QUESTION FOR BIG DATA PROCESSING USING AWS EMR-CLUSTER WITH HADOOP AND SPARK

Q1/ I am running a mapreduce job using EMR-Cluster but it’s taking long and the job doesn’t finish, fail or stop what can be the cause and how to find it?

A/

Q2/ I want to run a streaming job for a client where the data is streamed from many external producers (IoT) devices.

1. How do I configure this scenario using Amazon Kinesis and EMR-Cluster and push the processed data to redshift for BI or analytics purpose?
2. If the streaming fails and the client complain that they cannot get result of their real-time analytics how do I get to know if the problem is from the client side?

Q3/ Is the information provided in yarn logs sufficient to trace down the source of problem if this is related to the external configuration of the network or storage services?

Q4/ In case you want to run real-time and batch processing as this is required by the client is it possible to configure this with one EMR-Cluster or we should use two different clusters to handle the jobs separately?

Q5/ In the presentation and labs we have covered how to run MapReduce, Spark Application and HDFS using EMR-Cluster but which resource can I use to learn more about the configuration of other services such AWS Glue, Athena, Kinesis, S3?

Q6/ When is it so required to use a replication factor. And if it’s used and it requires more storage space which makes the solution more expensive is there another way to solve this problem in cost effective manner?

Q7/ In HDFS, when a metadata configuration is compromised is there a way to restore it or what else can be done to continue using the same cluster to allow clients to read and write to data node blocks?

Q9/ When is it appropriate to use safe mode and what are the benefits of this practice?

Q10/ What is the best way to solve a problem of corrupt block replicat and missing bloc?