Volume-7,Issue-6, (June-17)
ISSN (O) :- 2349-3585

BIG DATA WITH HADOOP-MAPREDUCE

K.Pavithradevi[1],A.Naveen[2] ,P.Pavithra[3]
[1]Assistant Professor,[2]MCA Student,[3]MCA Student, MCA
Gnanamani College of Technology, Namakkal,India

Abtract :The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. Big data can be structured, unstructured or semi-structured, resulting in incapability of conventional data management methods. Big Data is a data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes. This paper is an effort to present the basic understanding of BIG DATA is and it's usefulness to an organization from the performance perspective. A number of application examples of implementation of BIG DATA across industries varying in strategy, product and processes have been presented. There is no hard and fast rule about exactly what size a database needs to be in order for the data inside of it to be considered "big." Instead, what typically defines big data is the need for new techniques and tools in order to be able to process it. In order to use big data, you need programs which span multiple physical and/or virtual machines working together in concert in order to process all of the data in a reasonable span of time.

Keywords : big data, MapReduce, HDFS and Apache Hadoop

Introduction

Big data means really a big data, it is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not merely a data, rather it has become a complete subject, which involves various tools, techniques and frameworks.

Big Data Technologies

Big data technologies are important in providing more accurate analysis, which may lead to more concrete decision-making resulting in greater operational efficiencies, cost reductions, and reduced risks for the business. To harness the power of big data, you would require an infrastructure that can manage and process huge volumes of structured and unstructured data in realtime and can protect data privacy and security.There are various technologies in the market from different vendors including Amazon, IBM, Microsoft, etc., to handle big data. While looking into the technologies that handle big data, we examine the following two classes of technology.

Operational Big Data

This include systems like MongoDB that provide operational capabilities for real-time, interactive workloads where data is primarily captured and stored.NoSQL Big Data systems are designed to take advantage of new cloud computing architectures that have emerged over the past decade to allow massive computations to be run inexpensively and efficiently. This makes operational big data workloads much easier to manage, cheaper, and faster to implement.Some NoSQL systems can provide insights into patterns and trends based on real-time data with minimal coding and without the need for data scientists and additional infrastructure.

Analytical Big Data

This includes systems like Massively Parallel Processing (MPP) database systems and MapReduce that provide analytical capabilities for retrospective and complex analysis that may touch most or all of the data. MapReduce provides a new method of analyzing data that is complementary to the capabilities provided by SQL, and a system based on MapReduce that can be scaled up from single servers to thousands of high and low end machines.

The Apache Hadoop framework modules

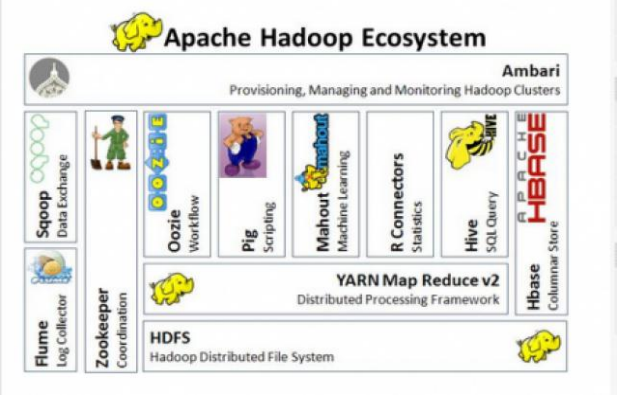
1.Hadoop Common: contains libraries and utilities needed by other Hadoop modules

2.Hadoop Distributed File System (HDFS): a distributed file-system that stores data on the commodity machines, providing very high aggregate bandwidth across the cluster

3.Hadoop YARN: a resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users' applications

4.Hadoop MapReduce: a programming model for large scale data processing

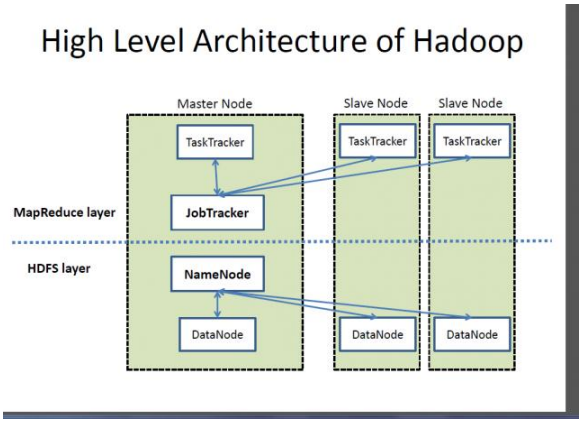
All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines, or racks of machines) are common and thus should be automatically handled in software by the framework. Apache Hadoop's MapReduce and HDFS components originally derived respectively from Google's MapReduce and Google File System (GFS) papers.Beyond HDFS, YARN and MapReduce, the entire Apache Hadoop "platform" is now commonly considered to consist of a number of related projects as well: Apache Pig, Apache Hive, Apache HBase, and others.



