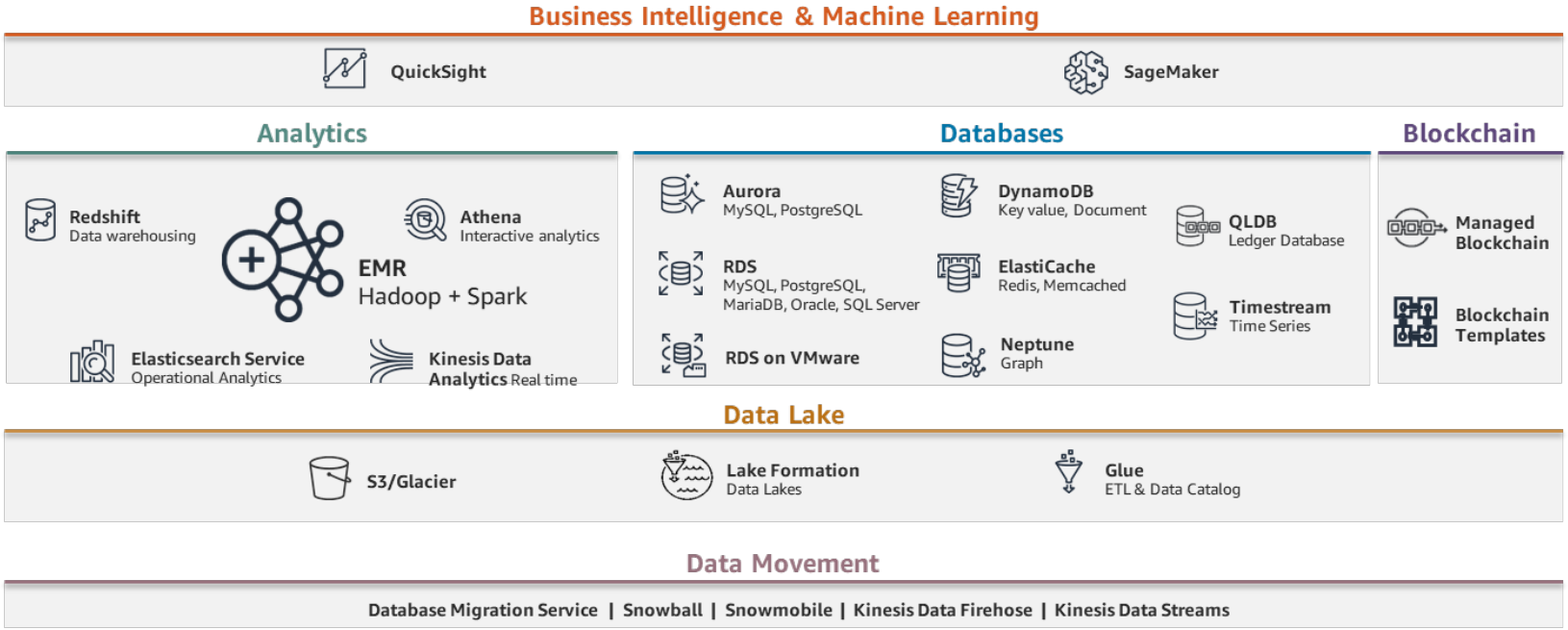
Introduction to Amazon EMR

Twann Atkins - Solutions Architect,
Mid-Atlantic Enterprise

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

near real-time analytics for 140M players

scales 3,000 transient clusters on a daily basis

powers the Predix solution processing 1,000,000 data executions/day

achieves costs savings of 55% when compared to on-demand pricing and 40% savings when compared to Reserved Instances

computes Zestimates on 100M +homes in hours instead of 1 day

aws

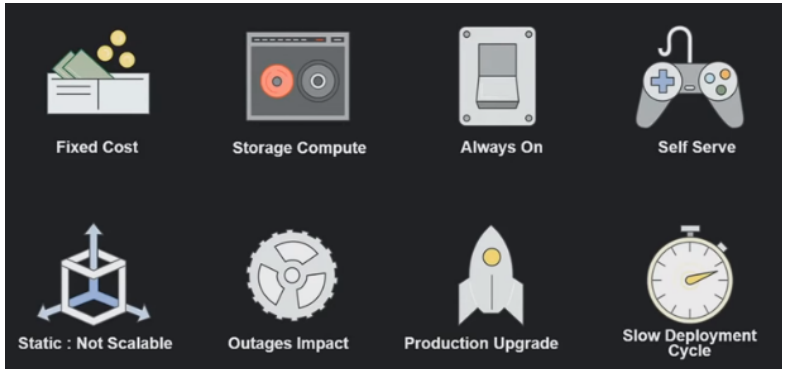
On-premises migrations to Amazon EMR

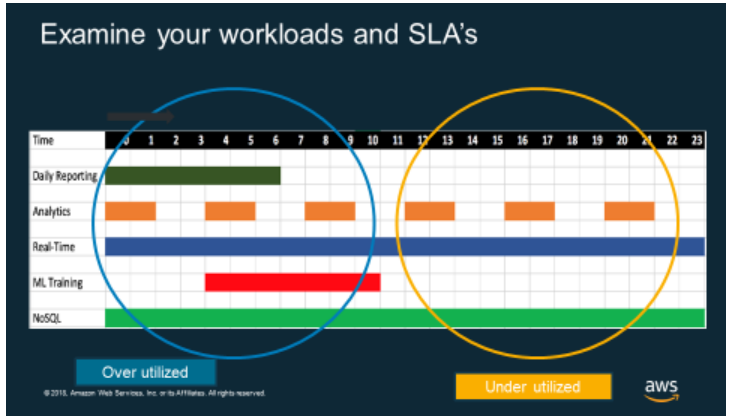
FINRA	Processes 135B events/day and have cost savings of 60% (~$20M)
Vanguard	decreased costs by $600k in less than 5 months
airbnb	reduced cost of operation and improved Spark performance 3x
Aol.	saves 75% and is 60% more efficient
Guardian	re-architects 1 monolithic pipeline into 3 purpose built clusters

