International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 12 (2015) pp. 30353-30358

http://www.ripublication.com

High Functioning and Scattering System For Huge Data Processing and Storage With The Help of Hadoop

M. Harika [1], B. Gurunadha Rao[2]
[1]*M.Tech Student, Department of CSE, Sree Vidyanikethan Engineering College, Tirupathi, A.P, India*
[2]*Assistant Professor, Department of CSE, Sree Vidyanikethan Engineering College, Tirupathi, A.P, India*

Abstract

Hadoop is a rapidly promising ecosystem of Google's file system work and Map Reduce algorithm components for executing Map Reduce algorithms in a feasible style and scattered on hardware services. Hadoop allows clients to preserve and develop huge volumes of information and analyze it in customs not earlier achievable with SQL-based approached or less feasible solutions. Outstanding developments in conservative calculate and storage resources assist make Hadoop clusters practical for various organizations. This thesis starts with the conversation of Big Data development and the prospect of Big Data depending on Gartner's Hype Cycle. We have clarified how HDFS (Hadoop Distributed File System) functions and its framework with appropriate design. Hadoop's Map Reduce model for allocating a assignment across various nodes in Hadoop is conferred with sets of sample data. The functioning of Map Reduce and Hadoop Distributed File System when they are placed jointly is described. At last the thesis lasts with a conversation on Big Data Hadoop sample utilize cases which explore how enterprises can increase a competitive advantage by being big data analytics premature adopters.

Introduction

Most of the present applications like as index banking transactions, social networking, web searches, recommendation engines genome exploitation in life sciences and machine learning manufacture massive amounts of information in the form of email, blogs, logs, and other technical structured and unstructured information streams.

These information needs to be stored, processed and associated to increase close view into today's business processes. Also, the require to keep both structured and unstructured data to fulfill the government regulations in certain industry sectors requires the storage, processing and analysis of large amounts of data. While a haze of

excitement often envelops the universal discussions of Big Data, a clear agreement has at least combined around the definition of the term. The term "Big Data" is typically considered to be a data collection that has grown so large it can't be affordably or effectively managed using conventional data management tools such as traditional RDBMS (relational database management systems) or conventional search engines, based on the task at hand. Another buzzing term "Big data Analytics" is where advanced analytic techniques are made to operate on big data sets. Thus, big data analytics is really about two things namely, big data and analytics and how the two have coalesced up to create one of the most philosophical trends in business intelligence (BI) today. There are several ways to store, process and analyze large volumes of data in a massively parallel scale. Hadoop is considered as a best example for a massively parallel processing system.

Problem Statement

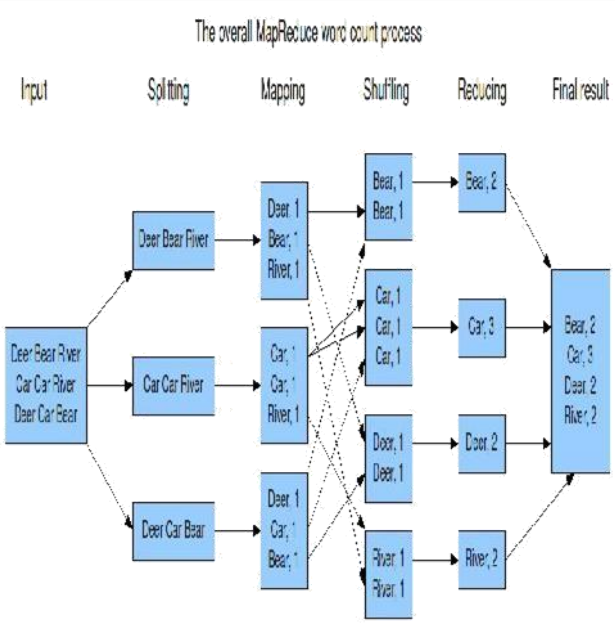
Maintaining a huge data is a very complex thing. There will be the entire time. Only one concerned and that is the security of the data. Just managing a complex application such as Hadoop can be challenging. A classic example can be seen in the Hadoop security model, which is disabled by default due to sheer complexity.

