Big Data - Concepts

19 December 2023

1. Introduction to Big Data

Lecture

- Big Data - Beyond the Hype,
- Big Data Skills and Sources of Big Data,
- Big Data Adoption,
- Research and Changing Nature of Data Repositories,
- Data Sharing and Reuse Practices and Their Implications for Repository Data Curation,
- Overlooked and Overrated Data Sharing,
- Data Curation Services in Action,
- Open Exit: Reaching the End of The Data Life Cycle,
- The Current State of Meta-Repositories for Data
- Curation of Scientific Data at Risk of Loss: Data Rescue And Dissemination

2. Introduction to Hadoop

Lecture

- A Brief History of Hadoop,
- Evolution of Hadoop,
- Introduction to Hadoop and its components
- Comparison with Other Systems,
- Hadoop Releases
- Hadoop Distributions and Vendors

3. Hadoop Distributed File System (HDFS)

Lecture

- Distributed File System,
- What is HDFS,
- Where does HDFS fit in,
- Core components of HDFS,
- HDFS Daemons,
- Hadoop Server Roles: Name Node, Secondary Name Node, and Data Node

4. HDFS Architecture

Lecture

- HDFS Architecture,
- Scaling and Rebalancing,
- Replication,
- Rack Awareness,
- Data Pipelining,
- Node Failure Management.
- HDFS High Availability NameNode

Lab-Assignment:

- Run the HDFS commands, and add a one liner understanding for each of the command.
- Execute the provided code using HDFS, step run and understand

How does Hadoop Work?

Investment on higher end configuration machines to handle large scale processing it very expensive, hence we can use the alternate solution to this is combining multiple commodity computers with single-CPU as a single functional distributed systems and practically the clustering of machines to read the dataset in parallel and provide a much higher throughout. This will be a cheaper solution rather than investing on high-end servers. " **Hadoop runs across clusters and low-cost machines** "

Hadoop perform the following tasks:-

- Data is initially divided into directories and files.
- Files are divided into uniform sized blocks of 128MB or 64MB (preferred is 128MB)
- These files are then distributed across various cluster nodes for further processing
- HDFS, on top of the local file systems, supervises the processing
- Blocks are replicated for handling hardware failures
- Checking whether the code was executed successfully
- Perform the sort which takes place between the map and reduce stages
- Sending the sorted data to a certain computer.
- Writes the debugging logs for each job.

Advantages of Hadoop

- It is an open source compatible with all the platforms since its java based.
- We can add or remove servers form the clusters dynamically & Hadoop continues to operate without any interruption
- Does not rely on H/w to provide FT and HA but Hadoop library itself has been designed inorder to detect and handle failures at the application layer itself.
- Allows users to quickly write and test distributed systems.
- IT is efficient, and it automatically distributes data and work across the machines and in turn, utilizes the underlying parallelism of the CPU Cores

Hadoop Setup [Environment Preparation]

Hadoop is supported on any GNU/Linux flavours. Install Linux OS (preferred is server editions for enterprise setup and for testing we can also use desktop machines). If there is an different OS other than Linux use any Virtualization engine (VMware workstation / Oracle VirtualBox software) and install Linux inside the virtual machine

```
useradd hadoop
passwd hadoop
mkdir /home/hadoop
chown hadoop.hadoop  /home/hadoop
su - hadoop
ssh-keygen -t rsa
cd .ssh
ls -la
cat id_rsa.pub  > authorized_keys
chmod 0600 authorized_keys
usermod -s /bin/bash hadoop
```

Download the jdk from the internet to any location of the local server

```
$ cd Downloads
$ tar -zxvf jdk-7u80-linux-x64.tar.gz
$ su
$ mv jdk1.7.0_80 /usr/local
$ cd /usr/local
$ chown -R root.root /usr/local/jdk1.7.0_80
$ chmod -R +x /usr/local/jdk1.7.0_80

$ nano /etc/bash.bashrc
export JAVA_HOME=/usr/local/jdk1.7.0_80
export PATH=$PATH:$JAVA_HOME/bin
<save and exit>

$ java -version
java version "1.7.0_80"
Java(TM) SE Runtime Environment (build 1.7.0_80-b15)
Java HotSpot(TM) 64-Bit Server VM (build 24.80-b11, mixed mode)

---- desired output ---
---installing java is successful ---

$ cd /opt
$ wget https://archive.apache.org/dist/hadoop/common/hadoop-2.4.1/hadoop-2.4.1.tar.gz
$ tar zxf hadoop-2.4.1.tar.gz
```

$ mv hadoop-2.4.1 /usr/local/hadoop
$ chown -R root.root /usr/local/hadoop

$ nano /etc/bash.bashrc
export JAVA_HOME=/usr/local/jdk1.7.0_80
Export HADOOP_HOME=/usr/local/hadoop
export PATH=$PATH:$JAVA_HOME/bin:$HADOOP_HOME/bin
<save and exit>

