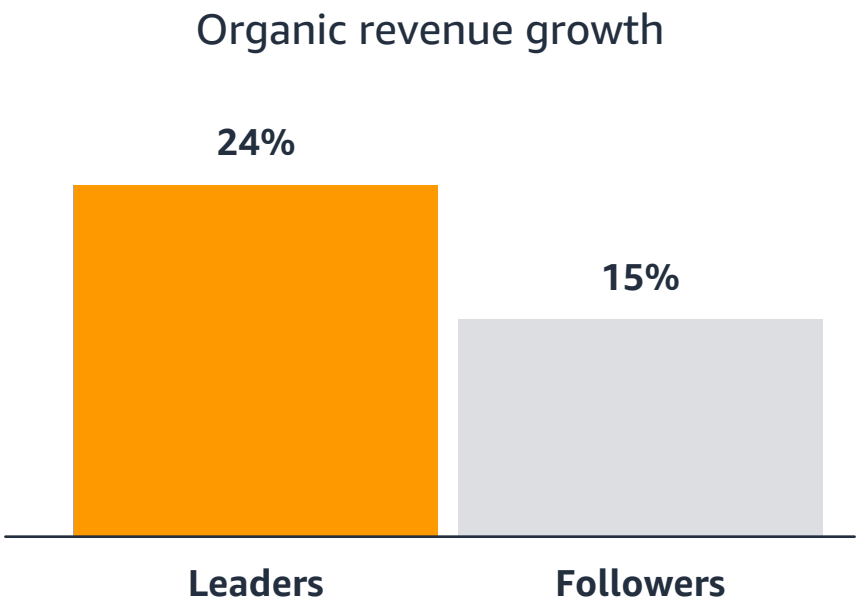
Big Data Processing with EMR

To Become a Leader, Data is Your Differentiator

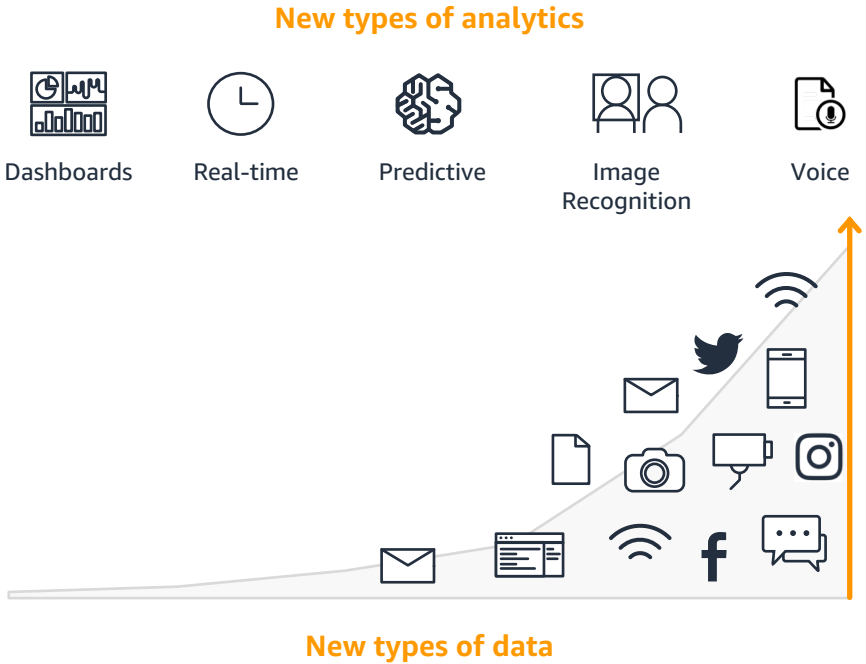


Organizations that successfully generate business value from their data, will outperform their peers. An Aberdeen survey saw organizations who implemented a Data Lake outperforming similar companies by 9% in organic revenue growth.*

*Aberdeen: Angling for Insight in Today's Data Lake, Michael Lock, SVP Analytics and Business Intelligence