aws

Common big data challenges

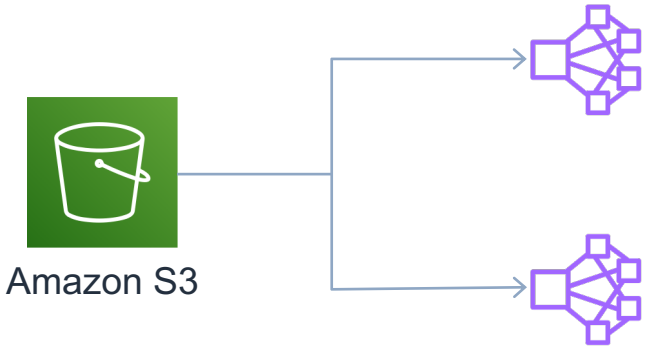
1. Fixed costs on-premises
2. Couples storage & compute
3. Always-on is inefficient
4. Lack of self-service choice
5. Static, not-scalable, & unused capacity
6. Outages impact
7. Production upgrades are difficult
8. Slow deployment cycles





aws

Cluster Types

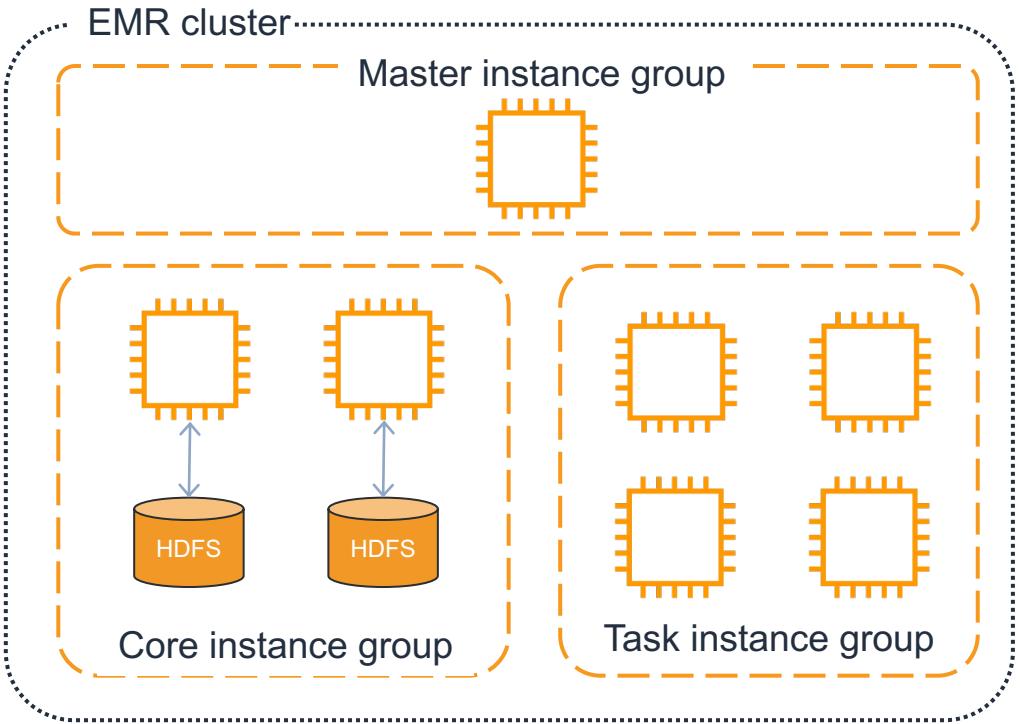


Persistent Cluster – Interactive Queries
(Spark-SQL | Presto)

