
# Big Data on Amazon AWS

Data Processing – [Day 2]

CARLOS BARBOSA

Head of Content & Instructor

# Amazon Glue

AWS Glue is a serverless data integration service that makes it easy to discover, prepare, and combine data for analytics, machine learning, and application development.

# Amazon Glue

AWS Glue is a serverless data integration service that makes it easy to discover, prepare, and combine data for analytics, machine learning, and application development.

**Discovery**

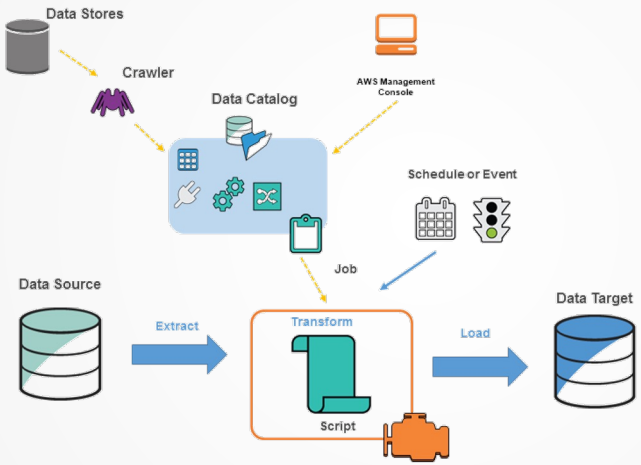
- Glue Data Catalog
- Metadata Store
- All Datasets
- Serverless

**Automatic Schema Discovery**

- AWS Glue Crawler
- Source or Target
- Schema for your Data
- Creates metadata in Glue Data Catalog
- Serverless

**Manage and Enforce schemas for data streams**

- AWS Glue Schema Registry
- Apache Avro Schemas
- Amazon MSK
- Amazon Kinesis Data Streams
- Apache Flink
- AWS Lambda
- Serveless




**Transformation**

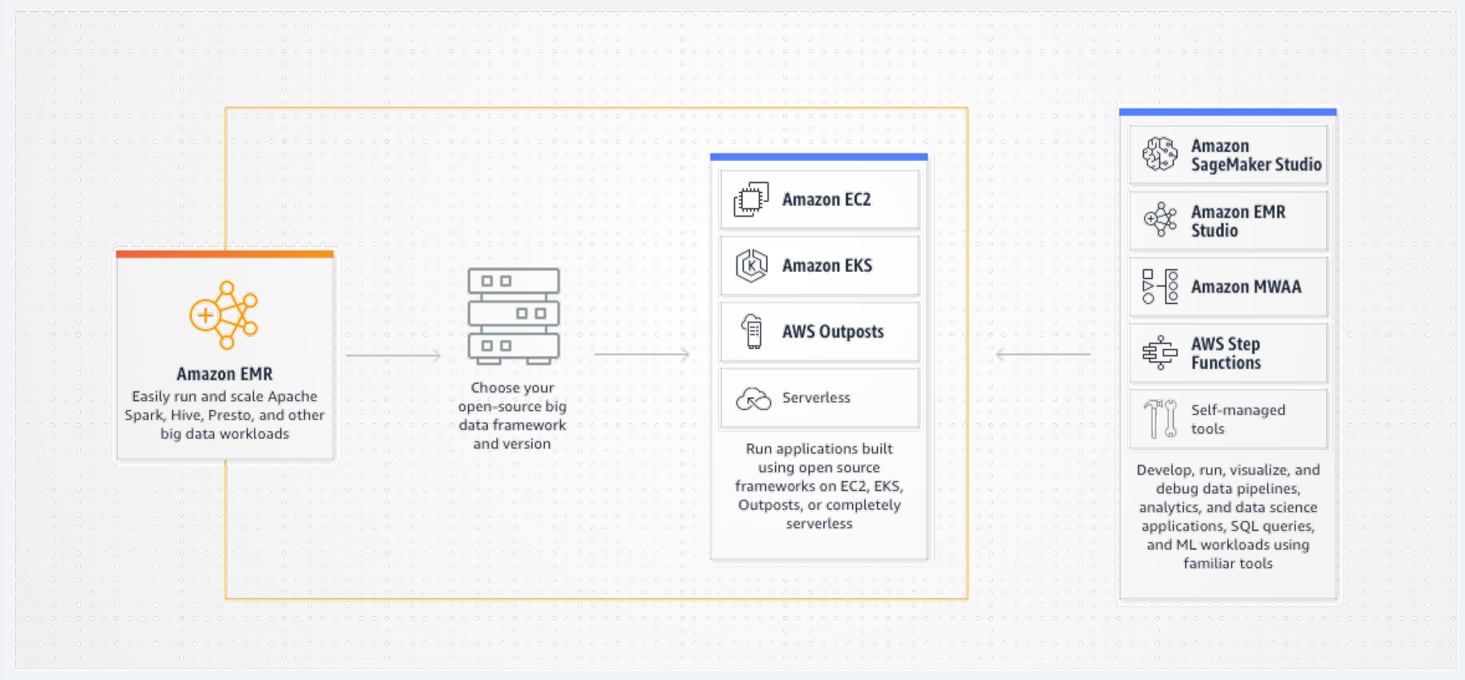
- AWS Glue Studio
- Highly Scalable ETL Jobs
- Apache Spark
- Scala or Python
- Serverless

