Hadoop Technology to Analyze Big Data

Garima Rani, Sunil Kumar
Department of Computer Engineering (SITE)
Department of Mathematics (FOS)
Swami Vivekanand Subharti University,Meerut, Uttar Pradesh, India

***Abstract* - Hadoop is not a type of database, but rather a software ecosystem that allows for massively parallel computing. It is an enabler of certain types NoSQL distributed databases (such as HBase), which can allow for data to be spread across thousands of servers with little reduction in performance. A staple of the Hadoop ecosystem is MapReduce, a computational model that basically takes intensive data processes and spreads the computation across a potentially endless number of servers (generally referred to as a Hadoop cluster). It has been a game-changer in supporting the enormous processing needs of big data; a large data procedure which might take 20 hours of processing time on a centralized relational database system, may only take 3 minutes when distributed across a large Hadoop cluster of commodity servers, all processing in parallel.**

***Key words* - Data management. Data visualization. Advanced analytics. SAS. Complex big data challenges.**

I. INTODUCTION

As the world becomes more information-driven than ever before, a major challenge has become how to deal with the explosion of data. These technologies demand a new breed of DBAs and infrastructure engineers/developers to manage far more sophisticated systems. Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.

II. COMPONENTS MAKE UP HADOOP

Currently, four core modules are included in the basic framework from the Apache Foundation:

- **Hadoop Common** – the libraries and utilities used by other Hadoop modules.
- **Hadoop Distributed File System (HDFS) –** the Java-based scalable system that stores data across multiple machines without prior organization.
- **MapReduce** – a software programming model for processing large sets of data in parallel.
- **YARN** – resource management framework for scheduling and handling resource requests from distributed applications. (YARN is an acronym for Yet Another Resource Negotiator.)

GETTING DATA INTO HADOOP

Here are just a few ways to get your data into Hadoop.

- Load files to the system using simple Java commands. HDFS takes care of making multiple copies of data blocks and distributing them across multiple nodes.
- If you have a large number of files, a shell script that runs multiple “put” commands in parallel will speed up the process. You don’t have to write MapReduce code.
- Create a cron job to scan a directory for new files and “put” them in HDFS as they show up. This is useful for things like downloading email at regular intervals.
- Mount HDFS as a file system and copy or write files there.
- Use Sqoop to import structured data from a relational database to HDFS, Hive and HBase. It can also extract data from Hadoop and export it to relational databases and data warehouses.
- Use Flume to continuously load data from logs into Hadoop.
- Use third-party vendor connectors (like SAS/ACCESS® or SAS Data Loader for Hadoop).

III. WHAT IS HADOOP USED FOR?

Going beyond its original goal of searching millions (or billions) of web pages and returning relevant results, many organizations are looking to Hadoop as their next big data platform. Popular uses today include:

1. **Low-cost storage and active data archive.** The modest cost of commodity hardware makes Hadoop useful for storing and combining data such as transactional, social media, sensor, machine, scientific, click streams, etc. The low-cost storage lets you keep information that is not deemed currently critical but that you might want to analyze later.
2. **Staging area for a data warehouse and analytics store.** One of the most prevalent uses is to stage large amounts of raw data for loading into an enterprise data warehouse (EDW) or an analytical store for activities such as advanced analytics, query and reporting, etc. Organizations are looking at Hadoop to handle new types of data (e.g., unstructured), as well as to offload some historical data from their enterprise data warehouses.
3. **Data lake.** Hadoop is often used to store large amounts of data without the constraints introduced by schemas commonly found in the SQL-based world. It is used as a low-cost compute-cycle platform that supports processing ETL and data quality jobs in parallel using hand-coded or commercial data management technologies. Refined results can then be passed to other systems (e.g., EDWs, analytic marts) as needed.
4. **Sandbox for discovery and analysis.** Because Hadoop was designed to deal with volumes of data in a variety of shapes and forms, it can run analytical algorithms. Big data analytics on Hadoop can help your organization operate more efficiently, uncover new opportunities and derive next-level competitive advantage. The sandbox approach provides an opportunity to innovate with minimal investment.
5. **Recommendation systems.** One of the most popular analytical uses by some of Hadoop's largest adopters is for web-based recommendation systems. Facebook – people you may know. LinkedIn – jobs you may be interested in. Netflix, eBay, Hulu – items you may be interested in. These systems analyze huge amounts of data in real time to quickly predict preferences before customers leave the web page.

