Amazon Elastic Map Reduce: the concepts

What to expect

- Amazon EMR
- Amazon S3 as HDFS
- Core node and Task nodes
- Elastic clusters
- Beyond Map Reduce: Spark
- Q&A

What Hadoop is good at

Semi-structured/unstructured data

Disparate Data Sets

ETL at scale

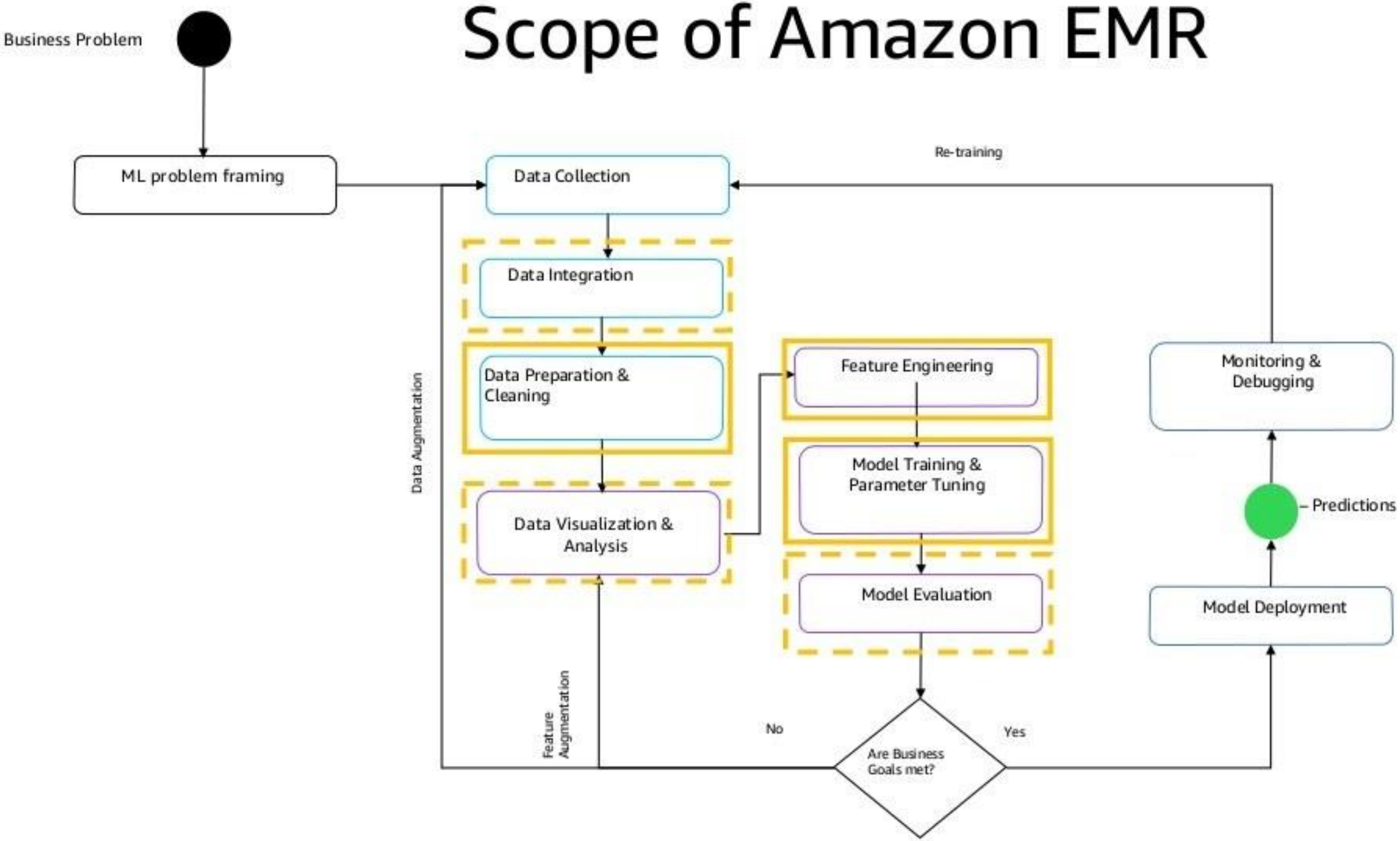
Batch Analytics

Log Processing & Aggregation

Amazon EMR – Hadoop, Spark, Presto in the Cloud

- Managed platform
- Launch a cluster in minutes
- Leverage the elasticity of the cloud
- Baked in security features
- Pay by the hour and save with Spot
- Flexibility to customize

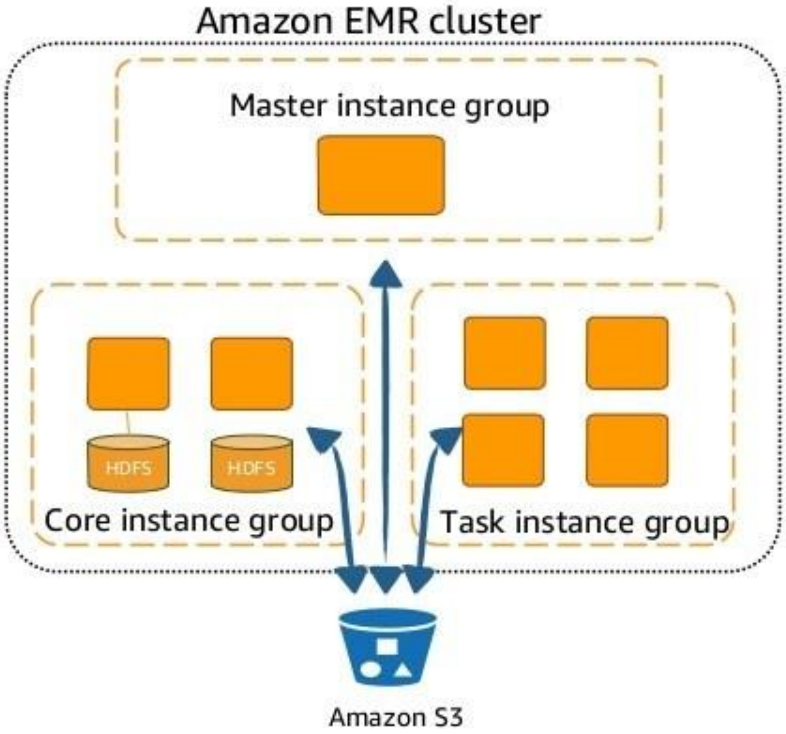
Scope of Amazon EMR



Amazon S3 as HDFS

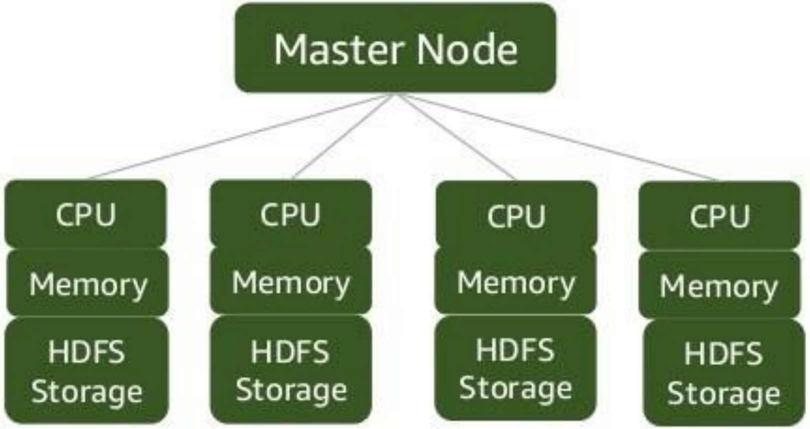
Amazon S3 as HDFS

- Use Amazon S3 as your permanent data store
- HDFS for temporary storage data between jobs
- No additional step to copy data to HDFS