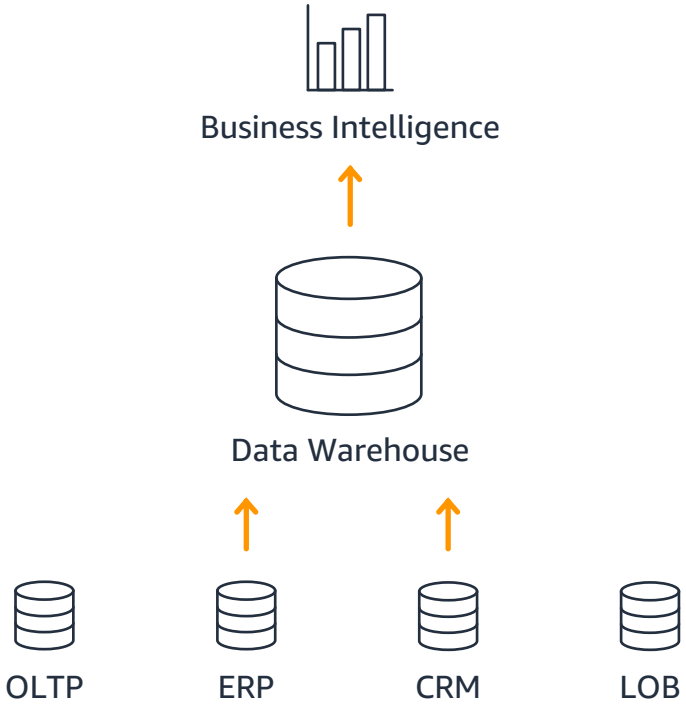
aws

For Data to Be a Differentiator, Customers Need to Be Able to…



- Capture and store new non-relational data at PB-EB scale in real time
- New type of analytics that go beyond batch reporting to incorporate real-time, predictive, voice, and image recognition
- Democratize access to data in a secure and governed way

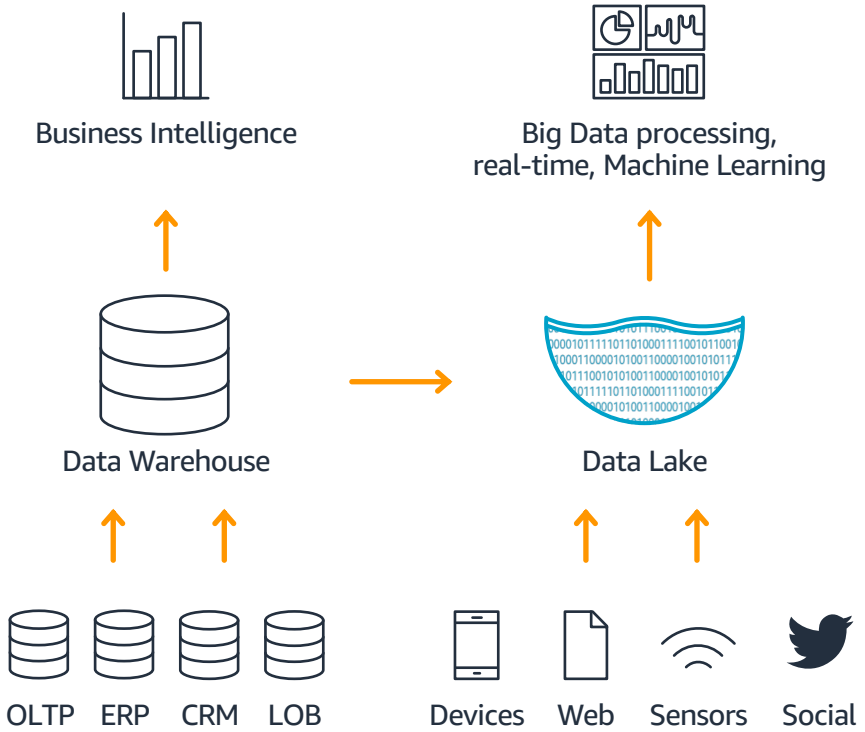
Traditionally, Analytics Used to Look Like This



