Lecture 6: Setting up EMR and Spark Intro

Contents

Gittu George, March 6 2024

6.1. Objectives

- Learn how to set up your own EMR cluster.
- Discuss the various option available in EMR.
- Summarize AWS cloud services and compare them with other vendors.
- Explore the different ways you can interact with the EMR cluster and Spark.
- Understand Spark architecture and terminologies.
- Learn about the various Spark deployment modes.

6.2. Refresher questions

- What property is considered to be the heart and soul of Hadoop?
- What is the difference between Hadoop and Spark? - Why is this a wrong question?
- What is the importance of YARN in Hadoop?
- What is the difference between the local file system and HDFS?
- Can you name a cloud-based Hadoop distribution?
- What are the different types of nodes in an EMR cluster?
- What is the difference between a core node and a task node in an EMR cluster?

- What property is considered to be the heart and soul of Hadoop?
  - Data locality
- What is the difference between Hadoop and Spark?
  - Wrong question. You can compare map-reduce and spark.
- What is the importance of YARN in Hadoop?
  - YARN is the resource manager for Hadoop.
- What is the difference between the local file system and HDFS?
  - HDFS is a distributed file system, and the local file system is a single-node file system.
- Can you name a cloud-based Hadoop distribution?
  - AWS EMR, Azure HDInsight, Google Cloud Dataproc

- What are the different types of nodes in an EMR cluster?
  - Master node, core node, task node
- What is the difference between a core node and a task node in an EMR cluster?
  - Core nodes are used for processing, and task nodes are used for data processing.

6.3. Setting up EMR - DEMO

- After setting up an EMR cluster, do you know how to:
  - Terminate
  - Clone
  - Resize your cluster
  - use AWS CLI

I will be explaining the purpose of each component, and make sure you complete the below thoughts/discussion while I do so;

Thoughts/Discussion

- What is the version of the EMR we selected?
- What is the purpose of option steps? And why is it important?
- How many nodes and cores are in the instance we set up?
- Why do we go with the On-demand purchasing option?
- What are bootstrap actions?
- What is the name of the key pair that you selected while setting up the cluster?
- What happens behind the scenes when we set up an EMR cluster? (like how the servers are set up, what are the different components, etc.)

Warning

Don't check "termination protection" while creating the cluster.

6.4. What we learned today?

- Learned how to spin an EMR cluster with the elements we want from the Hadoop ecosystem.
- Various ways to connect to the EMR cluster.
- Various options available in EMR
- List the advantages of using Spark.
- Spark components and their functions.
- Difference between various spark deployment modes and when to use what.