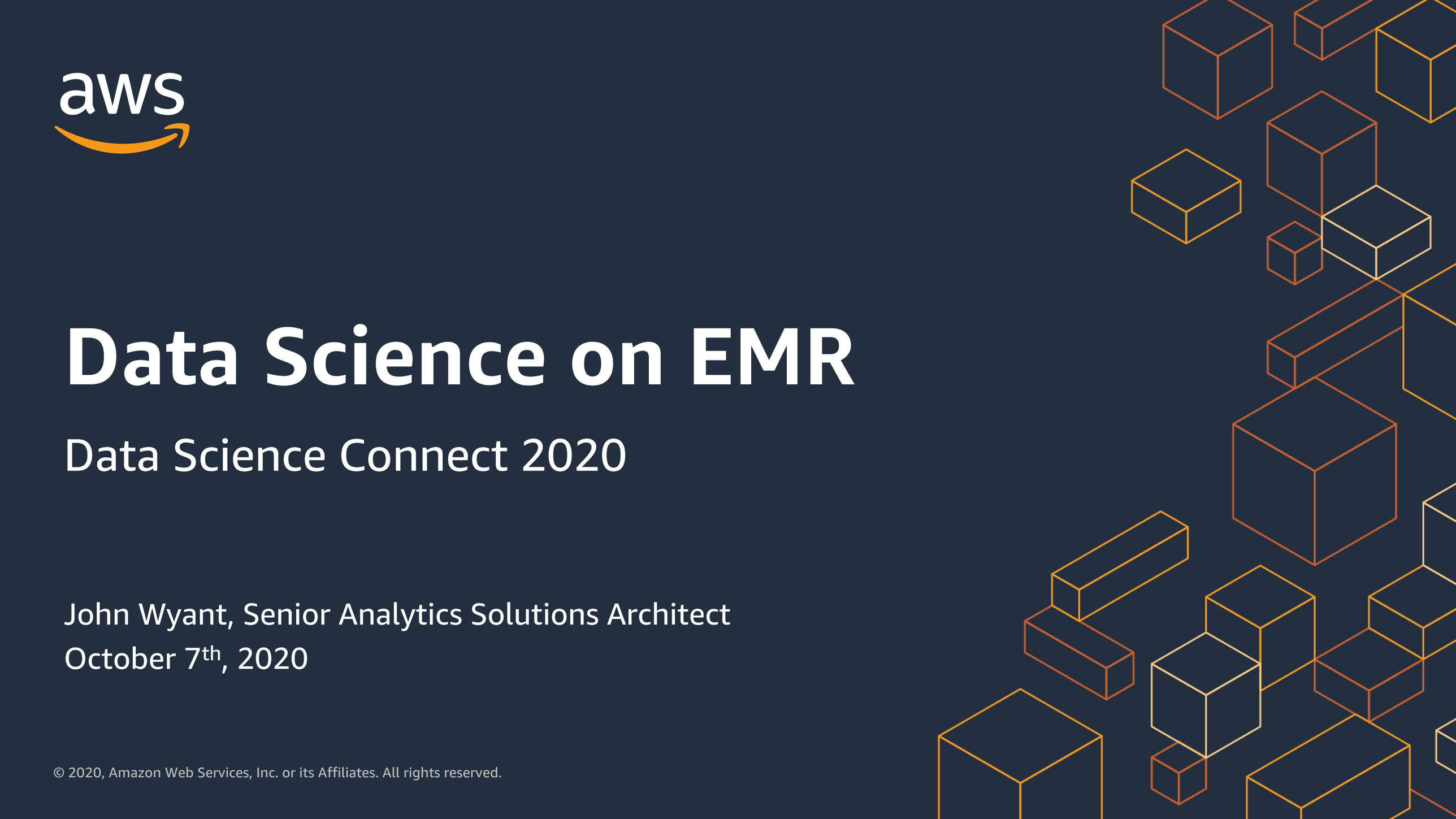


Data Science on EMR

Data Science Connect 2020

John Wyant, Senior Analytics Solutions Architect

October 7th, 2020

Table of contents

Most comprehensive

Broadest and deepest portfolio, purpose-built for builders

Visualization & Machine Learning

- Dashboards
- Predictive Analytics

Analytics

- Data Warehousing
- Big Data Processing
- Serverless Data processing
- Interactive Query
- Operational Analytics
- Real time Analytics

Data Lake Infrastructure & Management

- Infrastructure
- Security & Management
- Data Catalog & ETL

Data Movement

Migration & Streaming Services

aws

The broadest choice of analytics services

Discover

AWS Glue (Crawlers)

Ingest

AWS Glue

Kinesis Streams & Data Firehose

Database Migration Service (DMS)

Snowball

Snowmobile

Direct Connect

Managed Streaming for Kafka (MSK)

Store

S3

S3 Glacier

RDS

Aurora

DynamoDB

Secure

Identity and Access Management (IAM)

Key Management Service (KMS)

Macie

Cloudtrail

Cloudwatch

Catalog

AWS Glue (Data Catalog)

EMR (Hive Metastore)

Prepare

AWS Glue ETL (Serverless Apache Spark)

EMR (Apache Spark & Hadoop)

Analyze

Sagemaker

Redshift

Athena

Kinesis Data Analytics

EMR (Apache Spark & Hadoop)

Elasticsearch

AI services

Visualize

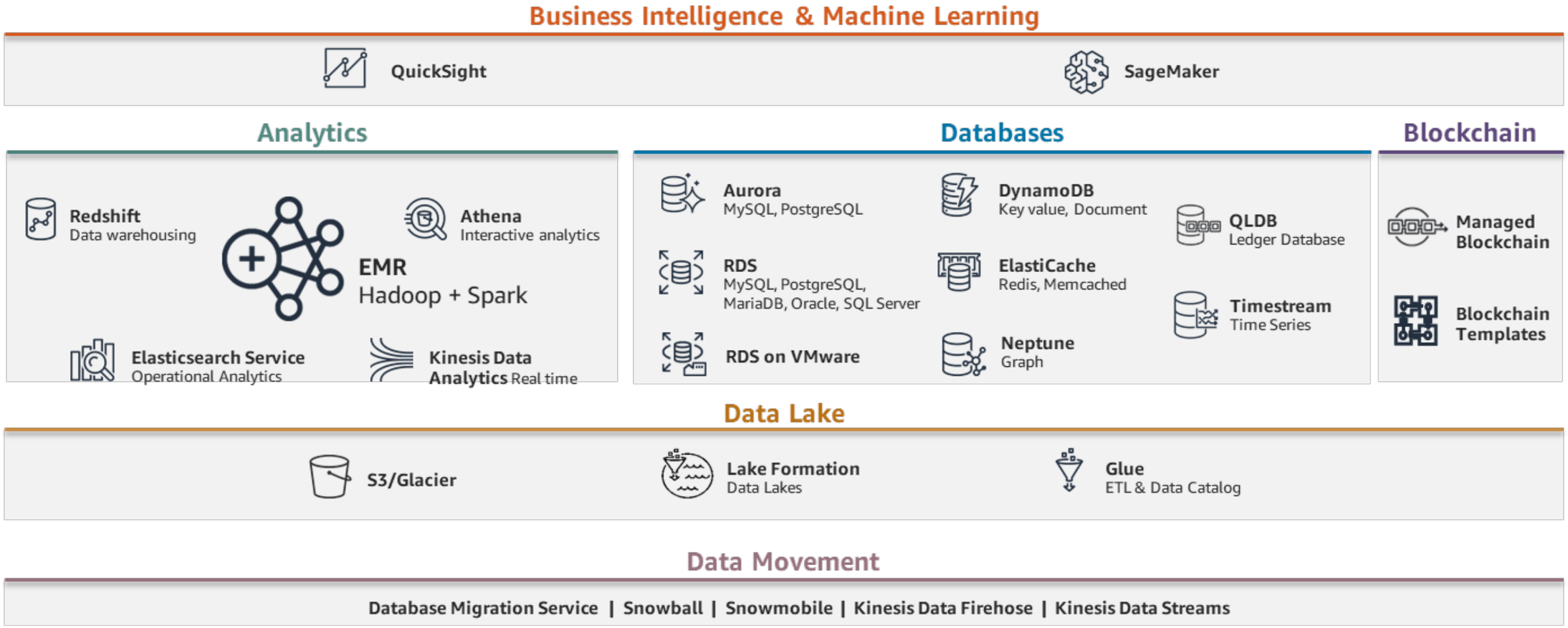
Quicksight

EMR Notebooks

aws

What is EMR & where is it in the Analytics stack?

Amazon EMR is an enterprise-grade Spark and Hadoop managed service empowering businesses, researchers, data analysts, and developers to easily process and analyze vast amounts of data. EMR solves complex technical and business challenges such as clickstream and log analysis along with real-time and predictive analytics.