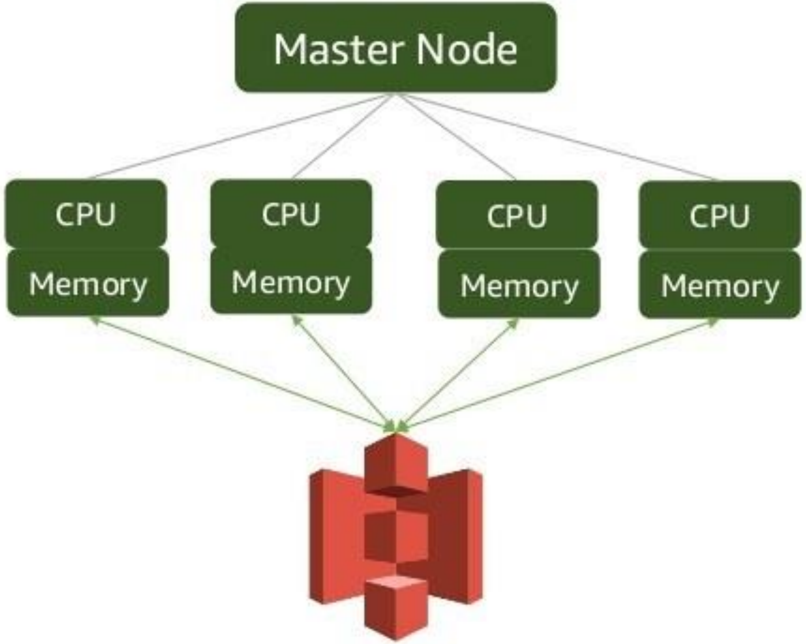
Understanding Decoupled Storage & Compute

Old Clustering / Localized Model



HDFS = 3X Replication
500 TB Dataset = 1.5 PB cluster with replication

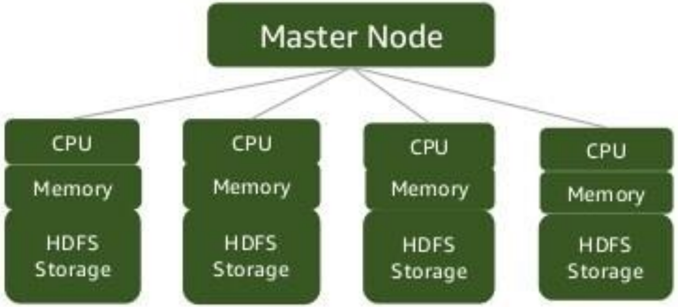
Amazon EMR Decoupled Model



S3 as Streaming HDFS via EMRFS

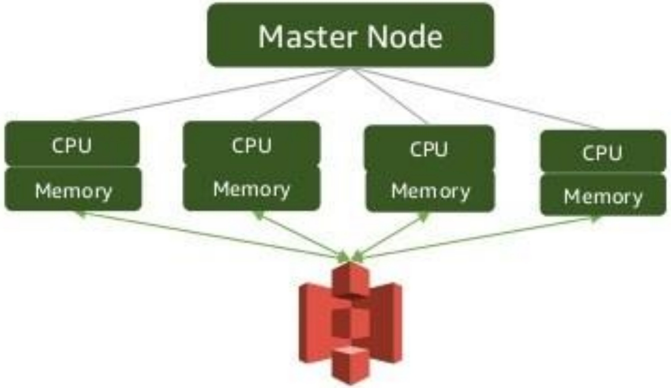
Understanding Decoupled Storage & Compute

Hadoop on EC2



- 3X HDFS Replication of Data across Nodes
- Protected against data loss from node failure
- Single AZ replication
- Paying for storage units via EC2 ephemeral drives or EBS volumes
- Data stored locally so cluster = 24x7 by default

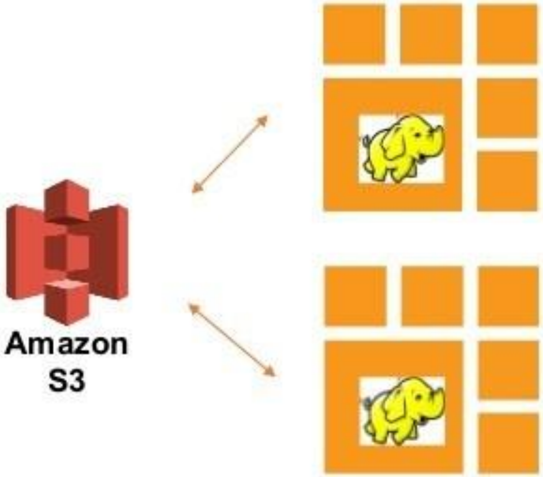
EMR



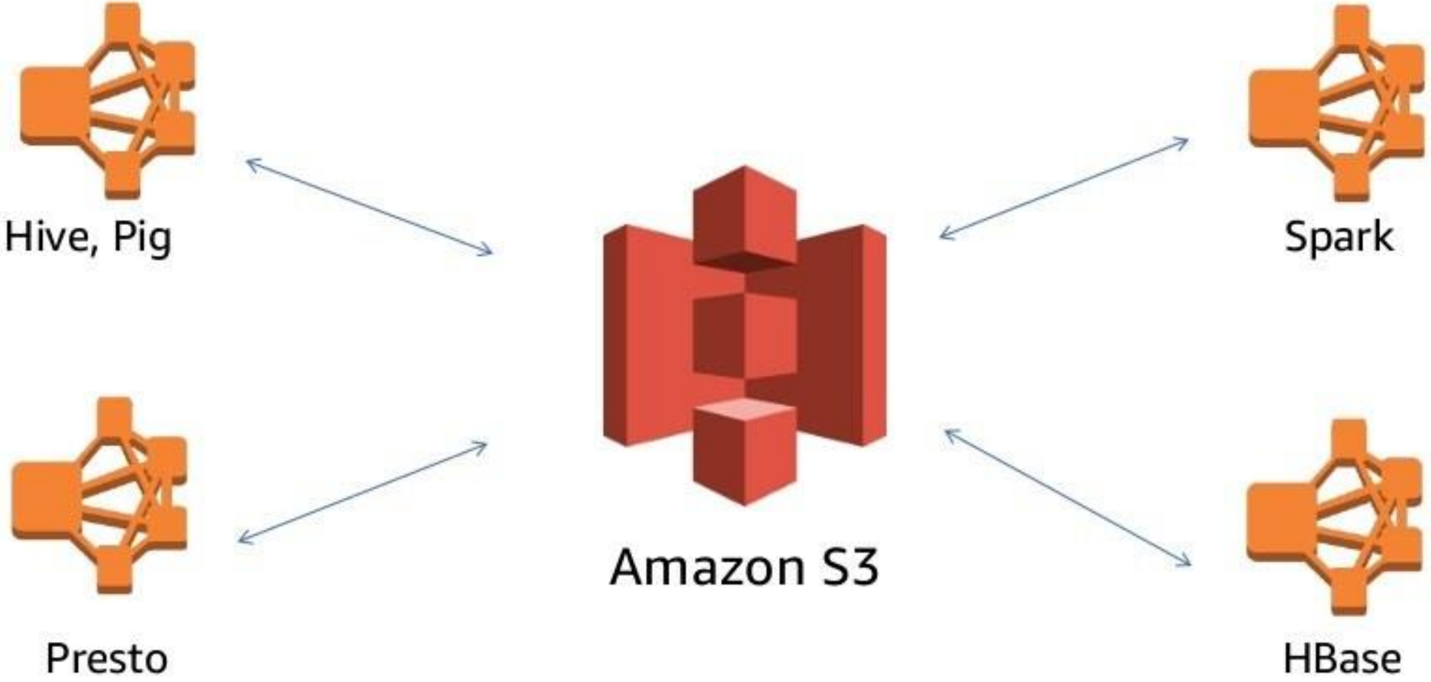
- Native S3 multi-AZ replication
- Protected against data loss from node failure, cluster failure or AZ failure
- Multiple physical facility replication
- Paying for storage units via S3 (cheap!!!)
- Spin clusters up and down or turn off!
- Stream multiple EMR clusters from same dataset in S3

