International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 10, pp 182-185, October-2024

https://doi.org/10.47001/IRJIET/2024.810025

Hadoop Environment Setup for Big Data

[1]Pooja S. Gadhave, [2]Sanika D. Pangul, [3]Tejaswini J. Bhande, [4]Dr. Shilpa B. Sarvaiya

[1,2,3]MCA-II, Department of MCA, Vidya Bharti Mahavidyalaya, Amravati, Maharashtra, India

[4]Head, Department of MCA, Vidya Bharti Mahavidyalaya, Amravati, Maharashtra, India

Authors E-mail: [1]gadhawepooja2@gmail.com, [2]sanupangul6@gmail.com, [3]bhandeteju@gmail.com, [4]sarvaiya.shilpa@gmail.com

***Abstract* - Hadoop is a powerful open-source framework designed for the distributed storage and processing of large data sets across clusters of computers, making it a vital tool in the era of big data. Big data refers to vast volumes of data generated at high velocity and from various sources, presenting significant challenges in storage, analysis, and management. This paper outlines the installation steps necessary to set up a Hadoop environment on a Linux operating system, which provides a stable and efficient platform for running distributed applications. By offering a comprehensive overview of the installation and operational aspects of Hadoop, this research serves as a practical guide for beginners and practitioners, facilitating efficient data processing and enhancing the understanding of big data management in a Linux ecosystem.**

Keywords: Apache Hadoop, Big Data, Hadoop Distributed File System, Hadoop Installation Linux Commands for Hadoop, Linux Operating System for Big Data, NameNode, Yarn.

I. Introduction

In the era of big data, organizations face significant challenges in processing and analyzing vast amounts of information, especially as traditional systems struggle to scale. Apache Hadoop addresses these challenges by enabling distributed storage and processing of large datasets across commodity hardware. Hadoop's core components include the Hadoop Distributed File System (HDFS) for storage and the MapReduce model for parallel processing, while YARN (Yet Another Resource Negotiator) manages cluster resources. Hadoop's adoption has transformed data-intensive industries by offering a cost-effective solution for big data analytics. This paper explores the installation of Hadoop on Linux and the essential commands for managing HDFS. Understanding these fundamentals is crucial for harnessing Hadoop's power for large-scale data processing and analysis.[1] [2]

A) Hadoop Environment setup

Step 1: Install Java

1. Update package repository:

```
//sql
sudo apt-get update
```

2. Install Java:

```
//arduino
sudo apt-get install openjdk-8-jdk -y
```

3. Set up JAVA_HOME in .bashrc:

```
//bash
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PATH=$PATH:$JAVA_HOME/bin
source ~/.bashrc
```

Step 2: Install and Configure Hadoop

1. Download and extract Hadoop:

```
//bash
Wget https://downloads.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
tar -xzvf hadoop-3.3.6.tar.gz
sudo mv hadoop-3.3.6 /usr/local/hadoop
```

2. Set Hadoop environment variables in .bashrc:

```
//bash
export HADOOP_HOME=/usr/local/hadoop
export PATH=$PATH:$HADOOP_HOME/bin
source ~/.bashrc
```

Step 3: Configure Hadoop Files

1. Set JAVA_HOME in hadoop-env.sh:

```
//javascript
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
```

2. Configure core-site.xml:

```
//xml
<property>
<name>fs.defaultFS</name>
<value>hdfs://localhost:9000</value>
</property>
```

3. Configure hdfs-site.xml:

```
//xml
<property>
<name>dfs.replication</name>
<value>1</value>
```


International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 8, Issue 10, pp 182-185, October-2024
https://doi.org/10.47001/IRJIET/2024.810025

</property>

4. Configure mapred-site.xml and yarn-site.xml.

Step 4: Setup Passwordless SSH

1. Install SSH and generate key:

```
//javascript
sudo apt-get install openssh-server openssh-client
ssh-keygen -t rsa -P ""
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
ssh localhost
```

