aws INNOVATE

DATA EDITION

19 August 2021

Move Apache Spark , Hadoop & other big data applications to the cloud with Amazon EMR

Melody Yang

Sr Analytics Specialist Solution Architect
Amazon Web Services

Agenda

01 Challenges of on-premises clusters

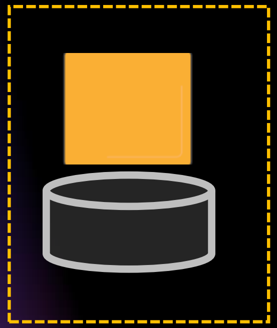
02 Migrate workload to Amazon EMR

03 Demo

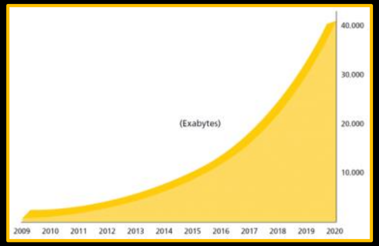
04 Future-proof architecture

Customer’s voice about the challenges

Challenge #1: missing business SLAs



Tightly coupled +



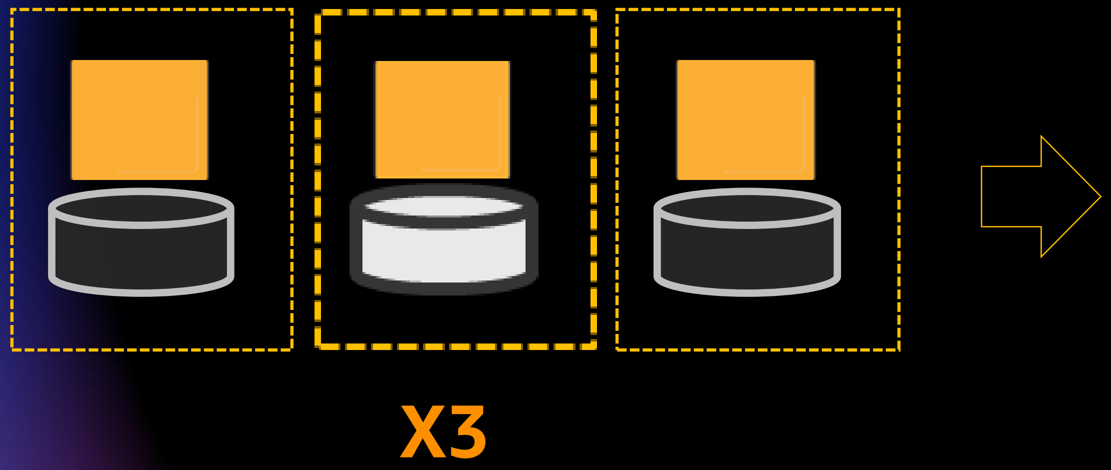
Business Impact

Job Failure

Slow processing

Missing SLAs

Challenge #2: Replication adds to cost & risk



- Data is replicated several times
- Typically only in one data center

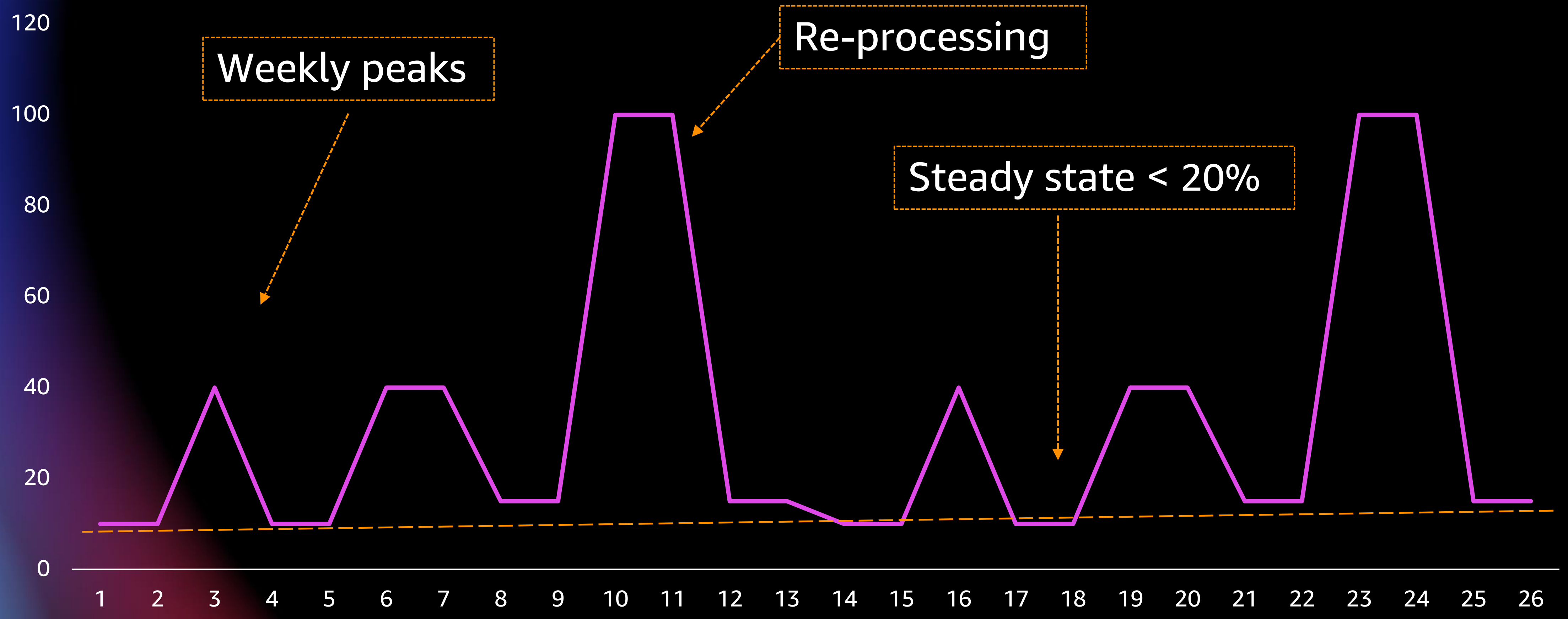
Business Impact

Cost for unnecessarily compute growth

Risk of non-compliant