- Relational data
- TBs–PBs scale
- Schema defined prior to data load
- Operational reporting and ad hoc
- Large initial CAPEX + $10K–$50K/TB/Year

aws

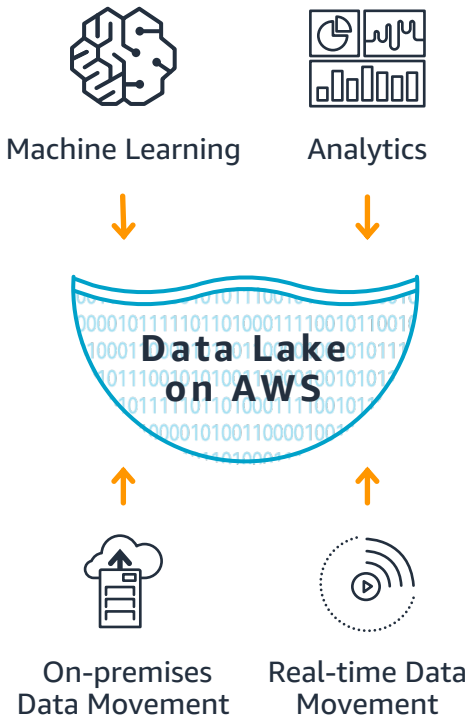
Data Lakes Extend the Traditional Approach



- Relational and non-relational data
- TBs–EBs scale
- Diverse analytical engines
- Low-cost storage & analytics

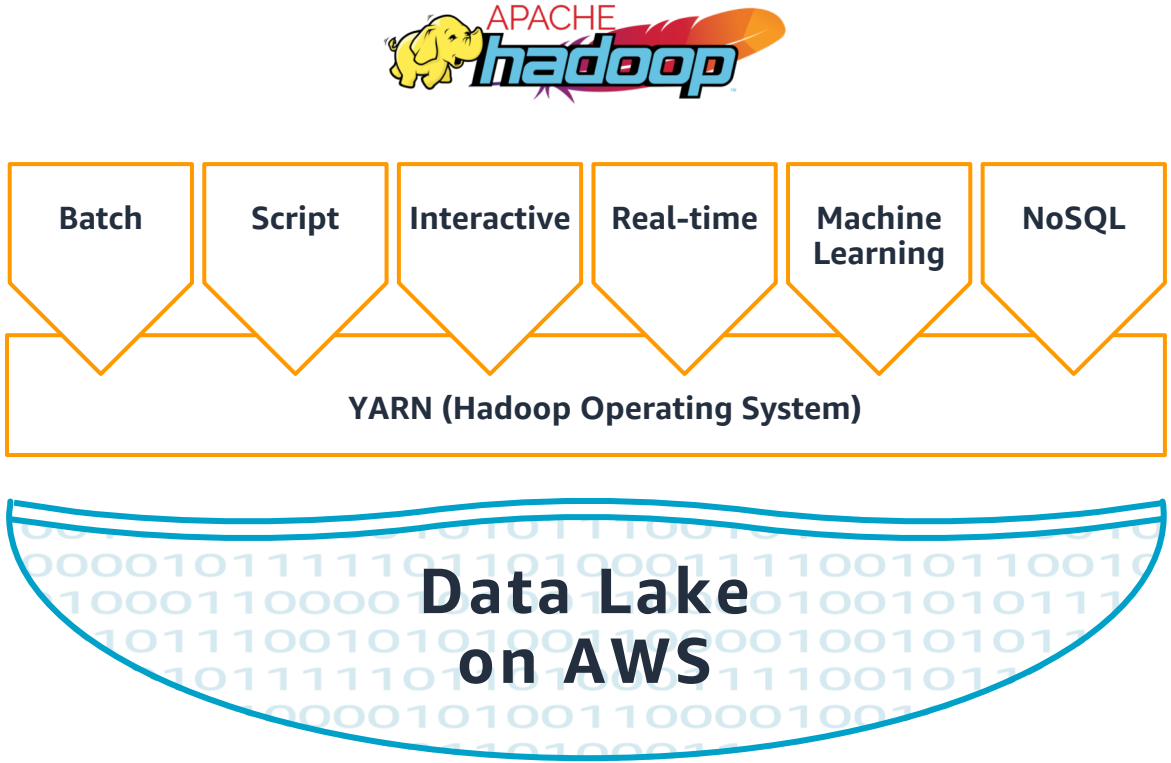
aws

Data Lakes and Analytics from AWS



- Open and comprehensive
- Secure
- Scalable and durable
- Cost-effective

Apache Hadoop on the Data Lake



- Distributed processing
- Diverse analytics
  - Batch/Script (Hive/Pig)
  - Interactive (Spark, Presto)
  - Real-time (Spark)
  - Machine Learning (Spark)
  - NoSQL (HBase)
- For many use-cases
  - Log and clickstream analysis
  - Machine Learning
  - Real-time analytics
  - Large-scale analytics
  - Genomics
  - ETL

aws

Hadoop is best deployed in the cloud

"The public cloud for big data analytics is in the future for most companies. It's time to decide how far you will go and how fast."