Amazon S3 as your persistent data store

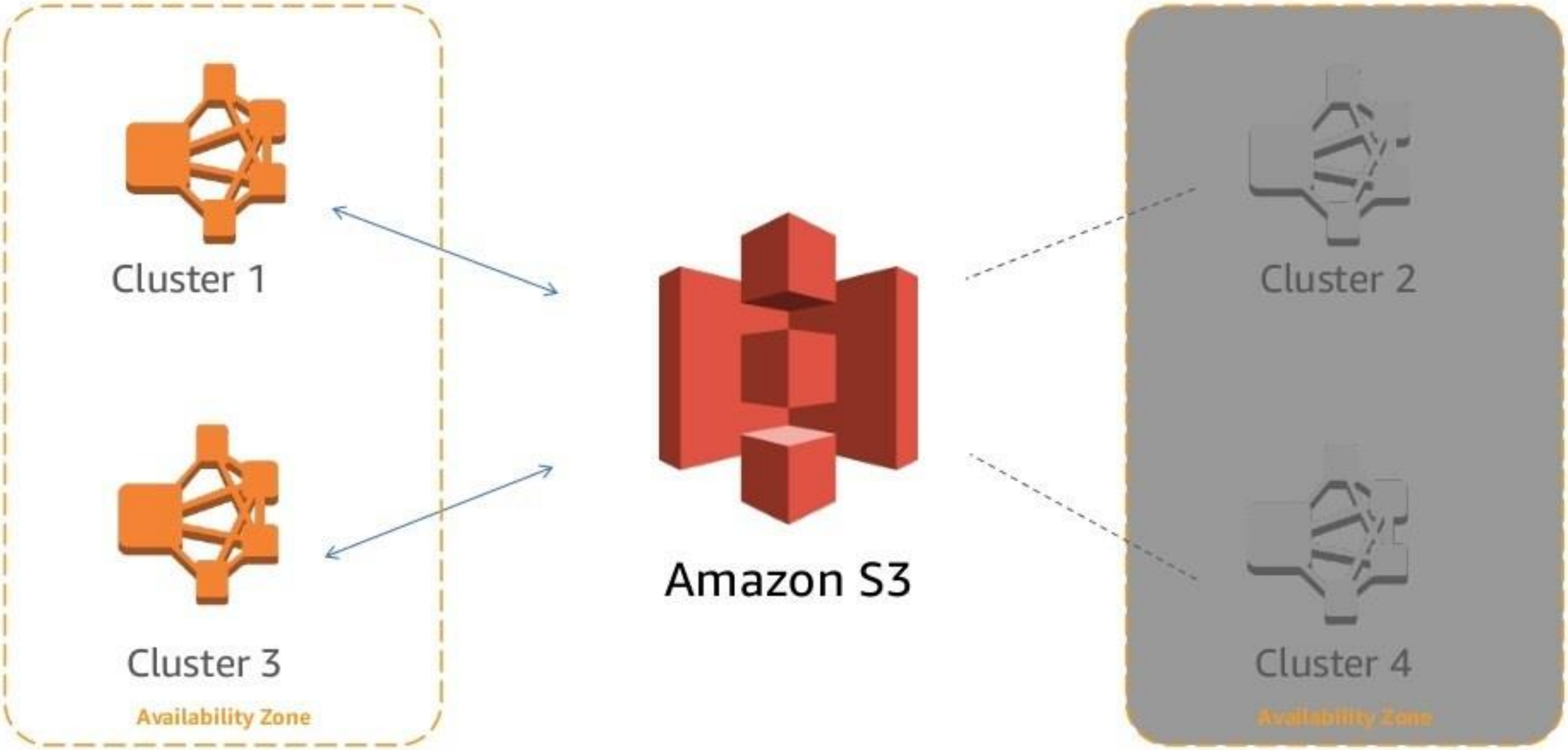
- Designed for 99.999999999% durability
- Separate compute and storage
- Resize and shut down clusters with no data loss
- Point multiple clusters at the same data in S3
- Easily evolve your analytics infrastructure as technology evolves



EMR with Amazon S3



Disaster Recovery built in



Benefits: Amazon S3 as HDFS

- No need to scale HDFS
  - Capacity?
  - Replication?

- Amazon S3 scales with your data
  - Both in IOPS and storage
  - Your data can grow independent of CPU and RAM

Benefits: Amazon S3 as HDFS

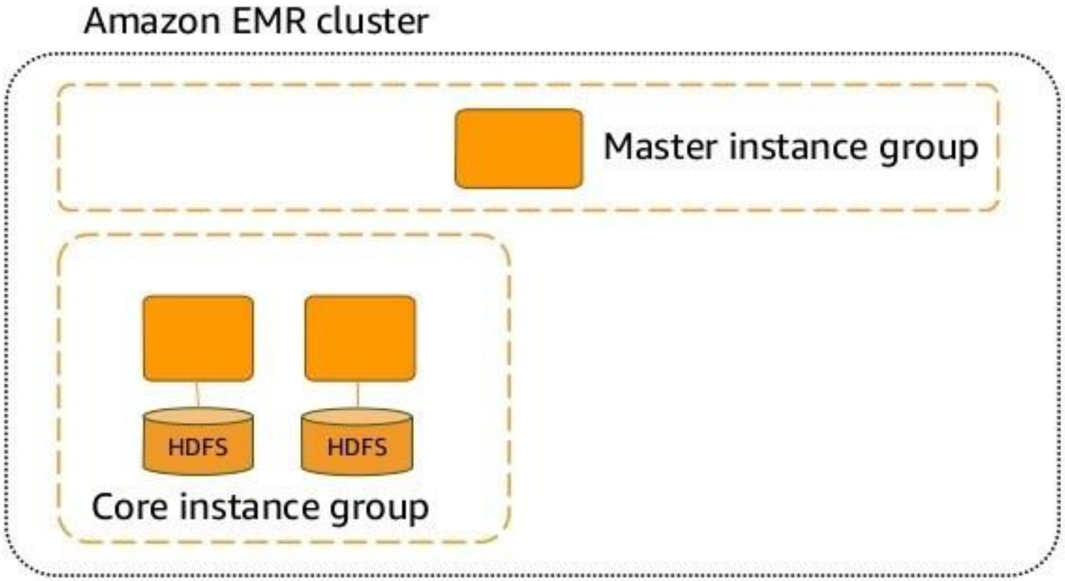
- Take advantage of S3 features
  - Server-side encryption
  - Lifecycle policies
  - Versioning to protect against corruption
- Build elastic clusters
  - Add nodes to read from Amazon S3
  - Remove nodes with data safe on Amazon S3

Core nodes and task nodes

Amazon EMR Core nodes

Run TaskTrackers
(Compute)

Run DataNode
(HDFS)