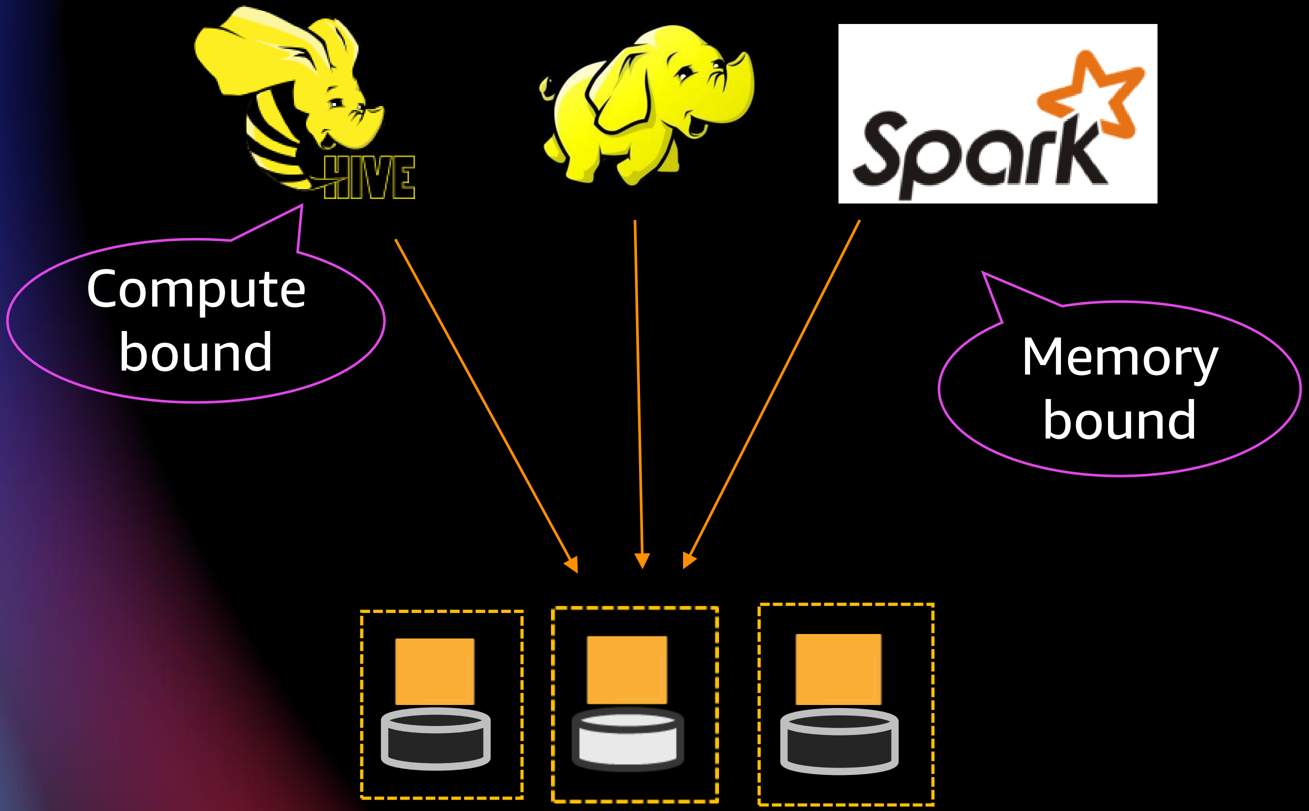
Challenge #3: Underutilized resources

“Managing the infrastructure for scaling of computations is time consuming and labour intensive”



Challenge #4: Contention for the same resources

“The platform is just designed for Data Scientists, it is not a user-friendly unified enterprise level platform”



Compute bound

Memory bound

Business Impact

Single tenancy

Conflicting SLA

Limiting innovation

Challenge #5: Limited on operation and support

- End-of-life support:
  - CDH6.3 in March 2022
  - HDP3.1 in December 2021

- Security update/patching/upgrade for:
  - Data Processing: Pig ,Hive, Flink, Spark
  - Queries: Impala, Spark SQL, Presto
  - Machine Learning: Spark ML, MxNet, Tensorflow
  - Notebooks: Jupyter, Zeppelin
  - NoSQL: HBase

If we want to scale at the rate we want, we need a different strategy.