For the end-users, though MapReduce Java code is common, any programming language can be used with "Hadoop Streaming" to implement the "map" and "reduce" parts of the user's program. Apache Pig and Apache Hive, among other related projects, expose higher level user interfaces like Pig latin and a SQL variant respectively. The Hadoop framework itself is mostly written in the Java programming language, with some native code in C and command line utilities written as shell-scripts.

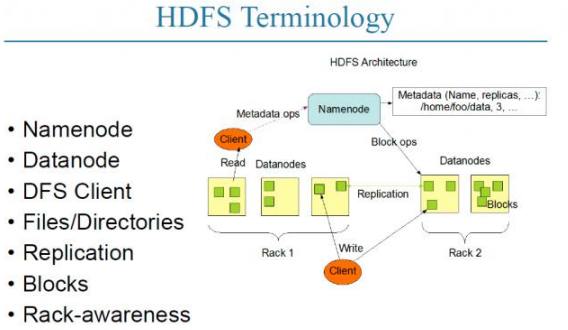
HDFS and MapReduce

There are two primary components at the core of Apache Hadoop 1.x: the Hadoop Distributed File System (HDFS) and the MapReduce parallel processing framework. These are both open source projects, inspired by technologies created inside Google.



Hadoop Distributed File System

The Hadoop distributed file system (HDFS) is a distributed, scalable, and portable file-system written in Java for the Hadoop framework. Each node in a Hadoop instance typically has a single name node, and a cluster of data nodes form the HDFS cluster. The situation is typical because each node does not require a data node to be present. Each data node serves up blocks of data over the network using a block protocol specific to HDFS. The file system uses the TCP/IP layer for communication. Clients use Remote procedure call (RPC) to communicate between each other.

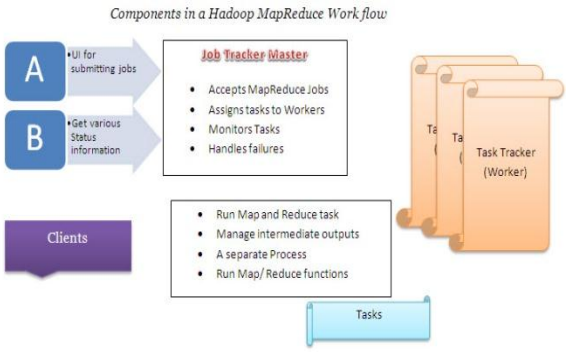


HDFS stores large files (typically in the range of gigabytes to terabytes) across multiple machines. It achieves reliability by replicating the data across multiple hosts, and hence does not require RAID storage on hosts. With the default replication value, 3, data is stored on three nodes: two on the same rack, and one on a different rack. Data nodes can talk to each other to rebalance data, to move copies around, and to keep the replication of data high. HDFS is not fully POSIX-compliant, because the requirements for a POSIX file-system differ from the target goals for a Hadoop application. The tradeoff of not

having a fully POSIX-compliant file-system is increased performance for data throughput and support for non-POSIX operations such as Append. HDFS added the high-availability capabilities for release 2.x, allowing the main metadata server (the Name Node) to be failed over manually to a backup in the event of failure, automatic fail-over. The HDFS file system includes a so-called secondary name node, which misleads some people into thinking that when the primary name node goes offline, the secondary name node takes over. In fact, the secondary name node regularly connects with the primary name node and builds snapshots of the primary name node's directory information, which the system then saves to local or remote directories. These check pointed images can be used to restart a failed primary namenode without having to replay the entire journal of file-system actions, then to edit the log to create an up-to-date directory structure. Because the namenode is the single point for storage and management of metadata, it can become a bottleneck for supporting a huge number of files, especially a large number of small files. HDFS Federation, a new addition, aims to tackle this problem to a certain extent by allowing multiple name-spaces served by separate namenodes. An advantage of using HDFS is data awareness between the job tracker and task tracker. The job tracker schedules map or reduce jobs to task trackers with an awareness of the data location. For example, if node A contains data (x, y, z) and node B contains data (a, b, c), the job tracker schedules node B to perform map or reduce tasks on (a,b,c) and node A would be scheduled to perform map or reduce tasks on (x,y,z). This reduces the amount of traffic that goes over the network and prevents unnecessary data transfer. When Hadoop is used with other file systems, this advantage is not always available. This can have a significant impact on job-completion times, which has been demonstrated when running data-intensive jobs. HDFS was designed for mostly immutable files and may not be suitable for systems requiring concurrent write-operations. Another limitation of HDFS is that it cannot be mounted directly by an existing operating system. Getting data into and out of the HDFS file system, an action that often needs to be performed before and after executing a job, can be inconvenient. A filesystem in Userspace (FUSE) virtual file system has been developed to address this problem, at least for Linux and some other Unix systems. File access can be achieved through the native Java API, the Thrift API, to generate a client in the language of the users' choosing (C++, Java, Python, PHP, Ruby, Erlang, Perl, Haskell, C#, Cocoa, Smalltalk, or OCaml), the command-line interface, or browsed through the HDFS-UI web app over HTTP.