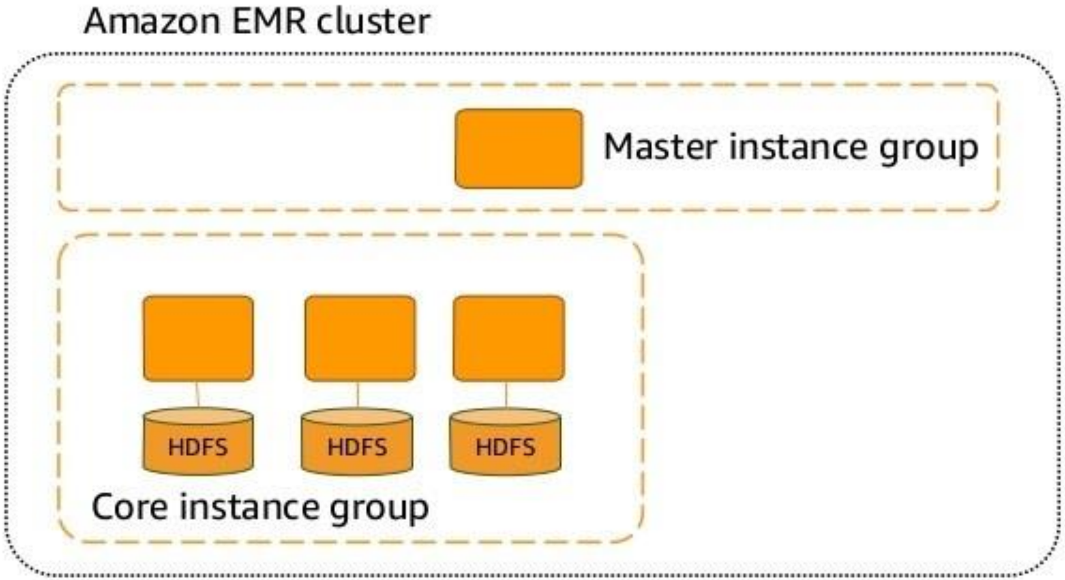


Amazon EMR Core nodes

Can add core nodes

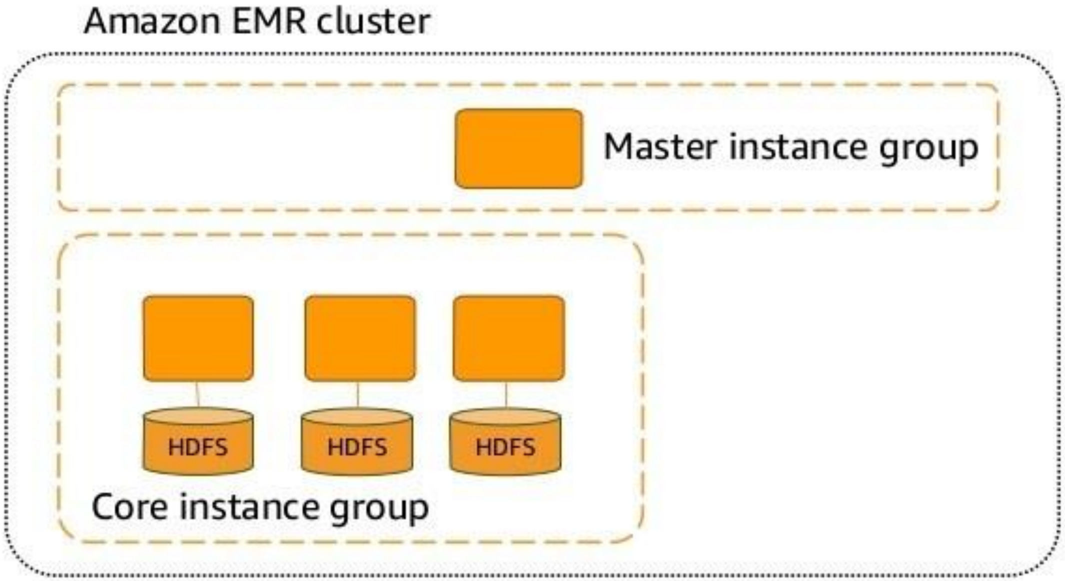
More HDFS space

More CPU/memory



Amazon EMR Core nodes

Can’t remove core nodes because of HDFS

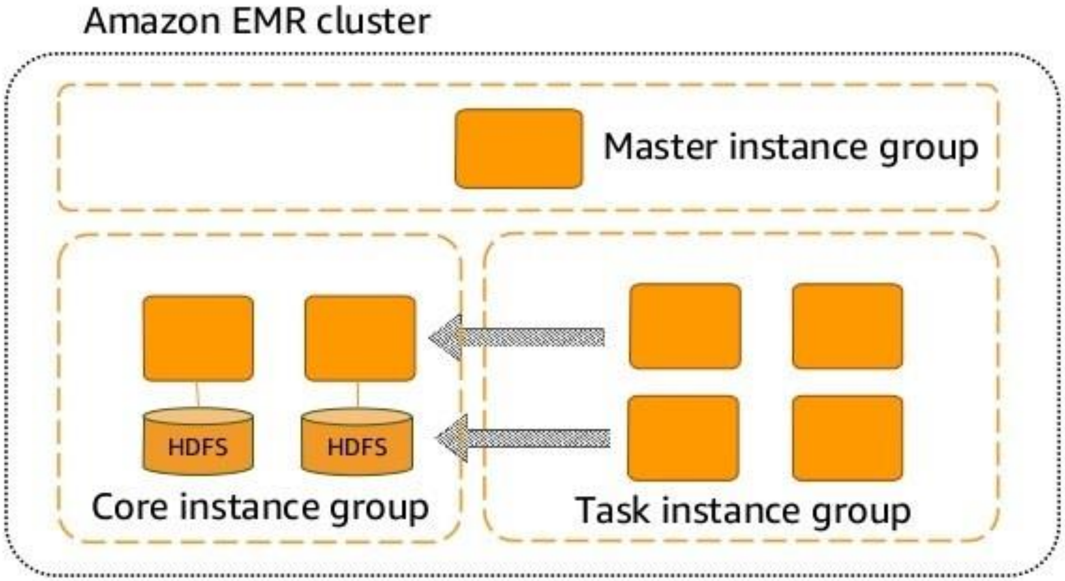


Amazon EMR Task nodes

Run TaskTrackers

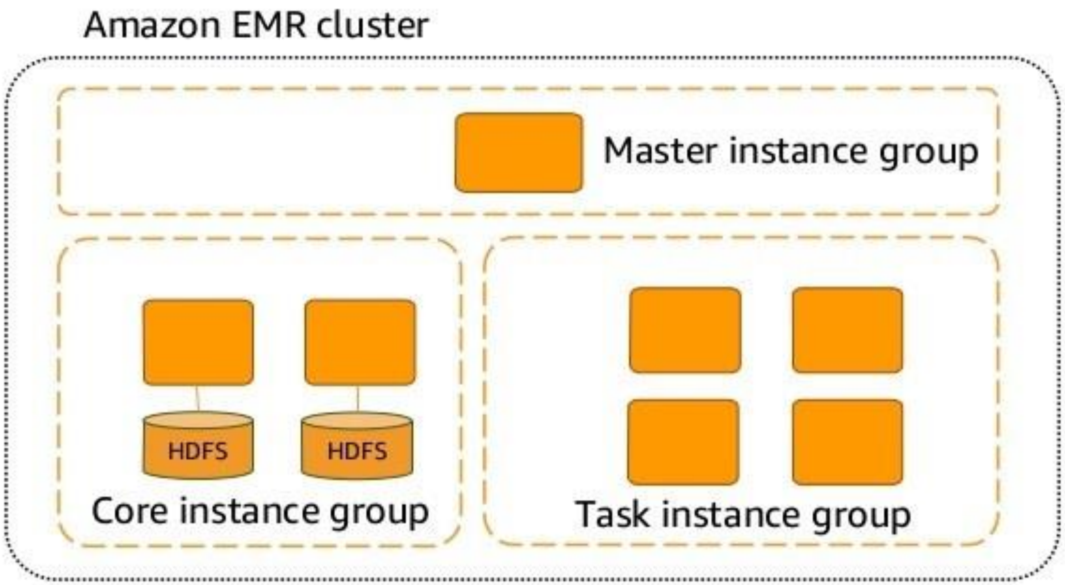