The current strategy of being on-prem and adding hardware is not going to work

Amazon EMR

The best place to run your analytics workloads

Amazon EMR

EASILY RUN SPARK, HADOOP, HIVE, PRESTO, HBASE, AND OTHER BIG DATA FRAMEWORKS

Automate provisioning, configuring, and tuning
Easy setup, management, and monitoring

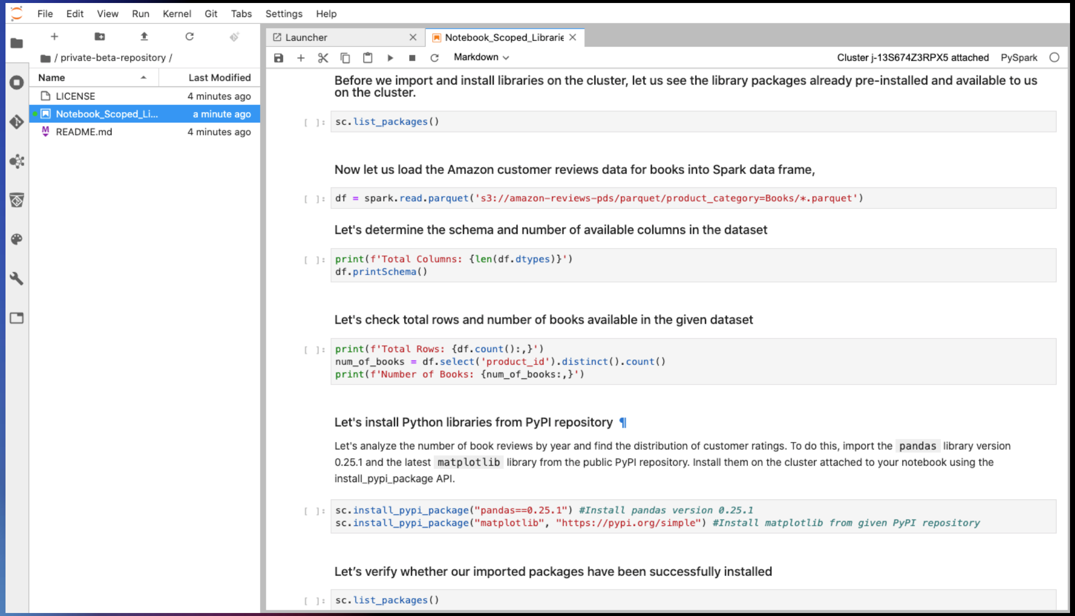
Get the latest, stable, open-source releases
Latest open-source framework updates within 30 days

Automatically scale up and down
Manage cluster size based on utilization to reduce costs

Simple and predictable pricing
Per-second pricing, and save 50%–80% with Amazon EC2 Spot and Reserved Instances

Amazon EMR Studio

FULLY MANAGED NOTEBOOKS WITH SSO INTEGRATION



Before we import and install libraries on the cluster, let us see the library packages already pre-installed and available to us on the cluster.

```
[ ]: sc.list_packages()
```

Now let us load the Amazon customer reviews data for books into Spark data frame,

```
[ ]: df = spark.read.parquet('s3://amazon-reviews-pds/parquet/product_category=Books/*.parquet')
```

Let's determine the schema and number of available columns in the dataset

```
[ ]: print(f'Total Columns: {len(df.dtypes)}')
     df.printSchema()
```

Let's check total rows and number of books available in the given dataset

```
[ ]: print(f'Total Rows: {df.count():,}')
     num_of_books = df.select('product_id').distinct().count()
     print(f'Number of Books: {num_of_books:,}')
```

Let's install Python libraries from PyPI repository ¶

Let's analyze the number of book reviews by year and find the distribution of customer ratings. To do this, import the `pandas` library version 0.25.1 and the latest `matplotlib` library from the public PyPI repository. Install them on the cluster attached to your notebook using the install_pypi_package API.

```
[ ]: sc.install_pypi_package("pandas==0.25.1") #Install pandas version 0.25.1
     sc.install_pypi_package("matplotlib", "https://pypi.org/simple") #Install matplotlib from given PyPI repository
```

Let's verify whether our imported packages have been successfully installed

```
[ ]: sc.list_packages()
```

Use corporate identity to log directly into notebooks

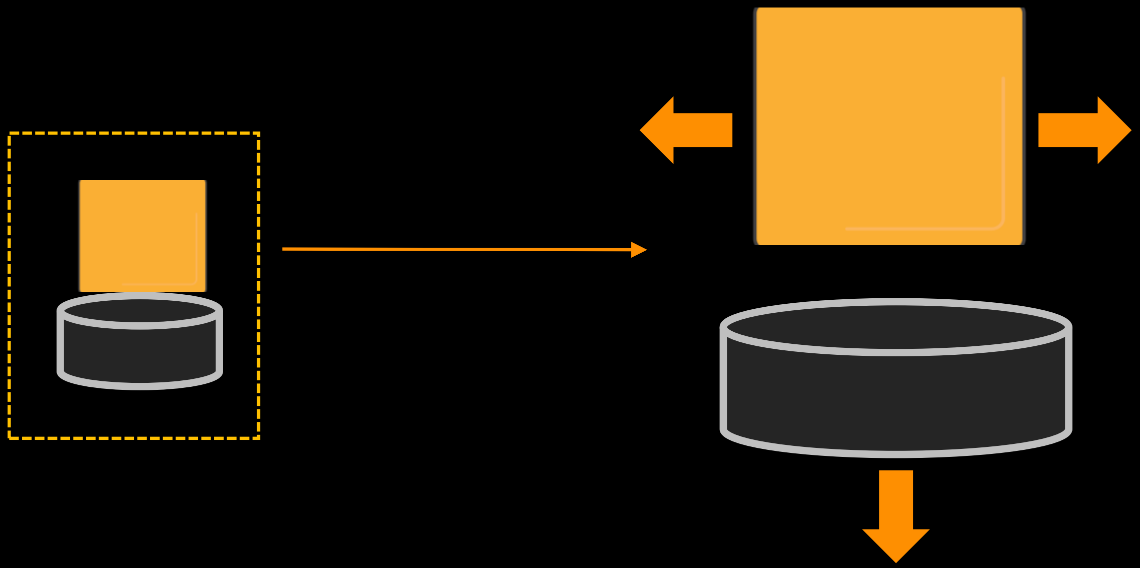
Develop, visualize, debug, and optimize analytics and ML apps