Transient Cluster - Batch Jobs
(X hours nightly) – Add/Remove Nodes

aws

EMR Node Types



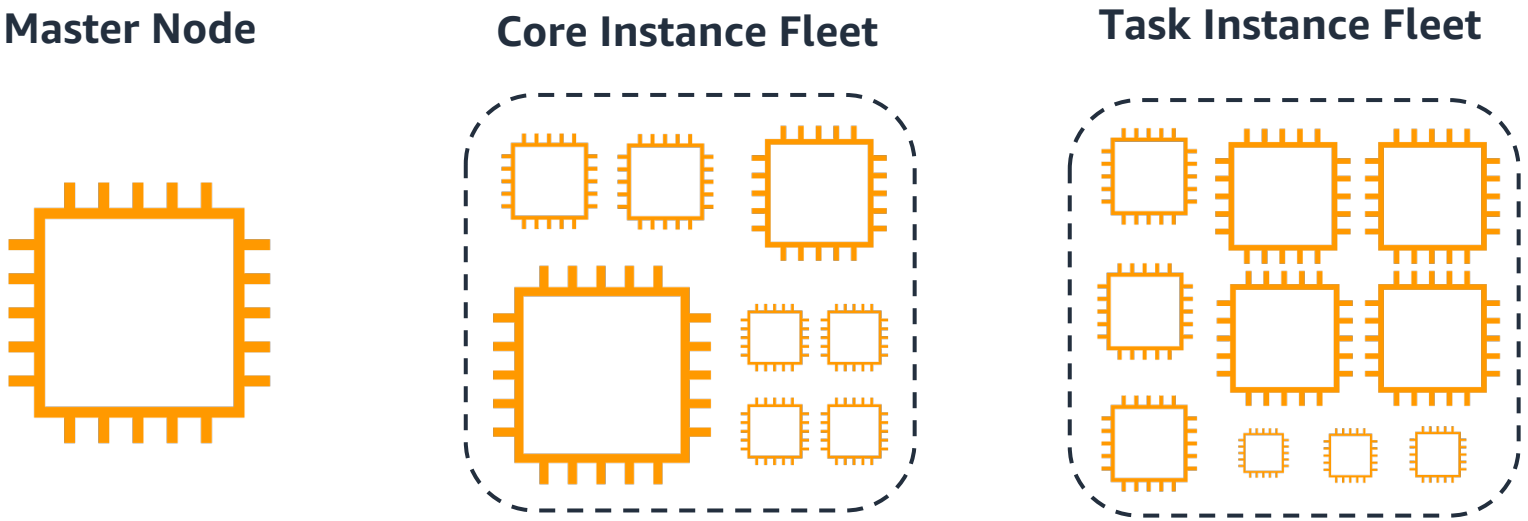
Master Node must keep running

Core nodes can be added and removed gracefully

Cluster can tolerate loss of task nodes.

aws

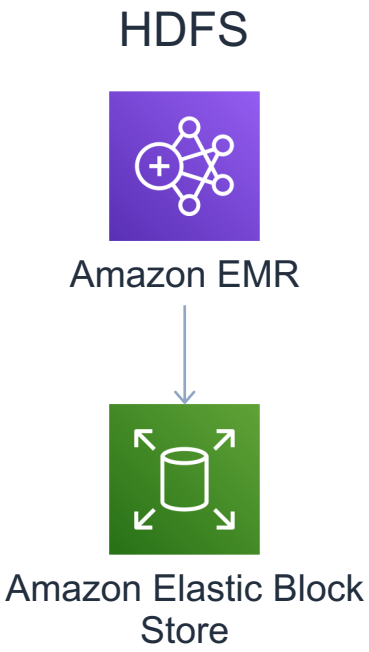
Instance fleets for advanced Spot provisioning

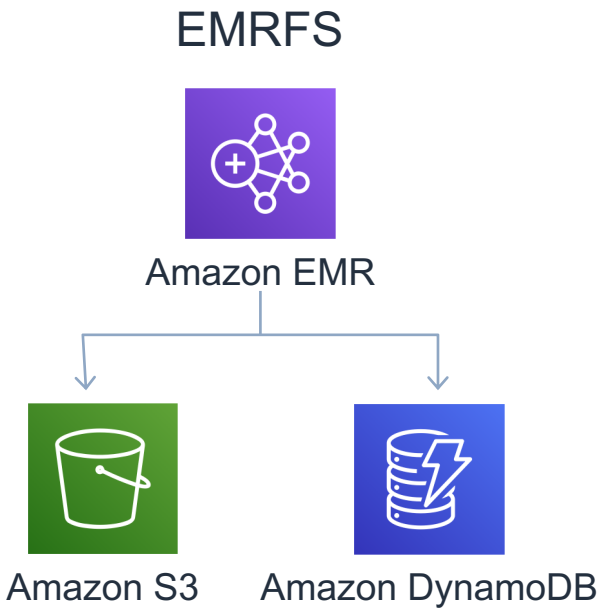


- Provision from a list of instance types with Spot and On-Demand
- Launch in the most optimal Availability Zone based on capacity/price