No HDFS

Reads from core nodes



Amazon EMR Task nodes

Can add task nodes

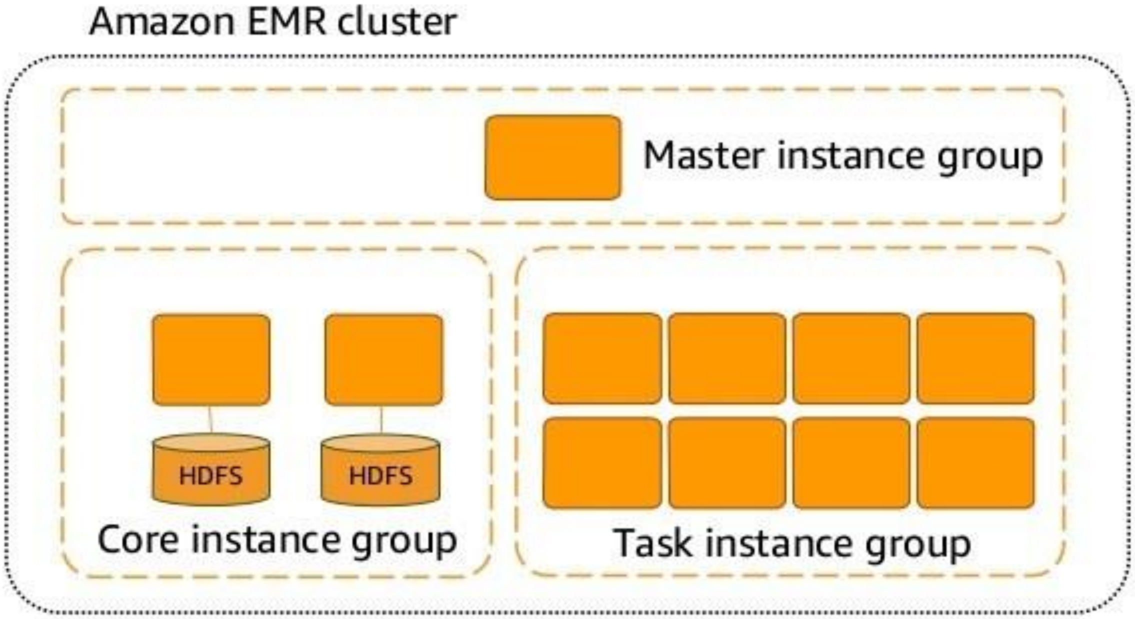


Amazon EMR Task nodes

More CPU power

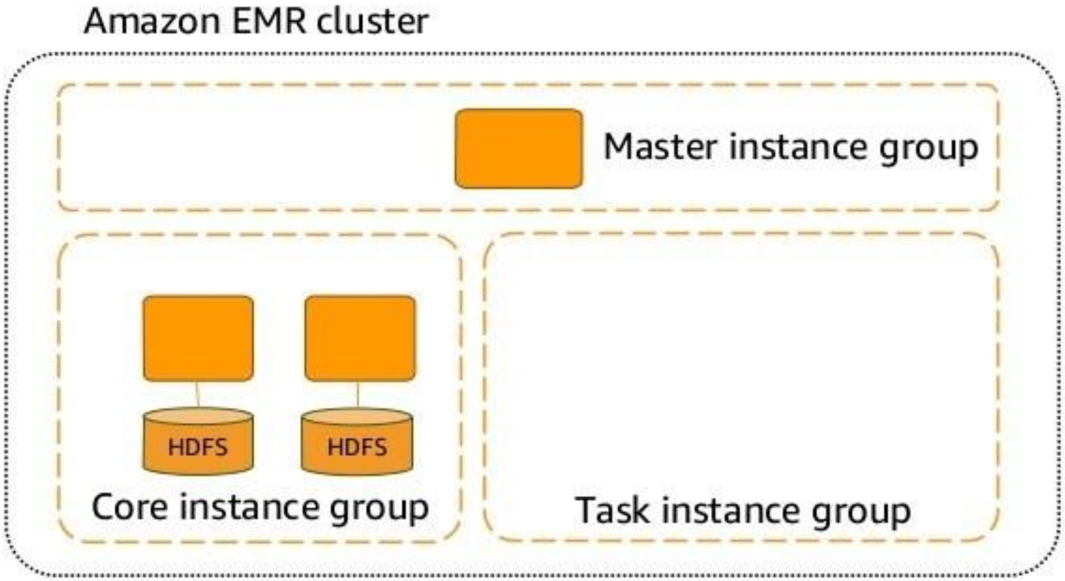
More memory

Jobs done faster



Amazon EMR Task nodes

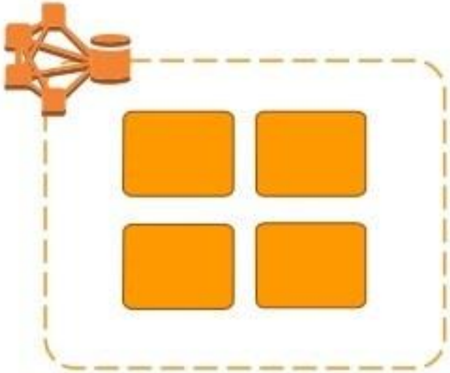
You can remove task nodes when processing is completed



