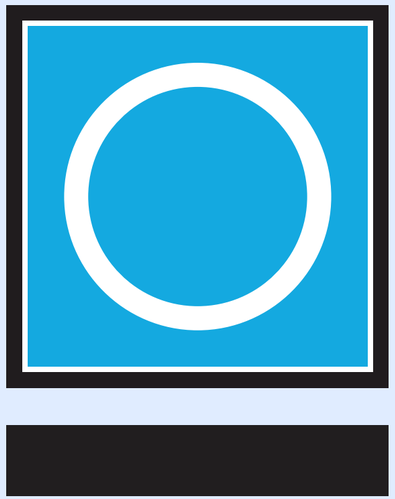
Beginner From Zero To Hero

AWS EMR FULL COURSE



Johnny Chivers

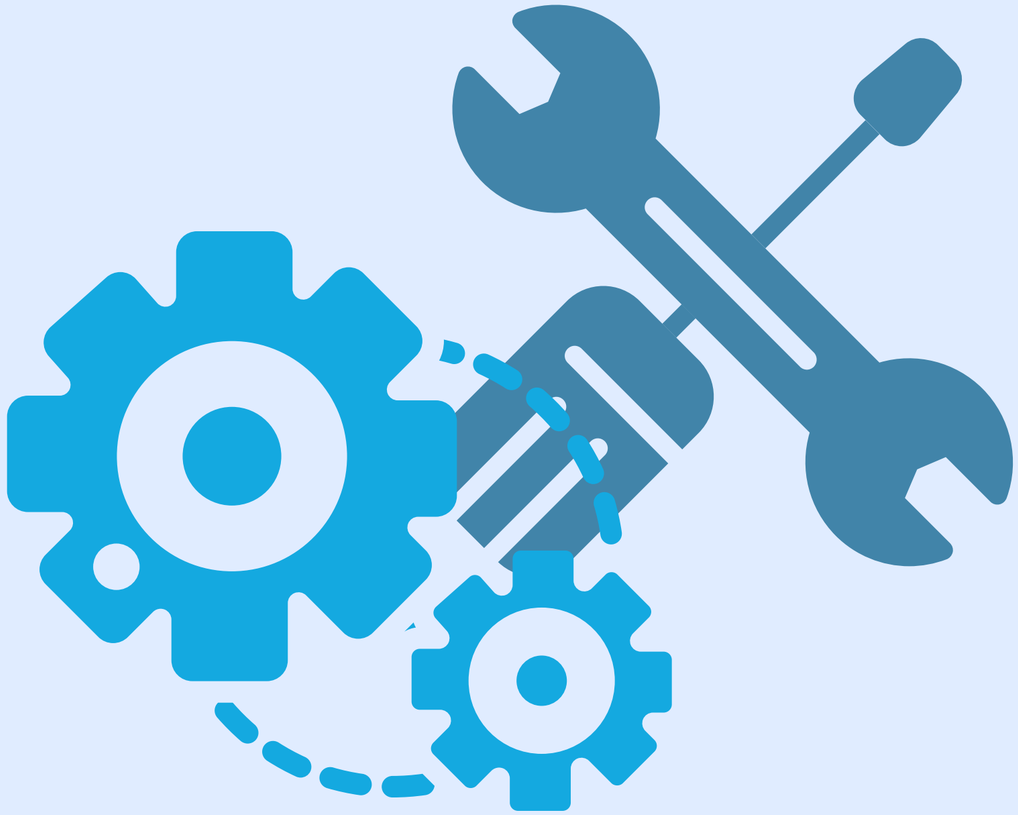
Clusters, Spark, Notebooks and more

TABLE OF CONTENTS

Presentation Outline

- Set Up work
- What is EMR?
- Spark ETL
- Apache Hive
- Apache PIG
- AWS Step Functions
- EMR Autoscaling