Collaborate with peers by sharing notebooks via GitHub

Build pipelines using orchestration services like Apache Airflow

The value of cloud-based Apache Hadoop & Spark with Amazon EMR

Decouple storage and compute



Amazon S3 is your persistent data store

Amazon Simple Storage Service (Amazon S3)

11 9’s of durability

Low cost

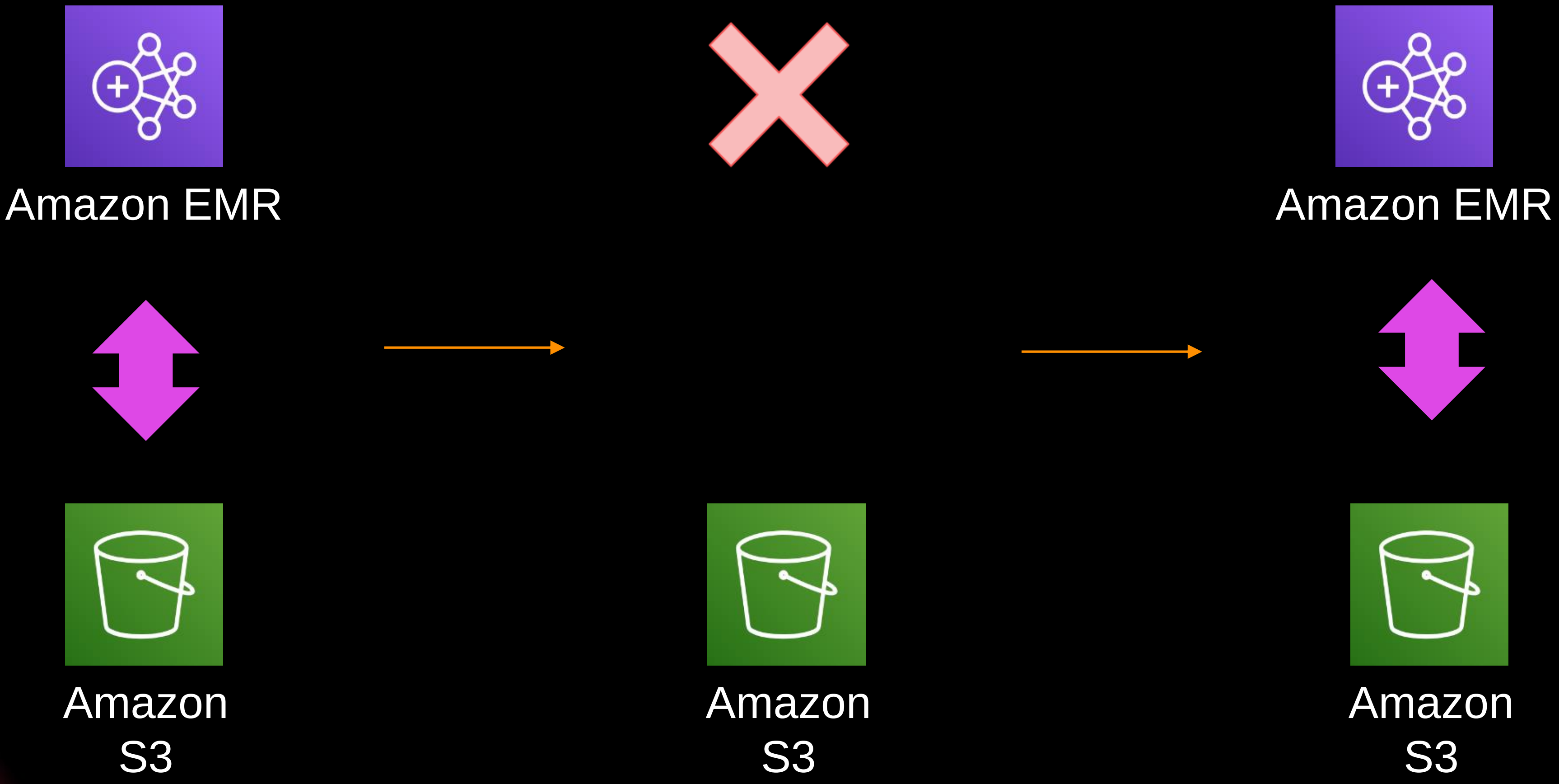
Life Cycle Policies

Versioning

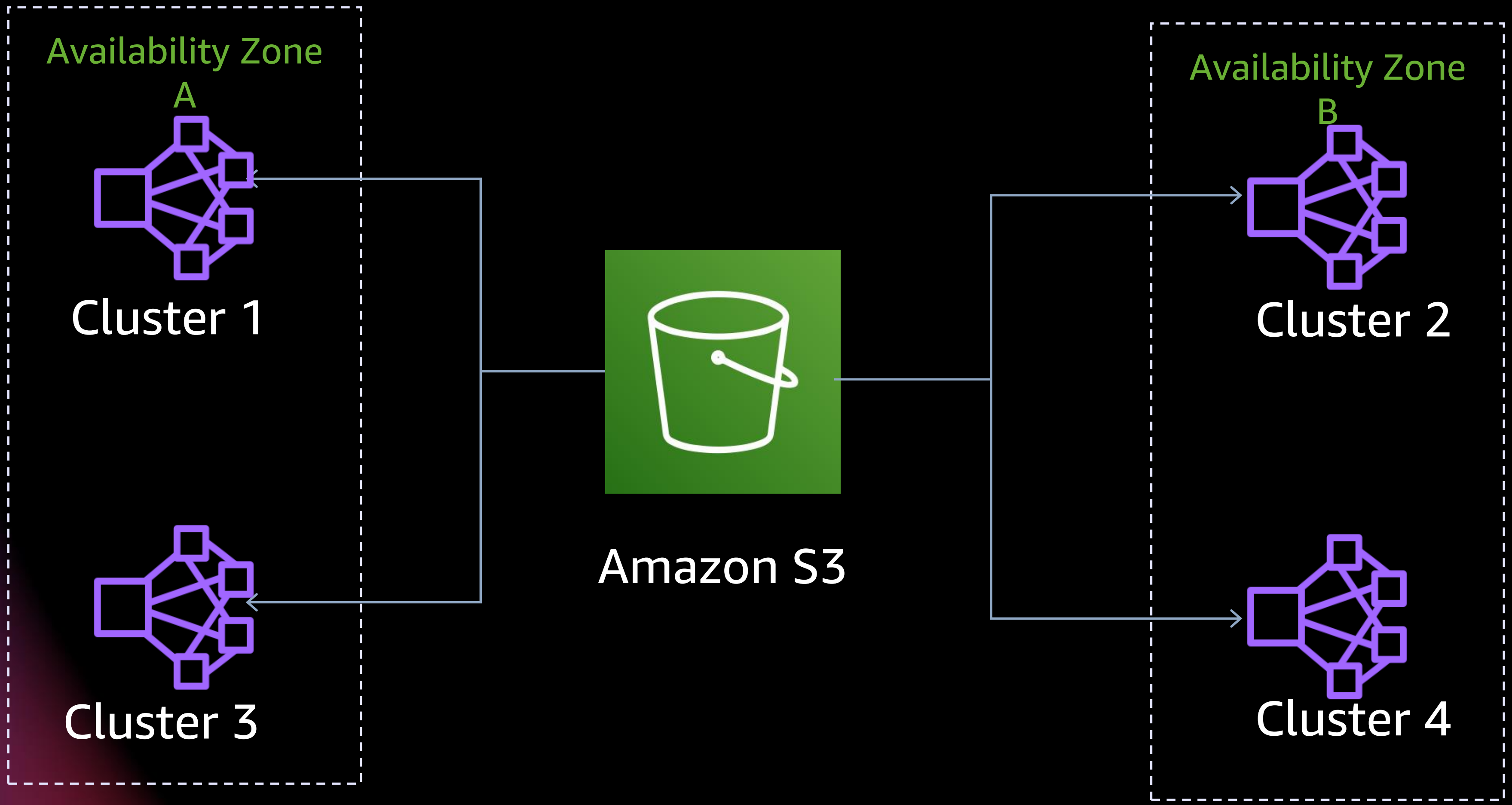
Distributed by default