Elastic clusters

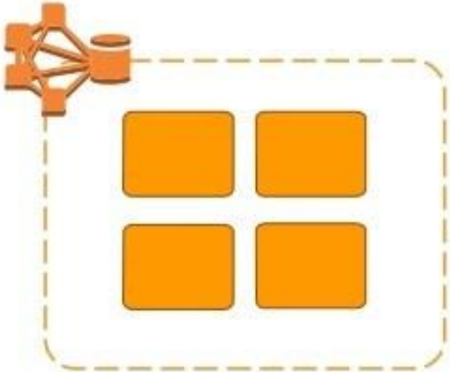
Elastic Clusters

1. Start cluster with certain number of nodes



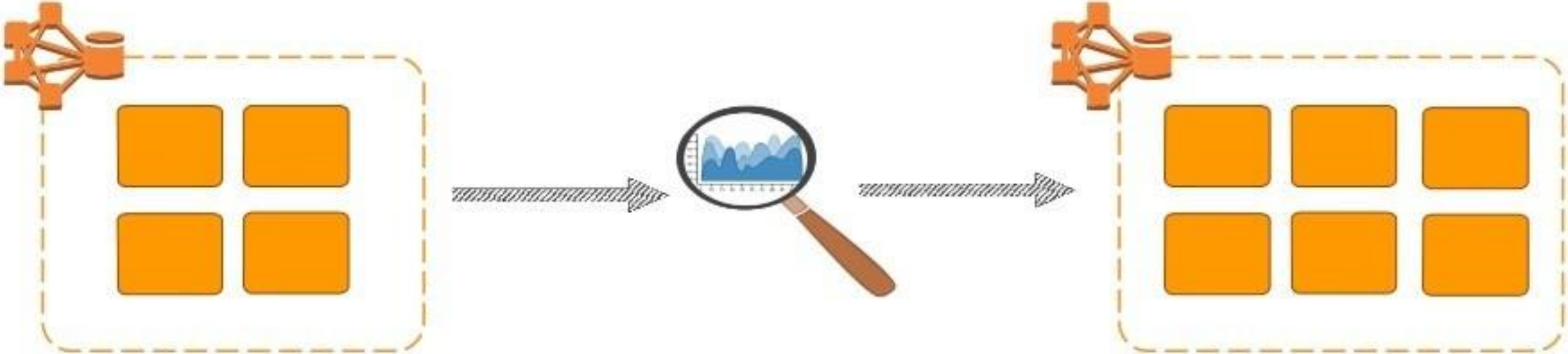
Elastic Clusters

2. Set Auto-Scaling policies in CloudWatch



Elastic Clusters

3. Leverage EMR Auto-Scaling to add nodes or gracefully remove nodes based on CloudWatch policies

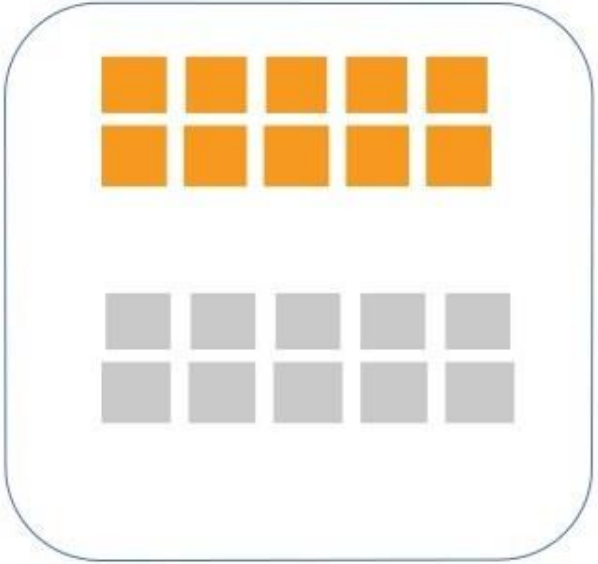


Scale up with Spot instances

10 node cluster running for 14 hours

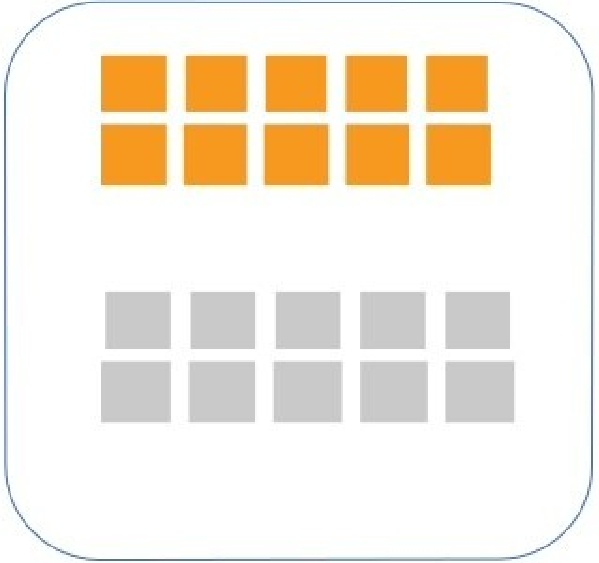
Cost = $1 * 10 * 14 = $140

Resize nodes with Spot instances



Add 10 more nodes on Spot

Resize nodes with Spot instances



20 node cluster running for 7 hours

Cost = $1 * 10 * 7 $70

+ $0.5 * 10 * 7 $35

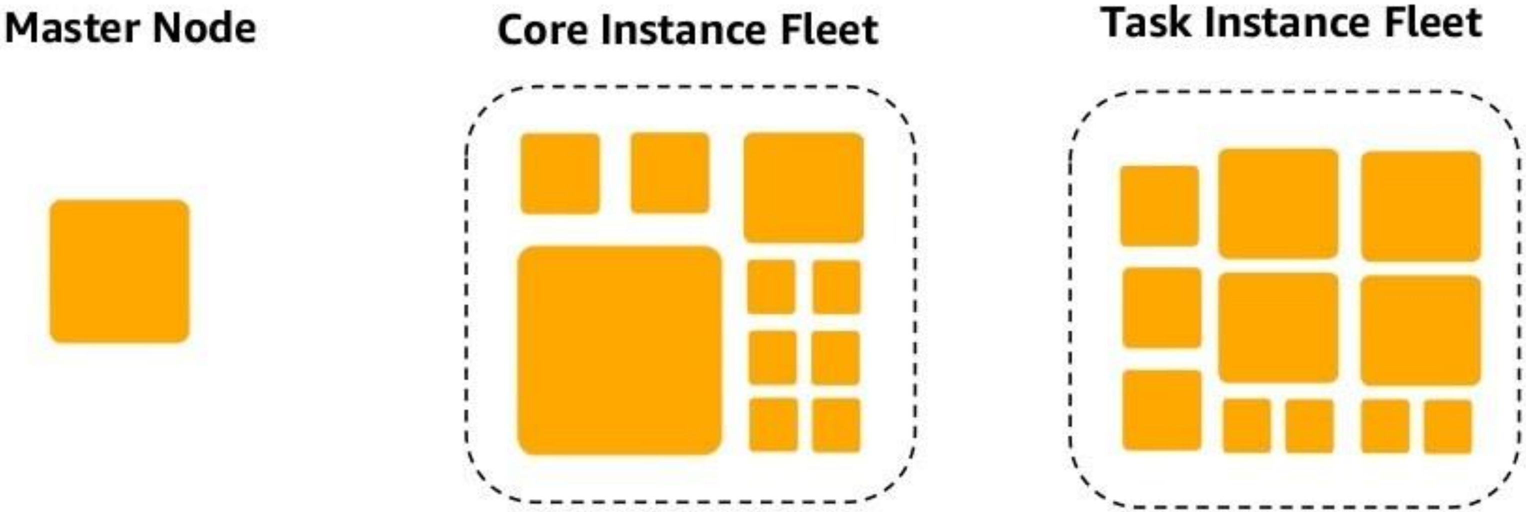
Total $105

Resize nodes with Spot instances

50% less run-time: 14 → 7 hours

25% less cost: $140 → $105

Instance fleets for advanced Spot provisioning



Master Node

Core Instance Fleet

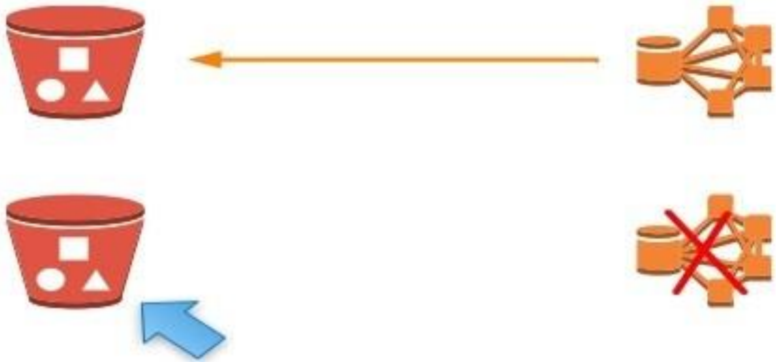
Task Instance Fleet

- Provision from a list of instance types with Spot and On-Demand
- Launch in the most optimal Availability Zone based on capacity/price
- Spot Block support (Defined Duration)

https://aws.amazon.com/blogs/aws/new-amazon-emr-instance-fleets/

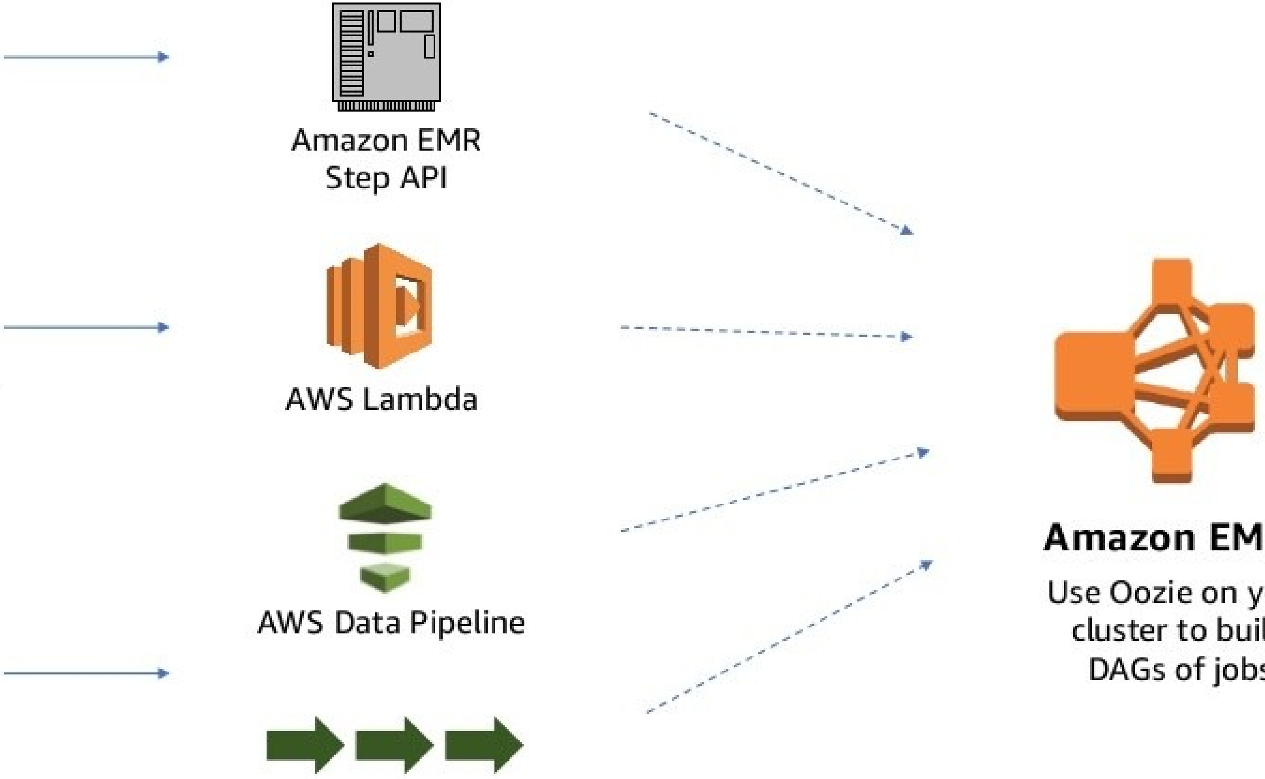
EMR Enables Transient Clusters

- Cluster lives for the duration of the job
- Shut down the cluster when the job is done
- Input and output data persists on Amazon S3



Data on Amazon S3

Options to submit jobs



Submit a application

Use AWS Lambda to submit applications to EMR Step API or directly on your cluster