Brian Hopkins and **Mike Gualteri**, Forrester
The Cloudy Future of Hadoop, March, 2017

- Scale to any number of resources
- Spin up/down clusters in minutes
- Costs based on actual utilization
- Reduce time to analyze large datasets

aws

Introducing Amazon EMR

Managed Hadoop and Spark in the cloud at 1/8th the cost

Enterprise-grade

Easy

Lowest cost

aws

Introducing Amazon EMR

Managed Hadoop and Spark in the cloud at 1/8th the cost

Enterprise-grade

Easy

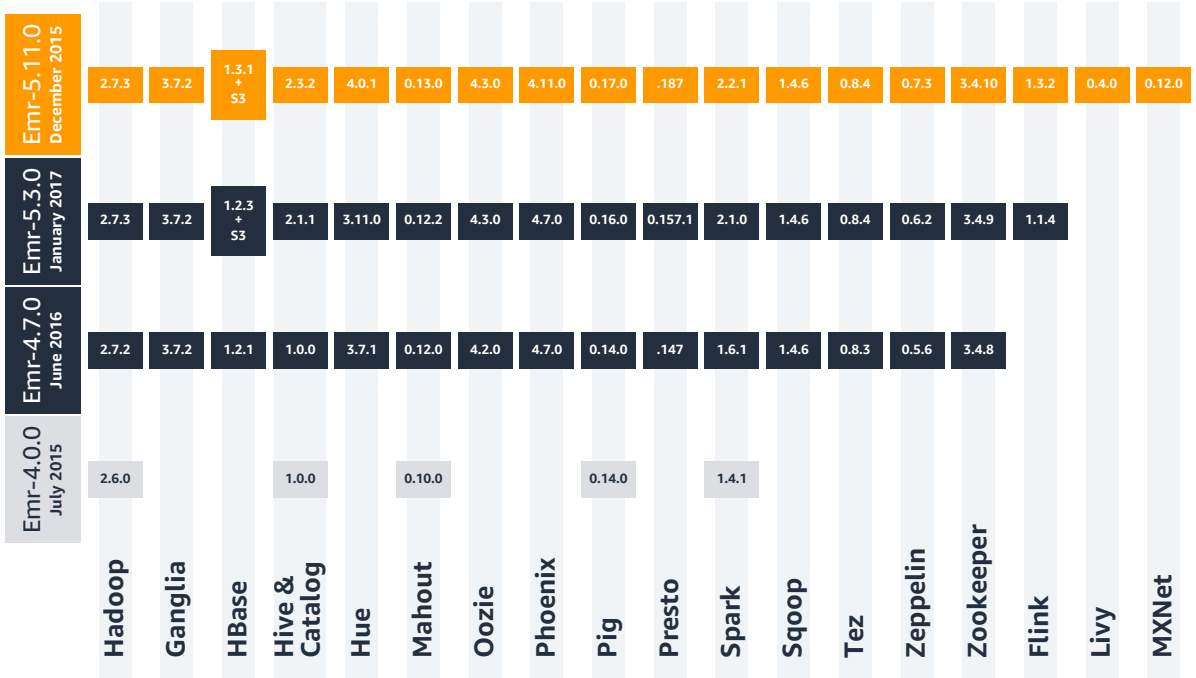
Lowest cost

aws

Enterprise-grade Hadoop & Spark

Deploy latest releases in Hadoop and Spark ecosystems

EMR Releases

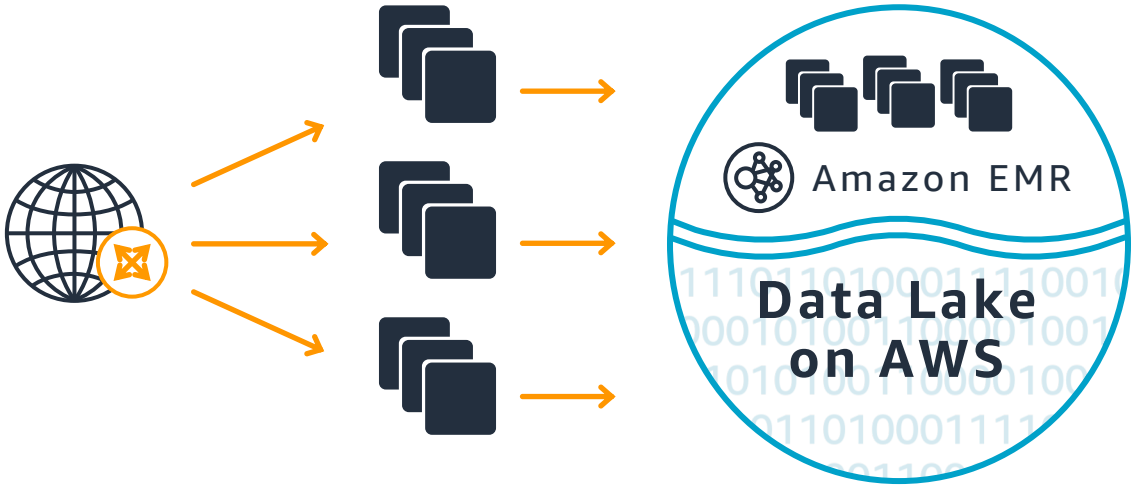


- 19 open-source projects: Apache Hadoop, Spark, HBase, Presto, and more
- Updated with the latest open source frameworks within 30 days of release

aws

Enterprise-grade Hadoop & Spark

Scale to any size



- Scale compute (EMR) & storage (S3) independently
- Store, and process any amount of data—PB to EBs
- Provision one, hundreds, or thousands of nodes
- Auto-scaling

aws

Enterprise-grade Hadoop & Spark

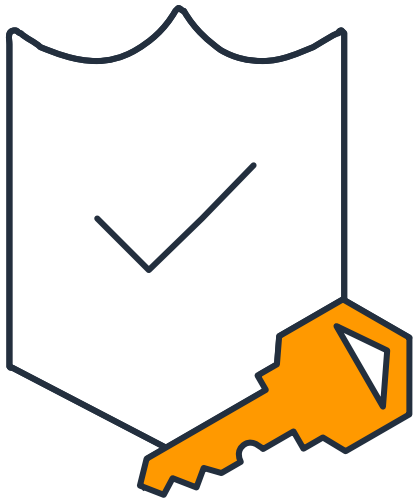
Highly available and durable

- S3 is designed to deliver 99.999999999% durability
- EMR monitors your cluster—replacing poorly performing & failed nodes, and restarting services
- Monitor your clusters using Amazon CloudWatch
- Built-in console to view job history & browse logs
- EMR has on-cluster HDFS for data persistence

aws

Enterprise-grade Hadoop & Spark

Highly secure



- Encryption of data at rest and in-transit
- ML-powered security with Amazon Macie
- Network isolation using Amazon VPC
- Access and permissions control with IAM policies
- Log, and audit activity with AWS CloudTrail
- Microsoft AD integration with Kerberos support

