Increase Your Data Engineering Productivity using Amazon EMR Notebooks

Vignesh Rajamani

Sr. Product Mgr. Amazon EMR

Date: 01/22/2019

Agenda

- What is EMR Notebooks?
- Why should I use it?
- How does it work?→ Demo
- How do I access it?
- Q&A

What are EMR Notebooks?

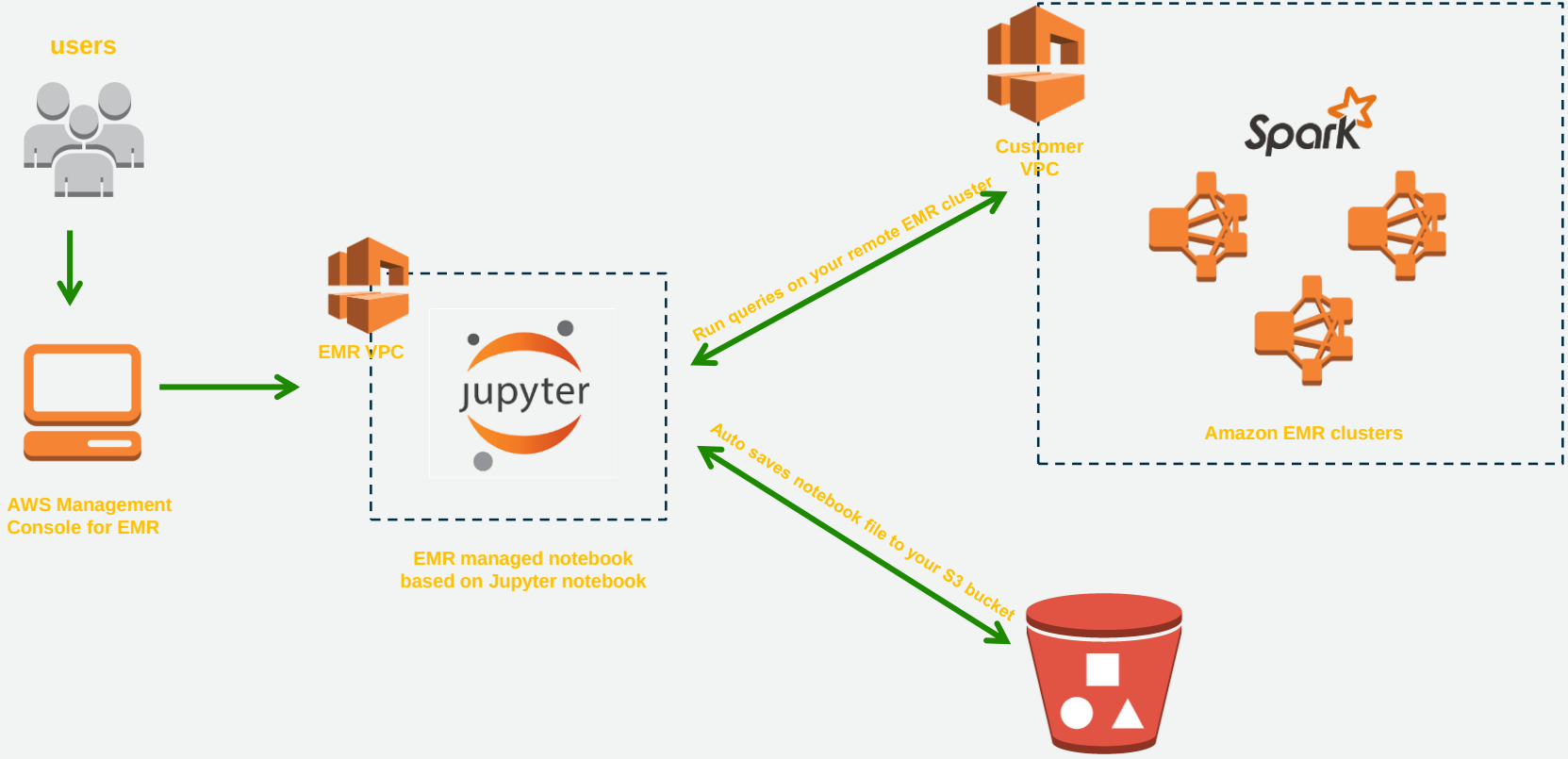
EMR Notebooks is a managed notebook environment, based on Jupyter Notebook, that allows data scientists and developers to experiment and build applications with Apache Spark, perform interactive analysis using EMR cluster, prepare and visualize data and collaborate with peers.

Amazon EMR

aws

EMR Notebooks

A managed analytics environment based on Jupyter Notebooks



Why should I use it?

As a Data Scientist or Developer

- Build applications, prepare and visualize data, collaborate with peers, and run interactive analysis through PySpark, Spark SQL, Spark R and Scala
- Create multiple notebooks instantly from the console, attach them to a minimum 1-node EMR cluster and immediately start experimenting with Apache Spark.
- Monitor the progress of your job from within the notebook with the integrated Spark monitor.
- Visualize your results in rich graphical plots using the pre-installed open-source libraries from Anaconda.
- Detach your notebook from clusters and re-attach to a different cluster suitable for your workload
- Durably persist your work to S3 and share with others. Easily retrieve saved work from the console.

Why should I use it?

As an IT Administrator

- Easily setup a multi-tenant cluster for your data-scientists and developers and make most use of your EMR cluster
- No need to deploy, maintain and upgrade software or notebook instances
- Provide secure access to notebooks without providing multi-user access to the master node.
- Safely terminate your cluster without fear of losing notebooks
- Set up fine grained access control to notebooks and clusters through IAM policies
- Track and audit your notebook users by enforcing user impersonation

What programming languages and libraries can I use?

- PySpark, SparkR, SparkSQL, Spark (Scala), and Python kernels
- Libraries found in the open-source Anaconda repositories
- Custom libraries on your EMR cluster with bootstrap or AMIs
- Attached to EMR clusters running EMR release 5.18.0 or later. With Spark and Livy installed

What is the cost of using EMR Notebooks?

- No additional charge to Amazon EMR customers.
- No. of notebooks that can be attached to cluster depend on size of the master node.
- Charged as usual for the attached EMR clusters in your account
- Find out more about the pricing for your cluster: https://aws.amazon.com/emr/pricing/

Region: US East (Ohio)

	Amazon EC2 Price	Amazon EMR Price
General Purpose - Current Generation		
m5.xlarge	$0.192 per Hour	$0.048 per Hour
m5.2xlarge	$0.384 per Hour	$0.096 per Hour
m5.4xlarge	$0.768 per Hour	$0.192 per Hour
m5.12xlarge	$2.304 per Hour	$0.270 per Hour
m5.24xlarge	$4.608 per Hour	$0.270 per Hour
m5a.xlarge	$0.172 per Hour	$0.043 per Hour
m5a.2xlarge	$0.344 per Hour	$0.086 per Hour

When should I use EMR Notebooks vs Sagemaker?

- Use Sagemaker for:
    - Machine learning model building, tuning, and management
    - General purpose ML outside of EMR
- Use EMR Notebooks for:
    - Spark developers
    - Big data applications on EMR

Now we Demo

aws

How do I get started with EMR Notebooks?

- Open the EMR console and choose **Notebooks** in the navigation pane
- Choose **Create Notebook**, enter a name for your notebook
- Choose an EMR cluster or instantly create a new one
- Provide a service role for the notebook to use
- Choose an S3 bucket where you want to save your notebook in ipynb file format
- Once **Ready**, choose **Open** to start the notebook editor

Q&A