IV. HADOOP DATA ANALYSIS TECHNOLOGIES

Let’s have a look at the existing open source Hadoop data analysis technologies to analyze the huge stock data being generated very frequently.

MAPREDUCE

- Powerfull model for parallelism.
- Based on rigid procedural structure.

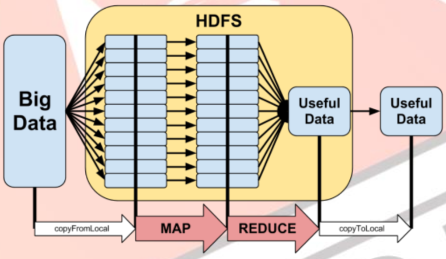


Figure 1: Map Reduce

PIG

- Procedural dataflow language.
- Used by programmers and researchers.

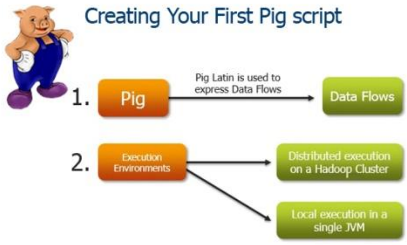
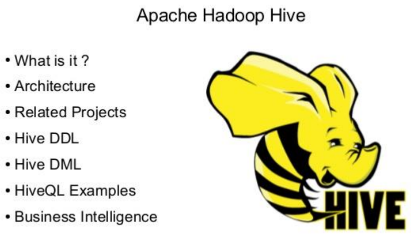

Creating Your First Pig script
1.
Pig
Pig Latin is used to express Data Flows
Data Flows
2.
Execution Environments
Distributed execution on a Hadoop Cluster
Local execution in a single JVM

Figure 2: Pig

HIVE

- Declarative SQLish Language
- Used by analysts for generating reports



Apache Hadoop Hive

- What is it ?
- Architecture
- Related Projects
- Hive DDL
- Hive DML
- HiveQL Examples
- Business Intelligence

Figure 3: Hive

NO SQL

NoSQL (commonly referred to as "Not Only SQL") represents a completely different framework of databases that allows for high-performance, agile processing of information at massive scale. In other words, it is a database infrastructure that as been very well-adapted to the heavy demands of big data. The efficiency of NoSQL can be achieved because unlike relational databases that are highly structured, NoSQL databases are unstructured in nature, trading off stringent consistency requirements for speed and agility. NoSQL centers around the concept of distributed databases, where unstructured data may be stored across multiple processing nodes, and often across multiple servers. This distributed architecture allows NoSQL databases to be horizontally scalable; as data continues to explode, just add more hardware to keep up, with no slowdown in performance.

The NoSQL distributed database infrastructure has been the solution to handling some of the biggest data warehouses on the planet – i.e. the likes of Google, Amazon, and the CIA.

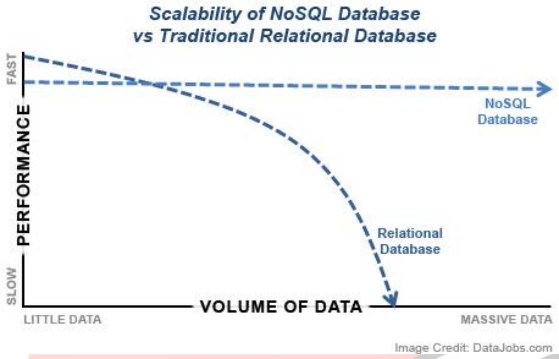


Image Credit: DataJobs.com

Figure 4: No Sql

BIG DATA ANALYSIS

Big data is mostly generated from social media websites, sensors, devices, video/audio, networks, log files and web, and much of it is generated in real time and on a very large scale. Big data analytics is the process of examining this large amount of different data types, or big data, in an effort to uncover hidden patterns, unknown correlations and other useful information.