Create a pipeline to schedule job submission or create complex workflows

Amazon EMR Step API

AWS Lambda

AWS Data Pipeline

Airflow, Luigi, or other schedulers on EC2

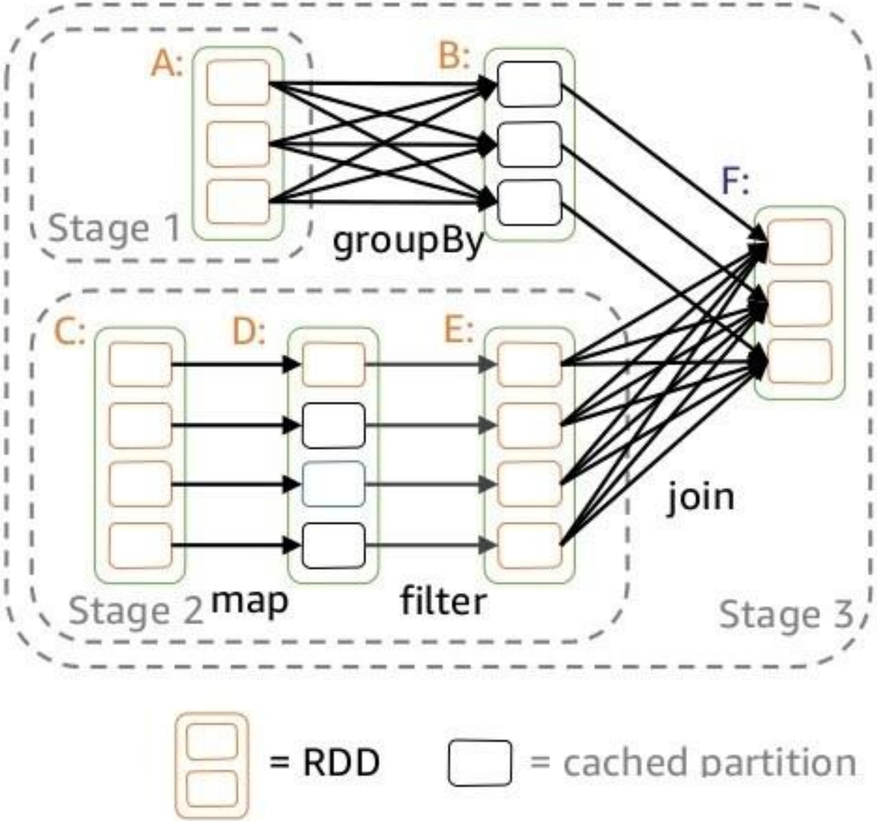
Amazon EMR

Use Oozie on your cluster to build DAGs of jobs

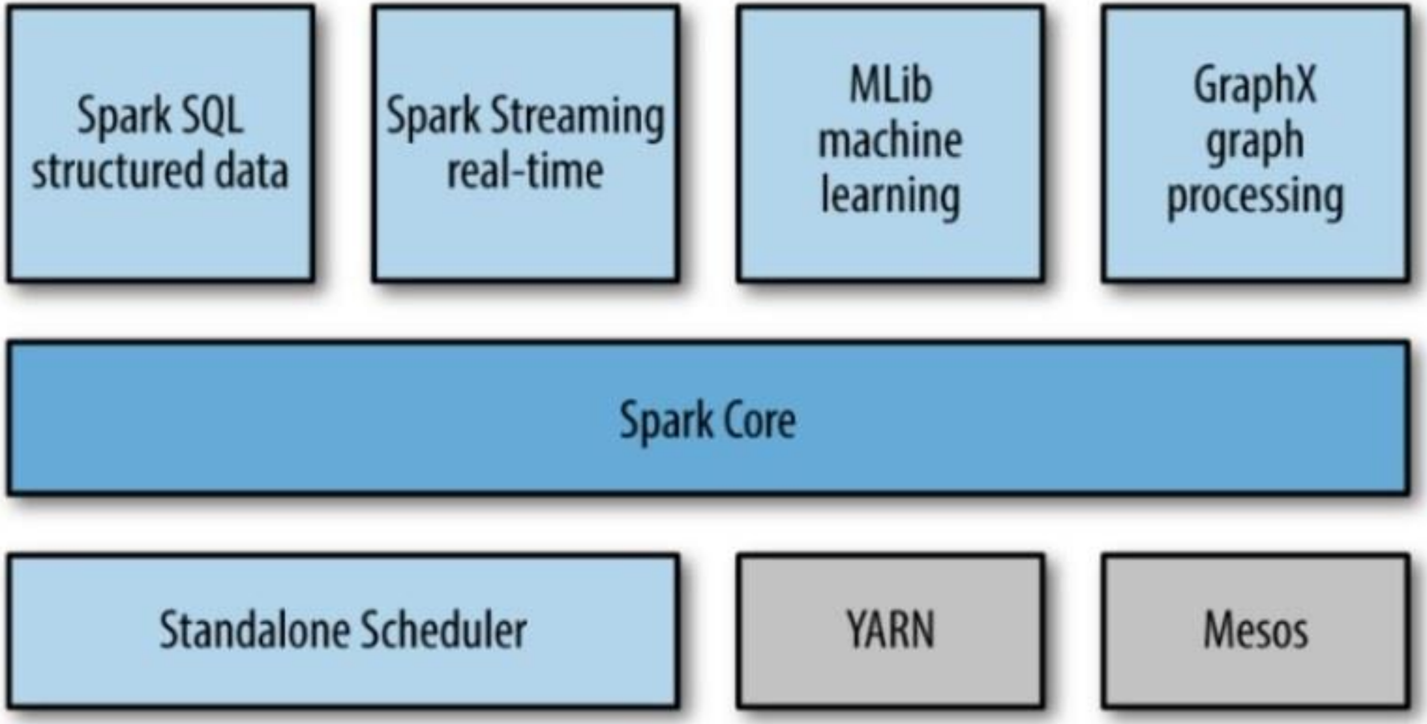
Beyond Map Reduce: Spark

Spark moves at interactive speed

- Massively parallel
- Uses DAGs instead of map-reduce for execution
- Minimizes I/O by storing data in memory
- Partitioning-aware to avoid network-intensive shuffle