High impact results with Amazon EMR

EPIC GAMES FORTNITE	near real-time analytics for 140M players
salesforce DMP	scales 3,000 transient clusters on a daily basis
GE POWER	powers the Predix solution processing 1,000,000 data executions/day
asurion	achieves costs savings of 55% when compared to on-demand pricing and 40% savings when compared to Reserved Instances
Zillow	computes Zestimates on 100M +homes in hours instead of 1 day

aws

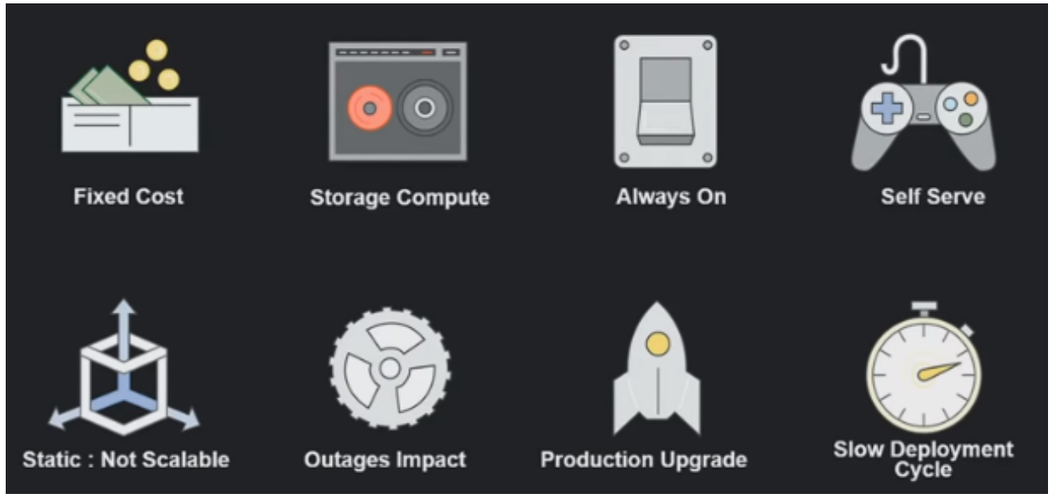
On-premises migrations to Amazon EMR

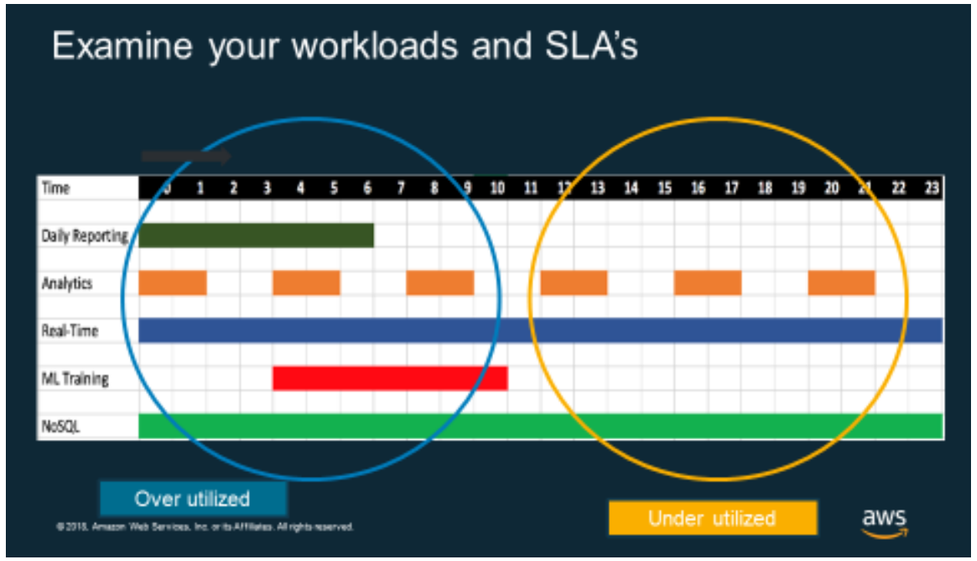
FINRA	Processes 135B events/day and have cost savings of 60% (~$20M)
Vanguard	decreased costs by $600k in less than 5 months
airbnb	reduced cost of operation and improved Spark performance 3x
Aol.	saves 75% and is 60% more efficient
Guardian	re-architects 1 monolithic pipeline into 3 purpose built clusters

aws

Common big data challenges

1. Fixed costs on-premises
2. Couples storage & compute
3. Always-on is inefficient
4. Lack of self-service choice
5. Static, not-scalable, & unused capacity
6. Outages impact
7. Production upgrades are difficult
8. Slow deployment cycles





Example EMR customers for Hadoop/Spark projects

Companies want more value from their data

Data is:

Growing exponentially

From new sources

Increasingly diverse

Used by many people

Analyzed by many applications

Complications:

Siloed approaches don’t work anymore

It’s too expensive and limiting to store data on-premises

Implication:

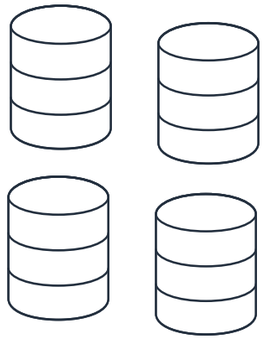
A new approach is needed to extract insights and value

aws

Traditional data infrastructures are impeding innovation

On-premises data infrastructures **do not scale** to meet variable and increasing volumes of data

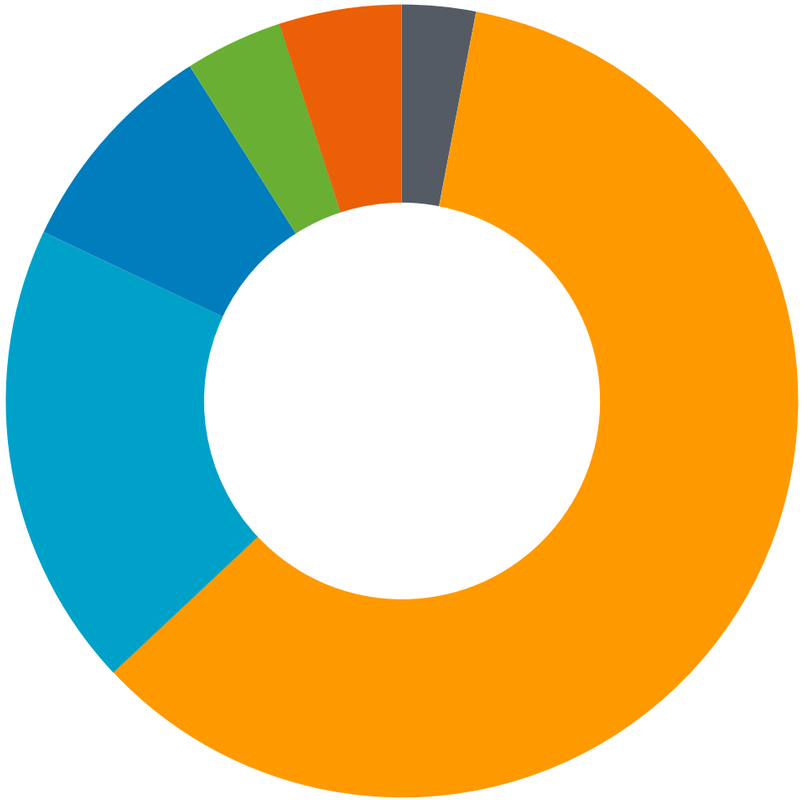
Rigid data schemas prevent access to source data and limit the use of advanced analytics and machine learning



Multiple disconnected data silos with inconsistent formats obscure data lineage and prevent a consolidated view of activity

The **high costs of legacy data warehouses** limit access to historical data

Data preparation accounts for ~80% of the work



- Building training sets
- Cleaning and organizing data
- Collecting data sets
- Mining data for patterns
- Refining algorithms
- Other

Data demands are driving next generation architectures for analytics and innovation

Structured data

Data that are highly normalized with common schema and stored in relational databases, powering transactional line-of-business applications

Semistructured data

Data that contain identifiers without conforming to a predefined schema

Unstructured data

Data that do not conform to a data model and are typically stored as individual files

Batch load

Extracts data from various data sources at periodic intervals and moves them to the data lake


Amazon EMR


AWS Glue

Streaming

Ingests data that are generated from multiple sources such as log files, telemetry, mobile applications, and social networks


Amazon Kinesis


Amazon MSK

Amazon S3 data lake

Cloud-scale centralized and scalable architecture that enables enterprise data science


Amazon S3

Analytics

Data Warehouses are repositories of normalized data and provide the foundational technology for BI


Amazon Redshift

And data stored in the data lake can also be made directly searchable and queryable


Amazon Athena


Amazon QuickSight

Machine Learning

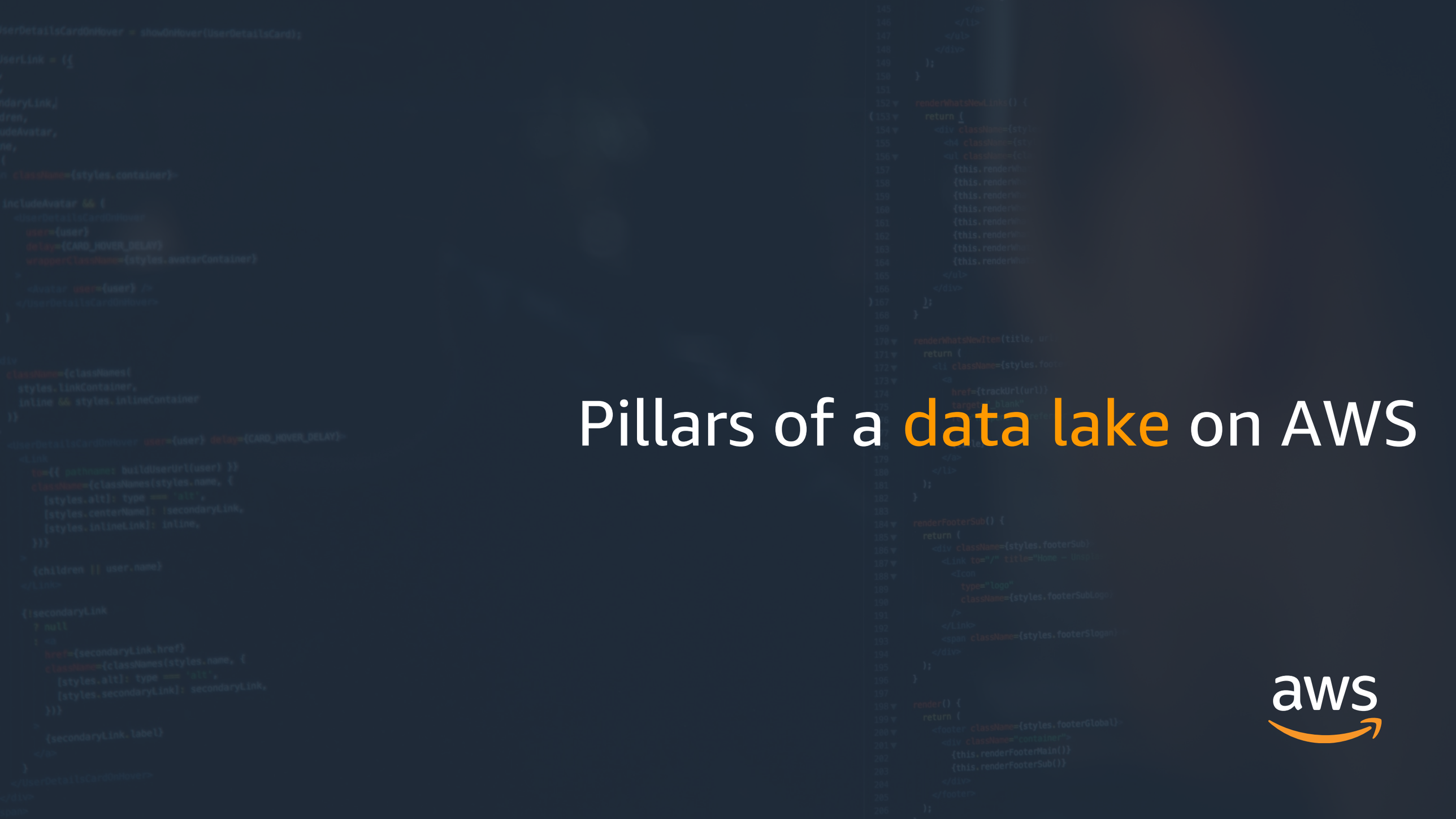
Storing data in an Amazon S3 data lake enables customers to leverage predictive or prescriptive analytics; perform ad-hoc analyses; and use AI/ML for automation and efficiency