aws

File Systems

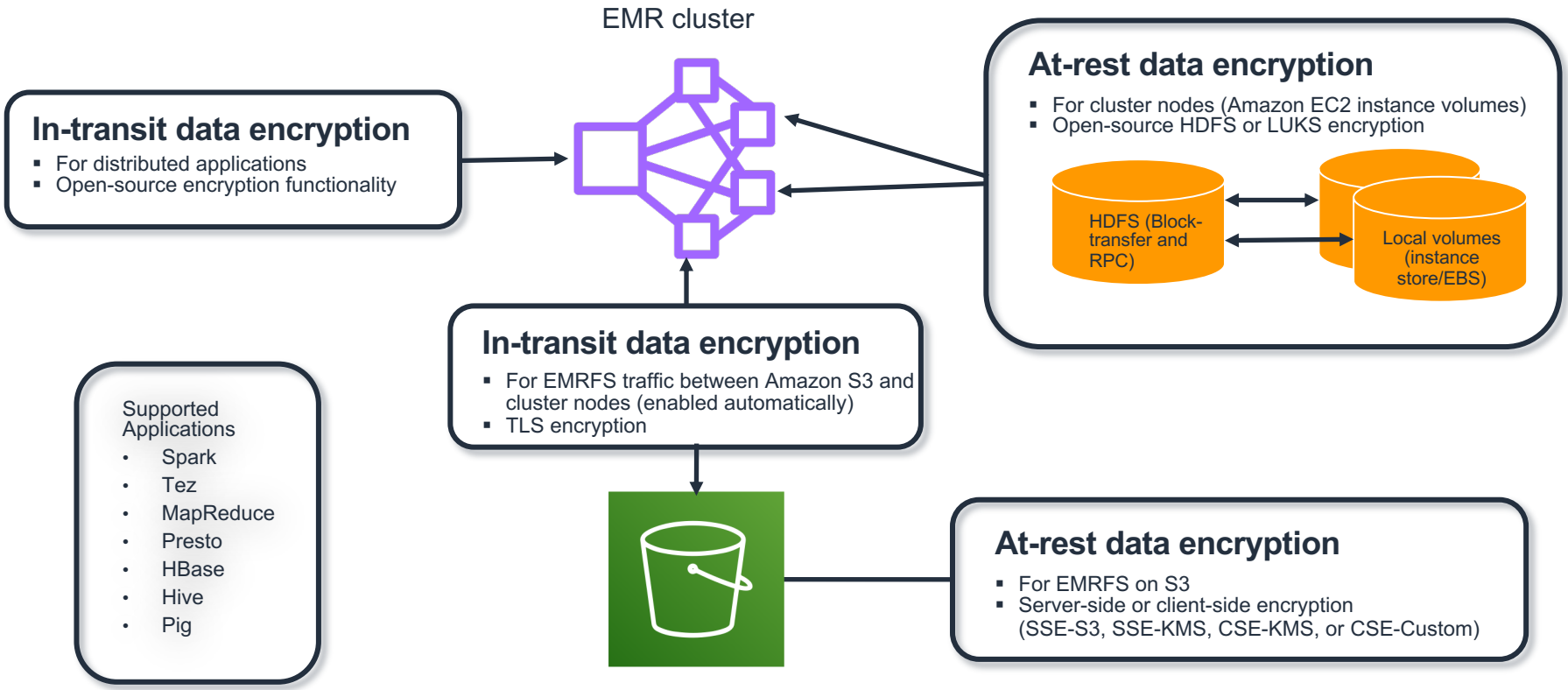




Security – Privilege Management

1. Identity and Access Management (IAM)
    1. IAM Identity-Based Policies
    2. IAM roles
    3. IAM roles for EMRFS requests to Amazon S3
2. Authentication with Kerberos
3. Secure Socket Shell (SSH)

Security – Encryption



aws

Security – Infrastructure Protection

1. Security Configurations
2. VPC and Private Subnets
3. S3 Endpoints
4. Security Groups
5. Updates to the default Amazon Linux AMI
6. Audit logs for S3 and VPC Flowlogs

Logging and Monitoring

Monitoring:

- CloudWatch Events and Metrics
- Cluster instances in Amazon EC2
- Cluster application metrics with Ganglia

Logging:

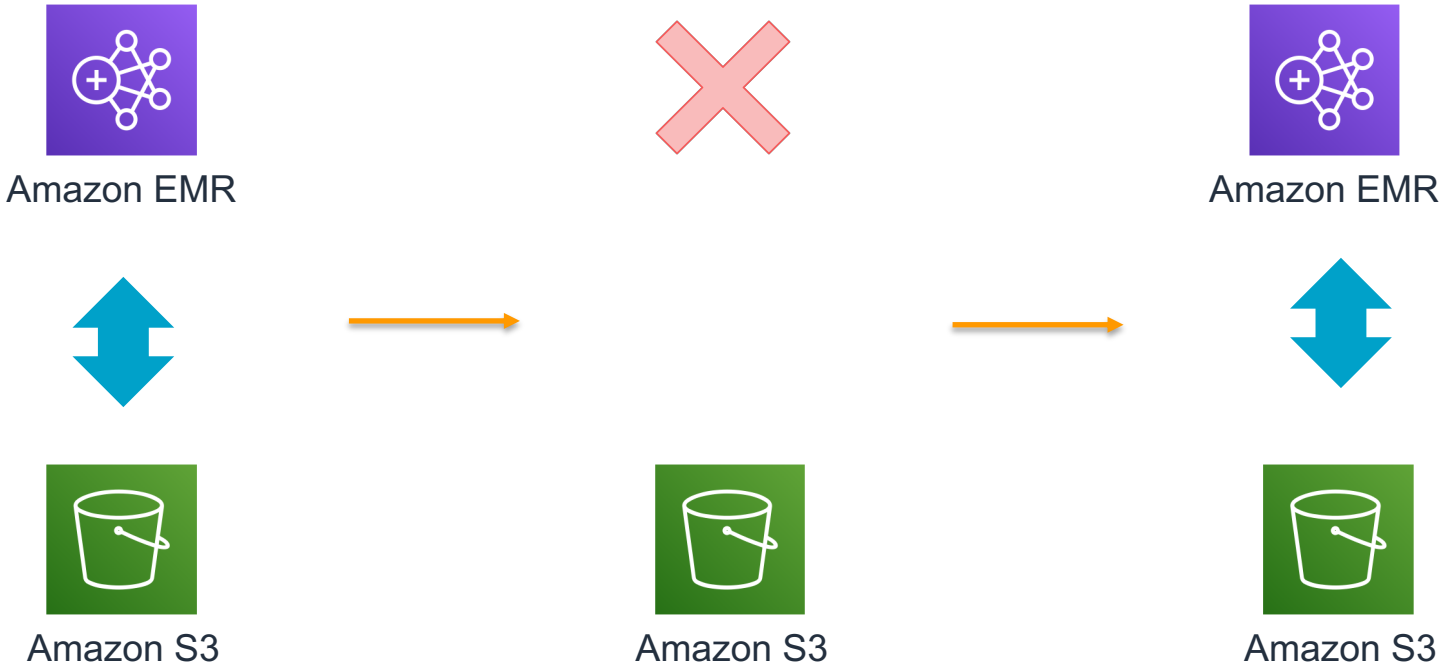
- Log Files on Master Node
- Log Files Archived to S3
- API Calls in AWS CloudTrail
- Native logs for AWS services (S3, VPCFlow logs, etc.)