JOHNNYCHIVERS.CO.UK



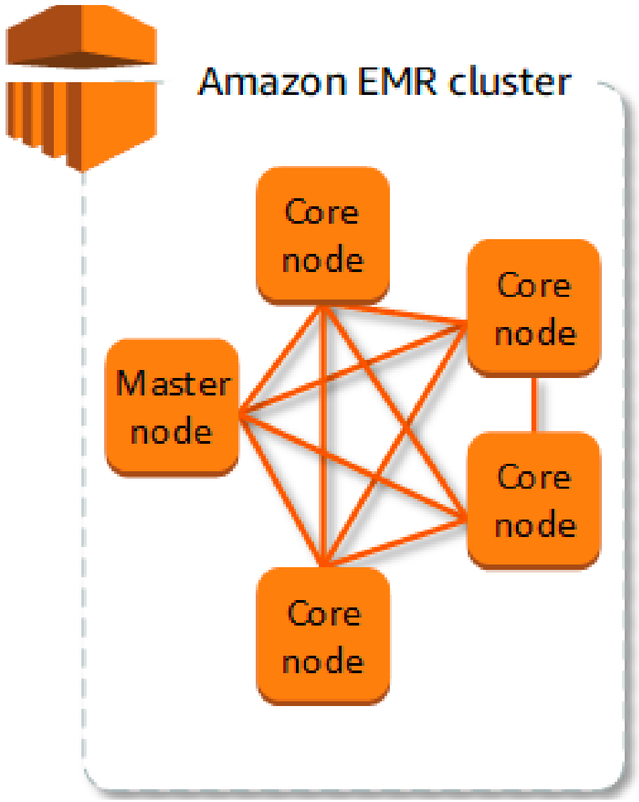
SET UP WORK

Just A Couple Of Things

JOHNNYCHIVERS.CO.UK

What is AWS EMR?

Managed cluster platform that simplifies running big data frameworks



Master node: A node that manages the cluster by running software components to coordinate the distribution of data and tasks among other nodes

Core node: A node with software components that run tasks and store data in the Hadoop Distributed File System (HDFS) on your cluster. Multi-node clusters have at least one core node.

Task node: A node with software components that only runs tasks and does not store data in HDFS. Task nodes are optional.

https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-overview.html

What is AWS EMR? continued...

Data Processing Frameworks

- Engine used to process and analyze data
- Different frameworks are available for different kinds of processing needs
- The main processing frameworks available for Amazon EMR are Hadoop MapReduce and Spark

Cluster Resource Management

- The resource management layer is responsible for managing cluster resources and scheduling the jobs for processing data.
- By default, Amazon EMR uses YARN (Yet Another Resource Negotiator).

Storage

- Hadoop Distributed File System (HDFS) is a distributed, scalable file system for Hadoop.
- Using the EMR File System (EMRFS), Amazon EMR extends Hadoop to add the ability to directly access data stored in Amazon S3 as if it were a file system like HDFS.
- The local file system refers to a locally connected disk.

Spark ETL

What is Spark?

Apache Spark™ is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters.

Faster Processing

Spark contains Resilient Distributed Dataset (RDD) which saves time in reading and writing operations, allowing it to run almost ten to one hundred times faster than Hadoop.

In Memory Computing

Spark stores the data in the RAM of servers which allows quick access and in turn accelerates the speed of analytics.

Flexibility

Apache Spark supports multiple languages and allows the developers to write applications in Java, Scala, R, or Python.

https://catalog.us-east-1.prod.workshops.aws/workshops/2300137e-f2ac-4eb9-a4ac-3d25026b235f/en-US/lab-0-getting-started/on-my-own/1-launch-resources

Hive

What is Hive?

The Apache Hive ™ data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL. Structure can be projected onto data already in storage. A command line tool and JDBC driver are provided to connect users to Hive.

SQL Like Interface

Hive provides the necessary SQL abstraction to integrate SQL-like queries (HiveQL) into the underlying Java without the need to implement queries in the low-level Java API.

