloud-storage-difference, 2011.
30. M. Lenzerini, "Data integration: A theoretical perspective," in Proceedings of the 21st ACMSIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2002
31. E. Slack, "Storage infrastructures for big data workflows," Storage Switchland, LLC, Tech. Rep., 2012.
32. Zibin Zheng, Jieming Zhu, and Michael R. Lyu, "Service-generated Big Data and Big Data-as-a-Service: An Overview", June, 2013.
33. http://en.wikipedia.org/wiki/MapReduce
34. http://www.cloudera.com/content/cloudera/en/products-and-services/cdh/hdfs-and-mapreduce.html
35. http://www.qubole.com/resources/articles/what-is-hadoop/#sthash.Cnsov1wL.dpuf
36. http://www.qubole.com/resources/articles/big-data-cloud-database-computing/#sthash.p8s4FGVu.dpuf]
37. An Enterprise Architect's Guide to Big Data Reference Architecture Overview oracle enterprise architecture white paper | Feb 2015
38. https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html
39. http://hadoop.apache.org/hdfs
40. http://www.qubole.com/resources/articles/big-data-cloud-database-computing/#sthash.p8s4FGVu.dpuf