Benefits of EMR

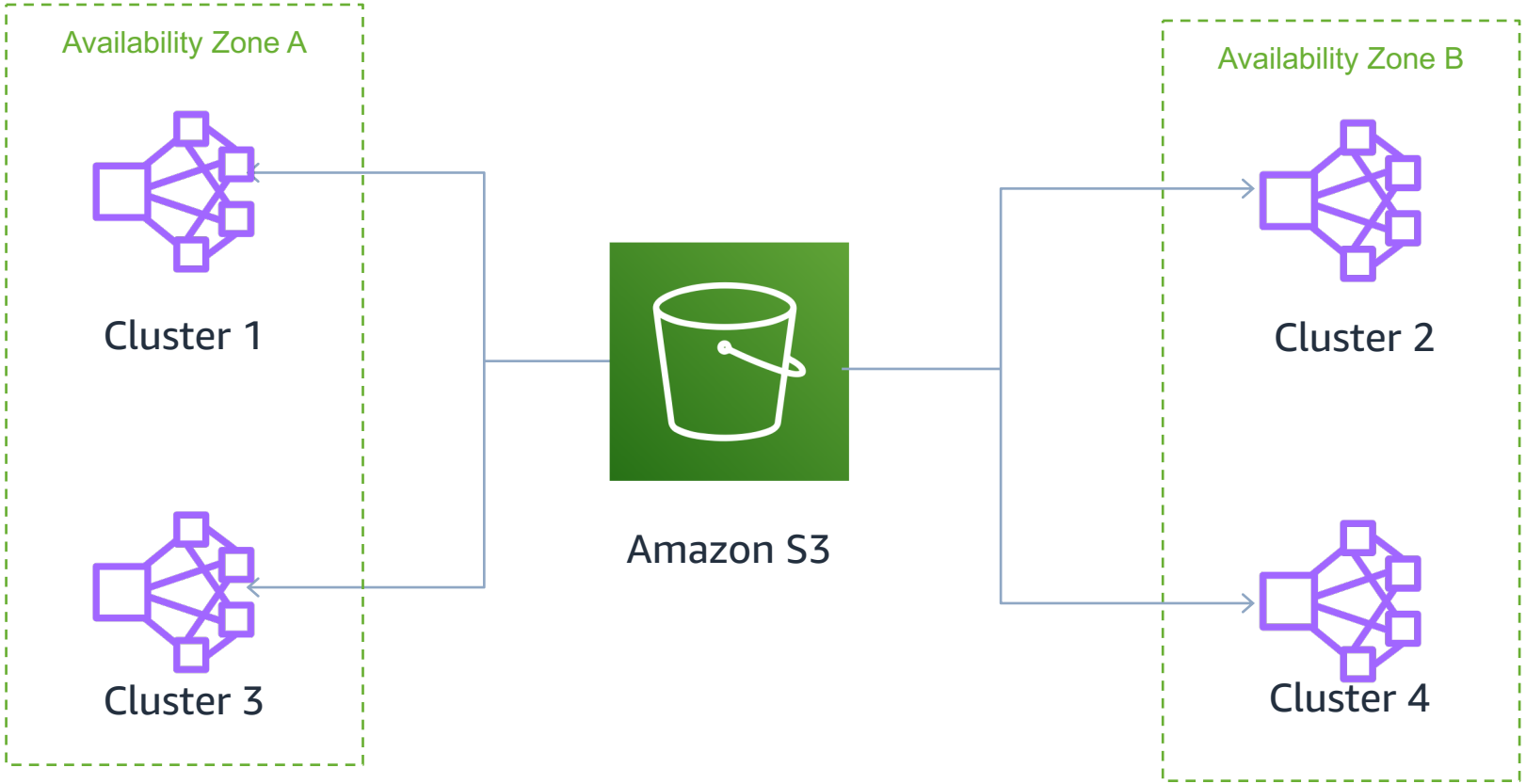
1. Decoupled compute & storage
2. Built-in disaster recovery
3. Turn off your clusters through transient clusters
4. Agility of Auto-scaling persistent clusters
5. Leverage EC2 Spot pricing
6. Self-service with AWS Service Catalog
7. Spark performance improvements
8. Fully-managed EMR Notebooks
9. Lowest TCO in the industry, analysts confirm
10. EMR is surrounded by the industry's broadest Analytics ecosystem

