Modelling Cloud Services for Big Data using Hadoop

Trapti Sharma
Research Scholar Manipal University Jaipur/SCIT Jaipur, India

Abstract. This Paper presents various cloud services for Big Data by using Hadoop. Today we live in the digital world. With increased digitization the volume of structured and unstructured data being created and stored is exploding. The data is being generated from various sources- transactions, social media, sensors, digital images, audios and videos. In addition to business and organizations individuals contribute to the data volume. Big data applications assist you make sense of your data and discover critical insights that drive business elaborations. But with large volumes of data, a lack of expertise, and large beforehand investments, your project costs can fastly spiral out of control. Cloud computing is a recently developed new technology for complex systems with massive scale service splitting, which is different from the capital sharing of the grid computing networks. Cloud reliability analysis and modeling are not easy tasks because of the complexity and large scale of the system. Cloud computing has aroused wide research interests and has been accepted by industry. Services are playing the essential role in cloud computing as cloud computing refers to "both the applications delivered as services over the Internet and the hardware and systems software in the datacenters that provide those services". Therefore, service-oriented architecture should play an important role in cloud computing. In addition, one of the characteristics of cloud computing is to make services available on demand. Given a group of services, different demands may involve different set of services and in different order. This is related to services reuse and composition.

Keywords: Big Data, cloud computing, hadoop, cloud services.

INTRODUCTION

Cloud computing is characterized as making services available on demand. Services are essential to cloud computing, in which all the applications are delivered as services and the beneath datacenter hardware and software are referred as cloud. Big data environments need clusters of servers to help the tools that process the large capacity, high velocity, and varied styles of big data. IT corporations are increasingly examining to cloud computing as the structure to assist their big data programmes. While organizations frequently keep their most sensitive data in-house, large volumes of data such as social media data may be revealed externally. Analyzing the data where it occupied—either in internal or public clouds—makes big data in the cloud more fascinating in terms of cost and obtaining faster insights. With the growth in the amount of unstructured data from social forum, more value can be separated from big data when structured data sets are combined and examined to gain total advantage. It is a fact that data that is very big to process is also too big to move anywhere, so it's just the analytical program which needs to be moved—not the data. This is feasible with public clouds, as almost all of the public data sets such as Facebook, Twitter, financial data, and aggregated industry-specific data live in the cloud and it becomes more cost-effective for the organizations to broken in this data in the cloud itself.

BIG DATA AND CLOUD COMPUTING

The rise of cloud computing and cloud data stores have been a precursor and facilitator to the emergence of big data. Cloud computing is the commodification of computing time and data storage by means of standardized technologies. Cloud computing employs visualization of computing resources to run numerous standardized virtual servers on the same physical machine. Cloud providers achieve with this economies of scale, which permit low prices and billing based on small time intervals . Typical big data projects focus on scaling or adopting Hadoop for data processing. MapReduce has become a de facto standard for large scale data processing. Tools like Hive and Pig have emerged on top of Hadoop which make it feasible to process huge data sets easily. Hive for example transforms SQL like queries to MapReduce jobs. It unlocks data set of all sizes for data and business analysts for reporting and greenfield analytics projects. Data can be either transferred to or collected in a cloud data sink like Amazon's S3, e.g. to collect log files or export text formatted data. Ideally a cloud service provider offers Hadoop clusters that scale automatically with the demand of the customer. This provides maximum performance for large jobs and optimal savings when little and no processing is going on. Amazon Web Services Elastic MapReduce, for example, allows scaling of Hadoop clusters. However, the scaling is not automatically with the demand and requires user actions. The scaling itself is not optimal since it does not utilize HDFS well and squanders Hadoop's strong point, data locality. This means that an Elastic MapReduce cluster wastes resources when scaling and has diminishing return with more instance.