Amazon EMR


Amazon SageMaker


AWS Deep Learning AMIs

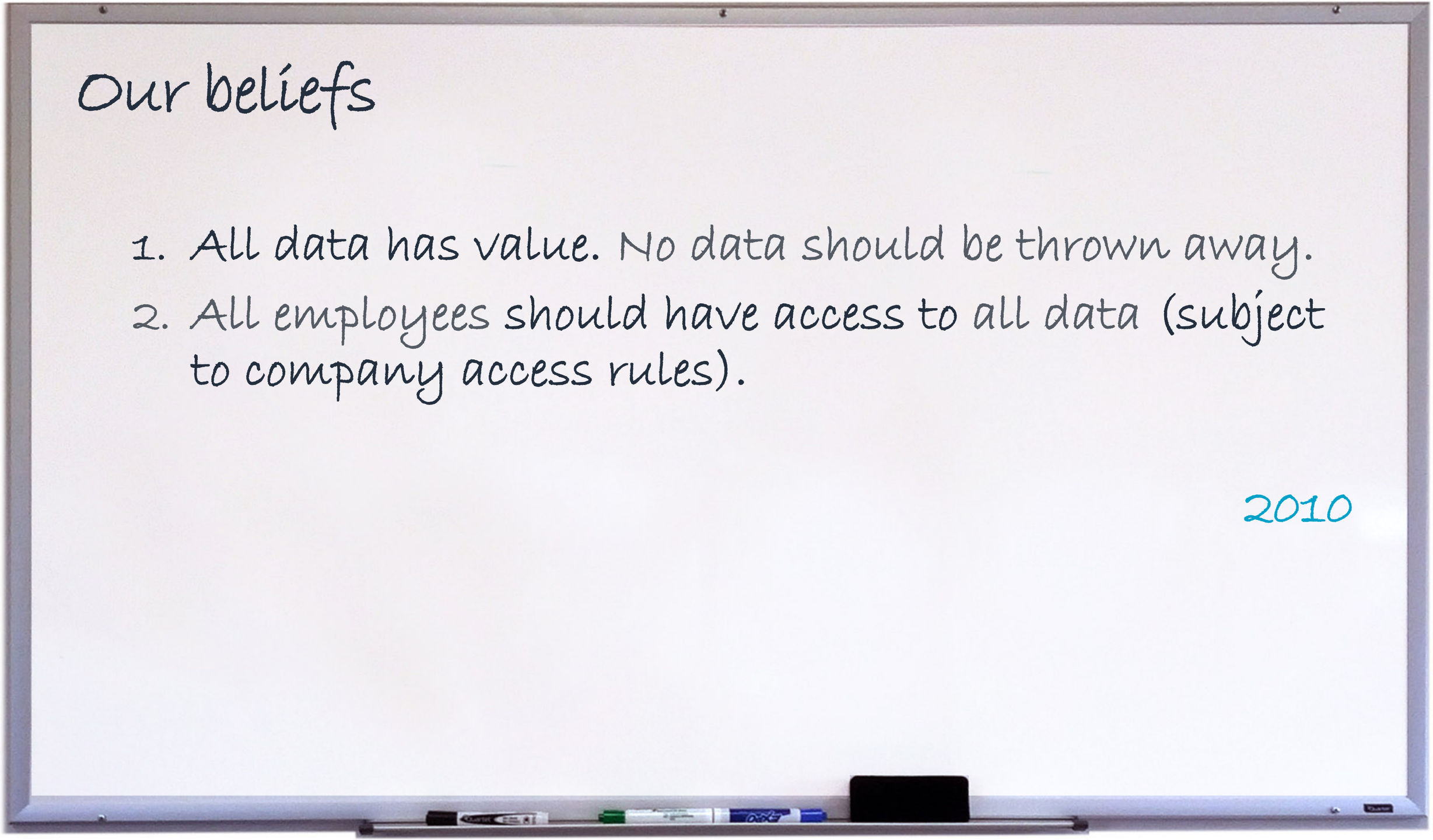
Pillars of a data lake on AWS



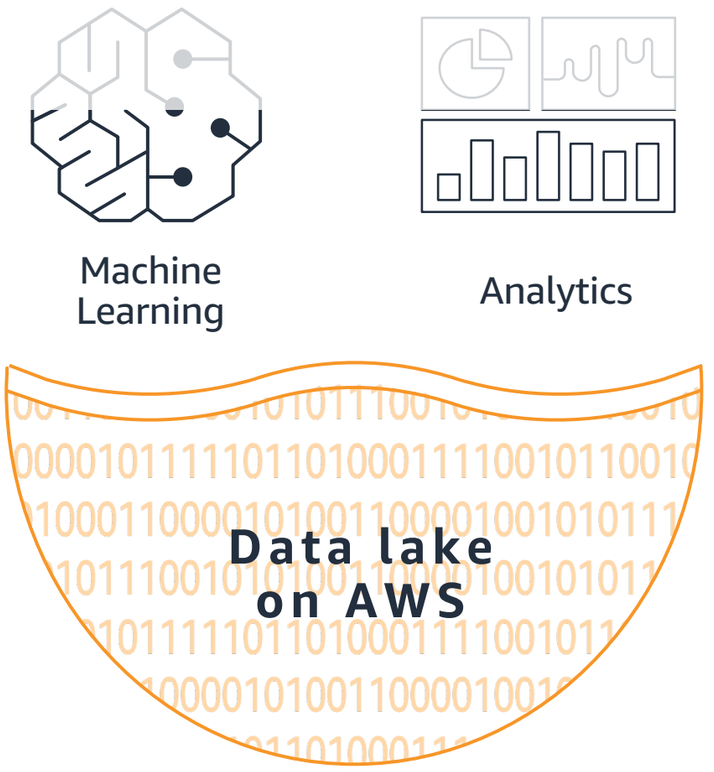
Our beliefs

1. All data has value. No data should be thrown away.
2. All employees should have access to all data (subject to company access rules).

2010



Step 1: Data movement



Data Movement

The first step to building data lakes on AWS is to move data to the cloud. AWS makes data transfer simple by providing the widest range of options to transfer data to the cloud.

On-premises data movement

- AWS Direct Connect
- AWS Storage Gateway
- AWS Snowball
- AWS Snowmobile

Real-time data movement

- AWS IoT Core
- Amazon Kinesis Data Firehose
- Amazon Kinesis Data Streams
- Amazon Kinesis Video Streams

Step 2: Data lake

Machine Learning

Analytics

On-premises Data Movement

Real-time Data Movement

Data lake

Once data is ready for the cloud, AWS makes it easy to store in any format, securely and at massive scale with Amazon S3 and Amazon Glacier. To make it easy for end users to discover the relevant data to use in their analysis, AWS Glue automatically creates a single searchable and queryable catalog.

Object Storage

Amazon S3

Amazon S3 Glacier Deep Archive

Amazon S3 Object Lock

Amazon S3 Intelligent Tiering

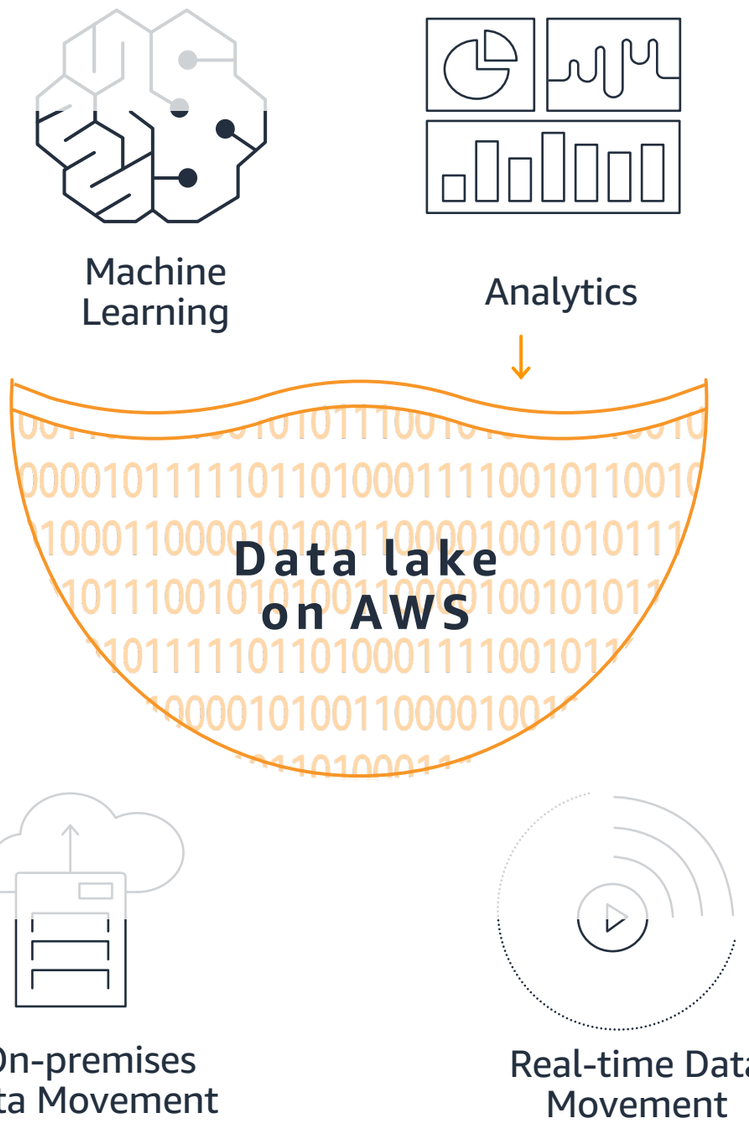
Backup and Archive

Amazon Glacier

Data Catalog

AWS Glue

Step 3: Perform Analytics



Analytics

AWS provides the broadest, and most cost-effective set of analytic services that run for data lakes.