Step 5: Format Namenode

```
//lua
hdfs namenode -format
```

Step 6: Start Hadoop

1. Start HDFS and YARN:

```
//sql
start-dfs.sh
start-yarn.sh
jps
```

Step 7: Access Hadoop Web UI

- NameNode: http://localhost:9870/
- ResourceManager: http://localhost:8088/

Stopping Hadoop

1. Stop YARN and HDFS:

```
//arduino
stop-yarn.sh
stop-dfs.sh[3] [4]
```

B) Hadoop Installation On Ubuntu:

Step 1: Install Java (OpenJDK 8):

```
//bash
sudo apt-get update
sudo apt-get install openjdk-8-jdk -y
```

Step 2: Set JAVA_HOME in .bashrc:

```
//bash
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PATH=$PATH:$JAVA_HOME/bin
source ~/.bashrc
```

Step 3: Download and Install Hadoop:

```
//bash
Wget https://downloads.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
tar -xzvf hadoop-3.3.6.tar.gz
sudo mv hadoop-3.3.6 /usr/local/hadoop
```

Step 4: Set Hadoop Variables in .bashrc:

```
//bash
export HADOOP_HOME=/usr/local/hadoop
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
source ~/.bashrc
```

Step 5: Configure Hadoop:

a) hadoop-env.sh: Set JAVA_HOME.
b) core-site.xml: Set fs.defaultFS to hdfs://localhost:9000.
c) hdfs-site.xml: Set dfs.replication to 1.

Step 6: Setup Passwordless SSH:

```
//bash
sudo apt-get install openssh-server openssh-client
ssh-keygen -t rsa -P ""
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
ssh localhost
```

Step 7: Format Namenode:

```
//bash
hdfs namenode -format
```

Step 8: Start Hadoop:

```
//bash
start-dfs.sh
start-yarn.sh
```

Step 9: Access Web UI:

d) NameNode: http://localhost:9870
e) ResourceManager: http://localhost:8088

Step 10: Stop Hadoop:

```
//bash
stop-yarn.sh
stop-dfs.sh[5] [6]
```

C) Check Hadoop Installation Status

To check if Hadoop is successfully installed:

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 10, pp 182-185, October-2024

https://doi.org/10.47001/IRJIET/2024.810025

1. Verify Hadoop Version:

```
//bash
hadoop version
```

2. Format NameNode:

```
//bash
hdfs namenode -format
```

3. Start Services:

```
//bash
start-dfs.sh
start-yarn.sh
```

4. Check Running Daemons:

```
jps
```

Look for NameNode, DataNode, ResourceManager, and NodeManager.

5. Access Web UIs:

- HDFS NameNode: http://localhost:9870/
- YARN ResourceManager: http://localhost:8088 [7] [8]

D) Advantages of Hadoop Installation/Setup

1. Scalability: Easily scales by adding more nodes to accommodate growing datasets.

- Limitation: Requires manual effort to manage and configure additional nodes.

2. Cost-Effective: Open-source and uses commodity hardware, reducing overall costs.

- Limitation: Costs can increase with high maintenance and resource demands for large clusters.

3. Fault Tolerance: Replicates data across nodes, ensuring no data loss during node failures.

- Limitation: Data replication increases storage requirements, making it inefficient for small-scale setups.

4. Distributed Processing: Processes data in parallel across multiple nodes for faster performance.

- Limitation: High input/output overhead can slow down tasks, especially with small datasets.

5. Flexible Data Handling: Can process structured, semi-structured, and unstructured data.

- Limitation: Complex data structures may require additional tools and configurations for efficient processing. [9] [10]

II. Conclusion

Hadoop provides a robust, distributed platform for handling and processing vast amounts of data, making it essential for Big Data applications. Its core components—HDFS for storage, MapReduce for processing, and YARN for resource management—ensure a scalable and efficient architecture capable of managing modern data workloads. The installation process, though requiring careful attention to configurations, lays the groundwork for a reliable and high-performing environment.

With fault tolerance, Hadoop ensures that data remains accessible even in the event of system failures, while its horizontal scalability allows businesses to easily expand infrastructure as data grows. Additionally, Hadoop's extensive ecosystem of tools like Hive, Pig, and HBase adds flexibility for various analytics and data processing tasks. As an open-source solution running on commodity hardware, Hadoop offers a cost-effective approach to large-scale data management.