Data Processing Framework & MapReduce

The data processing framework is the tool used to process the data and it is a Java based system known as MapReduce. People get crazy when they work with it.



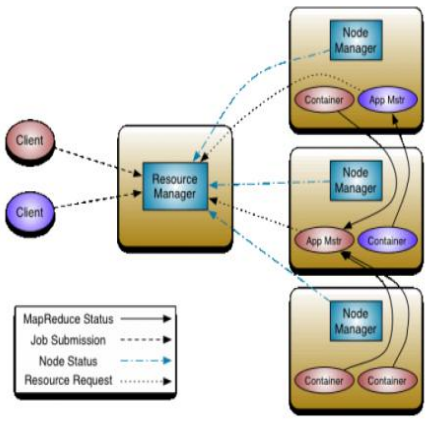
JobTracker and Task Tracker: The MapReduce Engine

Job Tracker Master handles the data, which comes from the MapReduce. Then it assigns tasks to workers, manages the entire process, monitors the tasks, and handles the failures if any. The JobTracker drives work out to available TaskTracker nodes in the cluster, striving to keep the work as close to the data as possible. As Job Tracker knows the architecture with all steps that has to be followed in this way, it reduces the network traffic by streamlining the racks and their respective nodes.

Apache Hadoop NextGen MapReduce (YARN) Yet Another Resource Negotiation

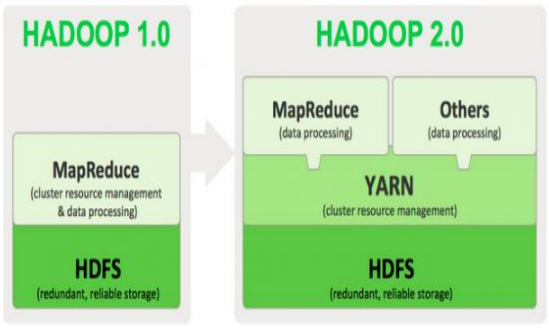
MapReduce has undergone a complete overhaul in hadoop-0.23 and we now have, what we call, MapReduce 2.0 (MRv2) or YARN. Apache Hadoop YARN is a sub-project of Hadoop at the Apache Software Foundation introduced in Hadoop 2.0 that separates the resource management and processing components. YARN was born of a need to enable a broader array of interaction patterns for data stored in HDFS beyond MapReduce. The YARN-based architecture of Hadoop 2.0 provides a more general processing platform that is not constrained to MapReduce.

ISSN:-2349-3585 |www.ijrdt.org

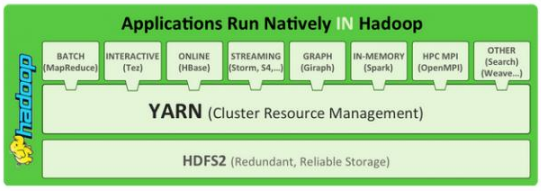


The fundamental idea of MRv2 is to split up the two major functionalities of the JobTracker, resource management and job scheduling/monitoring, into separate daemons. The idea is to have a global Resource Manager (RM) and per-application Application Master (AM). An application is either a single job in the classical sense of Map-Reduce jobs or a DAG of jobs.

The Resource Manager and per-node slave, the Node Manager (NM), form the data-computation framework. The Resource Manager is the ultimate authority that arbitrates resources among all the applications in the system.The per-application Application Master is, in effect, a framework specific library and is tasked with negotiating resources from the Resource Manager and working with the Node Manager(s) to execute and monitor the tasks.