aws

Introducing Amazon EMR

Managed Hadoop and Spark in the cloud at 1/8th the cost

Enterprise-grade

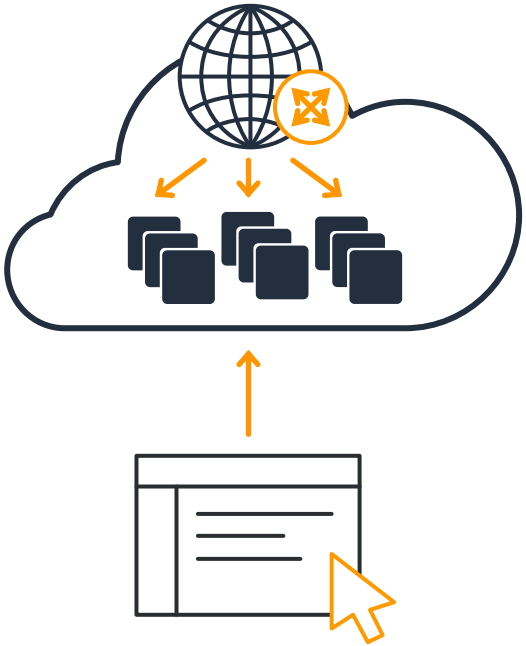
Easy

Lowest cost

aws

Easy

Spin up clusters in minutes

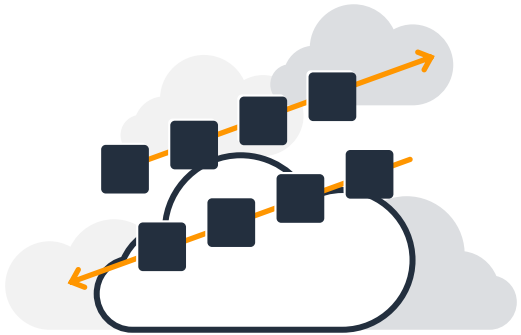


- Launch Hadoop & Spark clusters in minutes
- No need to install or maintain Hadoop
- Cluster tuning, and configuration done for you
- Latest Hadoop versions within 30 days of release

aws

Easy

Automatic and elastic scaling



- Automatically scale clusters based on policies
- Shut down cluster when processing completes
- Optimized for both transient and long-running clusters
- No manual intervention needed

