An Overview on Big Data: Technologies and their Applications

Dr. Kamal Gulati
Assistant Professor, Grade - III
Amity School of Insurance, Banking and Actuarial Science, Amity University, Noida, UP, India
Email:kgulati@amity.edu, drkamalgulati@gmail.com

Abstract: Big data might be petabytes (1,024 terabytes) or Exabyte of data consisting of billions to trillions of records from different sources (e.g. Web, sales, customer care, social media, mobile data and so on). Big Data refers to relatively large amounts of structured and unstructured data that is difficult to process using traditional database (RDBMS) and software techniques. This paper focuses on big data technologies, applications and commercially supported users. The Big Data technologies are Apache's Hadoop file systems, Map-Reduce model and NoSQL (NotOnlySQL) databases are discussed in this paper.

Keywords: Big data, HDFS, MapReduce, NoSQL Databases.

I. INTRODUCTION

Every few years, we come across the next big technological idea which radically transforms the businesses function by opening up new opportunities. Around 2.5 billion GB of data is generated every day, and more than 90 per cent of the data obtainable today has been invented in the earlier 3-4 years. This has primarily usage of mobile applications and social media. It's estimated that face book alone generates 15 Terabytes of data daily. For many years, enterprise organizations have hoarded growing stores of data. Growth in Big Data has led to significant infrastructure requirements to support the distributed processing of unstructured data. Big Data[1] technology mainly comprises Hadoop architecture that has a distributed file system, analytics and data storage platforms. Other than Hadoop, there are non-relational databases such as NoSQL (NotOnlySQL) databases and MPP systems that are scalable, network-oriented, semi-structured. Face book, Twitter, YouTube and Google are the vital examples.

Table 1 summarize the important characteristics of traditional and Big Data approach.

TABLE I. CHARACTERISTICS OF TRADITIONAL VS BIG DATA APPROACHES

Characteristics	Data Management	
	Relational Data	***Big Data***
Architecture	Centralized	Distributed
Data Volume	Tera-bytes	Peta-bytes to Exa-bytes
Data Relational	Known relation	Unknown/complex relation
Data Model	Static or schema-based	Dynamic or Schema-less
Data veracity	Related Data	Messy/imprecise
Data Velocity	Generated by human interaction	Machine generated data sensor, medical and FSI
Data source	Known	Heterogeneous
Scalability	Nonlinear	Liner
Data processing	Interactive	Batch and stream processing

II. BIG DATA TECHNOLOGIES

The data management in big data comprises the legacy systems as well as Hadoop-based systems[4] and NoSQL databases. Legacy systems comprises databases that repost and achieved structured data, i.e., RDBMS to store and evaluate structured data, and MPP systems to upgrade for large structured datasets. The dominant Big Data technologies are Apache's Hadoop file systems and No-SQL databases.

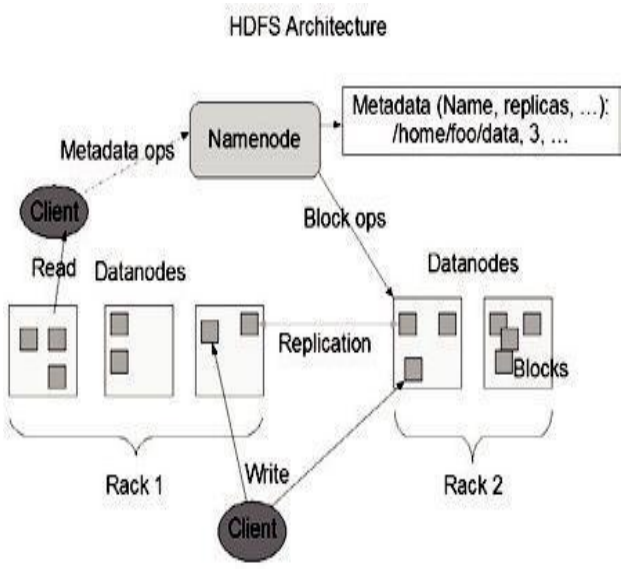
[1] Apache's Hadoop file systems

Hadoop is an open source software framework to support distributed applications. Hadoop is written in the Java, operating system is cross platform and speculators are APACHE[14] software foundation. Hadoop consists of the Hadoop Common package(contains the necessary Java Archive (JAR) files and scripts needed to start Hadoop) which provides file system and OS level abstractions, a

MapReduce engine (either MapReduce or YARN) and the Hadoop Distributed File System (HDFS)[5].The Hadoop framework transparently provides reliability to applications. Hadoop has two major components. They are Hadoop File System (HDFS) and Map-Reduce.

Hadoop Distributed File System (HDFS)

The Hadoop File System (HDFS)[6] are highly scalable, distributed, and portable which is written in Java for the Hadoop framework. The HDFS file system uses the TCP/IP layer for communication. Clients use Remote procedure call (RPC) to communicate between each other which was shown in the below figure 1.