Interactive analytics

Amazon Athena

Big data processing

Amazon EMR

Data warehousing

Amazon Redshift

Real-time analytics

Amazon Kinesis

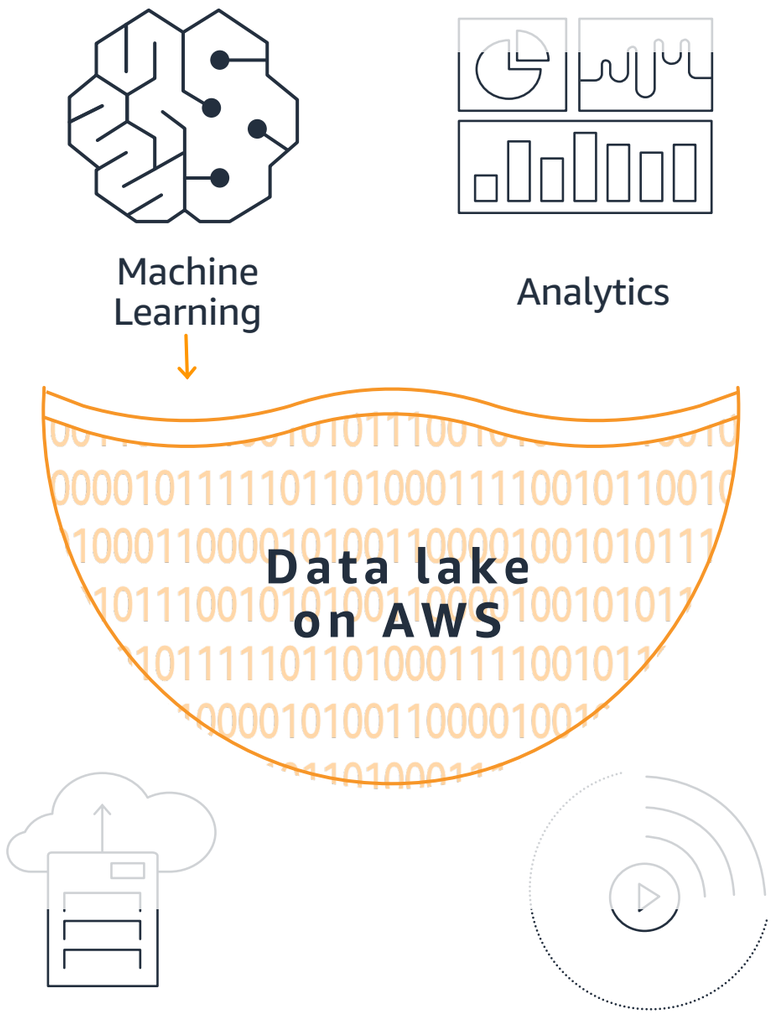
Operational analytics

Amazon Elasticsearch Service

Dashboards and visualizations

Amazon QuickSight

Step 4: Machine learning



Machine learning

For predictive analytics use cases, AWS provides a broad set of machine learning services, and tools that run on your AWS data lake.

Frameworks and interfaces

Amazon Deep Learning AMIs

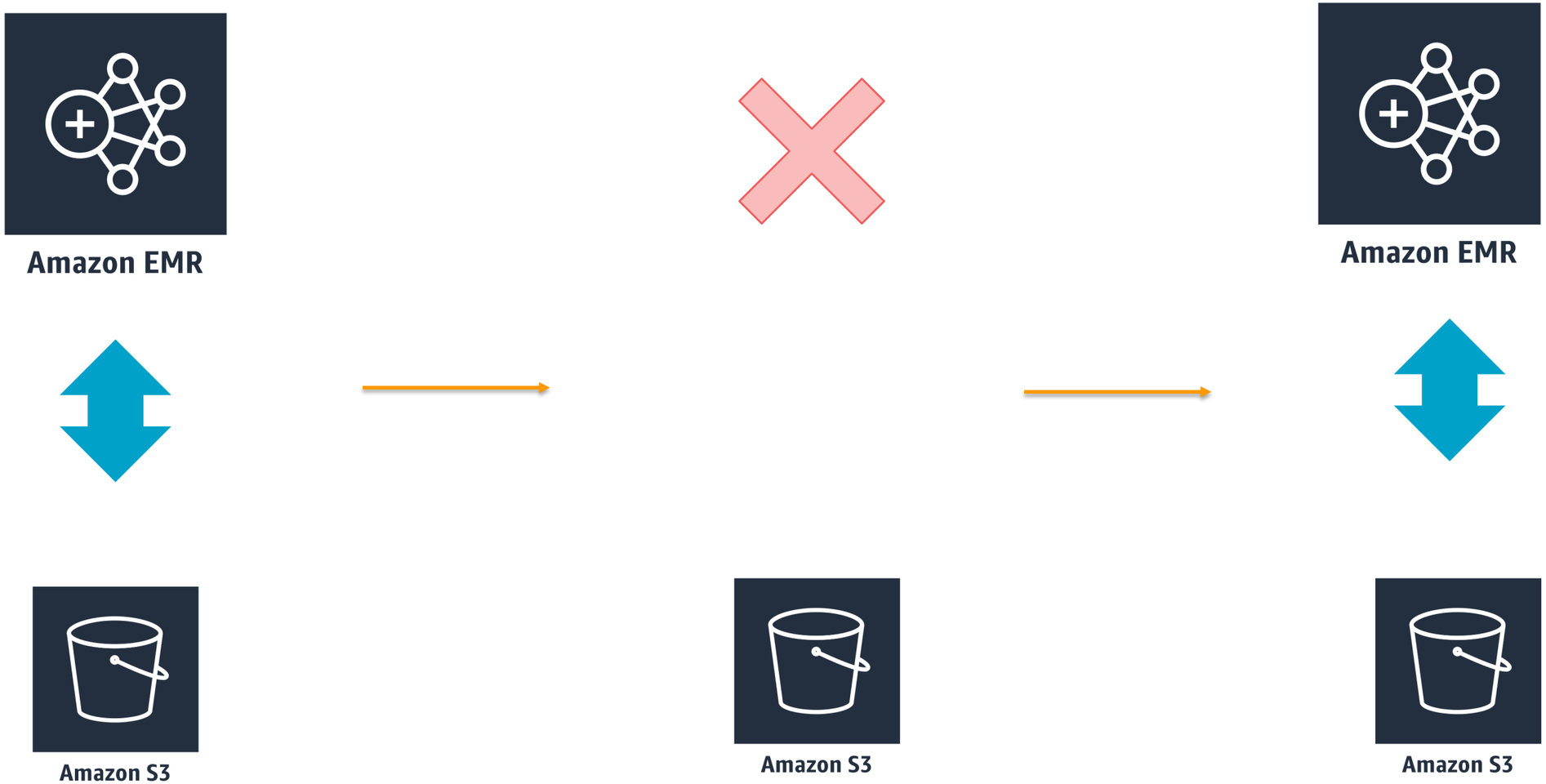
Platform services

Amazon SageMaker

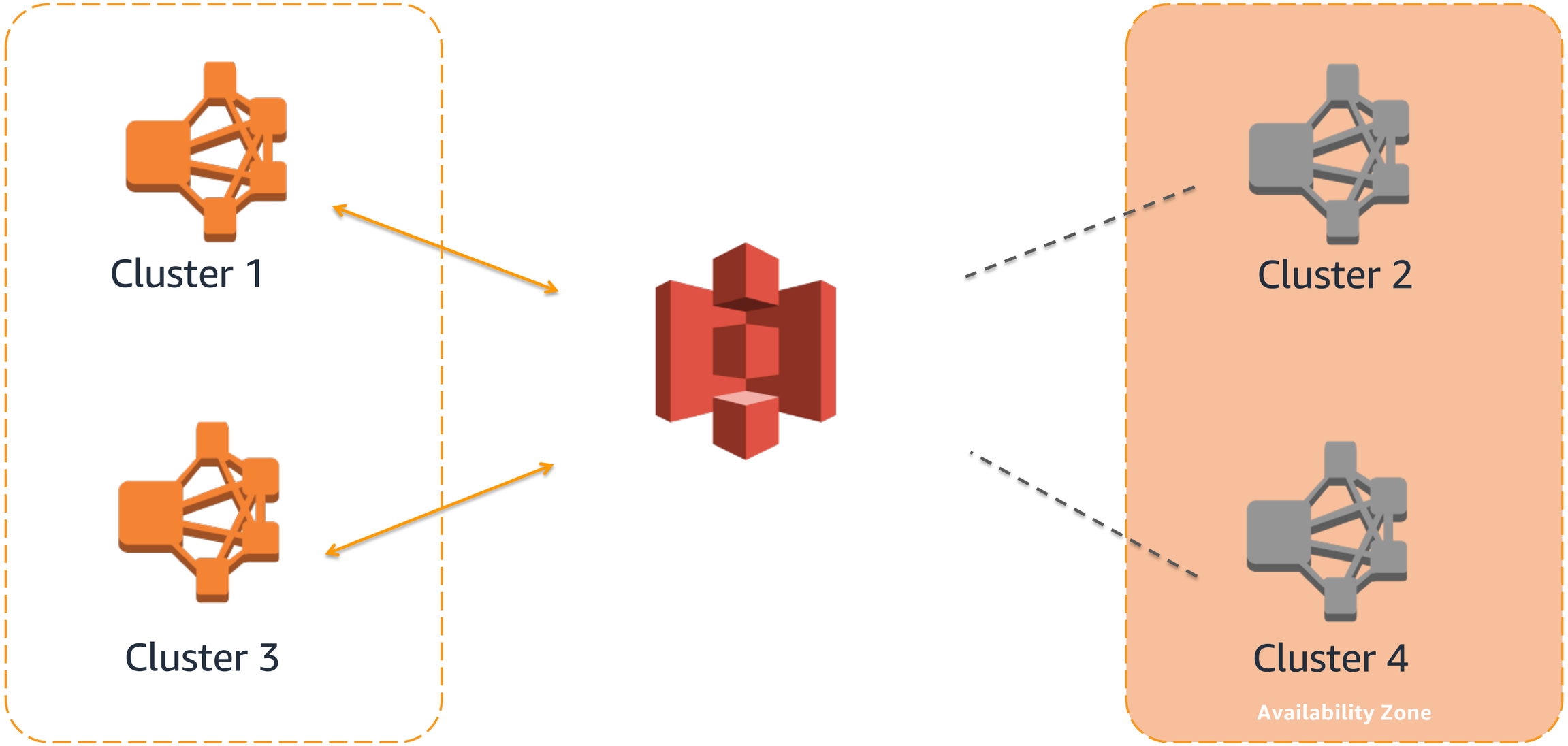
Application services

AWS provides solution-oriented APIs for computer vision and natural language processing