EMR FS

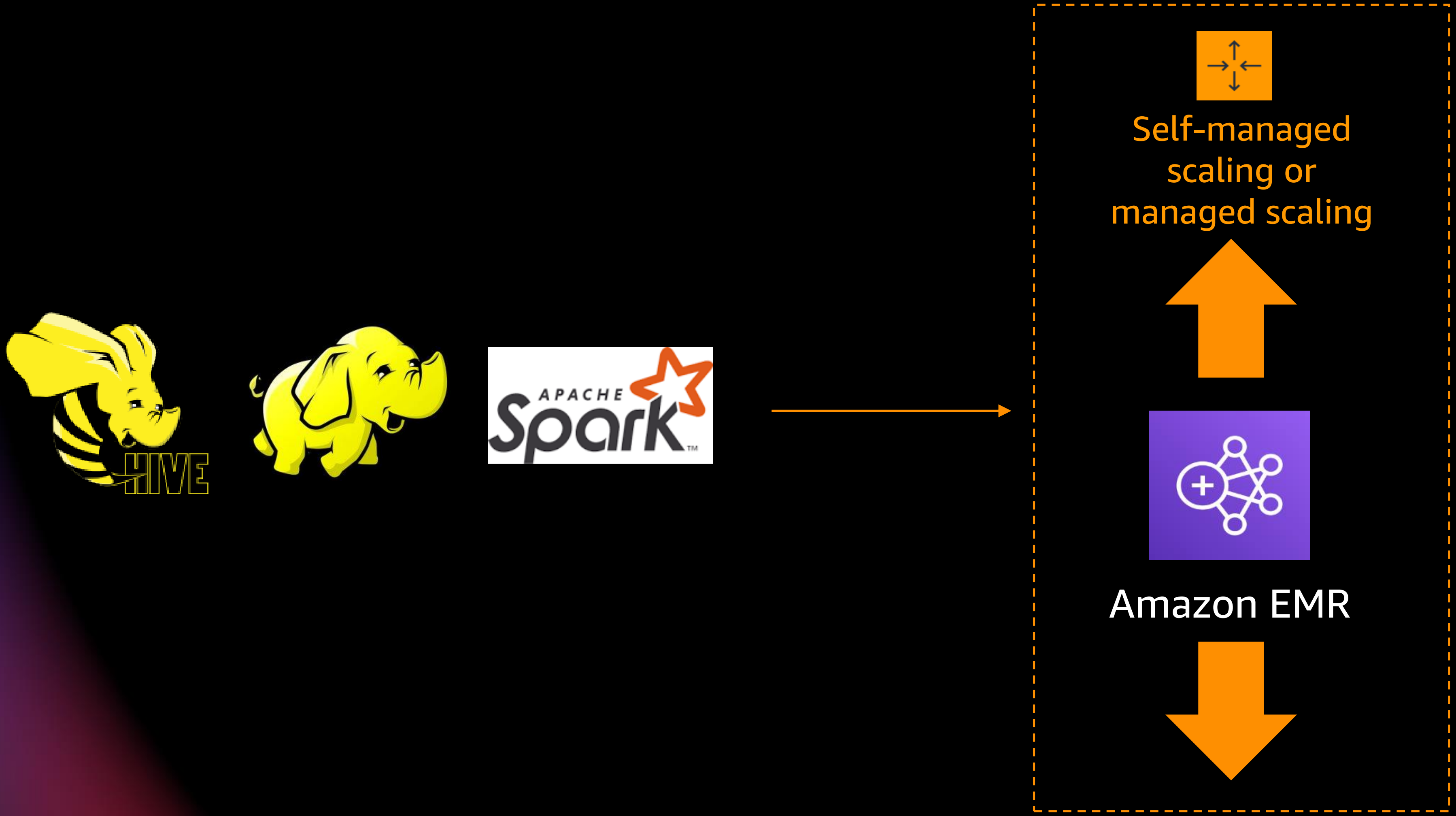
Benefit 1: Turn off clusters



Benefit 2: Built-in disaster recovery



Benefit 3: Auto-scaling persistent and transient clusters



Benefit 4: Self-service with AWS Service Catalog

Configure

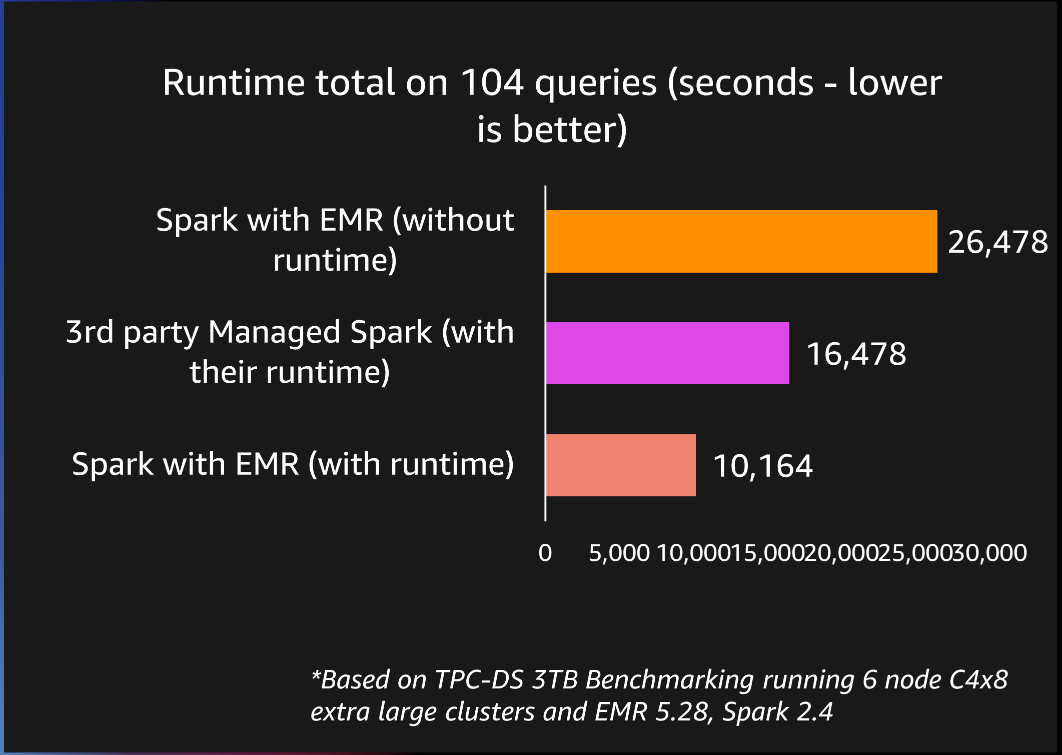
- Standardize
- Enforce consistency and compliance
- Limit Access
- Enforce tagging, security groups

Consume

- Developer autonomy
- One-stop shop
- Automate deployments
- Agile governance