Appendix

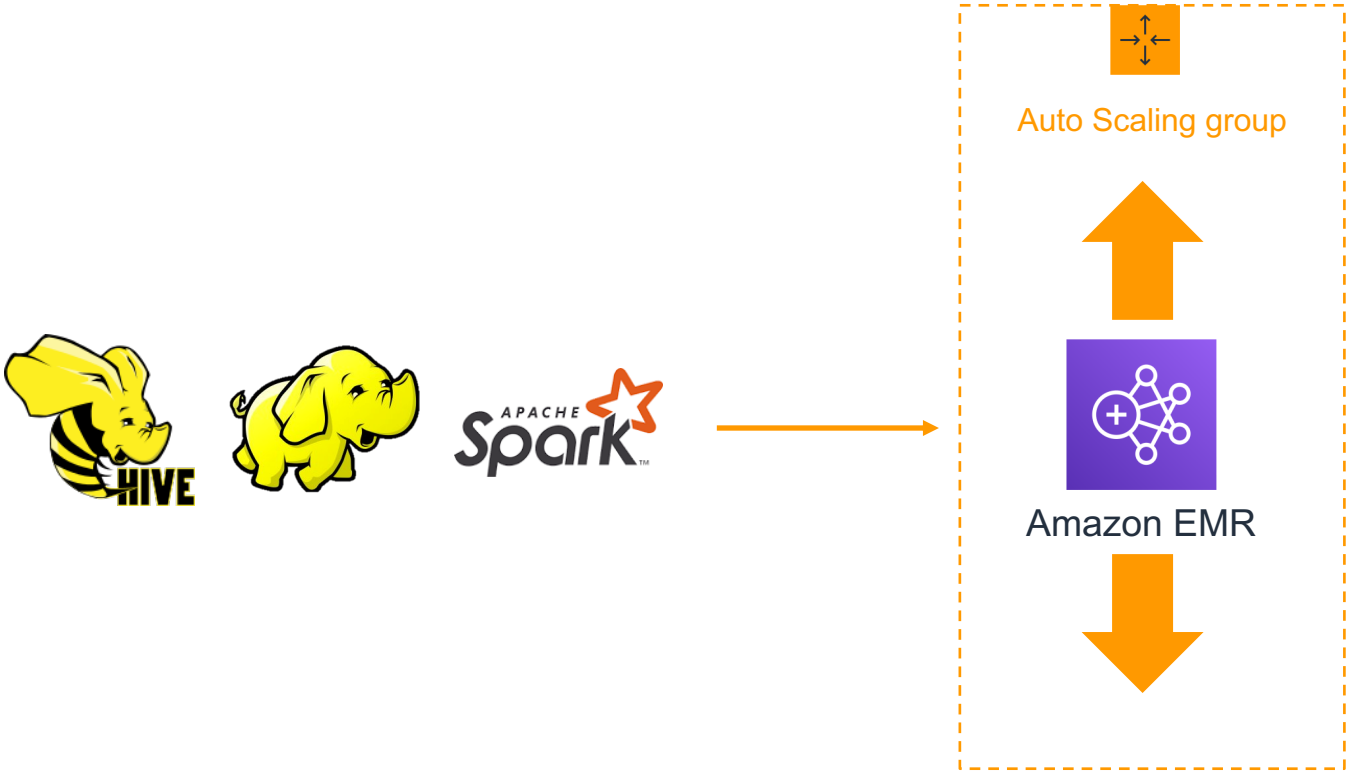
Benefit 1: Turn off clusters



Benefit 2: Built-in Disaster Recovery



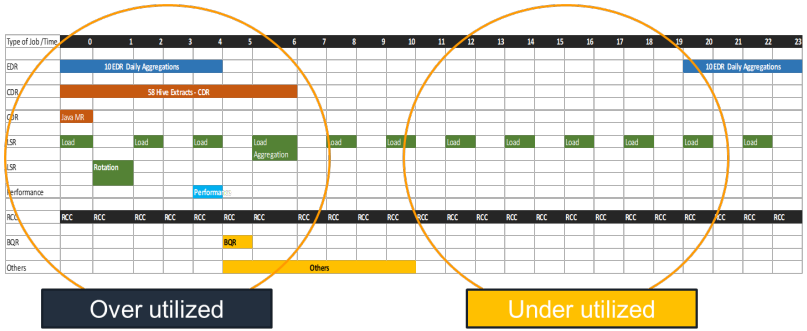
Benefit 3: Auto-scaling Persistent & Transient Clusters



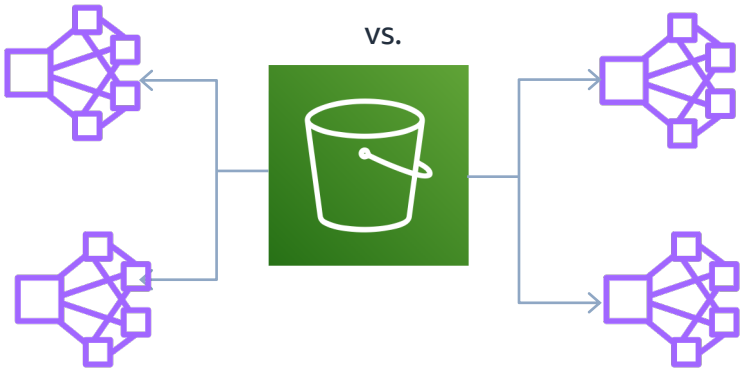
aws

Benefit 4: Logical separation of jobs/applications

Traditional Monolithic Cluster



Purpose-built Clusters



Re-architect Monolithic to Purpose-built clusters by:

