**Amazon Elastic MapReduce**

**Amazon Elastic MapReduce (EMR)** is a web service that provides a managed frameworkto run data processing frameworks such as Apache Hadoop, Apache Spark, and Presto in an easy, cost-effective, and secure manner.

It is used for data analysis, web indexing, data warehousing, financial analysis, scientific simulation, etc.

**How to Set Up Amazon EMR?**



Follow these steps to set up Amazon EMR:

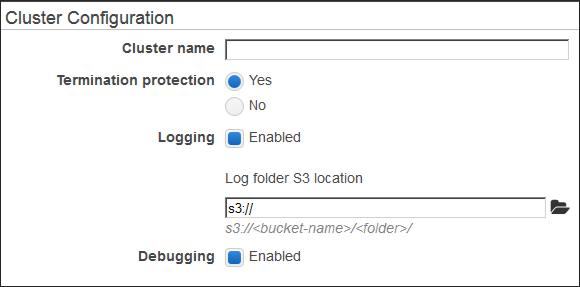
**Step 1**: Sign in to AWS account and select Amazon EMR on management console.

**Step 2**: Create Amazon S3 bucket for cluster logs & output data. (Procedure is explainedin detail in Amazon S3 section)

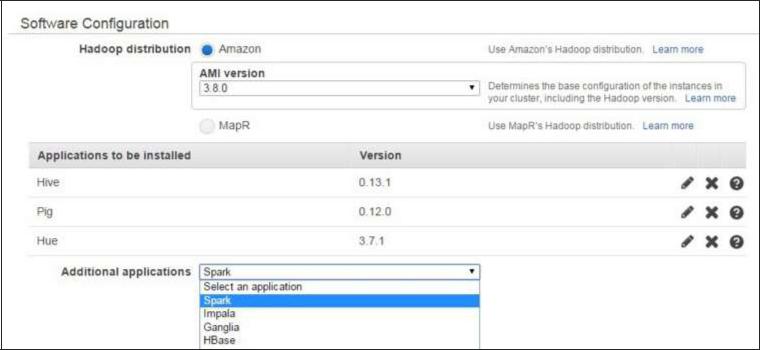
**Step 3**: Launch Amazon EMR cluster.

Following are the steps to create cluster and launch it to EMR.

* Use this link to open Amazon EMR console: <https://console.aws.amazon.com/elasticmapreduce/home>
* Select create cluster and provide the required details on Cluster Configuration page.

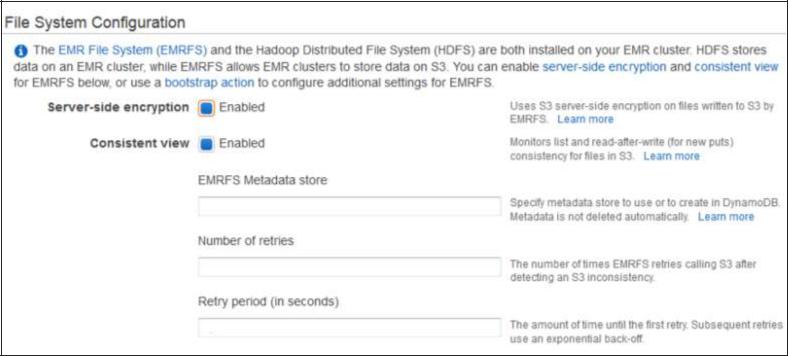


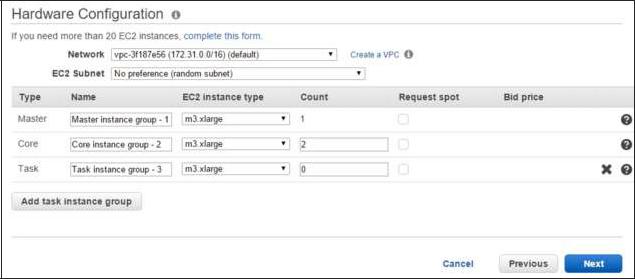
* Leave the Tags section options as default and proceed.
* On the Software configuration section, level the options as default.



* On the File System Configuration section, leave the options for EMRFS as set by

default. EMRFS is an implementation of HDFS, it allows Amazon EMR clusters to store data on Amazon S3.

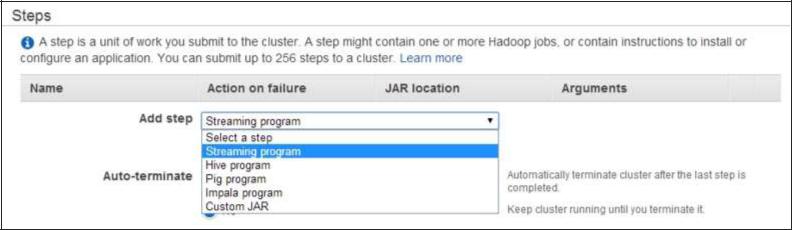


* On the Hardware Configuration section, select m3.xlarge in EC2 instance type field and leave other settings as default. Click the Next button.

* On the Security and Access section, for EC2 key pair, select the pair from the list in EC2 key pair field and leave the other settings as default.
* On Bootstrap Actions section, leave the fields as set by default and click the Add button. Bootstrap actions are scripts that are executed during the setup before Hadoop starts on every cluster node.
* On the Steps section, leave the settings as default and proceed.
* Click the Create Cluster button and the Cluster Details page opens. This is where we should run the Hive script as a cluster step and use the Hue web interface to query the data.

**Step 4**: Run the Hive script using the following steps.

* Open the Amazon EMR console and select the desired cluster.
* Move to the Steps section and expand it. Then click the Add step button.
* The Add Step dialog box opens. Fill the required fields, then click the Add button.



* To view the output of Hive script, use the following steps:

1. Open the Amazon S3 console and select S3 bucket used for the output data.
2. Select the output folder.
3. The query writes the results into a separate folder. Select **os\_requests**.
4. The output is stored in a text file. This file can be downloaded.

**Benefits of Amazon EMR**



Following are the benefits of Amazon EMR:

* **Easy to use**: Amazon EMR is easy to use, i.e. it is easy to set up cluster, Hadoopconfiguration, node provisioning, etc.
* **Reliable**: It is reliable in the sense that it retries failed tasks and automaticallyreplaces poorly performing instances.
* **Elastic**: Amazon EMR allows to compute large amount of instances to process dataat any scale. It easily increases or decreases the number of instances.
* **Secure**: It automatically configures Amazon EC2 firewall settings, controls networkaccess to instances, launch clusters in an Amazon VPC, etc.
* **Flexible**: It allows complete control over the clusters and root access to everyinstance. It also allows installation of additional applications and customizes your cluster as per requirement.
* **Cost-efficient**: Its pricing is easy to estimate. It charges hourly for every instanceused.