1. Mention Hadoop distribution? Difference between CDH and CDP

Sunday, August 11, 2024 10:27 AM

Hadoop Distribution: Refers to the specific version or package of Hadoop provided by vendors like Cloudera, Hortonworks, or MapR. These distributions often include additional tools and services for easier deployment and management.

CDH (Cloudera Distribution Hadoop): An older distribution by Cloudera that provided Hadoop and other related big data tools. It's stable but has been mostly phased out.

CDP (Cloudera Data Platform): The next-generation platform that unifies Cloudera and Hortonworks' technologies. CDP offers both on-premise (CDP Private Cloud) and cloud-native (CDP Public Cloud) solutions, with enhanced security, data governance, and easier management.

2. Explain Hadoop Architecture

Sunday, August 11, 2024 10:30 AM

HDFS (Hadoop Distributed File System): The storage layer, splitting data into blocks and distributing them across multiple nodes.
MapReduce: The processing layer, where jobs are divided into tasks and executed across the cluster.
YARN (Yet Another Resource Negotiator): The resource management layer, which handles job scheduling and resource allocation.

3. Configuration files used during hadoop installation

Sunday, August 11, 2024 10:30 AM

core-site.xml: Contains core configuration settings like default filesystem, I/O settings.

hdfs-site.xml: Configurations related to HDFS, such as replication factor and block size.

mapred-site.xml: Settings related to MapReduce jobs.

yarn-site.xml: Configurations for YARN, such as resource manager address and node manager settings.

4. Difference Between hadoop fs and hdfs dfs

Sunday, August 11, 2024 10:31 AM

hadoop fs: A generic filesystem shell command that can interact with any Hadoop-supported filesystem, including local, HDFS, S3, etc.
hdfs dfs: Specifically designed for HDFS operations. It’s optimized for HDFS and includes all HDFS-related commands.

5. Difference Between Hadoop 2 and Hadoop 3

Sunday, August 11, 2024 10:33 AM

Hadoop 2: Introduced YARN for better resource management, allowing non-MapReduce jobs. It also supported High Availability (HA) for NameNode.
Hadoop 3: Introduced Erasure Coding for more efficient storage, improved NameNode memory management, and native Docker support for YARN. It also supports multiple NameNodes.

6. What is replication factor ? why its important

Sunday, August 11, 2024 10:34 AM

Replication Factor: Determines how many copies of a data block are stored across the cluster. It's essential for fault tolerance and data availability. A typical default replication factor is 3.

7. What if DataNode Fails?

Sunday, August 11, 2024 10:35 AM

If a DataNode fails, the NameNode detects the failure and replicates the blocks stored on that DataNode to other DataNodes to maintain the replication factor.

8. What if NameNode Fails?

Sunday, August 11, 2024 10:53 AM

In Hadoop 2 and 3, NameNode High Availability (HA) is used to handle NameNode failure. It allows for a standby NameNode that can take over in case of failure.

9. Why is Block Size 128 MB? What If I Increase or Decrease the Block Size?

Sunday, August 11, 2024 10:53 AM

Block Size: 128 MB is chosen to reduce the overhead of metadata management while balancing the I/O operations for large datasets. Increasing the block size reduces metadata overhead but may lead to underutilized blocks for small files. Decreasing it may lead to more overhead for NameNode in managing metadata.

10. Small File Problem

Sunday, August 11, 2024 10:54 AM

Small File Problem: Refers to the inefficiency in HDFS when dealing with a large number of small files, as each file creates metadata overhead, burdening the NameNode.

11. Rack Awareness

Sunday, August 11, 2024 10:56 AM

Rack Awareness: Hadoop’s ability to understand the rack topology of the cluster and place replicas of data blocks on different racks to ensure fault tolerance and reduce network traffic.

12. SPOF (Single Point of Failure) & Its Resolution

Sunday, August 11, 2024 10:57 AM

SPOF: Refers to a component whose failure can halt the entire system. NameNode was originally a SPOF in Hadoop 1. In Hadoop 2 and 3, High Availability (HA) with standby NameNodes resolves this issue.

13. Zookeeper

Sunday, August 11, 2024 10:58 AM

Zookeeper: A centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. It’s used in Hadoop for managing the distributed systems' coordination and configuration.

14. Difference Between -put and -copyFromLocal

Sunday, August 11, 2024 10:58 AM

-put: Uploads files from the local filesystem to HDFS.
-copyFromLocal: Also uploads files from the local filesystem to HDFS but is more explicit about the source being local.

15. Erasure Coding

Sunday, August 11, 2024 10:59 AM

Erasure Coding: A method introduced in Hadoop 3 to reduce the storage overhead of the replication factor by breaking data into fragments and adding parity blocks. It provides the same level of fault tolerance with less storage overhead.

16. Speculative Execution

Sunday, August 11, 2024 11:00 AM

Speculative Execution: A technique where Hadoop runs duplicate instances of the same task on different nodes. The first instance to finish is used, and the others are killed. This helps in reducing the impact of slow-running tasks (stragglers).

17. YARN Architecture

Sunday, August 11, 2024 11:03 AM

ResourceManager: Manages resources across the cluster and schedules jobs.
NodeManager: Manages resources on a single node.
ApplicationMaster: Manages the lifecycle of an individual application (job), negotiating resources with the ResourceManager.

18. How does ApplicationManager and Application Master differ

Sunday, August 11, 2024 11:05 AM

ApplicationManager: Part of the ResourceManager, responsible for managing all the applications on the cluster.
ApplicationMaster: A per-application process that manages the application lifecycle and task execution.

19. Explain Mapreduce working?

Sunday, August 11, 2024 11:06 AM

Map Phase: Splits the input data into key-value pairs and processes them.
Shuffle and Sort Phase: The intermediate data is shuffled and sorted based on keys.
Reduce Phase: The sorted key-value pairs are processed to produce the final output.

20. How Many Mappers Are Created for a 1 GB File?

Sunday, August 11, 2024 11:07 AM

Number of mappers = Number of input splits. Typically, one mapper is created per block of data, so for a 1 GB file with a block size of 128 MB, you would get 8 mappers.

21. How Many Reducers Are Created for a 1 GB File?

Sunday, August 11, 2024 11:07 AM

The number of reducers is not directly related to the file size but is determined by the job configuration. It is often set by the developer or based on the job's requirements.

22. What is combiner? How does it work and provide performance gain? Where did you use it

Sunday, August 11, 2024 11:08 AM

Combiner: An optional component that runs on the Map output before sending it to the reducer. It reduces the volume of data transferred across the network. It’s often used in word count jobs to sum the occurrences of words locally before sending them to the reducer.

23. What is partitioner? How does it work and provide performance gain? Where did you use it

Sunday, August 11, 2024 11:12 AM

Partitioner: Determines how the intermediate keys produced by the mapper are distributed to the reducers. It ensures that all keys that are processed by a single reducer are grouped together, improving performance. It is commonly used when you want to control the distribution of data across reducers based on custom logic.