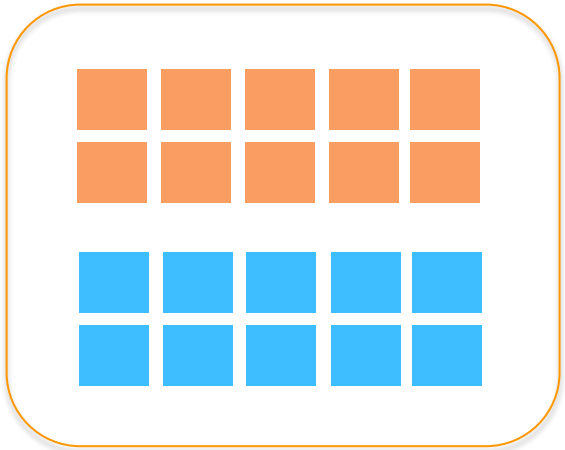
- Creating Transient and/or Persistent clusters
- Separating clusters by Application
- Separating clusters by Application Version
- Isolating Department specific clusters

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)

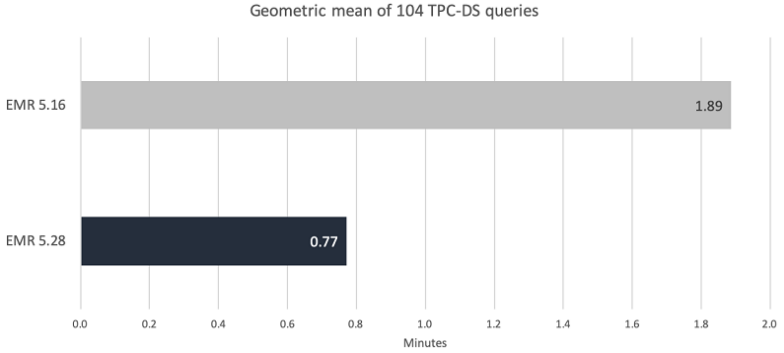
25% less cost ($140 → $105)

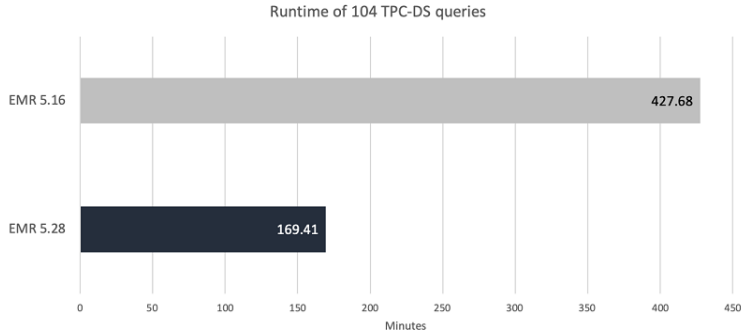
aws

Benefit 6: EMR Self-service with AWS Service Catalog

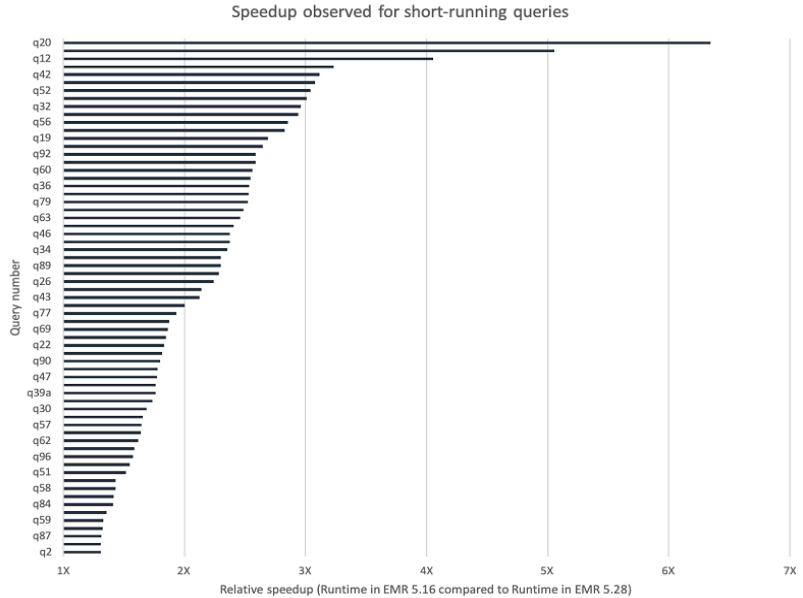
Configure	Consume
Standardize	Developer Autonomy
Enforce Consistency and Compliance	One-Stop Shop
Limit Access	Automate Deployments
Enforce Tagging, Security Groups	Agile Governance