$ source /etc/bash.bashrc
$ hadoop version

```
root@mainserver1:/etc# hadoop version
Hadoop 2.4.1
Subversion http://svn.apache.org/repos/asf/hadoop/common -r 1604318
Compiled by jenkins on 2014-06-21T05:43Z
Compiled with protoc 2.5.0
From source with checksum bb7ac0a3c73dc131f4844b873c74b630
This command was run using /usr/local/hadoop/share/hadoop/common/hadoop-common-2.4.1.jar
```

Hadoop Versions

Apache Hadoop, an open-source framework for distributed storage and processing of large datasets, has gone through several major releases

1. Hadoop 0.20.x Series: The first initial stable release, which laid the foundation for Hadoop's distributed file systems and mapreduce programming tool
2. Hadoop 1.x Series: Enhancements to HDFS , job scheuling and fault tolerance. Among these series Hadoop 1.0.0 was more popular, new features and improvements were added
3. Hadoop 2.x Servies (YARN): Yet Another Resource Negotiator. YARN decouples the resource management and job scheduling functions of mapreduce and hence allowed hadoop to support multiple programming models. Hadoop 2.2.0 are subsequent versions.
4. Hadoop 3.x series: the latest series - further improvements and optimizations were done. Includes support for erasure coding in HDFS, enhancements to resource management and got certain improvements in performance and scalability.

For detailed information you can visit https://hadoop.apache.org/ site

Download the Apache Hadoop form the following location (version choice is yours)
https://archive.apache.org/dist/hadoop/common/hadoop-2.4.1/

Or use the following command
$ cd /opt
$ wget https://archive.apache.org/dist/hadoop/common/hadoop-2.4.1/hadoop-2.4.1.tar.gz

Hadoop distribution - Vendors

Several vendors offer Hadoop distributions, providing users with pre-packaged versions of the Apache Hadoop framework along with additional tools, management features, and support. These distributions are designed to simplify the deployment, configuration, and management of Hadoop clusters. As of my last knowledge update in January 2022, here are some of the popular Hadoop distributions and vendors:

1. Cloudera:
   - Distribution: Cloudera Distribution Including Apache Hadoop (CDH)
   - Key Features: Cloudera Manager for cluster management, Cloudera Navigator for data management, and Cloudera Impala for interactive SQL queries.

2. Hortonworks:
   - Distribution: Hortonworks Data Platform (HDP)
   - Key Features: Ambari for cluster provisioning and management, SmartSense for diagnostic analytics, and Apache NiFi for data integration.

3. MapR:
   - Distribution: MapR Converged Data Platform
   - Key Features: MapR File System (MapR-FS) for distributed file storage, MapR-DB for NoSQL data, and MapR Event Store for event streaming.

4. Amazon EMR (Elastic MapReduce):
    - Managed Service: Amazon EMR is a cloud-based managed Hadoop and Spark service provided by Amazon Web Services (AWS).
    - Key Features: Integration with other AWS services, simplified cluster provisioning, and auto-scaling capabilities.

5. Microsoft Azure HDInsight:
    - Managed Service: Azure HDInsight is a cloud-based managed Hadoop and Spark service provided by Microsoft Azure.
    - Key Features: Integration with Azure services, enterprise-grade security, and support for various Hadoop ecosystem components.

6. IBM InfoSphere BigInsights:
    - Distribution: IBM BigInsights
    - Key Features: Integration with IBM Db2 and IBM Watson, Big SQL for SQL querying, and additional tools for analytics and data processing.

7. Pivotal HD:
    - Distribution: Pivotal HD (now part of VMware Tanzu)
    - Key Features: Integration with Pivotal Greenplum for analytics, support for HAWQ (SQL query engine), and tools for data science.

8. Apache Hadoop (Vanilla):
    - Distribution: Users can also choose to deploy the open-source Apache Hadoop distribution directly, without any additional vendor-specific enhancements. This is often referred to as the "vanilla" distribution.

It's important to note that the landscape of Hadoop distributions and vendors may have evolved since last update, and new players or changes in offerings may have occurred. Additionally, the trend in the industry has shifted towards broader data platforms that incorporate both traditional big data processing and newer technologies like Apache Spark. Organizations should carefully evaluate their specific requirements and preferences when selecting a Hadoop distribution or considering alternative technologies for their data processing and analytics needs

Overlooked and Overrated Data Sharing

The practice of data sharing is crucial for advancing research, fostering collaboration, and promoting transparency. However, there are instances where certain aspects of data sharing may be overlooked or overrated. Here are some considerations regarding overlooked and overrated aspects of data sharing:

Overlooked Aspects of data sharing

1. Ethical and privacy concerns
2. Data documentations and metadata
3. Long-Term data preservation
4. Data quality and validation

Overrated Aspects of Data Sharing

1. Data quantity over data quality
2. Openness without context
3. Universal applicability of data
4. Data sharing as a one-time activity
5. Assuming universal data reusability

Balancing the benefits of data sharing with ethical considerations, privacy safeguards, and a commitment to data quality is essential for responsible and impactful research. By addressing both the overlooked and overrated aspects of data sharing, the scientific community can ensure that shared data contributes meaningfully to knowledge advancement and collaboration