HDFS has master/slave architecture. An HDFS cluster consists of a single Name Node and number of Data Nodes, usually one per node in the cluster. Internally, a file is split into one or more blocks and these blocks are stored in a set of Data Nodes. The Name Node executes file system operations (opening, closing, rename). The Data Nodes are responsible for read and write requests from the file system clients. The Data Nodes also perform block creation, deletion, and replication upon instruction from the Name Node. The Name Node and Data Node are typically run a GNU/Linux operating system (OS). HDFS stores large files (gigabytes to petabytes) across multiple machines. It achieves reliability by replicating the data across multiple hosts, and hence does not oblige RAID storage on hosts.

Limitations Of HDFS(Hadoop Distributed File System)

HDFS was designed for mostly endurable files and may not be suitable for concurrent write-operations. Another imperfection of HDFS is that it cannot framed directly by an existing operating system. It is inconvenient to executing a job in HDFS file system. To address this problem, a File system in User Space (FUSE) virtual file system has been developed for Linux / Unix systems.

B. *MapReduce*

MapReduce [2] framework architecture provides a parallel processing to the huge amount of data. With MapReduce, queries are fragmented and distributed across parallel nodes and processed (the Map step). The results are then realized and dispensed (the Reduce step). An implementation of MapReduce framework was adopted by an Apache open source project named Hadoop**.** MapReduce can take advantage of processing data paralleling.

"Map" step

The master node divides the input into smaller sub-problems, and distributes them to worker nodes. Again a worker node may do this leading to a multi-level tree structure. Thus the worker node processes the smaller problem, and forwards the answer back to its master node.

"*Reduce" step*

The master node collects the answers from all the sub-problems and syndicates them to form an output.

The Map and Reduce function

The *–Map, Reduce"* are both defined with respect to data structured in (key, value) pairs. *–Map"* assemble one set of data in one data domain, and returns a list of pairs in a distinct domain. The *–Reduce"* function is then functional in parallel to each group and delivers a collection of values in the same domain. Thus the MapReduce framework transforms the list of values.

Uses Of Map reduce

MapReduce is useful in applications like distributed pattern-based searching, document clustering, machine learning and statistical machine translation. Besides that, MapReduce model has been revised to numerous computing environments like cloud environments and mobile environments. MapReduce's inputs and outputs are usually stored in a distributed file system. The transitory data is usually stored on local disk and procured remotely by the reducers.

III. NOSQL (NOT ONLY SQL)

A related new style of database called NoSQL (Not Only SQL) has surfaced like Hadoop process of multi-structured data. For in case of point HBase is a popular NoSQL [7] database model employed on top of HDFS(the Hadoop Distributed File System) to specify low-latency and advanced results in Hadoop. At present NoSQL databases traded with ACID property for extreme performance and scalability. The enhanced NoSQL databases are HBase, Cassandra, MongoDB DynamoDB, Couch base, Riak, Redis, Accumulo, Datatomic, Aerospike and. Out of these, HBase and Accumulo are more diligently tied to Hadoop.

A. HBase

HBase [9] is a distributed, column-oriented database. HBase provides big table-like capabilities on top of Hadoop. HBase does not support for complex transactions but it offers high

read and write performance in several large applications. HBase is an convenient for fast retrieved queries and updates but storage in HDFS is anticipated for usage with Pig, Hive, or other MapReduce-based tools.

B. Cassandra

Cassandra is the most admired NoSQL database for very large data sets. It is a clustered database that permits column-oriented storage and redundant storage for scalability in both data sizes and read/write transaction.

C. MongoDB

MongoDB is a document-oriented NoSQL database has a fruitful, Java script based query language where each record is a JSON(Java Script Object Notation) document.

D. DynamoDB

DynamoDB [16] is Amazon's highly scalable and existing NoSQL database.

E. Couch base

Couch base is a NoSQL database, appropriate for mobile applications where a copy of a data set is tenant on many devices, in which changes can be functioned on any copy like**,** how an mechanism of email client works with local copies of email history and corresponding email servers.

F. Redis

Redis supports basic data structures as values, strings, hash maps, lists and sets. Redis[15] is often called a data structure server.

G. Datomic

Datomic is a novel technique in the NoSQL landscape with a unique data model making historical reestablishment of events. Many standard database operations like joins and ACID transactions are assisted.

H. Riak

Riak is a fault-tolerant, scattered, key-value NoSQL[8] database. It is designed for large-scale implementations in cloud or hosted environments. It is flexible against the failure of multiple nodes and nodes can be added or deleted efficiently. Riak[13] is also optimized for read and write transactions.

IV. OTHER APPLICATIONS

The HDFS file systems are not limited to MapReduce jobs. It can be used for other appliances, many of which are further development at Apache like HBase database, Apache Mahout Machine learning systems, and the Apache Hive Data Warehouse systems. The other industrial applications of Hadoop included are:

A. Marketing analytics
B. Machine learning
C. Image processing
D. XML Processing
E. Web crawling
F. Text processing

V. PROMINENT USERS

A. Yahoo!

The Yahoo! Search Web map is a Hadoop operation that turns on a more than 10,000 core Linux cluster and that is used in every Yahoo! Web search query.