Spark components to match your use case



Spark speaks your language

Scala

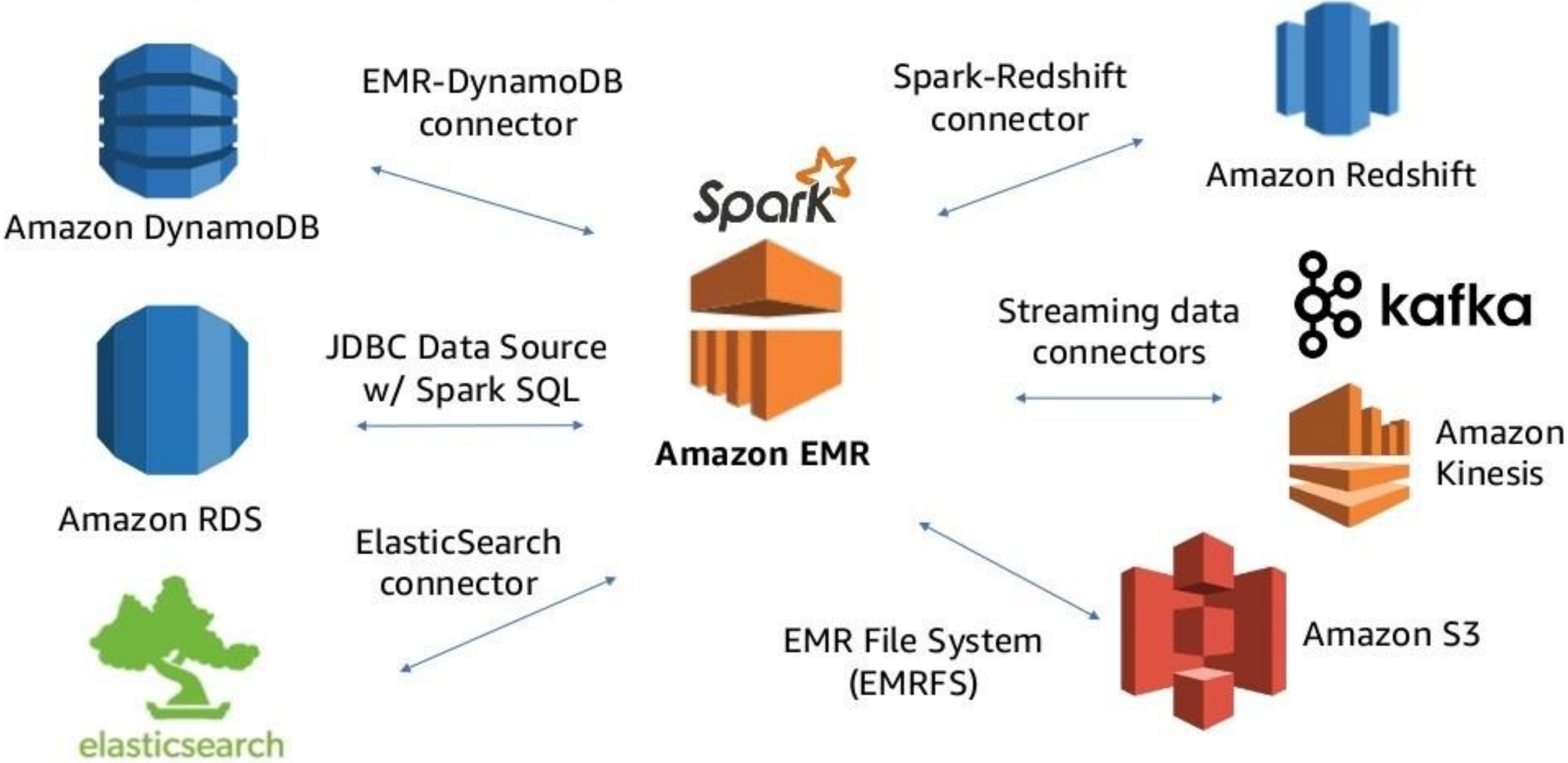
python

Java

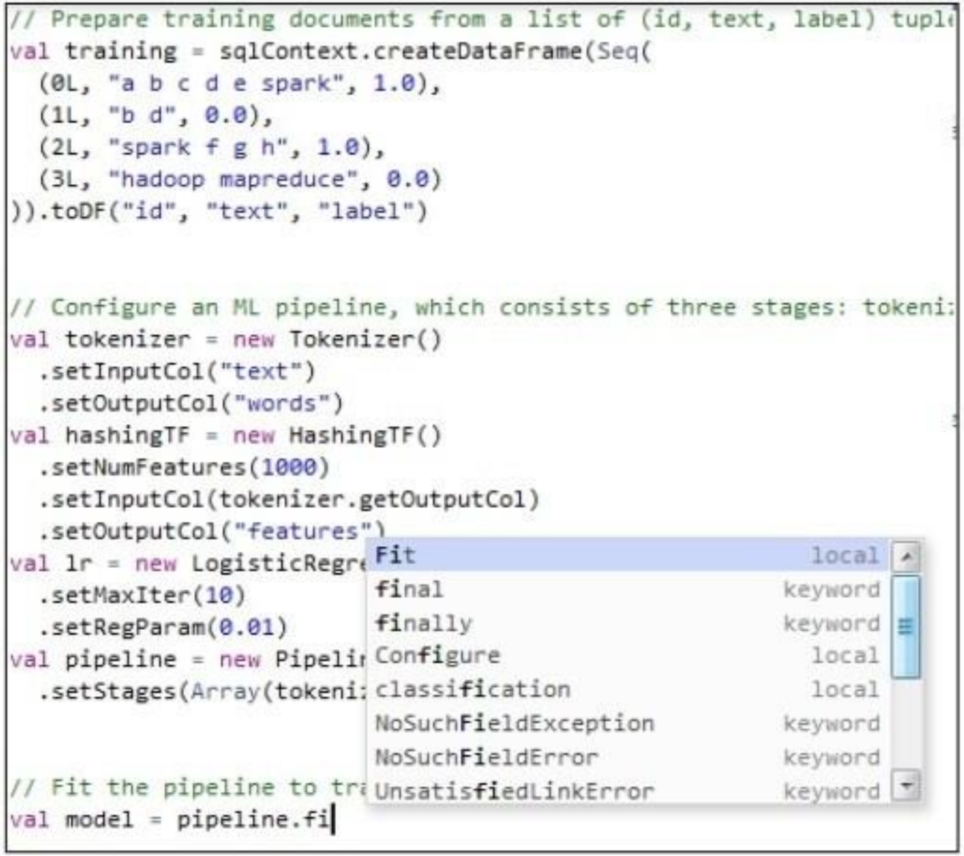
SQL

R

Many storage layers to choose from



Use DataFrames for machine learning



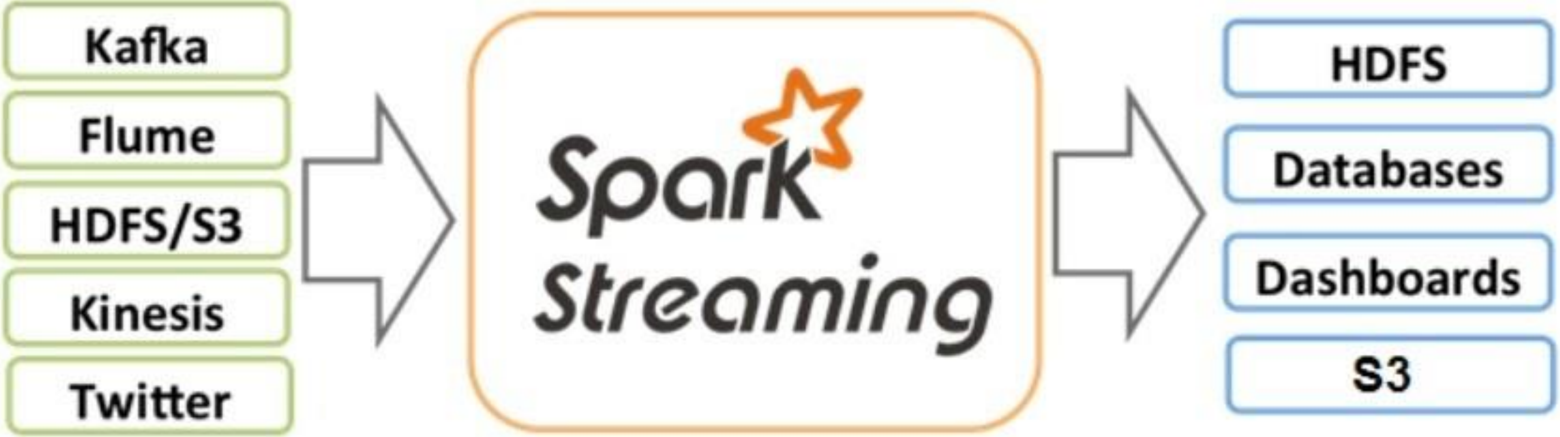
```
// Prepare training documents from a list of (id, text, label) tupl
val training = sqlContext.createDataFrame(Seq(
  (0L, "a b c d e spark", 1.0),
  (1L, "b d", 0.0),
  (2L, "spark f g h", 1.0),
  (3L, "hadoop mapreduce", 0.0)
)).toDF("id", "text", "label")


// Configure an ML pipeline, which consists of three stages: tokeni
val tokenizer = new Tokenizer()
  .setInputCol("text")
  .setOutputCol("words")
val hashingTF = new HashingTF()
  .setNumFeatures(1000)
  .setInputCol(tokenizer.getOutputCol)
  .setOutputCol("features")
val lr = new LogisticRegre
  .setMaxIter(10)
  .setRegParam(0.01)
val pipeline = new Pipeli
  .setStages(Array(tokeni


// Fit the pipeline to tra
val model = pipeline.fi
```

- Spark ML (replacing MLlib) uses DataFrames as input/output for models
- Create ML pipelines with a variety of distributed algorithms

Create DataFrames on streaming data



Amazon EMR

Easy to Use
Launch a cluster in minutes

Low Cost
Pay an hourly rate

Elastic
Easily add or remove capacity

Reliable
Spend less time monitoring

Secure
Manage firewalls

Flexible
Customize the cluster

Resources

https://aws.amazon.com/emr/

https://aws.amazon.com/blogs/bigdata

https://aws.amazon.com/whitepapers/

https://medium.com/@julsimon

Thank you!

Data Pipeline

- A web service for scheduling regular data movement and data processing activities in the AWS cloud.

- Data Pipeline integrates with on-premise and cloud-based storage systems.

- A managed ETL (Extract-Transform-Load) service.

- Native integration with S3, DynamoDB, RDS, EMR, EC2 and Redshift.

Features

- You can quickly and easily provision pipelines that remove the development and maintenance effort required to manage your daily data operations, letting you focus on generating insights from that data.

- Data Pipeline provides built-in activities for common actions such as copying data between Amazon Amazon S3 and Amazon RDS, or running a query against Amazon S3 log data.

- Data Pipeline supports JDBC, RDS and Redshift databases.

Components

- A **pipeline definition** specifies the business logic of your data management.

- A **pipeline** schedules and runs tasks by creating EC2 instances to perform the defined work activities.

- **Task Runner** polls for tasks and then performs those tasks. For example, Task Runner could copy log files to S3 and launch EMR clusters. Task Runner is installed and runs automatically on resources created by your pipeline definitions. You can write a custom task runner application, or you can use the Task Runner application that is provided by Data Pipeline.

Pipeline Definition

- From your pipeline definition, Data Pipeline determines the tasks, schedules them, and assigns them to task runners.

- If a task is not completed successfully, Data Pipeline retries the task according to your instructions and, if necessary, reassigns it to another task runner.

- If the task fails repeatedly, you can configure the pipeline to notify you.

Pipeline Definition

- A pipeline definition can contain the following types of components

- **Data Nodes** – The location of input data for a task or the location where output data is to be stored.

- **Activities** – A definition of work to perform on a schedule using a computational resource and typically input and output data nodes.

- **Preconditions** – A conditional statement that must be true before an action can run.

Pipeline Definition

And,

Scheduling Pipelines, Resources, Actions, Schedules

Task Runners

•When Task Runner is installed and configured, it polls Data Pipeline for tasks associated with pipelines that you have activated.

•When a task is assigned to Task Runner, it performs that task and reports its status back to Data Pipeline.