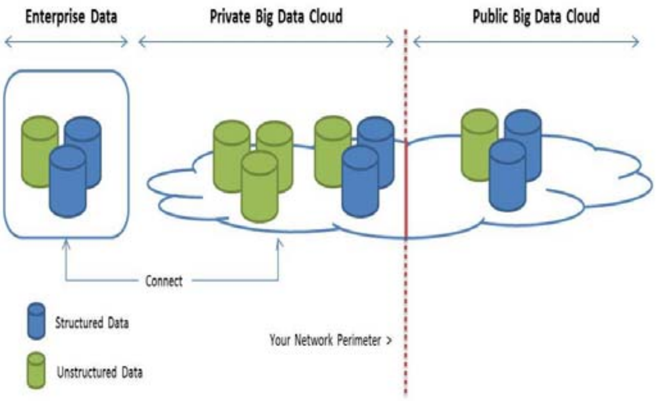

Fig.1 Big Data and Cloud

Cloud Services

Cloud services means services made available to users on demand via the Internet from a cloud computing provider's servers as opposed to being provided from a company's own on-premises servers. Cloud services are designed to provide easy, scalable access to applications, resources and services, and are fully managed by a cloud services provider. A cloud service can dynamically scale to meet the needs of its users, and because the service provider supplies the hardware and software necessary for the service, there's no need for a company to provision or deploy its own resources or allocate IT staff to manage the service. Examples of cloud services include online data storage and backup solutions, Web-based e-mail services, hosted office suites and document collaboration services, database processing, managed technical support services and more. Cloud Services provides a staging environment for testing new releases without impacting the existing one, reducing the chances of unwelcome customer downtime. When you're ready to deploy the new release to production, just swap the staging environment into production.

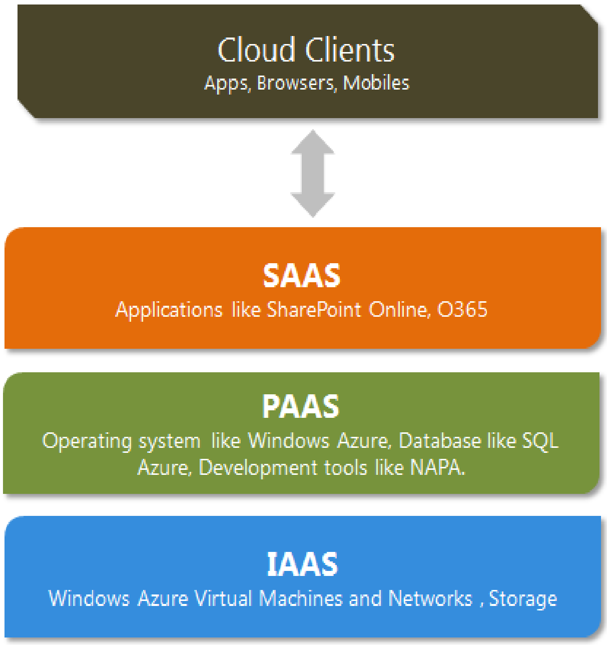
Cloud Clients
Apps, Browsers, Mobiles
SAAS
Applications like SharePoint Online, O365
PAAS
Operating system like Windows Azure, Database like SQL Azure, Development tools like NAPA.
IAAS
Windows Azure Virtual Machines and Networks , Storage

Fig.2 Cloud Services

Cloud Computing for Big Data Storage and Processing

Cloud computing and big data are the two key emerging disruptive technologies; these are accelerating business innovation and enabling new, disruptive solutions. Enterprises are seeking answers to some of their key business imperatives through Big data analysis like – modeling true risk, flexible supply chains, loyalty pricing etc. To address these business requirements a "Data Cloud" with an elastic and adaptive infrastructure such as public and private cloud platform for enterprise data warehousing and business intelligence functions is being considered. Cloud computing and Big Data together allows analysts and decision makers to discover new insights for intelligence analysis, as demonstrated by Google, Yahoo, and Amazon. A large set of data now exists in the cloud (private, public and hybrid) space and many organizations and applications are also making their way into cloud platforms every day. Benefits of Big Data in cloud: low cost by using infrastructure via utility cloud model there by reducing infrastructure costs, Fast turn is the around time for infrastructure that is On-demand provisioning of cloud infrastructure.

