Lab 9 - AWS EMR

Introduction:

Welcome to Lab 9! This simple exercise is designed to introduce you to AWS EMR

**General Instructions:**

1. **Watch the demo video before starting the experiment**
2. **Execute all 3 Tasks by referring to “AWS getting started with EMR” tutorial (link provided after EMR Workflow diagram below).**
3. Evaluation for the labs will be mixed, few labs will involve uploading screenshots and few might be auto-evaluated.
4. All screenshots must be uploaded to Edmodo before stipulated deadlines and as mentioned under the “deliverables” section for each task.
5. If a task has some questions, these will not be evaluated but are extremely important in understanding the concepts.
6. Try to solve the tasks by yourselves (by watching the demo video), all relevant information to complete the tasks have been provided. In case you are stuck and not able to solve the issue, feel free to send an email to [**pesu\_cc\_lab\_support@googlegroups.com**](mailto:pesu_cc_lab_support@googlegroups.com)

**Deliverables: 4 screenshots and 1 csv file** (9a-9d and 9e.csv)

Reading - 30 mins

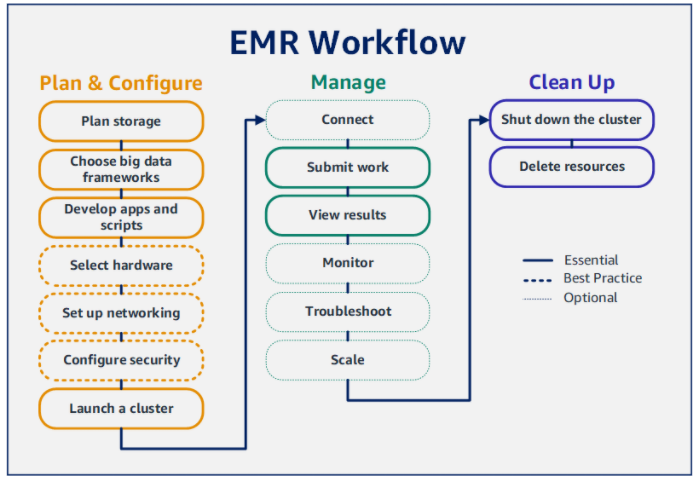
**What is AWS EMR?**

Amazon EMR is the industry-leading cloud big data platform for processing vast amounts of data using open source tools such as [Apache Spark](https://aws.amazon.com/emr/features/spark/), [Apache Hive](https://aws.amazon.com/emr/features/hive/), [Apache HBase](https://aws.amazon.com/emr/features/hbase/), [Apache Flink](https://aws.amazon.com/blogs/big-data/use-apache-flink-on-amazon-emr/), [Apache Hudi](https://aws.amazon.com/emr/features/hudi/), and [Presto](https://aws.amazon.com/emr/features/presto/). Amazon EMR makes it easy to set up, operate, and scale your big data environments by automating time-consuming tasks like provisioning capacity and tuning clusters. With EMR you can run petabyte-scale analysis at [less than half of the cost](https://pages.awscloud.com/Gated-IDC-The-Economic-Benefits-of-Migrating-Apache-Spark-and-Hadoop-to-Amazon-EMR.html) of traditional on-premises solutions and [over 3x faster](https://aws.amazon.com/blogs/big-data/amazon-emr-introduces-emr-runtime-for-apache-spark/) than standard Apache Spark. You can run workloads on Amazon EC2 instances, on Amazon Elastic Kubernetes Service (EKS) clusters, or on-premises using EMR on AWS Outposts.

Overview of AWS EMR : https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-overview.html

Read about Spark : https://spark.apache.org/

**EMR Workflow**



**LAB Reference**: Please use this detailed reference for executing your lab keeping in mind the configurations mentioned below. **You do not have to perform the optional steps in this tutorial.** [**https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs.html**](https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs.html)

**Watching the demo video will help you to complete the following 3 tasks quickly: (This is the last experiment that will require or use your AWS Credit)**

**TASK 1: Plan and Configure an Amazon EMR Cluster**

**Name of s3 bucket: [your-srn-bucket]**

**Upload both .csv and .py files into S3 bucket and take a screenshot with uploaded files required (9a)**

**Launching cluster**

**Cluster name : my-cluster-[yoursrn]**

**Use the following configurations in your cluster:**

**Release label: emr-6.2.0**

**Application: Spark**

**Instance type: m5.xlarge**

**Instance count: 3 (1 master and 2 workers)**

**Screenshot of running cluster (9b)**

**TASK2: Managing Amazon EMR Clusters**

**Screenshot of spark job completed (9c)**

**Screenshot of output folder in S3 (9d)**

**Download the output csv to your local machine and upload csv (and NOT the screenshot of the csv) to the drive ( 9e.csv)**

**NOTE: In case “Add step” task in Cluster Management fails, make sure you have added both .csv file and .py file in your S3 bucket and then repeat the “Add step” task. At this point, you can add a new step or clone the existing (failed) step but make sure all the configurations are correct for the spark application before you re-run the step.**

## **TASK3: Clean Up Amazon EMR Cluster Resources (Do not forget)**

**Go to EMR, select your cluster and click “Terminate”.**