As part of Hadoop 2.0, YARN takes the resource management capabilities that were in MapReduce and packages them so they can be used by new engines. This also streamlines MapReduce to do what it does best, process data. With YARN, you can now run multiple applications in Hadoop, all sharing a common resource management. Many organizations are already building applications on YARN in order to bring them IN to Hadoop.



As part of Hadoop 2.0, YARN takes the resource management capabilities that were in MapReduce and packages them so they can be used by new engines. This also streamlines MapReduce to do what it does best, process data. With YARN, you can now run multiple applications in Hadoop, all sharing a common resource management. Many organizations are already building applications on YARN in order to bring them IN to Hadoop. When enterprise data is made available in HDFS, it is important to have multiple ways to process that data. With Hadoop 2.0 and YARN organizations can use Hadoop for streaming, interactive and a world of other Hadoop based applications.

Scattered Across the Cluster

Here, the data is distributed on different machines and the work trends is also divided out in such a way that data processing software is housed on the another server. On a Hardtop cluster, the data stored within HDFS and the MapReduce system are housed on each machine in the cluster to add redundancy to the system and speeds information retrieval while data processing.

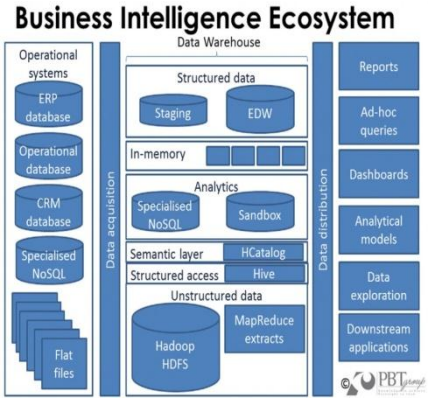
Big Data Analytics

Big data is massive and messy, and it's coming at you uncontrolled. Data are gathered to be analyzed to discover patterns and correlations that could not be initially apparent, but might be useful in making business decisions in an organization. These data are often personal data, which are useful from a marketing viewpoint to understand the desires and demands of potential customers and in analyzing and predicting their buying tendencies.

Organizational Architecture Need for an Enterprise

You can benefit by the enterprise architecture that scales effectively with development – and the rise of Big Data analytics means that this issue required to be addressed more

urgently. IDC believes that these below use cases can be best mapped out across two of the Big Data dimensions – namely velocity and variety as outlined below.



Put Big Data Value in the Hands of Analysts

Business Inefficiencies Identified: Analysts to view end-to-end processing of business transactions in an organization

Business Inefficiencies Rectified: Analysts to rectify end-to-end processing of business transactions in an organization

Knowledge Enhancement: Provide the analyst team additional operational and business context

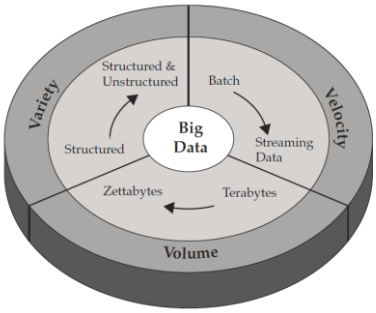
Store Terabytes of Data: Provide analysts visibility into the whole infrastructure

Enable More Data Usages: Cartel device, system, and application data to bring business operational views of IT professionals in an organization.

Enhance Value: Pinpoint and implement newfangled opportunities that would otherwise be impossible to view and act upon.

Categorize of Personal Data

This can be categorized as volunteered data, Observed data, and Inferred data. For any enterprise to succeed in driving value from big data, volume, variety, and velocity have to be addressed in parallel.



Engineering Big Data Platforms

Big data platforms need to operate and process data at a scale that leaves little room for mistake. Big data clusters should be designed for speed, scale, and efficiency. Many businesses venturing into big data don't have knowledge building and operating hardware and software, however, many are now confronted with that prospect. Platform consciousness enterprises will boost their productivity and churn out good results with big data.

Optimize Aspects of Business

Many enterprises are operating their businesses without any prior optimization of accurate risk analysis. Therefore, more risk analysis is required to tackle these challenges. There is a continuum of risk between aversion and recklessness, which is needed to be optimized. To some extent, risk can be averse but BI strategies can be a wonderful tool to mitigate the risk. For handling big data, companies need to revamp their data centers, computing systems and their existing infrastructure. Be prepared for the next generation of data handling challenges and equip your organization with the latest tools and technologies to get an edge over your competitors.