System Development

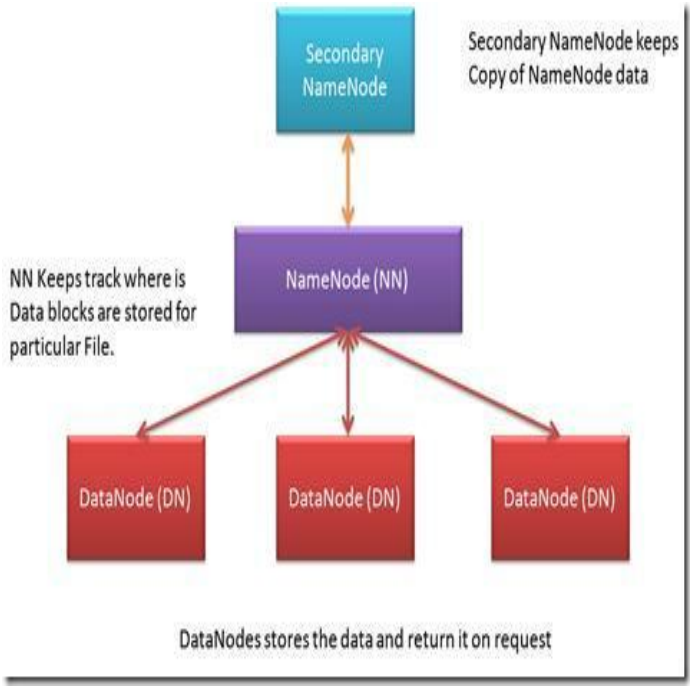
Big data is a term which portrays a situation where the variety of data, velocity and volume surpass an organization's database or calculate capacity for perfect and well-timed decision making. Some of this data is detained in transactional data stores, the byproduct of fast-growing online activity. Machineto- machine communications, such as metering, call detail records, environmental sensing and RFID systems, generate their own tidal waves of data. Data is pouring in from every feasible direction from transactional and operational systems, from facilities management systems and scanning from outbound and inbound customer contact nodes, from the Web and mobile media. All these forms of data are expanding, and that is together with fast-growing streams of unstructured and semi structured data from social media. Organizations are swamped with data – petabytes and terabytes of it. To keep it in perception, 1 TB contains 2,000 hours of CD-quality music and 10 TB could store the entire collection of US Library of Congress print. Yottabytes, zettabytes and Exabytes definitely are on the prospect. According to IDC, in 2011, the amount of information produced and replicated will go beyond 1.8 zettabytes (1.8 trillion GB), growing by a factor of nine in just five years. That is nearly as many bits of information in the digital universe as stars in the physical universe.

The Hype Cycle for Emerging Technologies statement is the greatest functioning annual Hype Cycle, offering a cross-industry viewpoint on the trends and technologies that senior executives, strategists, innovators, CIOs, business developers and technology planners should regard as in mounting promising technology portfolios. "It is the widespread joint featuring technologies that are the concentrate of attention because of predominantly sky-scraping levels of hype, or those that Gartner

considers have the capability for major impact", says Jackie Fenn, vice president, Gartner group and Gartner fellow. Gartner just released the new Hype Cycle for Emerging Technologies of 2013. Gartner's Hype Cycle Special statement offers strategists and planners with an evaluation of the future direction, business benefit and maturity of more than two thousand technologies, clustered into 98 areas. Hype Cycles guesstimates how long technologies and trends will acquire to reach maturity and helps organizations make a decision when to implement. It characterizes the new technology adoption of five stages and initiates with a Technology Trigger: a new innovation or invention. It goes next all the way up to a "peak of inflated expectations" and after that down to a "trough of disillusionment". Successful innovations possibly will climb the "slope of enlightenment" and, finally attain "plateau of productivity". One of the significant items on this year's Hype Cycle for Emerging Technologies is that Gartner tones down the expectations of big data [3]. In the 2012 version of the Hype Cycle for Emerging Technologies, Gartner predicted that it would take 2-5 years before big data would reach the plateau of productivity. In this year's edition Gartner adjusted that prediction to 5-10 years. The intimately connected trend of the Internet of Things still necessitate over 10 years to reach the plateau of productivity, just like it was predicted last year. To really understand how it is possible to scale a Hadoop cluster to hundreds and thousands of nodes, we should start with HDFS. Hadoop consist of two basic components: a distributed file system and the computational framework. In the first component of above two, data is stored in Hadoop Distributed File System (HDFS). Hadoop Distributed File System (HDFS) uses a write-once, read-many model that breaks data into blocks that it spreads across many nodes for fault tolerance and high performance. Hadoop and HDFS make use of master-slave architecture. HDFS is written in Java language, with an HDFS cluster consisting of a primary Name Node – a master server that manages the file system namespace and also controls access to data by clients.



There is also a Secondary Name Node which maintains a copy of the Name Node data to be used to restart the Name Node when failure occurs, although this copy may not be current and so some data loss is still likely to occur. In addition to it, there are a number of Data Nodes; generally, there is a one-to-one relationship between a Data Node and a physical machine. Each Data Node manages the storage attached to the boxes that it runs on.