**Replication**

- AWS Glue Elastic Views
- Data Stores and Materialize Views
- Supports DynamoDB, RedShift, OpenSearch, Amazon S3 & RDS.

# Amazon Elastic Map Reduce

Amazon EMR is a cloud big data platform for running large-scale distributed data processing jobs, interactive SQL queries, and machine learning (ML) applications using open-source analytics frameworks such as Apache Spark, Apache Hive, and Presto.



# Amazon Elastic Map Reduce

AWS Glue is a serverless data integration service that makes it easy to discover, prepare, and combine data for analytics, machine learning, and application development.

**Big Data as a Service**

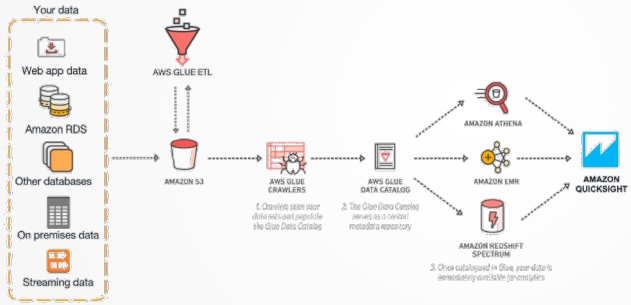
- Provision clusters in minutes
- Easily scale resources
- Low cost
- EC2 Spot Integration
- Amazon S3 Integration

**Big Data Tools**

- Apache Spark
- Apache Hive
- Apache HBASE
- Apache Flink
- Apache Hudi
- Apache Oozie

**Data exploration**

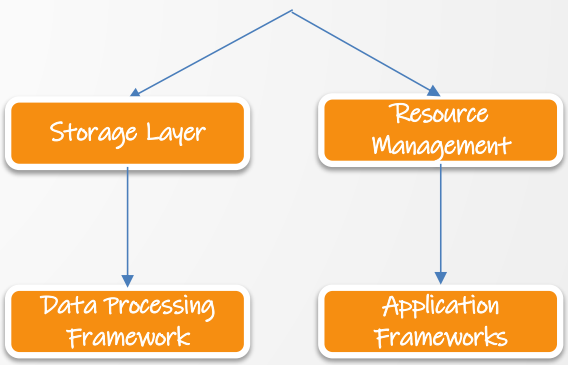
- Apache Livy
- Apache Jupyter Notebook
- Apache Zeppelin
- Apache Spark trought EMR Notebooks