Hadoop

Hadoop consists of Hadoop Map Reduce and Hadoop Distributed File System (HDFS). Hadoop Map Reduce is an implementation of Map Reduce designed for large clusters, while HDFS is a distributed file system designed for batch-oriented workloads. Each job in Map Reduce has two phases. First, users specify a map function that processes the input data to generate a list of intermediate key-value pairs. Second, a user-defined reduce function is called to merge all intermediate values associated with the same intermediate key . HDFS is used to store both the input to the map and the output of the reduce, but the intermediate results, such as the output of the map, are stored in each node's local file system. A Hadoop implementation contains a single master node and many worker nodes. The master node, called the Job-Tracker, handles job requests from user clients, divide these jobs into multiple tasks, and assign each task to a workerThe Apache Hadoop project develops an opensource platform for reliable, scalable, distributed computing. It consists of many subprojects such as Hadoop Common, Chukwa, HBase, HDFS, etc. Hadoop Common provides common utilities for the other subprojects. Through making use of a number of nodes, Hadoop establishes a super distributed computational system. Hadoop provides a distributed file system, which stores application data in a replicated way, and also gives user high throughput ability of accessing the data on HDFS. As a MapReduce system, it runs jobs very fast by using the aggregated power. There are two web interfaces along with Hadoop. User can browse the HDFS and track jobs through these two interfaces in a web browser. The efficiency of Hadoop depends on the file size, number of files, the number of nodes in the cluster, bandwidth connecting the nodes, etc. Especially, Hadoop is not good at dealing with big amount of small files. Hadoop is implemented in Java, and it has been widely tested in production. Hadoop was originally designed for processing batch oriented processing jobs, such as creating web page indices or analyzing log data. Hadoop is not used for Online Transaction Processing workloads and Online Analytical Processing or Decision Support System workloads where data are randomly or sequentially on structured data like a relational data to generate reports that provide business intelligence. However being reliable, (both in terms of computation and data), fault tolerant, scalable and powerful, Hadoop is now widely used by Yahoo!, Amazon, eBay, Facebook, IBM, Netflix, and Twitter to develop and execute large-scale analytics or applications for huge data sets.