Figure 5: Big Data Analysis

V. CHALLENGES OF USING HADOOP

1. **MapReduce programming is not a good match for all problems.** It’s good for simple information requests and problems that can be divided into independent units, but it's not efficient for iterative and interactive analytic tasks. MapReduce is file-intensive. Because the nodes don’t intercommunicate except through sorts and shuffles, iterative algorithms require multiple map-shuffle/sort-reduce phases to complete. This creates multiple files between MapReduce phases and is inefficient for advanced analytic computing.
2. **There’s a widely acknowledged talent gap.** It can be difficult to find entry-level programmers who have sufficient Java skills to be productive with MapReduce. That's one reason distribution providers are racing to put relational (SQL) technology on top of Hadoop. It is much easier to find programmers with SQL skills than MapReduce skills. And,

Hadoop administration seems part art and part science, requiring low-level knowledge of operating systems, hardware and Hadoop kernel settings.

3. **Data security.** Another challenge centers around the fragmented data security issues, though new tools and technologies are surfacing. The Kerberos authentication protocol is a great step forward for making Hadoop environments secure.
4. **Full-fledged data management and governance.** Hadoop does not have easy-to-use, full-feature tools for data management, data cleansing, governance and metadata. Especially lacking are tools for data quality and standardization.

VI. BENEFITS OF HADOOP

One of the top reasons that organizations turn to Hadoop is its ability to store and process huge amounts of data – any kind of data – quickly. With data volumes and varieties constantly increasing, especially from social media and the Internet of Things, that's a key consideration. Other benefits include:

- **Computing power.** Its distributed computing model quickly processes big data. The more computing nodes you use, the more processing power you have.
- **Flexibility.** Unlike traditional relational databases, you don't have to preprocess data before storing it. You can store as much data as you want and decide how to use it later. That includes unstructured data like text, images and videos.
- **Fault tolerance.** Data and application processing are protected against hardware failure. If a node goes down, jobs are automatically redirected to other nodes to make sure the distributed computing does not fail. And it automatically stores multiple copies of all data.
- **Low cost.** The open-source framework is free and uses commodity hardware to store large quantities of data.
- **Scalability**. You can easily grow your system simply by adding more nodes. Little administration is required.

VII. CONCLUSION

Hadoop and the MapReduce programming paradigm already have a substantial base in the bioinformatics community, especially in the field of next-generation sequencing analysis, and such use is increasing. This is due to the cost-effectiveness of Hadoop-based analysis on commodity Linux clusters, and in the cloud via data upload to cloud vendors who have implemented Hadoop/HBase; and due to the effectiveness and ease-of-use of the MapReduce method in parallelization of many data analysis algorithms.

REFERENCES

1. "Hadoop Releases". apache.org. Apache Software Foundation. Retrieved 2014-12-06.
2. "Hadoop Releases". Hadoop.apache.org. Retrieved 2015-07-29.
3. "Welcome to Apache™ Hadoop®!". hadoop.apache.org. Retrieved 2015-09-20.
4. "What is the Hadoop Distributed File System (HDFS)?". ibm.com. IBM. Retrieved 2014-10-30.
5. Malak, Michael (2014-09-19). "Data Locality: HPC vs. Hadoop vs. Spark". datascienceassn.org. Data Science Association. Retrieved 2014-10-30.
6. "Resource (Apache Hadoop Main 2.5.1 API)". apache.org. Apache Software Foundation. 2014-09-12. Retrieved 2014-09-30.
7. Murthy, Arun (2012-08-15). "Apache Hadoop YARN – Concepts and Applications". hortonworks.com. Hortonworks. Retrieved 2014-09-30.
8. "Continuuity Raises $10 Million Series A Round to Ignite Big Data Application Development Within the Hadoop Ecosystem". finance.yahoo.com. Marketwired. 2012-11-14. Retrieved 2014-10-30.
9. "Hadoop-related projects at". Hadoop.apache.org. Retrieved 2013-10-17.
10. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. John Wiley & Sons. 2014-12-19. p. 300. ISBN 9781118876220. Retrieved 2015-01-29.