B. Face book

Face book had the largest Hadoop cluster in the world with 21 PB of storage. Recently they heralded that warehouse grows by roughly half a PB per day.

C. Other Users

Beyond to Face book and Yahoo!, many other organizations are using Hadoop to run large distributed computations. Some of the significant users are Amazon.com, Google, IBM, NetApp and Twitter etc.

VI. COMMERCIALLY SUPPORTED HADOOP RELATED PRODUCTS

There are a number of companies subscribed for commercial implementations and granted endorsement for Hadoop.

A. Syncsort

It provides an ETL Solution, which extends the competencies of Hadoop[12], into a highly scalable, affordable integrated environment.

B. sqrrl

It offers sqrrl innovativeness, which extends Hadoop and combines the features of several data stores (Column + Document + Graph).

C. Pivotal

It offers a distribution of Hadoop that consists of HAWQ, with 100% ANSI SQL compatibility.

D. Cloudera

It offers CDH (Cloudera's Distribution including Apache Hadoop) and Cloudera Enterprise[10].

E. Silicon Graphics International

Hadoop enhanced solutions based on the SGI[11] Rackable and Cloud Rack server lines with implementation services.

F. Google added AppEngine-MapReduce

It is used to support successively Hadoop 0.20 programs on Google App Engine[3].

VII. CONCLUSION

As data volumes have grown, and as the convolution of data that is collected and analyzed has increased, new novel software architectures have occurred. New systems using big data will tender and possibly replace, our traditional DBMS. We have also seen that there is a definite problem relating to big data management because of the usage of Face book, twitter and YouTube. In this paper we debated about the big

data storage technologies and their architectures. The eminent technologies are Hadoop Distributed file system (HDFS) and NoSQL databases. The end goal is to improve security decision-making based on ordered, illegal insight derived from monitoring big data environments and recognize when an progressive targeted attack has avoided conventional security controls and perceived the organization.

REFERENCES

[1] "IBM What is big data? — Bringing big data to the enterprise". www.ibm.com. Retrieved 2013-08-26.

[2] Solaimurugan Vellaipandiyan, Resource Management in Big Data Analytics : Integrating and Running diverse framework, presented for Proceedings of International Conference on Computing, Cybernetics and Intelligent Information Systems (CCIIS) 2013.

[3] Impetus white paper, March 2011, "Planning Hadoop/NoSQL Projects for 2011" by Technologies
Available:http://www.techrepublic.com/resource-library/whitepapers/planning-hadoop-nosql-projects-for-2011/

[4] Big Data Study – The 10 Key Findings by TCS
http://sites.tcs.com/big-data-study/kinds-of-digital

[5] Webster, John. "MapReduce: Simplified Data Processing on Large Clusters", "Search Storage", 2004. Retrieved on 25 March 2013.

[6] Google added –AppEngine-MapReduce to support running Hadoop 0.20 programs‖ on Google App Engine

[7] Intel released its own –Hadoop distribution that takes advantage of capabilities in Intel Xeon chips, such as its processor instructions for accelerating AES encryption‖.

[8] The –Hadoop wiki‖ provides community input related to hadoop and HDFS.

[9] "Hadoop contains the distributed computing platform that was formerly a part of Nutch. This includes the Hadoop Distributed File system (HDFS) and an implementation of MapReduce."

[10] "NoSQL Relational Database Management System: Home Page". Strozzi.it. 2 October 2007. Retrieved 29 March 2010.

[11] The enterprise class – NoSQL database on djondb‖, Retrieved on 2013-09-18.

[12] George, Lars (September 20, 2011) –HBase: The Definitive Guide (1st ed.)‖, O'Reilly Media. p. 556. ISBN 978-1449396107.

[13] Cloudera offers –CDH (Cloudera's Distribution including Apache Hadoop) and Cloudera Enterprise‖.

[14] Silicon Graphics International offers –Hadoop optimized solutions based on the SGI Rackable and Cloud Rack server lines with implementation services‖.

[15] David Leinweber (April 26, 2013). "Big Data Gets Bigger: Now Google Trends Can Predict The Market". Forbes. Retrieved August 9, 2013.

[16] "Riak: An Open Source Scalable Data Store". 28 November 2010. Retrieved 13 October 2011.

[17] http://wiki.apache.org/hadoop/

[18] http://en.wikipedia.org/wiki/Redis

[19] www.mongodb.com

About Author:

Dr. Kamal Gulati is Ph.D., MCA from National Institute of Electronics & Information Technology (NIELIT), M.Sc (Computer Science) from Sikkim Manipal University and MBA from Symbiosis, Pune, Specialized in CCNA (Cisco Certified Network Associate), Certification in MCP (Microsoft Certified Professional) and Certification in Brainbench of Advanced Excel, Visual Basic, Operating System- 8 / 7 / Vista XP / 2000 / 98 and Ms Office. Published 20 papers in reputed journals and conference proceedings and Presented 6 papers in International and National Seminar and Conferences. With an experience of Thirteen Years+ in the field of teaching in Academia and three years in IT Industry. Area of Interest: Big Data Analytics, DBMS, Networking and Advanced Excel with Visual Basic Macros.