HDFS makes use of a file system namespace that enables data to be stored in files. Each file is divided into one or more blocks, which are then shared across a set of Data Nodes. The Name Node is accountable for tasks such as opening, renaming, and closing files and data directories. It also deals with the job of mapping blocks to Data Nodes, which are then responsible for managing incoming I/O requests from clients. The Data Node looks after block replication, creation, and removal of data when instructed to do so by the Name Node.

Related Work

Another basic component of Hadoop is Map Reduce, which affords a computational framework for data processing. Map Reduce is a programming replica and an associated implementation for processing and generating large data sets [6]. Map Reduce programs are inherently parallel and thus very suitable to a distributed environment. Hadoop takes a cluster of nodes to run Map Reduce programs massively in parallel. A single Job Tracker schedules all the jobs on the cluster, as well as individual tasks. Here, each benchmark test is a job and runs by itself on the cluster. A job is split into a set of tasks that execute on the worker nodes. A Task Tracker

running on each worker node is responsible for starting tasks and reporting progress to the Job Tracker. As the name implies, a Map Reduce program consists of two major steps, namely, the Map step processes input data and the next step Reduce assembles intermediate results into a final result. Both use key-value pairs defined by the user as input and output. This allows the output of one job to provide directly as input for another. Map Reduce programs runs on local file system and local CPU for each cluster node. Data are broken into data blocks (usually in size of 64MB blocks), stored across. The local files of different nodes, and replicated for reliability and fault tolerance. The local files constitute the file system which is called as Hadoop Distributed File System being discussed above. The number of nodes in each cluster differs from hundreds to thousands of machines. Map Reduce is a massively scalable, parallel processing framework that works in connection with HDFS. With Map Reduce and Hadoop, compute is executed at the location of the data, rather than moving data to the compute location; data storage and computation coexist on the same physical nodes in the cluster. Map Reduce processes exceedingly large amounts of data without being affected by traditional bottlenecks like network bandwidth by taking advantage of this data proximity. Map Reduce divides workloads up into multiple tasks that can be executed in parallel.

Conclusion

One of the significant items on this year for Emerging Technologies is big data touch the point of "Trough of Disillusionment". It means the market starts to mature, becoming more realistic about how big data can be useful for organization. The central theme for emerging technologies 2014 hype cycle is Digital Business. We are in the peak time of Big Data. Every day, we generate 2.5 quintillion bytes of data showing that the data in the world today has been created in the last two years alone. In this paper we have highlighted the evolution and rise of big data using Gartner's Hype Cycle for emerging technologies. We have discussed how HDFS produces multiple replicas of data blocks and distributes them on compute nodes throughout a cluster to enable reliable, exceptionally fast computations. We have implemented Map Reduce concept that scales to large clusters of machines comprising thousands of machines. Finally the paper ends with the discussion of Real-World Hadoop use cases which helps in Business Analytics.

References

[1] Neil Raden, "Big Data Analytics Architecture - Putting All Your Eggs in Three Baskets", 2012

[2] IDC study 2011, EMC. http://www.emc.com/collateral/analystreports/ idc-extracting-value-from-chaos-ar.pdf

[3] Gartner's Report, "Gartner's 2013 Hype Cycle for Emerging Technologies Maps Out Evolving Relationship Between humans and Machines", 2013 http://www.gartner.com/newsroom/id/2575515

[4] Gartner's Report, "Hype Cycle for Big Data, 2012", 2012
[5] Ghemawat, Sanjay, Howard Gobioff, and Shun-Tak Leung, "The Google File System" SOSP'03, Oct 19-23, Dean, Jeffrey and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", OSDI 2004
[6] HDFS Architecture Guide, http://hadoop.apache.org/docs/r1.0.4/hdfs_design.html
[7] What is Apache Hadoop? http://hadoop.apache.org/
[8] Cloudera Customer Case Study, "Streamlining Healthcare Connectivity with Big Data", 2012.
[10] Cloudera Customer Case Study, "Nokia: Using Big Data to Bridge the Virtual & Physical Worlds", 2012.
[11] Intel Case Study, "China Mobile Guangdong Gives Subscribers Real-Time Access to Billing and Call Data Records", 2012
[12] Cloudera Customer Case Study, "NetApp Improves Customer Support by Deploying Cloudera Enterprise", 2012.
[13] Cloudera Customer Case Study, "Joint Success Story: Major Retail Bank", 2012