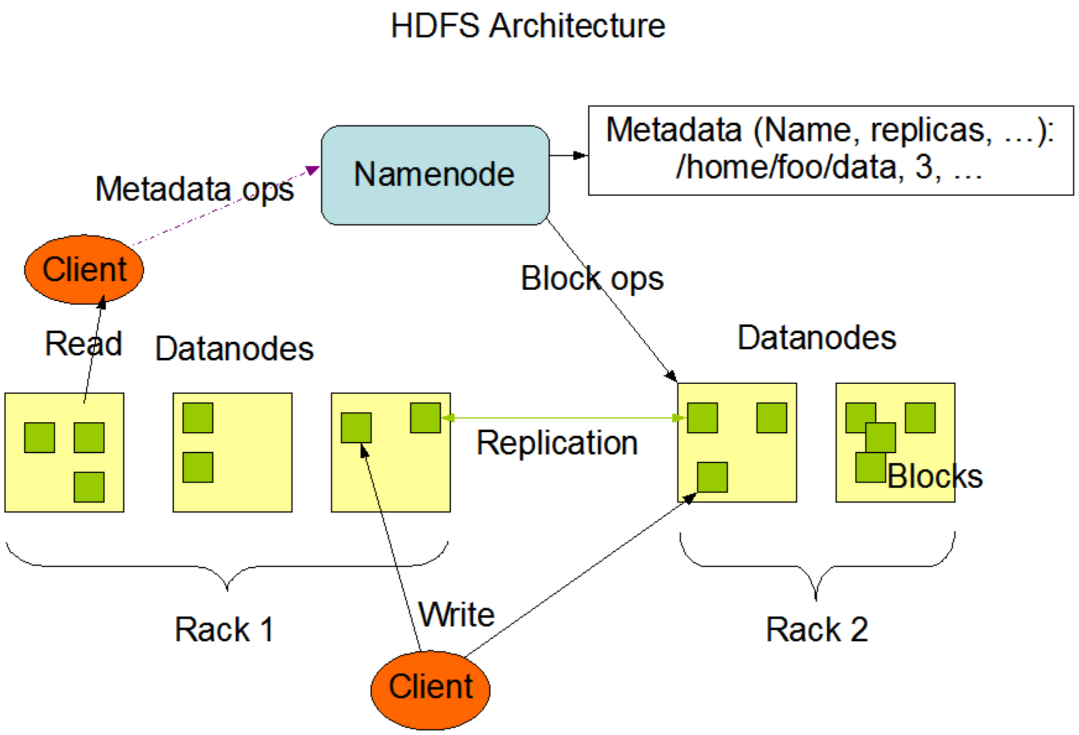

Fig.3 HDFS Architecture

CONCLUSION AND FUTURE WORK

Cloud computing has become a viable, mainstream solution for data processing, storage and distribution, but moving large amounts of data in and out of the cloud presented an insurmountable challenge for organizations with terabytes of digital content. A pioneer in the enablement of high-speed data-intensive workflows throughout the enterprise, Aspera has now unlocked the cloud for big data with its industry-leading high-speed transport solutions. Cloud computing provides enterprises cost-effective, flexible access to big data's enormous magnitudes of information. Big data on the cloud generates vast amounts of on-demand computing resources that comprehend best practice analytics. Both technologies will continue to evolve and congregate in the future.

REFERENCES

1. Global Forensic Data Analytics Survey 2014, Big risks require big data thinking.
2. Michael minelli, Michele chambers, Ambigadhiraj, Big data big analytics: emerging business intelligence and analytic trends for today's businesses.
3. Devesh Kumar Srivastava :Big Challenges in Big Data Research, CiiT International Journal of Data Mining and Knowledge Engineering , Vol 6, No 7(2014).
4. Gartner Business Intelligence & Analytics Summit, October 2014.
5. Eileen McNulty, "Understanding Big Data: The Seven V's", May 22, 2014.
6. Mike Barlow," Real-Time Big Data Analytics: Emerging Architecture", February 2013.
7. Australian Government Department of Finance and Deregulation, "Big Data Strategy –Issues Paper", March 2013.
8. Ericsson White Paper, Big Data Analytics, August 2013.
9. V. Laxmikanth, Managing Director, Broadridge Financial Solutions (India) Private Limited: "Getting Started with Greenplum for Big Data Analytics", October 2013.
10. Laney, Douglas. "The Importance of 'Big Data': A Definition". Gartner. Retrieved 21 June 2012.
11. Margaret Rouse, "Software as a Service BI (SaaS BI)," SearchBusinessAnalytics, June 2012.
12. Chris Stolte, Dan Jewett, and Pat Hanrahan, "A New Approach: Rapid Fire Business Intelligence", January 2011.
13. McKinsey Global Institute,"Big Data: The Next Frontier for Innovation, Competition and Productivity," June 2011.
14. The Search for Analysts to Make Sense of Big Data. Yuki Noguchi. National Public Radio, Nov. 30, 2011.
15. Dave Beulke, Big Data Impacts Data Management: The 5 Vs of Big Data, November 2011.
16. J Magoulas, Roger; Lorica, Ben "Introduction to Big Data" (February 2009).
17. Jacobs, A. (6 July 2009). "The Pathologies of Big Data". ACM Queue.
18. Jian Li IBM Research in Austin, Big Data System and Architecture.
19. Robin Bloor, Big Data Analytics- This Time It's Personal.
20. Laney, Douglas. "3D Data Management: Controlling Data Volume, Velocity and Variety". Gartner. Retrieved 6 February 2001.