Department of MCA

Topic: Hadoop Streaming

COURSE	UNIT - II	CLASS
19CA917 Big Data Analytics	Hadoop	III Semester / II MCA

Hadoop Streaming

•Hadoop is a powerful open-source software framework for distributed processing of large datasets across clusters of computers. One of the key features of Hadoop is its MapReduce programming model, which allows for the distributed processing of large datasets.

• However, MapReduce requires developers to write code in Java, which can be a barrier to entry for some organizations. To address this, Hadoop provides a feature called Hadoop Streaming which allows for the use of other programming languages, such as Python or Ruby, with Hadoop's MapReduce framework.

Hadoop Streaming

- Hadoop Streaming works by allowing users to write MapReduce jobs in other languages and then pass them to Hadoop for execution.
- This is achieved through the use of standard input and output streams, which are used to pass data between the Map and Reduce tasks.

Hadoop Streaming

- To use Hadoop Streaming, developers must first write a MapReduce job in a language other than Java. They can then pass the job to Hadoop using the Hadoop Streaming API.
- The API takes care of the details of passing data between the Map and Reduce tasks using standard input and output streams.

Hadoop Streaming

- One of the main benefits of Hadoop Streaming is its flexibility. By allowing developers to use other programming languages with Hadoop's MapReduce framework, it opens up Hadoop to a wider range of developers with different skill sets.
- This makes it easier for organizations to adopt Hadoop as a solution for big data analytics.

Hadoop Streaming

- Another benefit of Hadoop Streaming is its performance. Because the MapReduce jobs are executed using standard input and output streams, there is very little overhead associated with passing data between the Map and Reduce tasks. This allows for faster processing times and more efficient use of resources.
- Hadoop Streaming also provides businesses with a cost-effective solution for big data analytics. Because it allows developers to use programming languages other than Java, businesses can leverage their existing expertise and tools, reducing the need to hire specialized personnel or invest in new.

Hadoop Streaming

•’s closer at Hadoop Streaming works in practice. Suppose a business wants to analyze log data from their website to gain insights into user behavior. They have a team of developers who are skilled in Python, but not Java. They want to use Hadoop to process the log data, but don’t want to write the MapReduce job in Java.

•To use Hadoop Streaming, the business would first write the MapReduce job in Python. They would then pass the job to Hadoop using the Hadoop Streaming API. The API takes care of the details of passing data between the Map and Reduce tasks using standard input and output streams.

Hadoop Streaming

- The Python MapReduce job would read in the log data, perform any necessary transformations or aggregations, and output the results to standard output. Hadoop would then take care of passing the data between the Map and Reduce tasks using standard input and output streams.
- As the business generates more log data, they can scale out their Hadoop cluster to handle the increased workload. This allows the business to analyze large amounts of log data quickly and efficiently.

Hadoop Streaming

•However, there are some challenges associated with using Hadoop Streaming. One of the biggest challenges is managing the increased complexity that comes with using different programming languages with Hadoop's MapReduce framework. Businesses need to ensure that they have the proper infrastructure in place to manage the increased workload and ensure that the cluster is running smoothly.

Hadoop Streaming

- Another challenge is ensuring that the data is properly formatted for use with Hadoop Streaming.
- Because Hadoop Streaming uses standard input and output streams to pass data between the Map and Reduce tasks, businesses need to ensure that the data is properly formatted and that any necessary transformations or aggregations are performed before the data is passed to Hadoop.

Hadoop Streaming

- Hadoop Streaming is a powerful feature of Hadoop that allows businesses to use other programming languages, such as Python or Ruby, with Hadoop's MapReduce framework.
- This makes it easier for organizations to adopt Hadoop as a solution for big data analytics, as it allows them to leverage their existing expertise and tools. While there are some challenges associated with using Hadoop Streaming, the benefits of this feature make it an attractive option for businesses of all sizes.

References

- Tom White, “ Hadoop: The Definitive Guide” Third Edition, O’reilly Media, 2012

https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

https://data-flair.training/blogs/hadoop-hdfs-tutorial/

https://www.simplilearn.com/tutorials/hadoop-tutorial/hdfs (video)

https://www.techbooktutorial.com/hadoop/hadoop-distributed-file-system

https://www.techbooktutorial.com/hadoop/hadoop-streaming