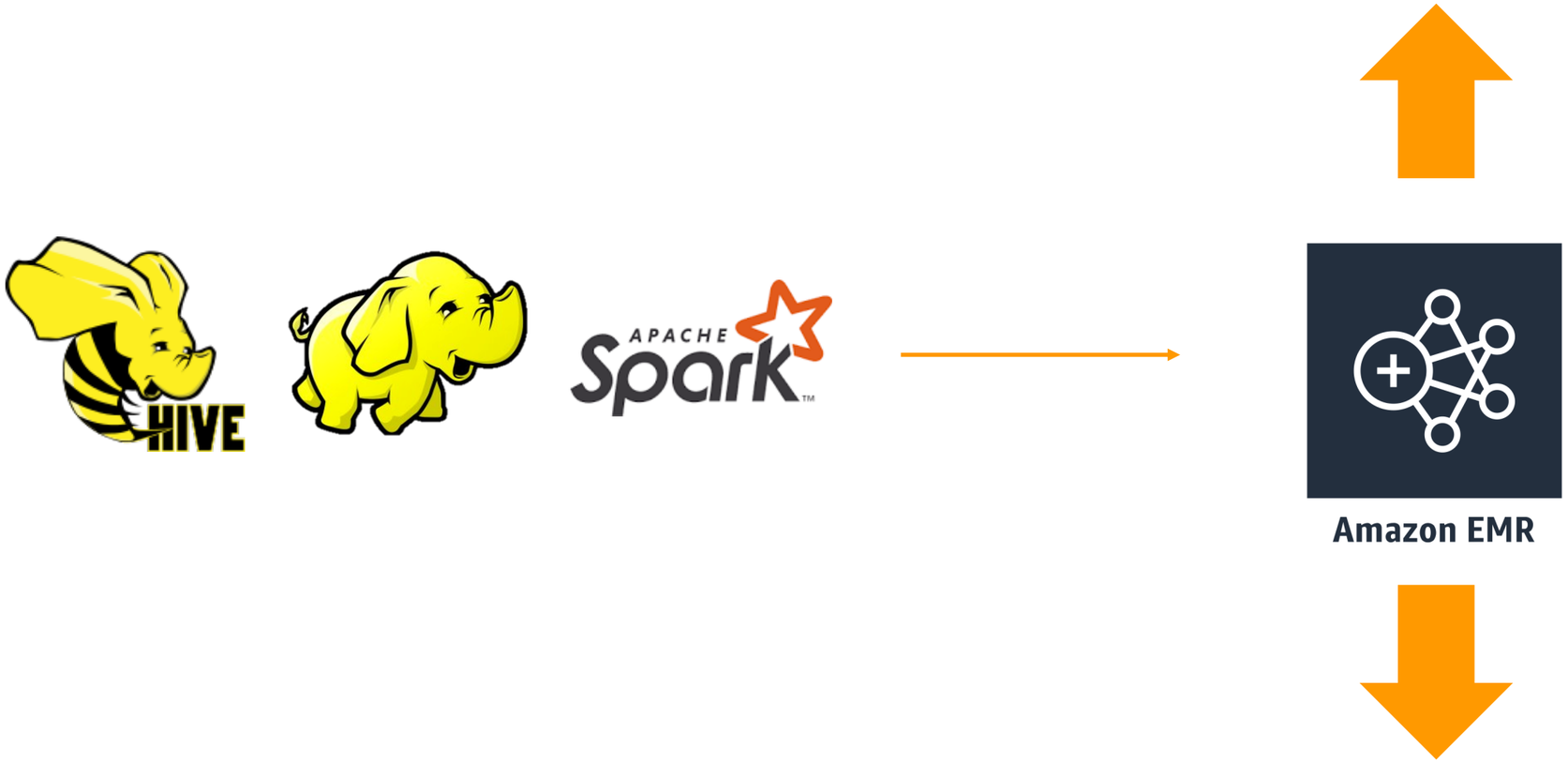
Benefit 1: Turn off clusters



Benefit 2: Built-in Disaster Recovery



Benefit 3: Auto-scaling Persistent & Transient Clusters



aws

NEW! EMR Managed Resize

Automatically resizes your EMR cluster for **best performance** at the **lowest possible cost**

Simplify - No metric specific auto-scaling policies

Reduce Cost – Up to 60% cost reduction compared to fixed-size clusters

Responsive – Evaluated every 5-10 seconds

Control –

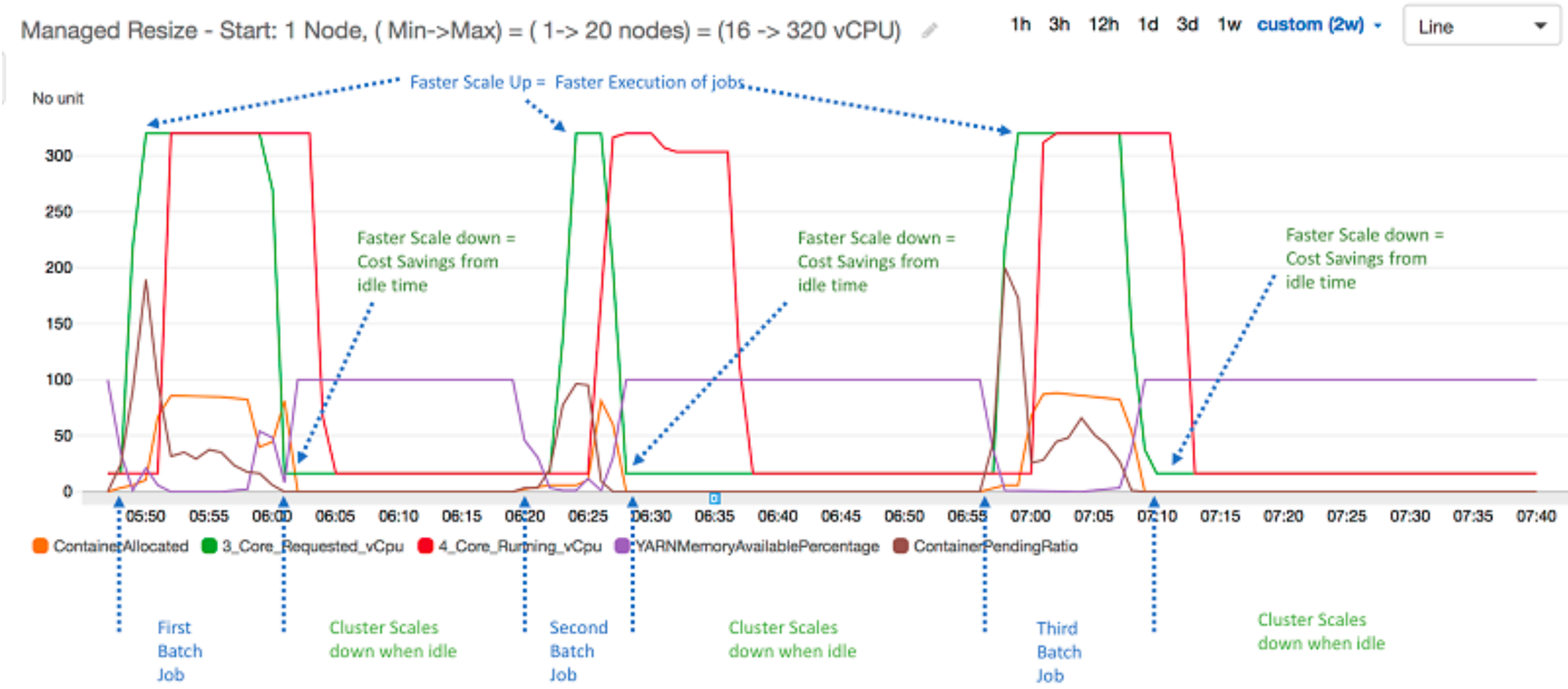
- Set minimum and maximum capacity limits
- Control Task/Core Node split
- Control On-Demand/Spot Node split

Amazon EMR Managed Scaling is available on Apache Spark, Apache Hive, and YARN-based workloads on **Amazon EMR version 5.30.1** and above.