Benefit 5: Apache Spark performance improvements

Performance-optimized runtime for Apache Spark, 2.6x faster performance



**Based on TPC-DS 3TB Benchmarking running 6 node C4x8 extra large clusters and EMR 5.28, Spark 2.4*

Runtime built on a optimized version of Apache Spark

Best performance

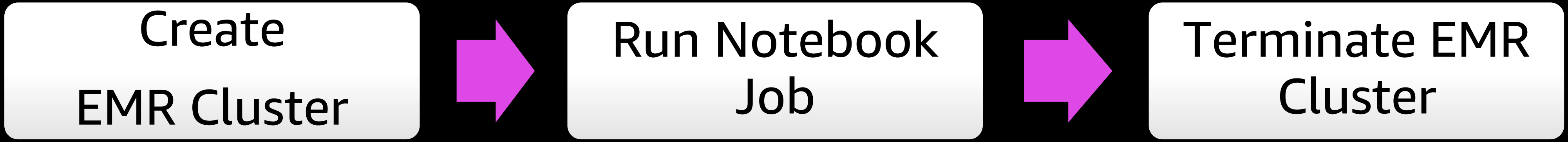
- **2.6x faster** than Spark with EMR without runtime
- **1.6x faster** than 3rd party Managed Spark (with their runtime)