Moreover, its support for both real-time and batch processing through tools like Apache Spark and Flink makes it adaptable to diverse business requirements. The availability of Hadoop commands empowers users to efficiently manage data and processing jobs, ensuring that organizations can harness the full potential of their data infrastructure.

Future Scope of Hadoop

1. Cloud Integration: Increased use of cloud platforms (AWS, Azure, Google Cloud) for scalable, cost-effective Hadoop services.
2. AI & Machine Learning: Expanded role in advanced analytics, integrating with ML frameworks like TensorFlow for real-time insights.
3. Real-time Processing: Enhanced real-time data handling, competing with Apache Spark for low-latency processing.
4. Serverless Hadoop: Transition to serverless models, reducing the need for manual infrastructure management.
5. Hadoop Ecosystem Growth: Greater use of ecosystem tools (e.g., Kafka, HBase) for specialized data processing.
6. Security Enhancements: Improved data security through tools like Apache Ranger and Sentry, crucial for regulated industries.
7. Edge Computing: Integration with edge computing to process data closer to the source, benefiting IoT applications.
8. Storage Efficiency: Better data compression and optimized storage, reducing costs for large datasets.[11] [12]

ISSN (online): 2581-3048
Volume 8, Issue 10, pp 182-185, October-2024
https://doi.org/10.47001/IRJIET/2024.810025

REFERENCES

[1] Forbes Welcome, https://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/#487d104413ae (Access on March 30, 2019).
[2] Hadoop, http://hadoop.apache.org (Access on March 30, 2019).
[3] Dean, J. and Ghemawat, S., MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), pp.107-113 (2008).
[4] Shah A., Padole M. (2019) Performance Analysis of Scheduling Algorithms in Apache Hadoop. In: Shukla R., Agrawal J., Sharma S., Singh Tomer G. (eds) Data, Engineering and Applications. Springer, Singapore.
[5] Shvachko, K., Kuang, H., Radia, S. and Chansler, R., 2010, May. The hadoop distributed file system. In MSST (Vol. 10, pp. 1-10).
[6] Vavilapalli, V.K., Murthy, A.C., Douglas, C., Agarwal, S., Konar,M., Evans, R., Graves, T., Lowe, J., Shah, H., Seth, S. and Saha, B., (2013). Apache hadoop yarn: Yet another resource negotiator. In Proceedings of the 4th annual Symposium on Cloud Computing (p.5). ACM.
[7] BaoRong Chang, Yo-Ai Wang, Yun-Da Lee, and Chien-Feng Huang, "Development of Multiple Big Data Analysis Platforms for Business Intelligence", Proceedings of the 2017 IEEE International Conference on Applied System Innovation.
[8] Chu-Hsing Lin, Jung-Chun Liu, Tsung-Chi Peng, "Performance Evaluation of Cluster Algorithms for Big Data Analysis on Cloud", Proceedings of the 2017 IEEE International Conference on Applied System Innovation.
[9] https://intellipaat.com/tutorial/hadooptutorial/introduction-hadoop/
[10] Apache Hadoop. http://hadoop.apache.org/
[11] Ms. Preeti Narooka, Dr. Sunita Choudhary, "Optimization of the Search Graph Using Hadoop and Linux Operating System", 2017 International Conference on Nascent Technologies in the Engineering Field (ICNTE-2017) IEEE-ICASI 2017.
[12] Yu-Sheng Su1, Ting-Jou Ding2, Jiann-Hwa Lue3, Chin-Feng Lai4, Chiu-Nan Su5,"Applying Big Data Analysis Technique to Students' Learning Behavior and Learning", Proceedings of the 2017 IEEE International Conference on Applied System Innovation IEEE-ICASI 2017.

Citation of this Article:

Pooja S. Gadhave, Sanika D. Pangul, Tejaswini J. Bhande, & Dr. Shilpa B. Sarvaiya. (2024). Hadoop Environment Setup for Big Data. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 8(10), 182-185. Article DOI https://doi.org/10.47001/IRJIET/2024.810025

****

