Assignment – 1

Disclaimer:

I have neither given nor received unauthorized assistance on this work.

Signed:
Hardik Jain (ID: 1001954448) Date: 12/05/2021
Anuhya Patibanda (ID: 1001969235)

1. What are the key differences between Hadoop and spark and their respective advantage?

⇒ Spark is top level project focused on the processing data. Spark approaches in the method called parallel across a cluster method. Hadoop is the open source which uses the algorithm called MapReduce algorithm and is resilient to the failure.

No	Spark	Hadoop
1.	Spark is the fastest cluster computing technology which extends the MapReduce model to efficiency with more types of computation.	Hadoop reads and writes from disk which causes the slowdown in the process.
2.	Spark handles real time data efficiently and can process real time data from application like twitter.	Hadoop handles the batch processing efficiently. In the process of using Hadoop MapReduce developer can only process data in batch only.
3.	Spark has low latency in the process of computing data.	Hadoop is a high latency computing framework.
4.	Spark is costly than Hadoop because it uses lot of RAM which increases the use of cluster hence, it makes it costlier.	Hadoop is a cheaper than compared to spark. Because it doesn't use RAM.
5.	Spark decreases the count of Read /Write cycle to disk and store data in memory.	Hadoop directly reads and writes from disk.
6.	Spark keeps intermediate data in memory.	Hadoop offers an interface called interactive MapReduce Hadoop.
7.	Spark rely on the HDFS for lager data.	Hadoop quickly scales to the demand when data grows rapidly hence very reliable
8.	Spark enhances security with authentication by logging into events, therefore less security than Hadoop.	Hadoop performs multiple authentications and has access to control techniques, making it more secure.
9.	Spark is easy to program because it works with high level of operators.	In this process of using Hadoop the developer has to code everything as operation which makes it difficult.
10.	Using Spark is easier because it enables RDD while using high level operators.	Using Hadoop MapReduce makes it difficult for two reasons: first, it's a low-level application, and second, each operation requires hand-written code.

Assignment – 1

Hence their respective advantages are Spark is easier to use and faster. Furthermore, spark implements multi -threading in its process and has high level operators and on the other hand Hadoop is cheaper and doesn't occupy lot of RAM. Moreover, Hadoop handles fault tolerance more efficiently also its an open source.

2. Discuss how to recover a failure task in Hadoop and Spark respectively.

⇒ Spark

If any task or processing data gets crashed, that results in fault of a cluster. They are types of recovering the failed task with concept of using the RDD. In one process data is replicated and recovered and in the other process the data is received but buffered for replication.

a) In this case one process is in such a way that if any failure occurs, we can retrieve the data for the further use by restarting.
b) In the case two, data will not be replicated the one and only way to get the data back is from the source.
c) In the case three, block is executed and given values are read and written directly after that log are useful to recover the data which is saved in it.
d) Concept of metadata, in this process the data is recovered from the metadata blocks, when the failure occurs due to incomplete tasks.
e) In this case, it gets data recovered from the metadata directly.

Hadoop

Processing the failed task in the Hadoop MapReduce, it uses the additional computation to recover the data

a) In the case one, Hadoop starts an additional computation resource to get the failed task done which additionally saves the time.
b) In the case two, the intermediate data is sent to the reducer in the constant intervals using a mapper and this process is known as fetch.
c) In this case, the data is recovered during the execution of the process as checkpoint.
d) In the final case, the data is recovered using the reducer like the node failure.