Lowest price

- 1/10th the cost of 3rd party Managed Spark (with their runtime)

100% compliant with Apache Spark API’s

Demo: Build and orchestrate a Spark Job



Create EMR Cluster

Run Notebook Job

Terminate EMR Cluster

Amazon EMR Studio

Interactive IDE with SSO

easy to job build and test Spark jobs

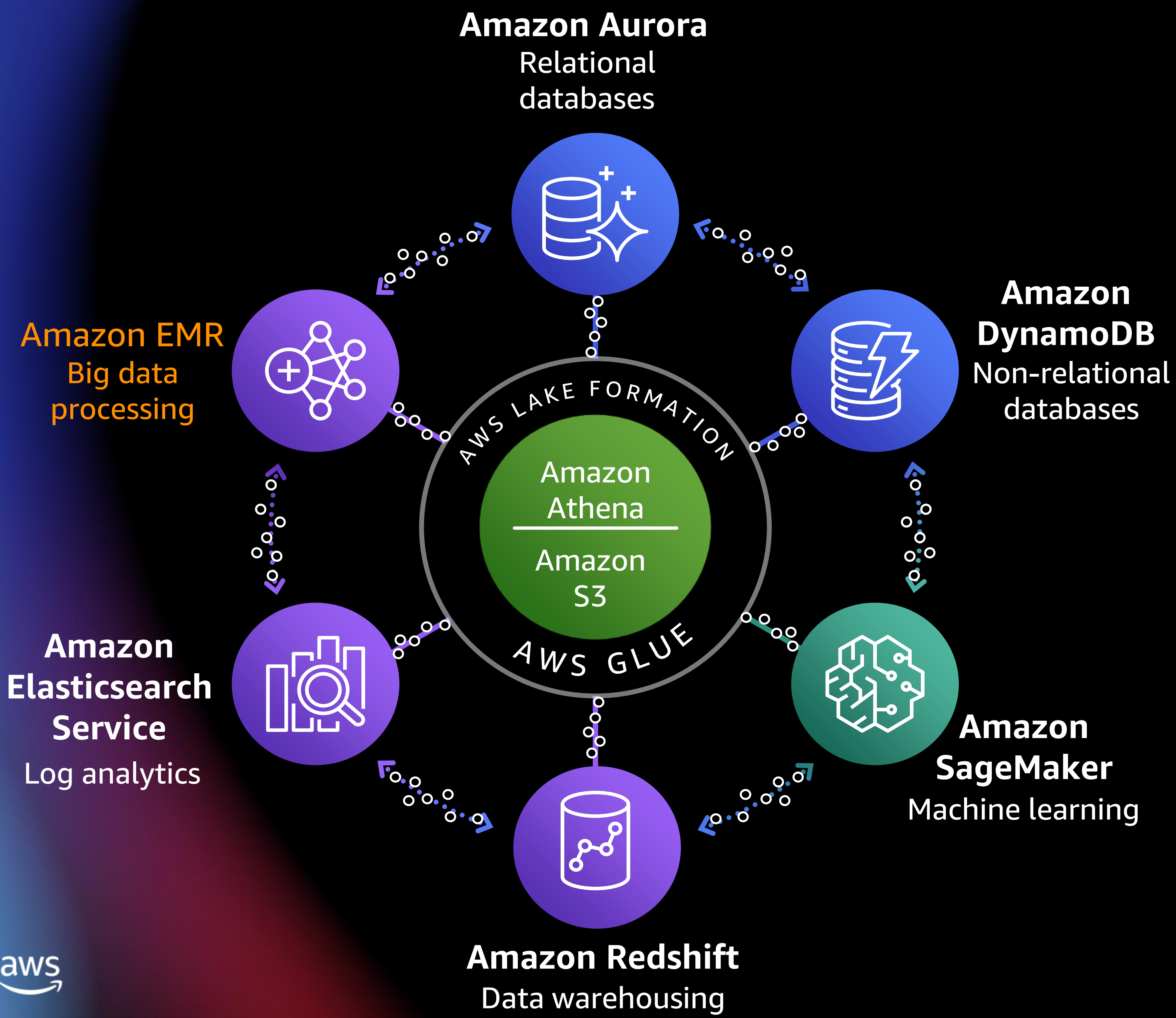
Amazon Managed Workflows for Apache Airflow (MWAA)