Roles and Responsibilities of Big Data Professionals

Big Data professionals work dedicatedly on highly scalable and extensible platform that provides all services like gathering, storing, modeling, and analyzing massive data sets from multiple channels, mitigation of data sets, filtering and IVR, social media, chats interactions and messaging at one go. The major duties include project scheduling, design, implementation and coordination, design and develop new components of the big data platform, define and refine the big data platform, understanding the architecture, research and experiment with emerging technologies, and establish and reinforce disciplined software development processes.

Big Data Benefits

Tremendous opportunities are there with big data as the challenges. Enterprises that are mastered in handling big data are reaping the huge chunk of profits in comparison to their competitors. The research shows that the companies, who has been taking initiatives through data directed decision making fourfold boost in their productivity; the proper use of big data goes beyond the traditional thinking like gathering and analyzing; it requires a long perspective how to make the crucial decision based on Big Data.

ISSN:-2349-3585 |www.ijrdt.org

Industries using Big Data

Big data has become a big deal in 2015 with 90% of world's data created in the last 2 years from social network posts, customer transactions, web browsing data trails, etc. If you Google for the term "Big Data" it will generate close to 8 million results under the news section and approximately 54 million results on regular in search. With 3 billion people online and 247 billion emails sent every day, a research estimates that 8 zettabytes of big data will be created in 2015.

As per big data industry trends, the hype of Big Data had just begun in 2011. In 2015, big data has evolved beyond the hype. 87% of companies using big data believe that within next 3 years big data analytics will redefine the competitive landscape of various industries. 89% of the companies using big data believe that companies that do not adopt big data analytics in the next year are likely to lose market share and momentum.

Conclusion

The paper describes the concept of Big Data. The paper also focuses on Big Data processing HDFS and MapReduce. These technical challenges must be addressed for efficient and fast processing of Big Data. The paper describes Hadoop which is an open source software used for processing of Big Data. Apache Hadoop is 100% open source, and pioneered a fundamentally new way of storing and processing data. With Hadoop, no data is too big. And in today's hyper-connected world where more and more data is being created every day, Hadoop's breakthrough advantages mean that businesses and organizations can now find value in data that was recently considered useless.

References

[1] M. A. Beyer and D. Laney, "The importance of "big data": A definition," Gartner, Tech. Rep., 2012.

[2] X. Wu, X. Zhu, G. Q. Wu, et al., "Data mining with big data," IEEE Trans. on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, January 2014.Rajaraman and J. D. Ullman, "Mining of massive datasets," Cambridge University Press, 2012.

[3] Shadi Ibrahim⋆ _ Hai Jin _ Lu Lu "Handling Partitioning Skew in MapReduce using LEEN" ACM 51 (2008) 107–113

[4] Dr.N.Muthumani, K.Pavithradevi, "Image Compression Using ASWDR and 3D – SPHIT Algorithms for Satellite Data "International Journal of Science and Engineering Research ", Volume 6, October 2015, PN – 289-296.

[5] C.Ganesh,B.Sathiyabama,T.Geetha"Fast Frequent Pattern Mining Using Vertical Data Format for Knowledge Discovery" International Journal of Emerging Research in Management and Technology",Vol 5,issue 5,2016.

[6] Elias, J., & Mehaoua, A. (2012, June). Energy-aware topology design for wireless body area networks. In Communications (ICC), 2012 IEEE International Conference on (pp. 3409-3410). IEEE.

[7] Ullah, S. and Shen, B. and Riazul Islam, SM and Khan, P. and Saleem, S. and Sup Kwak, K., 2009. A study of MAC protocols for WBANs: SENSOR.

[8]Moshaddique Al Ameen, NiamatUllah, —A power efficient MAC protocol for wireless bodyarea networks‖Al Ameen et al. EURASIP Journal on Wireless Communications and Networking 2012, 2012:33.

[9] R.Karthikeyan",Improved Apriori Algorithm for Mining rules",International Journal of Advanced Research in Biology EngineeringScienceandTectnology(IJARBEST)Vol.2,Issue.4, APR Pages71-77.2016.

[10]L.Gomathi ,K.Ramya " Data Mining Analysis using Query Formulation In Aggregation Recommendation"

,International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) Vol .2, Issue .1 October 2013.

[11]R.Karthikeyan,T.Geetha, K.Ramya, K.Pavithradevi, " A Survey of Sensor Networks" , International Journal for Research & Development in Technology (IJRDT) Vol.7,Issue.1, Januray 2017.