aws

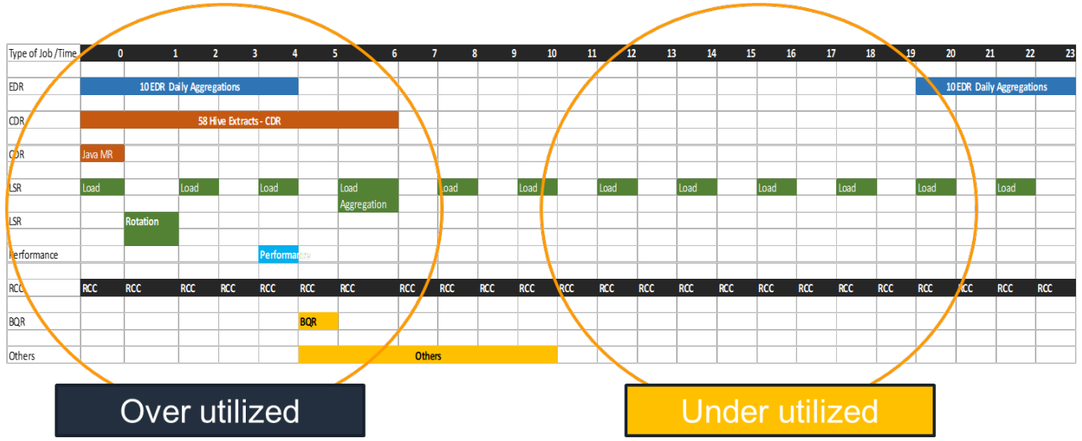
Managed Resize – Example Scenario

Scenario: Multiple parallel Spark jobs (from the TPC-DS benchmark) to the cluster at 30-minute intervals

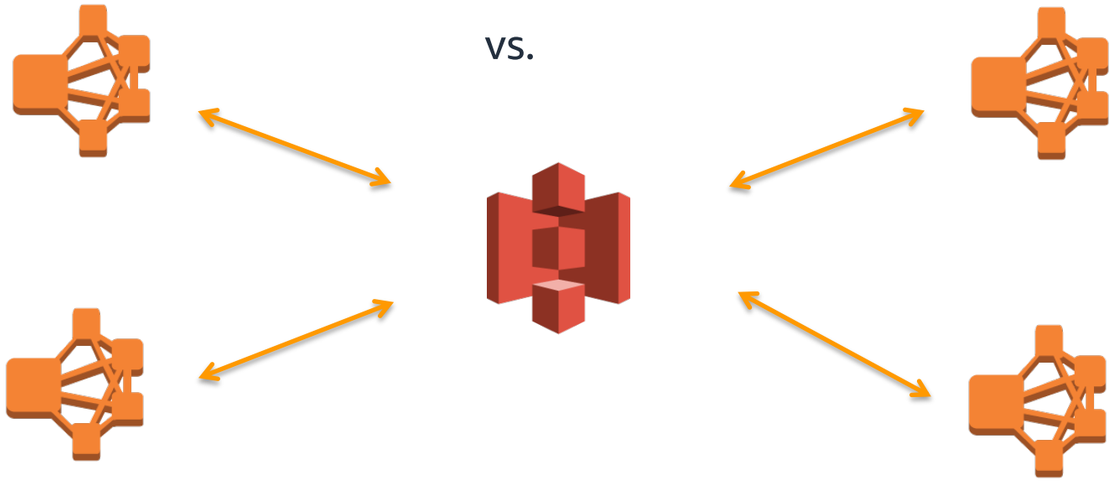


Benefit 4: Logical separation of jobs/applications

Traditional Monolithic Cluster



Purpose-built Clusters



Re-architect Monolithic to Purpose-built clusters by:

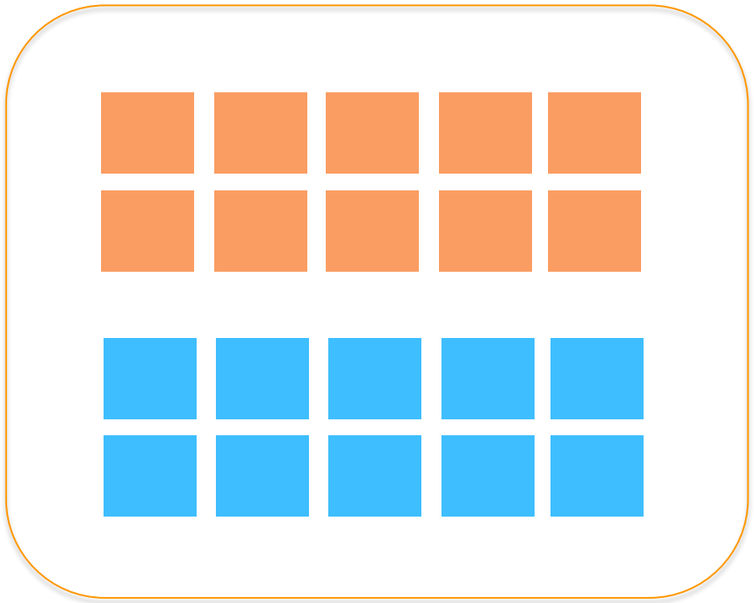
- Creating Transient and/or Persistent clusters
- Separating clusters by Application
- Separating clusters by Application Version
- Isolating Department specific clusters

Design consideration are given to:

- How do you submit jobs or build pipelines
- Persisting your data in S3
- Storing metadata off the cluster
- How long does the job run
- What applications are needed

Benefit 5: Auto-scale nodes with Spot instances

Parallelization on Spot can drastically reduce time-to-insight and cost.



Results:

50% less run-time (14hrs → 7hrs)

25% less cost ($140 → $105)

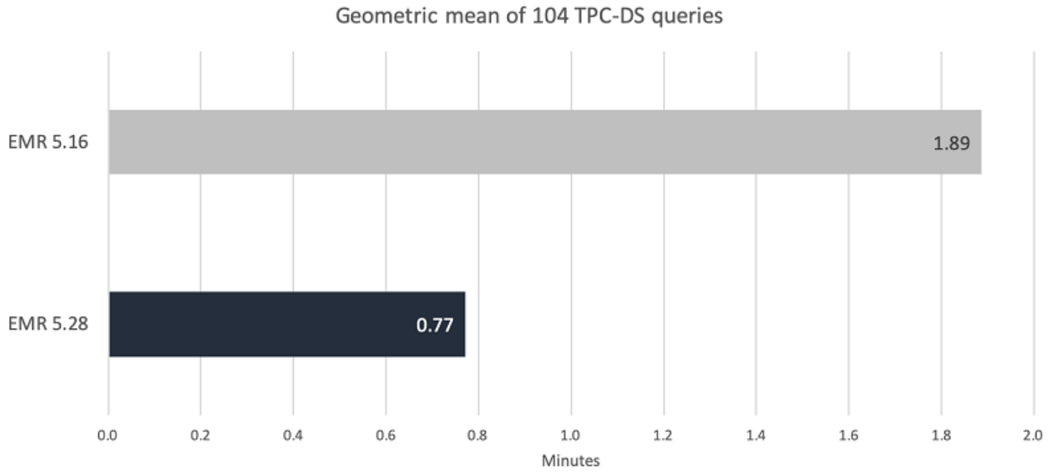
aws

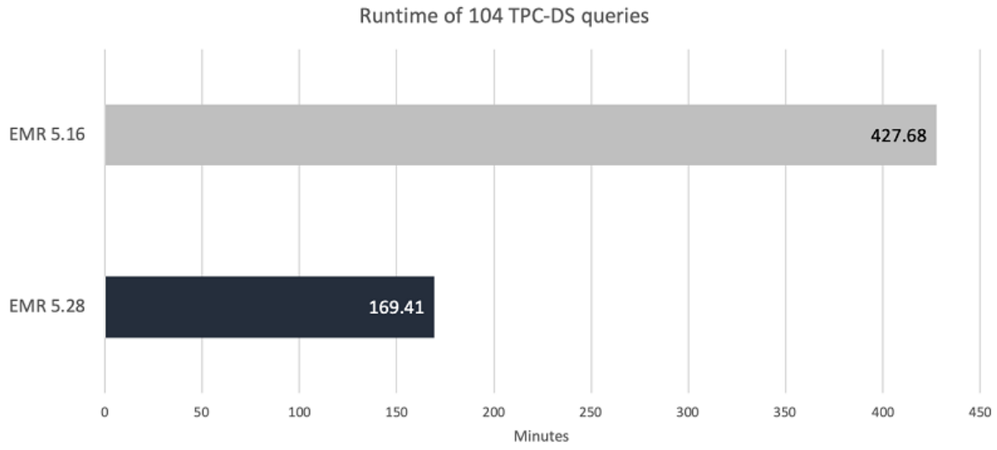
Benefit 6: EMR Self-service with AWS Service Catalog

Configure		Consume	
	Standardize		Developer Autonomy
	Enforce Consistency and Compliance		One-Stop Shop
	Limit Access		Automate Deployments
	Enforce Tagging, Security Groups		Agile Governance

