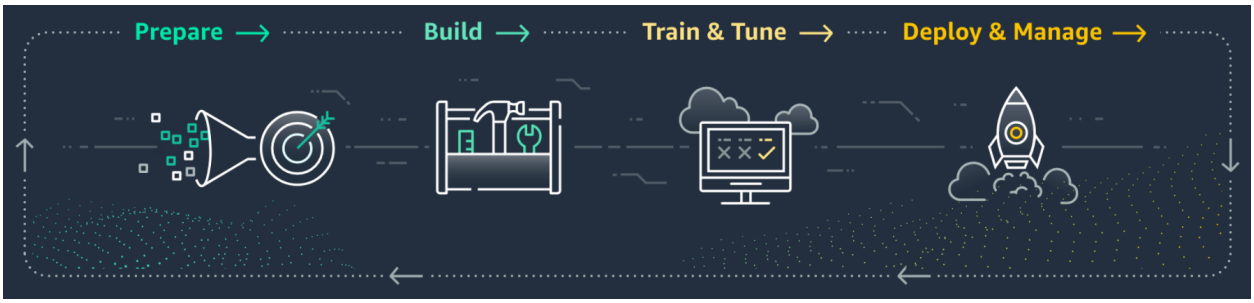
Setup AWS SageMaker Notebook with EMR Cluster

Introduction of AWS SageMaker

AWS SageMaker is a fully managed service that provides every developer and data scientist with the ability to prepare build, train, and deploy machine learning (ML) models quickly.



Introduction of AWS EMR

Amazon EMR is a managed cluster platform that simplifies running big data frameworks, such as Apache Hadoop and Apache Spark, on AWS to process and analyze vast amounts of data. By using these frameworks and related open-source projects, such as Apache Hive and Apache Pig, you can process data for analytics purposes and business intelligence workloads. Additionally, you can use Amazon EMR to transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB.

SageMaker notebooks manage Spark in EMR

One of the important parts of Amazon SageMaker is the powerful Jupyter notebook interface, which can be used to build models. You can enhance the Amazon SageMaker capabilities by connecting the notebook instance to an Apache Spark cluster running on Amazon EMR.

To facilitate a connection between an Amazon SageMaker notebook and a Spark EMR cluster, you will need to use Livy. Livy is an open source REST interface for interacting with Spark clusters from anywhere without the need for a Spark client.

Steps of build Amazon SageMaker notebooks backed by Spark in Amazon EMR.

1. Create AWS developer account, create development key "EC2-Tutorial"(my key)
2. Create AWS EMR cluster with following settings:
    a. Software configuration:pick release "erm-5.33.0", toggle "Spark 2.4.7" and "Livy 0.7.0" only.
    b. Hardware configuration: 3 nodes, 1 master, 2 slaves, m5.xlarge
    c. Security. Select EC2 key pair "EC2-Tutorial"(my key). Pick master security group ElasticMapReduce-Master.

    Then create cluster.
3. After the cluster is starting, click the cluster, click Hardware, click master node, copy down the private IP address, for my instance is 172.31.6.XXX
4. Go to aws→Network & Security→ Security group, edit ElasticMapReduce-master group

a. Open Custom TCP port 8998

5. Now create SageMake notebook, go to AWS→SageMaker→Create notebook, make sure your notebook is in the same VPC and region with EMR cluster.
6. Edit "lifecycle configurations"

```
set -e
#XXX is a fake value
EMR_MASTER_IP=172.31.6.XXX


cd /home/ec2-user/.sparkmagic

echo "Fetching Sparkmagic example config from GitHub..."
wget https://raw.githubusercontent.com/jupyter-incubator/sparkmagic/master/sparkmagic/example_config.json

echo "Replacing EMR master node IP in Sparkmagic config..."
sed -i -- "s/localhost/$EMR_MASTER_IP/g" example_config.json
mv example_config.json config.json

echo "Sending a sample request to Livy.."
curl "$EMR_MASTER_IP:8998/sessions"
```