Amazon EMR cluster provides a managed Hadoop framework that makes it fast, easy and cost effective to process the big data across dynamically scalable EC2 instances.

**STORAGE SYSTEM OF THE DATASET:**

Here I’m going to use S3 as my data store.

**Amazon Simple Storage Service S3:**

**Intro about S3:**

S3 stores the data in the form of objects. This enables to store, retrieve and analyze large data from multiple applications. These S3 objects are commonly referred as **buckets.**

**Some of the advantages of using S3:**

* Simple to use: easy to integrate with any platform.
* Enhanced security: Supports the transfer of data on SSL(Secure Socket Layer), automatic encryption of data upon uploading and configuration policies to manage permission granting. Effective monitoring system to track users accessing the data.
* Scalability: Easy to scale the storage.
* Integration: Easy to integrate with other AWS systems.

**My use case of S3 in this system:**

S3 is a distributed storage system and AWS’s equivalent to the hadoop’s HDFS.

Here I choose to use s3 as my data source because:

* The data is coming from a distributed file that can accessed by every node on the Spark Cluster.
* The spark application assumes that the data is coming from multiple sources rather than a single local hard disk because that will not scale. For example, we can’t store a terabyte of dataset in a single hard disk. So it is necessary to distribute the data.

Therefore, by saving my dataset into my S3 each spark node deployed in the EMR cluster will be able to read from the data source(S3).

**How to setup the S3 Service:**

* Click on Services in Aws console and select S3 from the Storage options
* Click on “Create bucket” and then provide the Bucket Name(car-price-bucket), Region to deploy the bucket and click on “Create”
* Select the newly created bucket name and click on “Upload” and select the dataset(in my case it’s the car price dataset)
* Once uploaded, the csv file appears in the bucket.

**Connecting spark with S3:**

* We must replace the input file path to S3 path so that each spark clusters reads the data from S3.

Path: “s3n://bucket-name/file-name.csv”

Eg: “s3n://car-price-bucket/car-price-dataset.csv”

**Uploading the python script to S3 bucket:**

* Click on the bucket name and then click on upload
* Select the python script on the file manager.

**SETTING UP CLUSTER ENVIRONMENT:**

**STEPS TO SETUP THE EMR SYSTEM:**

* Click “Services” and select “EMR”
* On the EMR page select **create cluster**
* Provide the name of the cluster.
* The logging will be automatically stored in the s3 bucket folder
* In the launch mode we choose the “**cluster**” mode
* Under the software configuration, Select “**Spark**” in the options of Applications
* The instance type determines the amazon ec2 instance type that the amazon emr initializes for the instance which runs in our cluster. We let it to the default value set by AWS.
* Under security and access, select one of the EC2 key pair.
* Click on “Create Cluster” to start the process of creating the cluster.

After setting up the EMR system, we need to modify the Security and access inorder to access the system from our local machine through ssh.

* Click on the **Security groups for Master.**
* In the Security group page, Select the one with “**Master group**” in the description
* Select **Inboud type** tab and click on  **Edit**
* Click on “**Add Rule**” and select “SSH” in type and select “Anywhere” in the Source column. This setting is only for testing purposes of this project.
* Click **Save** to save the new rule.

**CONNECTION AND RETRIEVAL OF DATA FROM SPARK MASTER:**

**Steps to login to our Spark Master machine:**

* Start Putty and click on **Load** then select the **key-pair** file named “pedro-key.ppk”
* Click on **Save private key**
* Copy the Host Name provided by the AWS and paste the Host Name under hostname box in putty.
* Select **SSH** as Connection type
* On the left panel of putty, go to Connection->SSH->Auth and the click on Browse and select the ppk file generated in the previous step.
* Click **Open.** This will open the console of the master machine of the EMR Cluster in our local machine.

**Steps to Run our file in EMR:**

* In the EMR Master console type the command “aws s3 cp s3://bucket-name/file-name.csv” to copy the files of the S3 bucket into our Spark EMR cluster
* Now run spark command “spark-submit python-file-name.py” to run our spark job.
* We get all the job output when we run the command in the previous step.

Through these steps, we can successfully execute the spark job on Amazon EMR cluster.