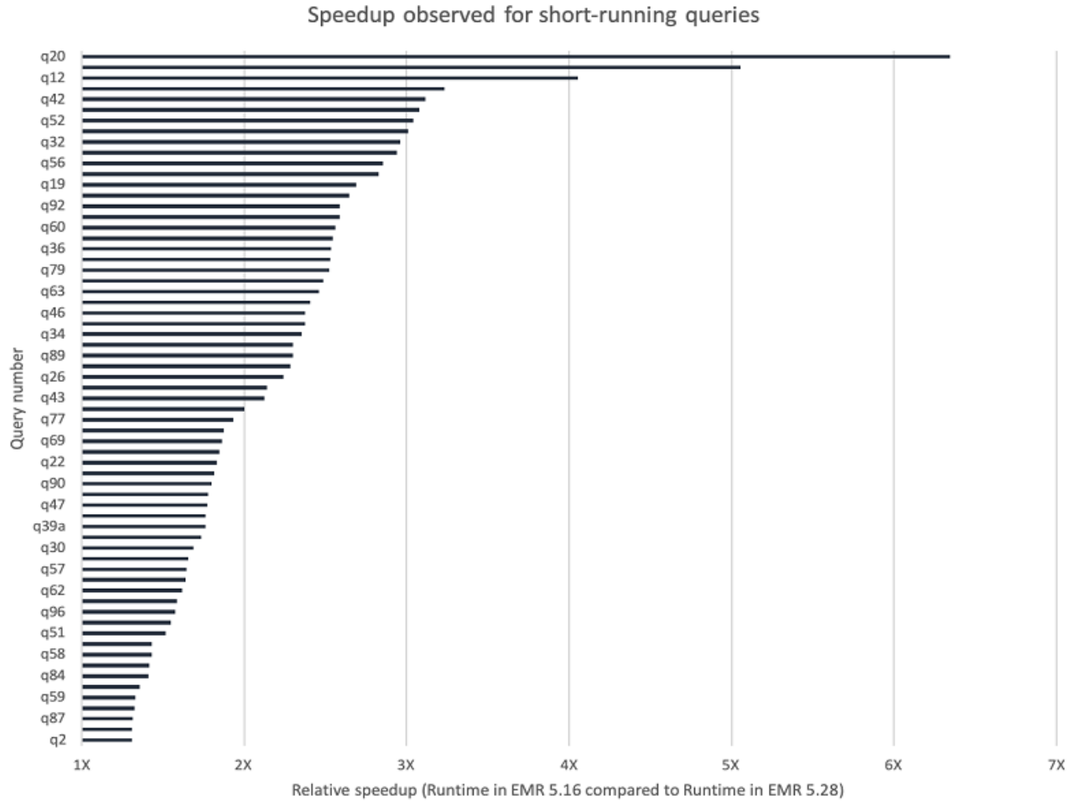
aws

Benefit 7: Spark Performance Improvements



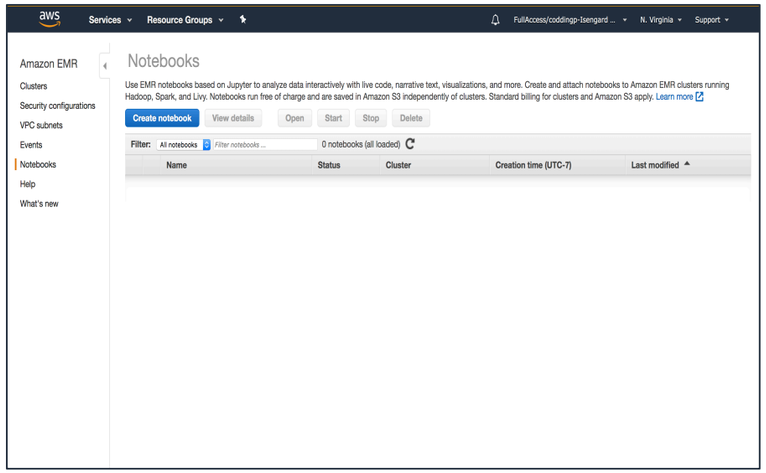


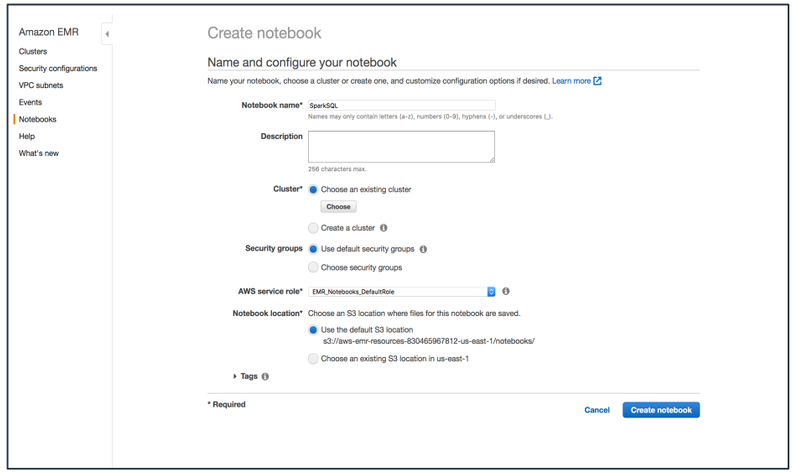
1. The EMR Runtime for Apache Spark available in EMR v5.28 realized Spark improvements of up to 32x against TPC-DS 3TB dataset in comparison to EMR v5.16. LINK to blog.
2. The EMR Runtime for Apache Spark maintains API compatibility with OSS Spark.
3. More coming every release.

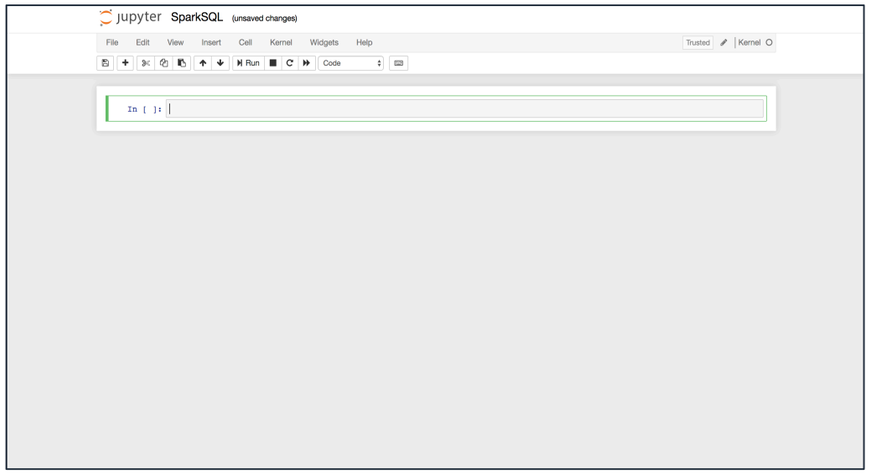


Benefit 8: Fully Managed EMR Notebooks

1. Provide an end to end data engineering and data science using EMR Notebooks which is based on the popular open source Jupyter Notebooks to build applications with Apache Spark
2. Attach / Detach from individual clusters; automatically backed up to S3
3. Tag-based Permissions
4. Support for PySpark, Spark SQL, Spark R, and Scala
5. NEW features include a visual experience to debug and monitor Spark jobs directly into the off-cluster, persistent, Apache Spark History Server using the EMR Console, associate Git repositories such as GitHub and Bitbucket, and compare and merge two different notebooks using the nbdime utility.



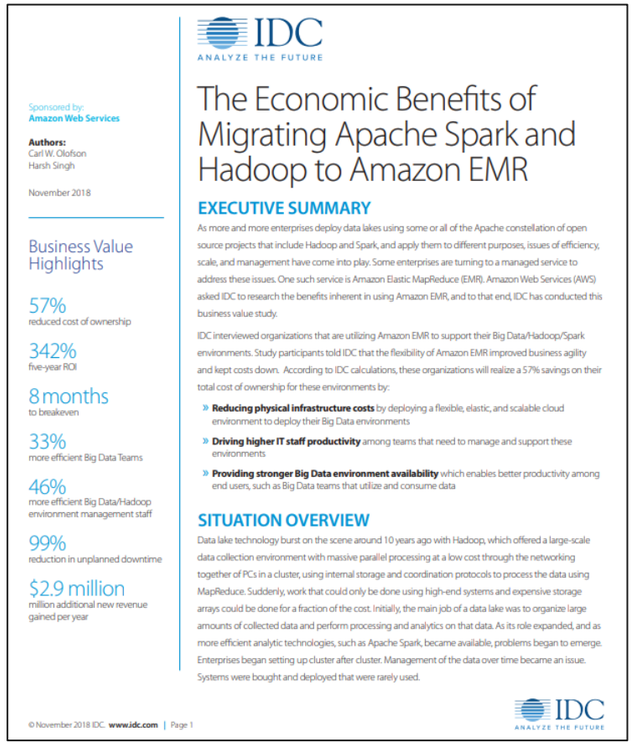




Benefit 9: Analysts confirm Lowest TCO in the Industry

Nov. 2018, **IDC** report confirms:

"EMR provides 57% reduced costs vs. on premise resulting in 342% ROI over 5 years."



IDC
ANALYZE THE FUTURE

Sponsored by:
Amazon Web Services

Authors:
Carl W. Olofson
Harsh Singh

November 2018

Business Value Highlights

57%
reduced cost of ownership

342%
five-year ROI

8 months
to breakeven

33%
more efficient Big Data Teams

46%
more efficient Big Data/Hadoop environment management staff

99%
reduction in unplanned downtime

$2.9 million
million additional new revenue gained per year

The Economic Benefits of Migrating Apache Spark and Hadoop to Amazon EMR

EXECUTIVE SUMMARY

As more and more enterprises deploy data lakes using some or all of the Apache constellation of open source projects that include Hadoop and Spark, and apply them to different purposes, issues of efficiency, scale, and management have come into play. Some enterprises are turning to a managed service to address these issues. One such service is Amazon Elastic MapReduce (EMR). Amazon Web Services (AWS) asked IDC to research the benefits inherent in using Amazon EMR, and to that end, IDC has conducted this business value study.

IDC interviewed organizations that are utilizing Amazon EMR to support their Big Data/Hadoop/Spark environments. Study participants told IDC that the flexibility of Amazon EMR improved business agility and kept costs down. According to IDC calculations, these organizations will realize a 57% savings on their total cost of ownership for these environments by:

- **Reducing physical infrastructure costs** by deploying a flexible, elastic, and scalable cloud environment to deploy their Big Data environments
- **Driving higher IT staff productivity** among teams that need to manage and support these environments
- **Providing stronger Big Data environment availability** which enables better productivity among end users, such as Big Data teams that utilize and consume data

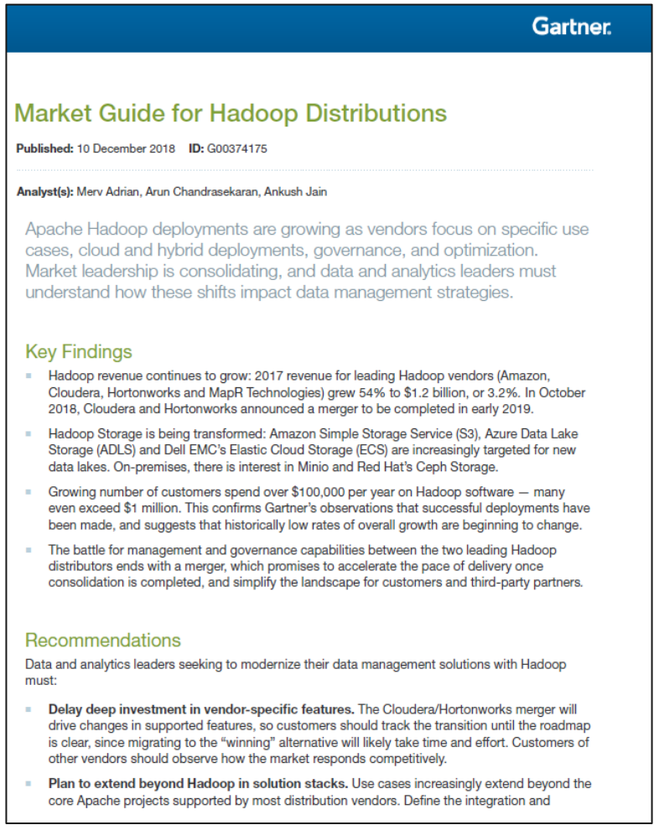
SITUATION OVERVIEW

Data lake technology burst on the scene around 10 years ago with Hadoop, which offered a large-scale data collection environment with massive parallel processing at a low cost through the networking together of PCs in a cluster, using internal storage and coordination protocols to process the data using MapReduce. Suddenly, work that could only be done using high-end systems and expensive storage arrays could be done for a fraction of the cost. Initially, the main job of a data lake was to organize large amounts of collected data and perform processing and analytics on that data. As its role expanded, and as more efficient analytic technologies, such as Apache Spark, became available, problems began to emerge. Enterprises began setting up cluster after cluster. Management of the data over time became an issue. Systems were bought and deployed that were rarely used.