AWS manages the underlying infrastructure

for scalability, availability, and security

The future proof ecosystem that supports you

AWS Lake house Architecture



SCALABLE DATA LAKES

PURPOSE-BUILT DATA SERVICES

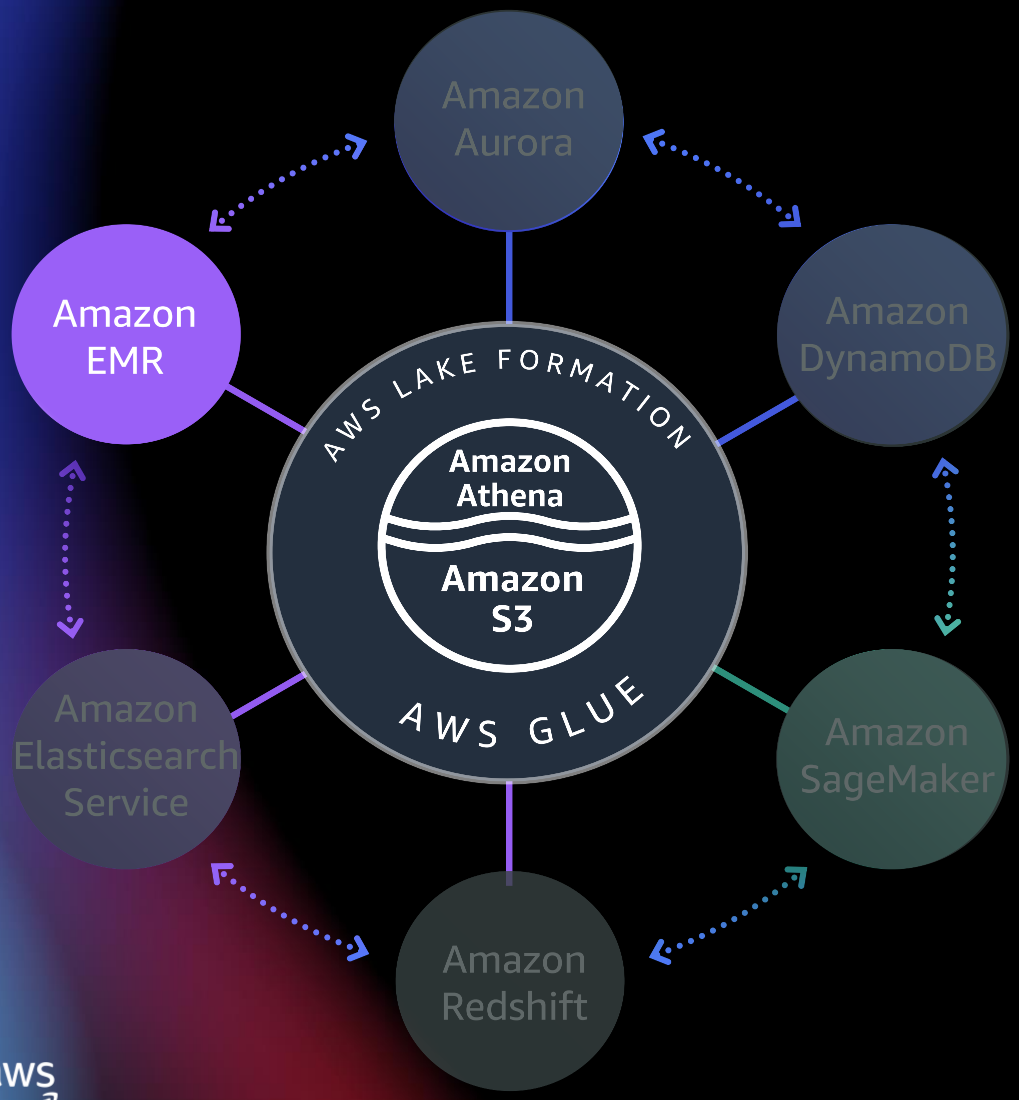
SEAMLESS DATA MOVEMENT

UNIFIED GOVERNANCE

PERFORMANT AND COST-EFFECTIVE

Amazon EMR

Easily Run Spark, Hive, Presto, HBase, Flink, and more big data apps on AWS



Automate provisioning, configuring, and tuning

Easy setup, management, and monitoring

Get the latest, stable, open-source releases

Latest open-source framework updates

Automatically scale up and down

Manage cluster size based on utilization to reduce costs

Simple and predictable pricing

Per-second pricing, and save 50%–80% with Amazon EC2 Spot and Reserved Instances

Resources

1. Amazon EMR Migration Program
2. Amazon EMR Migration Guide (2020)
3. Guide for Migrating to Apache HBase on Amazon S3 on Amazon EMR (2021)
4. Migrate to Amazon EMR - Best practices
5. Amazon EMR Service Catalog Lab
6. Amazon EMR Studio Lab

Visit the AWS Data Resource Hub

Dive deeper with these resources, get inspired and learn how you can use data to make better decisions and innovate faster.

- Building a winning data strategy
- The new leadership mindset for data & analytics
- Harness data to reinvent your organization
- Put your data to work with a modern analytics approach
- Breaking free from on-premises database constraints
- Cloud storage adoption: From cost optimization to agility & innovation
- A strategic playbook for data, analytics, and machine learning
- … and more!

https://tinyurl.com/aws-data-resource

Visit resource hub

AWS Training and Certification

Empower your teams with comprehensive training

By building skills with AWS Training and Certification, businesses and individuals can see the bigger picture understanding the reasoning behind every data point. As training progresses and teams become data-fluent, previously hidden insights come into view.

Leverage free digital training

Learn how to harness the world's most valuable resource: data. Access digital and virtual instructor-led courses on data analytics and databases built by the experts at AWS and start your learning journey to become data-driven.

Take a digital course »

Get certified

Earn industry-recognized credibility and set tangible goals for success with industry-recognized certifications, like *AWS Certified Data Analytics – Specialty*.

Learn more »

Ramp-up your skills

Deep dive into new topics and focus on knowledge gaps at your own pace with the *AWS Ramp-Up Guide: Database* and *AWS Ramp-Up Guide: Data Analytics*. With a wide range of whitepapers, blog posts, videos, webinars and peer resources available for data professionals to leverage for independent learning.

Download ramp-up guides »

Thank you for attending AWS Innovate – Data Edition

We hope you found it interesting! A kind reminder to **complete the survey.**
Let us know what you thought of today’s event and how we can improve the event experience for you in the future.

- aws-apj-marketing@amazon.com
- twitter.com/AWSCloud
- facebook.com/AmazonWebServices
- youtube.com/user/AmazonWebServices
- slideshare.net/AmazonWebServices
- twitch.tv/aws

Thank you!