Benefit 7: Spark Performance Improvements



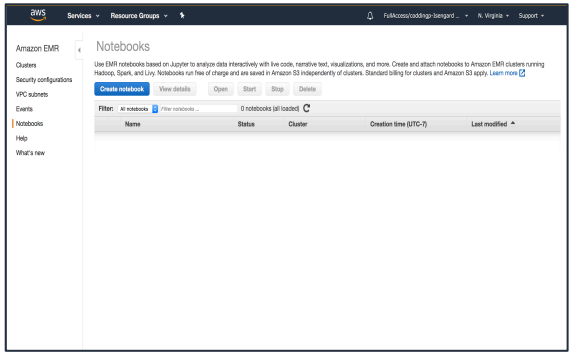


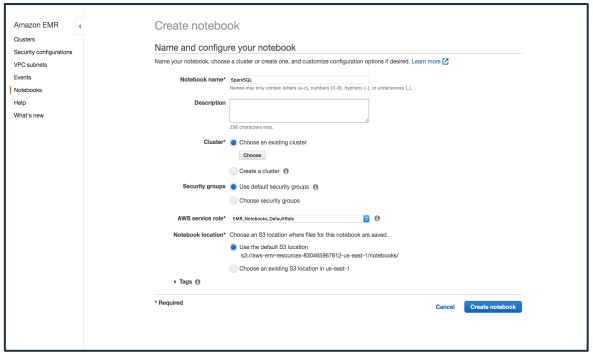
1. The EMR Runtime for Apache Spark available in EMR v5.28 realized Spark improvements of up to 32x against TPC-DS 3TB dataset in comparison to EMR v5.16. LINK to blog.
2. The EMR Runtime for Apache Spark maintains API compatibility with OSS Spark.
3. More coming every release.

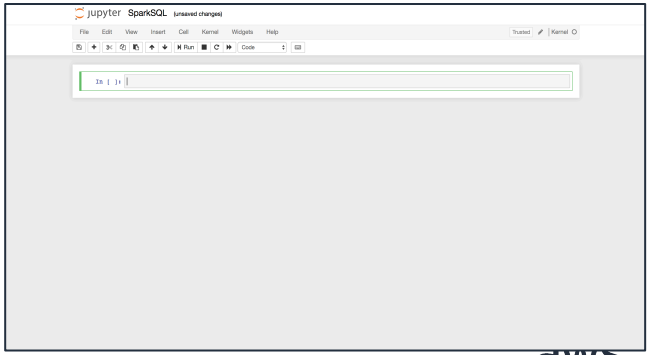


Benefit 8: Fully Managed EMR Notebooks

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.



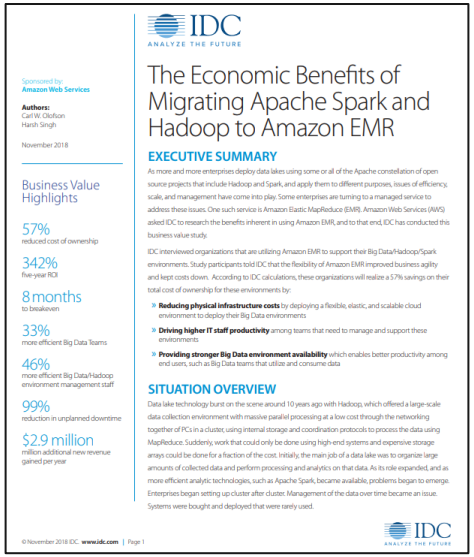




Benefit 9: Analysts confirm Lowest TCO in the Industry

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

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

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

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

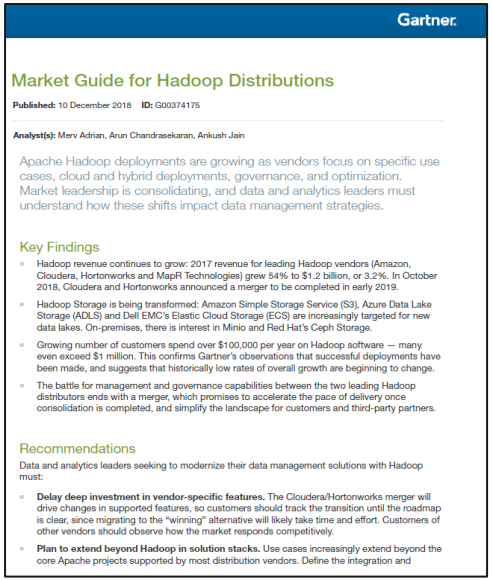
SITUATION OVERVIEW

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

Amazon EMR

Application Index

- Amazon EMR is one of the largest Spark and Hadoop service providers in the world, enabling customers to run ETL, machine learning, real-time processing, data science, and low-latency SQL at petabyte scale.
- See the EMR Release page for up to date schedules. Users can bootstrap applications not listed below.

Administrative

- Ganglia — Cluster monitoring
- Livy — REST API interacting with Spark
- Oozie — Workflow Scheduler
- ZooKeeper — Configure & Sync node

Machine Learning

- Mahout — Machine Learning
- MXNet — Deep Learning
- SparkML — Machine Learning
- TensorFlow — Deep Learning

Data Movement

- Sqoop — Relational DB connector

NoSQL

- HBase — Hadoop's non-relational DB

Data Processing

- Flink — Stream Processing
- Hive — Hadoop Data warehouse
- MapReduce — Batch Data Processing
- Presto — Distributed SQL for Big Data
- Spark — In-memory Data Processing & Machine Learning
- Tez — Interactive Data Processing
- Hudi — Update Data Lake Storage

Query Tools

- EMR Notebooks — Serverless notebook
- Hue — Visualization and Querying for Hadoop
- JupyterHub — Multi-user Juptyer Notebook (installed on cluster)
- Phoenix — Querying for HBase
- Pig — Scripting language for MapReduce jobs
- Zeppelin — Data Science notebook

aws

Resources

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

Getting started
https://aws.amazon.com/emr/getting-started/

EMR Migration Program
https://aws.amazon.com/emr/emr-migration

AWS Big Data Blog
https://aws.amazon.com/blogs/big-data/

aws