**Components**

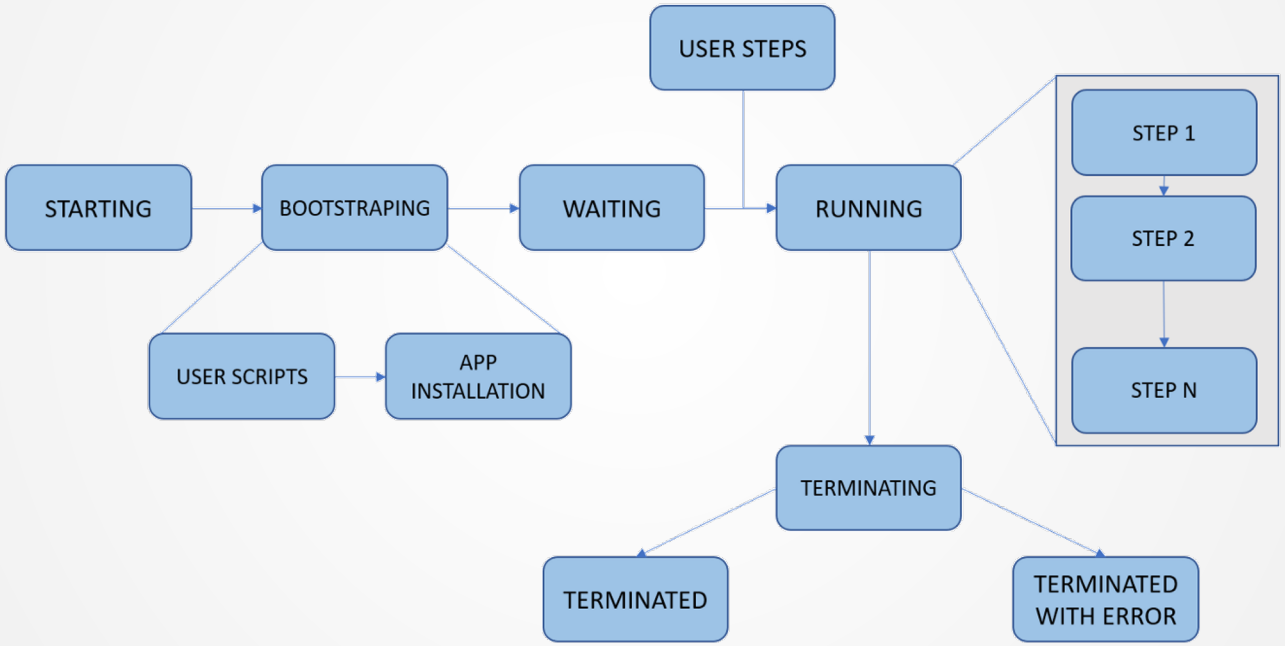
- Cluster
- Master Node
- Core Node
- Task Node