Storage

Different storage types such as plain text, RCFile, HBase, ORC, and others.

https://catalog.us-east-1.prod.workshops.aws/workshops/2300137e-f2ac-4eb9-a4ac-3d25026b235f/en-US/lab-0-getting-started/on-my-own/1-launch-resources

PIG

What is PIG?

Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

How Does It Work?

It is an abstraction of Map Reduce which integrates with the lower level java api which means parallel processing is easily achieved.

Storage

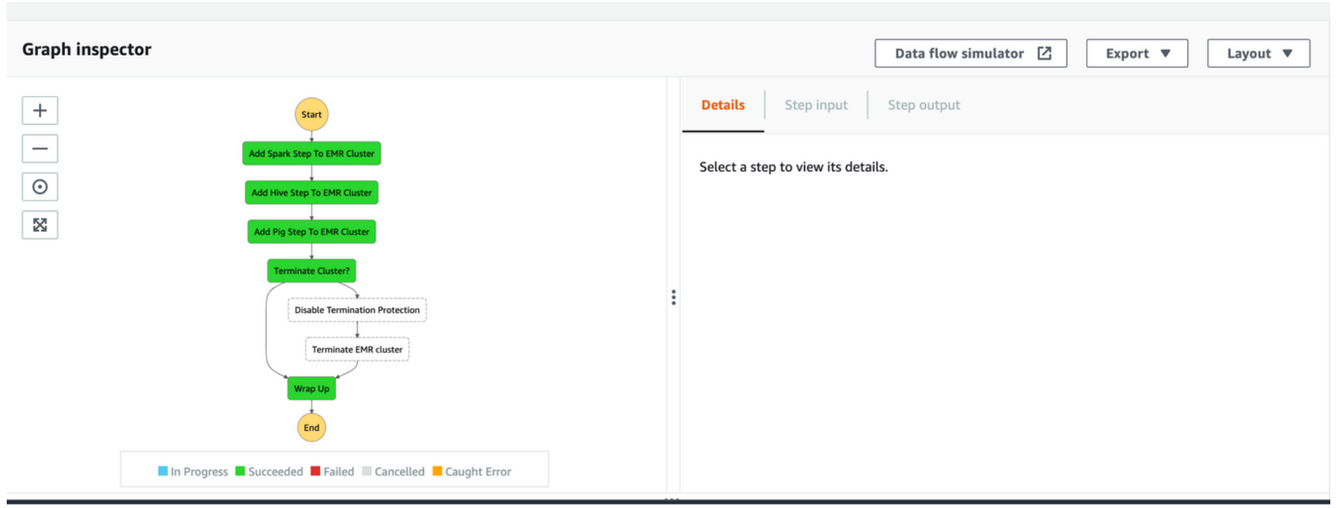
Different storage types such as plain text, RCFile, HBase, ORC, and others.

https://catalog.us-east-1.prod.workshops.aws/workshops/2300137e-f2ac-4eb9-a4ac-3d25026b235f/en-US/lab-0-getting-started/on-my-own/1-launch-resources

AWS Step Functions

What Are Step Functions?

AWS Step Functions is a low-code, visual workflow service that developers use to build distributed applications, automate IT and business processes, and build data and machine learning pipelines using AWS services. Workflows manage failures, retries, parallelization, service integrations, and observability so developers can focus on higher-value business logic.



https://catalog.us-east-1.prod.workshops.aws/workshops/2300137e-f2ac-4eb9-a4ac-3d25026b235f/en-US/lab-0-getting-started/on-my-own/1-launch-resources

EMR Autoscaling

What is Auto Scaling?

Autoscaling is a cloud computing feature that enables organizations to scale cloud services such as server capacities or virtual machines up or down automatically, based on defined situations such as traffic ir utilization levels.

How Does It Work?

Autoscaling policies are added to an EMR cluster which define how nodes should be added or removed. There options in terms of available RAM, disc, app running, apps pending etc.

https://catalog.us-east-1.prod.workshops.aws/workshops/2300137e-f2ac-4eb9-a4ac-3d25026b235f/en-US/lab-0-getting-started/on-my-own/1-launch-resources