IDC
ANALYZE THE FUTURE

Dec. 2018, **Gartner** suggests:

"AWS remains the largest Hadoop provider in terms of both revenue and user base."



Gartner.

Market Guide for Hadoop Distributions

Published: 10 December 2018 **ID:** G00374175

Analyst(s): Merv Adrian, Arun Chandrasekaran, Ankush Jain

Apache Hadoop deployments are growing as vendors focus on specific use cases, cloud and hybrid deployments, governance, and optimization. Market leadership is consolidating, and data and analytics leaders must understand how these shifts impact data management strategies.

Key Findings

- Hadoop revenue continues to grow: 2017 revenue for leading Hadoop vendors (Amazon, Cloudera, Hortonworks and MapR Technologies) grew 54% to $1.2 billion, or 3.2%. In October 2018, Cloudera and Hortonworks announced a merger to be completed in early 2019.
- Hadoop Storage is being transformed: Amazon Simple Storage Service (S3), Azure Data Lake Storage (ADLS) and Dell EMC's Elastic Cloud Storage (ECS) are increasingly targeted for new data lakes. On-premises, there is interest in Minio and Red Hat's Ceph Storage.
- Growing number of customers spend over $100,000 per year on Hadoop software — many even exceed $1 million. This confirms Gartner's observations that successful deployments have been made, and suggests that historically low rates of overall growth are beginning to change.
- The battle for management and governance capabilities between the two leading Hadoop distributors ends with a merger, which promises to accelerate the pace of delivery once consolidation is completed, and simplify the landscape for customers and third-party partners.

Recommendations

Data and analytics leaders seeking to modernize their data management solutions with Hadoop must:

- **Delay deep investment in vendor-specific features.** The Cloudera/Hortonworks merger will drive changes in supported features, so customers should track the transition until the roadmap is clear, since migrating to the "winning" alternative will likely take time and effort. Customers of other vendors should observe how the market responds competitively.
- **Plan to extend beyond Hadoop in solution stacks.** Use cases increasingly extend beyond the core Apache projects supported by most distribution vendors. Define the integration and

Feb. 2019, **Forrester** recognizes:

AWS EMR as the Cloud Hadoop/Spark (HARK) Leader.

The Forrester Wave™ is copyrighted by Forrester Research, Inc. Forrester and Forrester Wave™ are trademarks of Forrester Research, Inc. The Forrester Wave™ is a graphical representation of Forrester's call on a market and is plotted using a detailed spreadsheet with exposed scores, weightings, and comments. Forrester does not endorse any vendor, product, or service depicted in the Forrester Wave™. Information is based on best available resources. Opinions reflect judgment at the time and are subject to change.

Benefit 10: Leverage the broadest Analytic ecosystem

On-premises Data Movement

AWS Direct Connect
AWS Storage Gateway
AWS Snowball
AWS Snowmobile

Real-time Data Movement

AWS IoT Core
AWS Kinesis Firehose
AWS Kinesis Data Streams
AWS Kinesis Video Streams

Machine Learning

Amazon SageMaker
AWS Deep Learning AMIs
Amazon Rekognition
Amazon Lex
AWS DeepLens
Amazon Comprehend
Amazon Translate
Amazon Transcribe
Amazon Polly

Analytics

Amazon Athena
Amazon EMR
Amazon Redshift
Amazon Elasticsearch service
Amazon Kinesis
Amazon QuickSight

- Leverage S3 as the Data Lake
  - Data gravity to bring analytics to the data
  - Interoperability of data available in S3
  - Agility to use purpose-built services
- Analyze data with the broadest selection of analytics tools
  - Data warehousing
  - Interactive SQL queries
  - Big Data processing
  - Real-time analytics
  - Dashboards & Visualizations
  - Machine Learning
- Query in place without moving to a separate analytics system
- Up to 400% faster with S3 Select and Glacier Select
- Ensures you can meet existing and future use cases, minimizing risks
- Largest ISV ecosystem with built-in integration

aws

The Amazon ML Stack

Broadest and deepest set of capabilities

AI Services

VISION			SPEECH		LANGUAGE		CHATBOTS	FORECASTING	RECOMMENDATIONS
REKOGNITION IMAGE	REKOGNITION VIDEO	TEXTRACT	POLLY	TRANSCRIBE	TRANSLATE	COMPREHEND	LEX	FORECAST	PERSONALIZE

ML Services

Amazon SageMaker

Ground Truth	Notebooks	Algorithms + Marketplace	Reinforcement Learning	Training	Optimization	Deployment	Hosting

ML Frameworks + Infrastructure

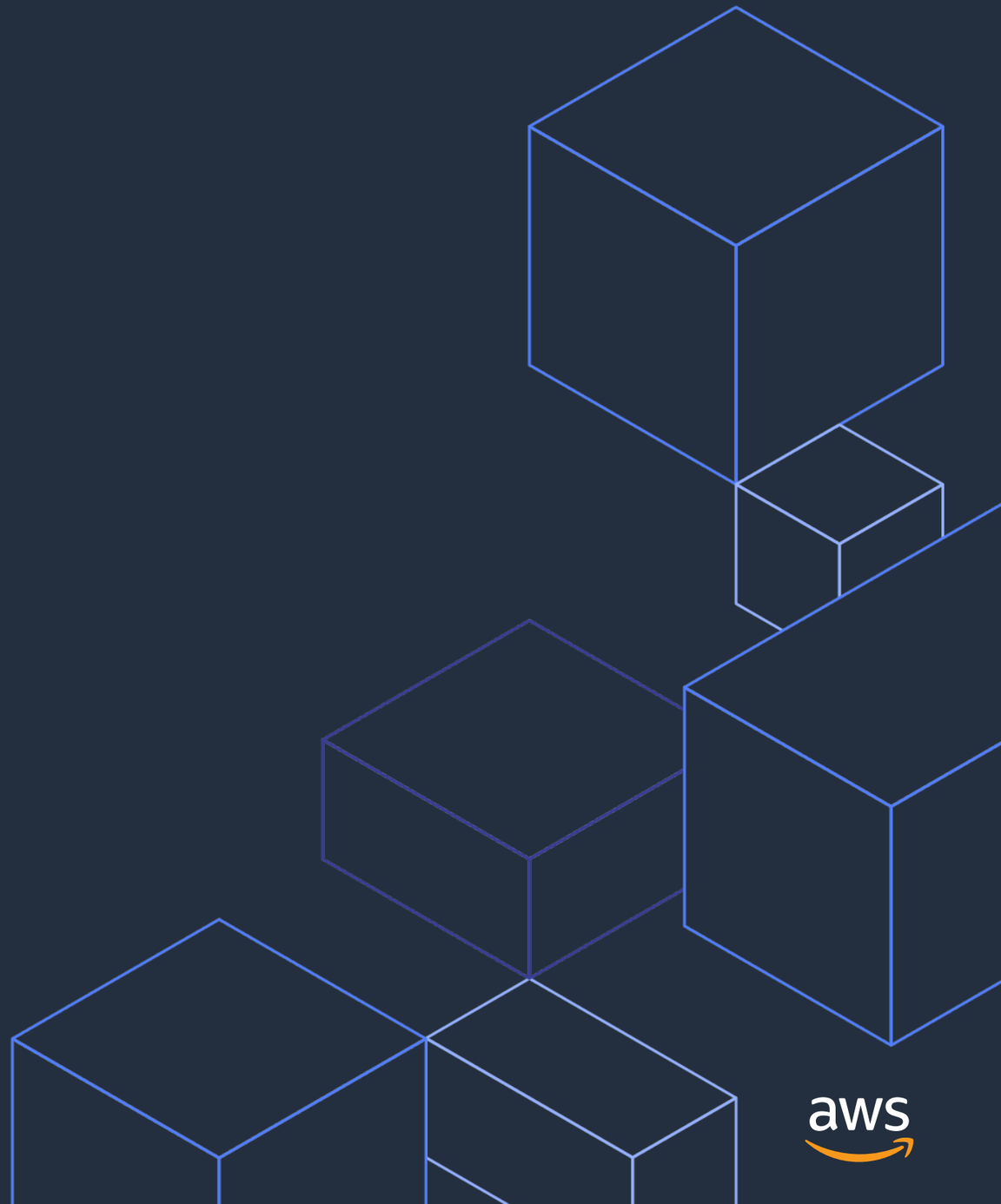
FRAMEWORKS	INTERFACES
TensorFlow, mxnet, PYTORCH	GLUON, Keras

INFRASTRUCTURE						
EC2 P3 & P3DN	EC2 G4	EC2 C5	FPGAS	GREENGRASS	ELASTIC INFERENCE	INFERENTIA

aws

Demo Time!

Running SageMaker from EMR Notebooks



Thank You!

jwyant@amazon.com

https://aws.amazon.com/big-data/datalakes-and-analytics/

Try it on your own: https://github.com/aws/amazon-sagemaker-examples/blob/master/advanced_functionality/working_with_redshift_data/working_with_redshift_data.ipynb