aws

Introducing Amazon EMR

Managed Hadoop and Spark in the cloud at 1/8th the cost

Enterprise-grade

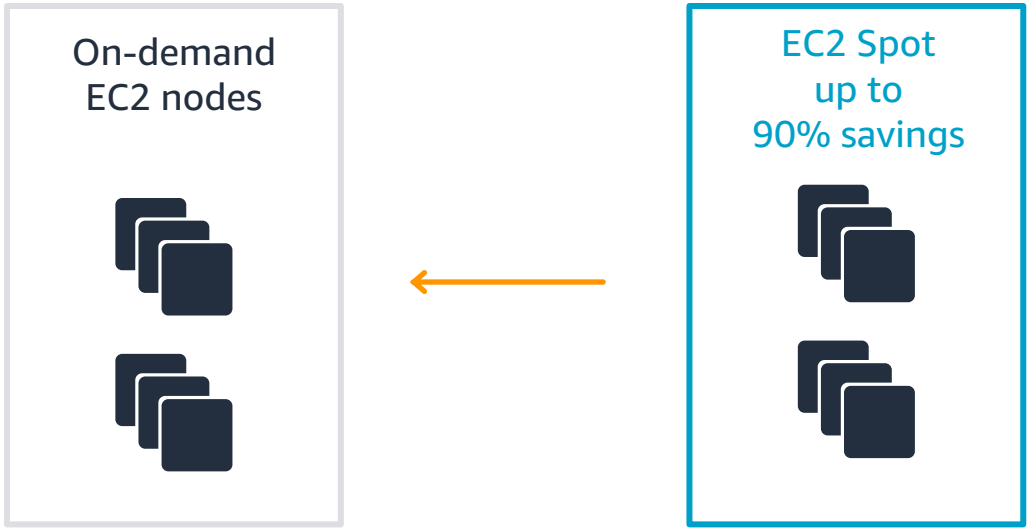
Easy

Lowest cost

aws

Lowest cost

Save 75–90% using Reserved Instances and Spot

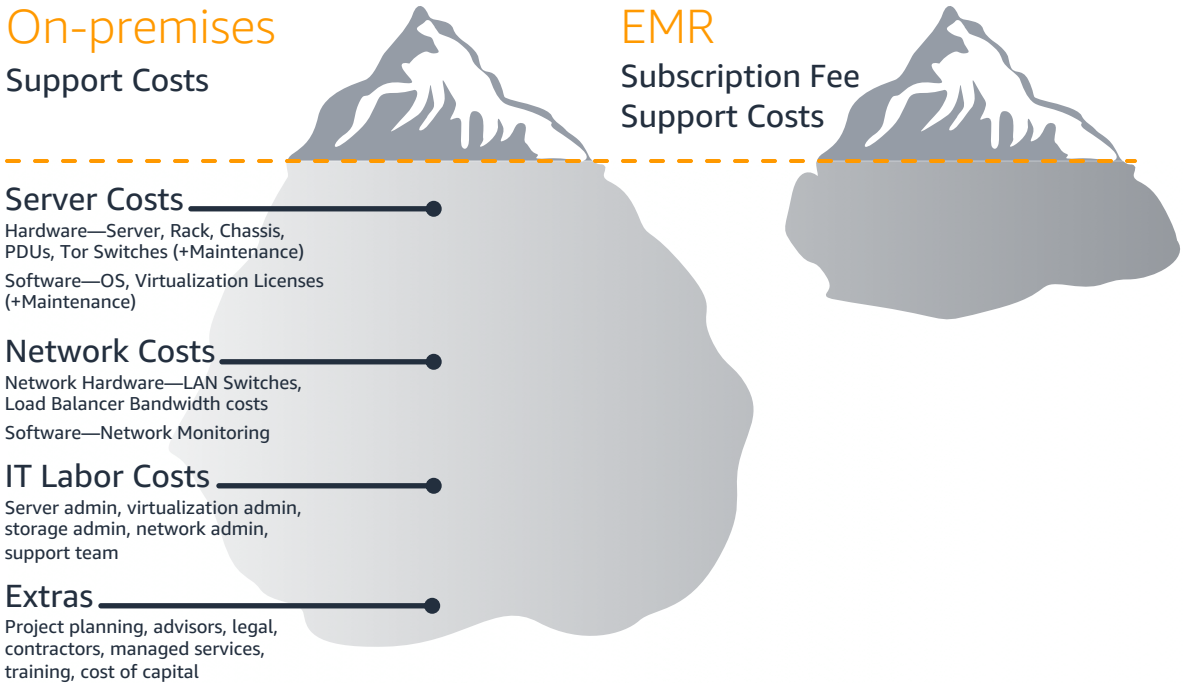


- Commit to a set term and save up to 75% with Reserved Instance
- Run on spare compute capacity and save up to 90% with Spot
  - Pay a fraction of on-demand price
  - Name your bid price. If it exceeds the market, you get the resource
  - Provision from a list of instance types with Spot and on-demand
  - Launch in the most optimal AZ based on capacity/price
  - Spot Block support

aws

Lowest cost

Lowest TCO

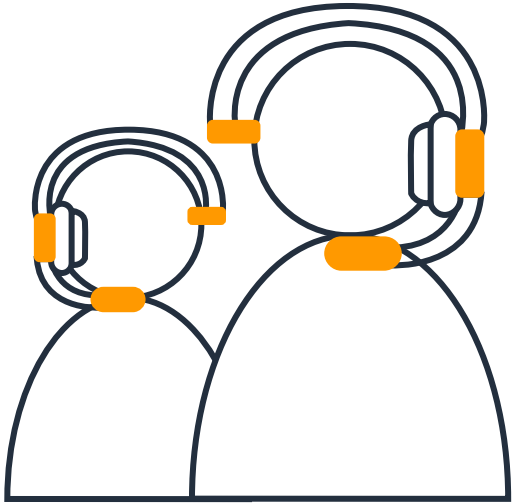


- Less admin time to manage, and support Hadoop clusters
- No up-front costs—hardware acquisition, installation
- Save on operating costs—data center space, power, cooling
- Business Value: Cost of delays, Risk Premium, Competitive abilities, governance, etc.

aws

Lowest cost

Hadoop & Spark support included



- Hadoop & Spark support included in AWS support
- Support for a 10-node cluster is $4.5K/year
- Up to 1/8th the cost of commercial Hadoop offerings

aws

EMR powers the most cloud Hadoop & Spark projects

aws

EMR Partner Solutions

North America

47 Lining

RELUS TECHNOLOGIES

accenture
High performance. Delivered.

TEKsystems
Our people make IT possible.

apps associates
extreme expertise

ONICA

clearscale

CORECOMPETE

NorthBay

Europe

BEEVA

capside
architects of the digital society

KCOM

tecRACER
Get Your Business Online

Asia Pacific

altis

softserve

aws

FINRA oversees > 3,000 securities firms doing business in the United States.

Challenge:

FINRA's legacy system did not scale well

- Up to 75 billion events per day
- Run complex surveillance queries over 20+ PB of data

Solution:

- Migrated their big data appliance to a S3 Data Lake and used EMR for ingestion and processing
- Migrated to RDS and testing Aurora

Get started

Service page
aws.amazon.com/emr

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

Contact us
https://aws.amazon.com/contact-us/