**EMR Architecture**



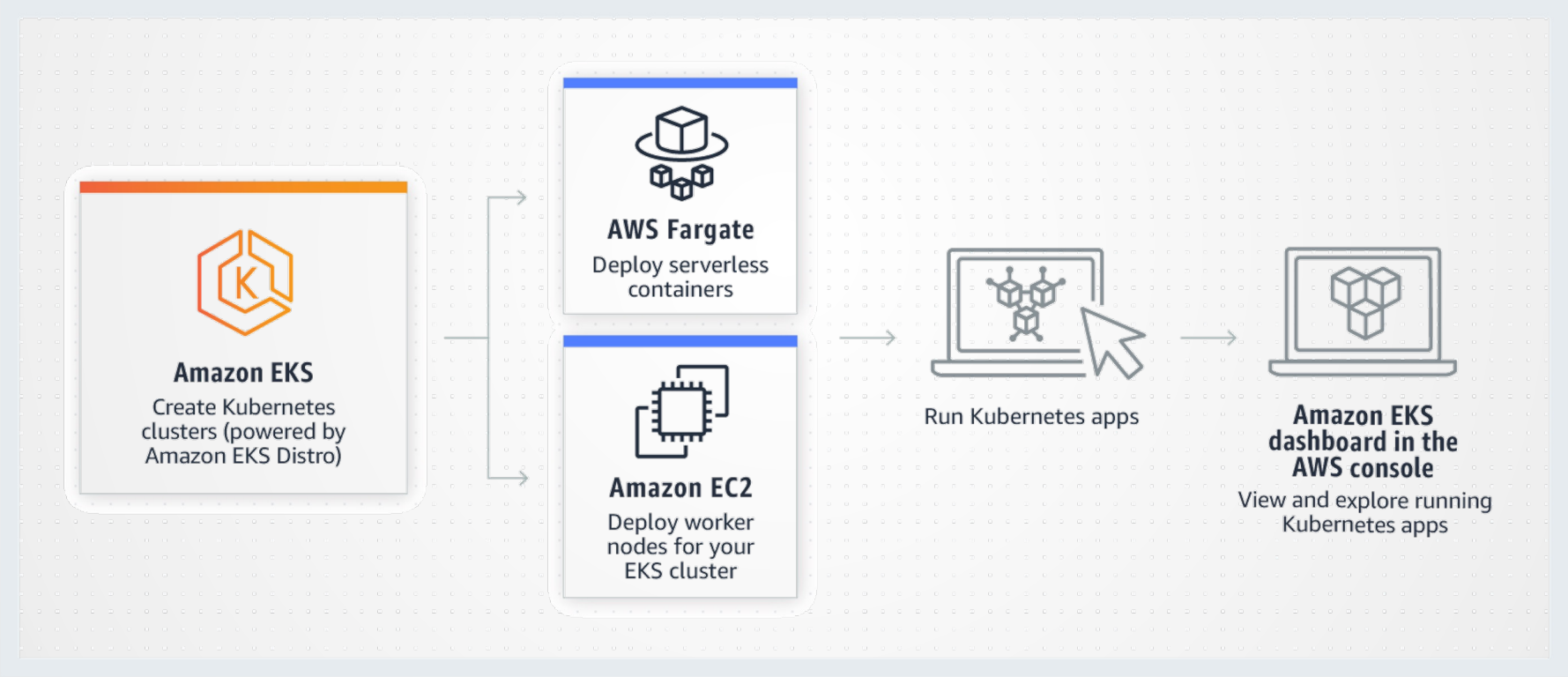
# Amazon Elastic Map Reduce

Amazon EMR is a cloud big data platform for running large-scale distributed data processing jobs, interactive SQL queries, and machine learning (ML) applications using open-source analytics frameworks such as Apache Spark, Apache Hive, and Presto.



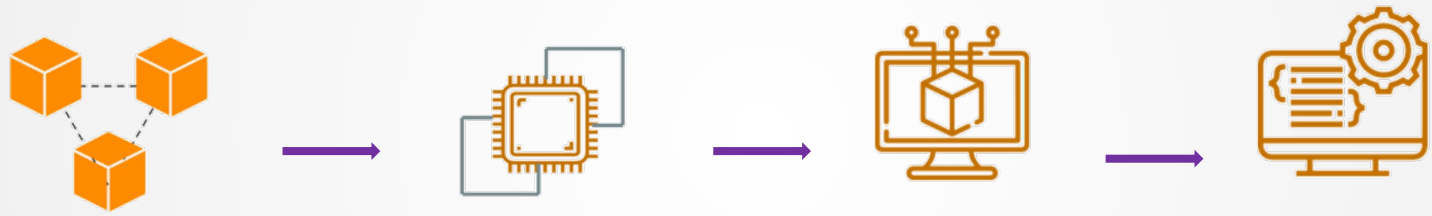
# Elastic Kubernetes Service

Amazon Elastic Kubernetes Service (Amazon EKS) is a managed container service to run and scale Kubernetes applications in the cloud or on-premises.



# Elastic Kubernetes Service

Amazon Elastic Kubernetes Service (Amazon EKS) is a managed container service to run and scale Kubernetes applications in the cloud or on-premises.



Provision an Amazon EKS cluster

Deploy worker nodes

Connect to EKS

Run Kubernetes apps on EKS cluster

# Elastic Kubernetes Service

Amazon Elastic Kubernetes Service (Amazon EKS) is a managed container service to run and scale Kubernetes applications in the cloud or on-premises.

**Managed Control Pane**

- Scalable and highly-available
- Across multiple AWS Availability Zones (AZs)
- Service Integrations
- Open-Source Compatibility

**Networking and Security**

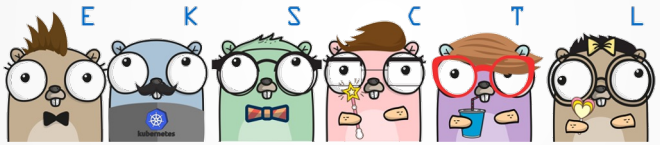
- VPC Native Networking
- IAM for Service Accounts
- Support for IPV6
- AWS Iam Authenticator
- IAM for Service Accounts

**Load Balancing**

- Network Load Balancer
- Ingress Controller

**Elastic Block Store**

- Storage SSD/HD
- Storage as a Service
- Native Integration with EKS
- Auto Scaling Disk




**Logging**

- AWS Cloudtrail
- AWS Cloudwatch

**Open-Source Compatibility**

- Spark Operator [Google]
- Strimzi Operator [Kafka]
- Minio Storage S3
- Airflow
- Trino
- Datahub
- ClickHouse

**Eksctl**

- Command Line Interface
- Windows / MacOS / Linux
- Cloudformation
- Golang
- Yaml

# Hands on (Day 2)

Use case using AWS EKS, Amazon EMR, AWS Glue.