



Data lakes and analytics on AWS: Turn data into insights

Shazli Mansor Bin Mohd Ghazali
Solutions Architect

Epic Games continually improves Fortnite for 250+ million players globally



Challenge

They needed a way to process and analyze over 100 PB of data (125M events/min) ingested from game clients and game servers to understand and adapt to player engagement.

Solution

Epic Games turned to AWS for an Amazon S3 data lake in combination with Amazon EMR, Amazon EC2, and Amazon Kinesis.

Benefits

The data provides a constant feedback loop for designers, and an up to the minute analysis of gamer satisfaction to drive gamer engagement.

Data is a strategic asset for every organization

“ The world’s most valuable resource is no longer oil, but **data**.* ”



*Copyright: The Economist, 2017, David Parkins

Customers want more value from their data



Growing
exponentially



From new
sources



Increasingly
diverse



Used by
many people



Analyzed by
many applications

Types of analytics users—which are you?



Architects

Application developers

Business intelligence (BI) analysts

CxO

Data engineers, operations

Data modelers

Data scientists

Data warehouse admins

Database admins (DBAs)

DevOps engineers

LOB knowledge workers

Product managers

IT operations

IT security and governance

VP/director analytics

Common analytics use cases—which do you need?



Data warehouse modernization

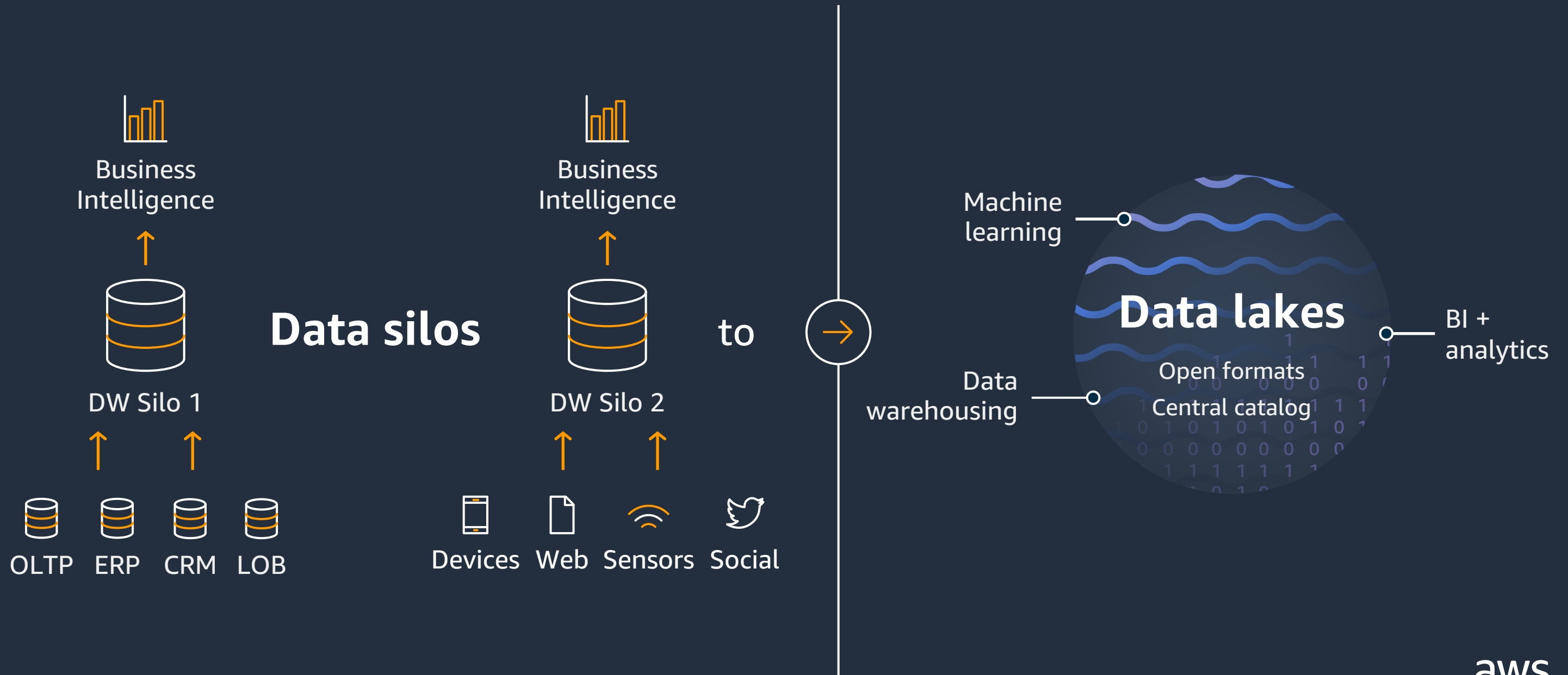
Big data and data lakes

Real-time streaming and analytics

Operational and search analytics

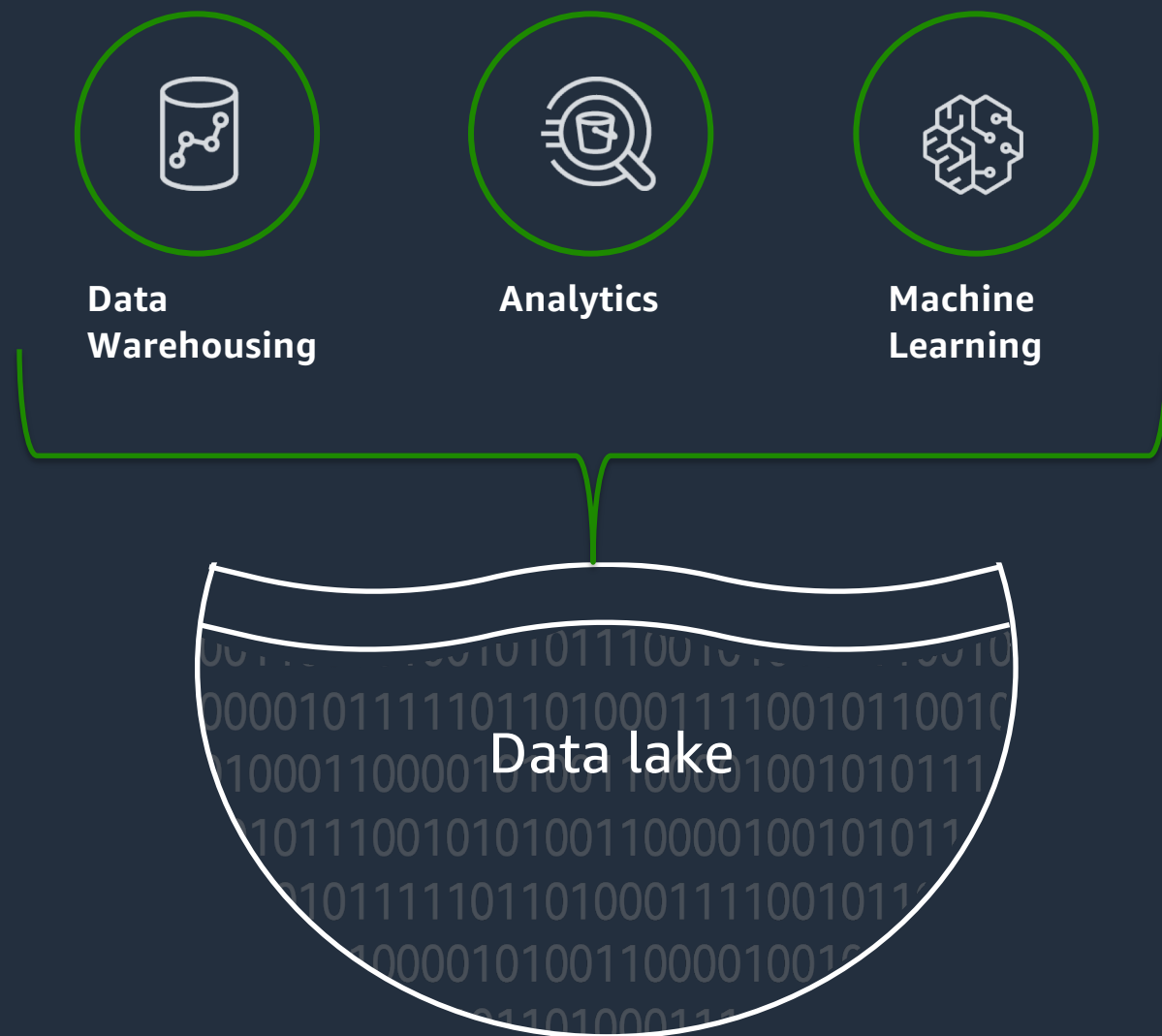
Self-service business analytics

Traditional data warehousing approaches don't scale



Customers moving to data lake architectures

Bringing together the best of both worlds



Extends or evolves DW architectures

Store any data in any format

Durable, available, and exabyte scale

Secure, compliant, auditable

Run any type of analytics from DW to Predictive

Why choose AWS for data lakes and analytics?

1



Easiest to build
data lakes and
analytics

2



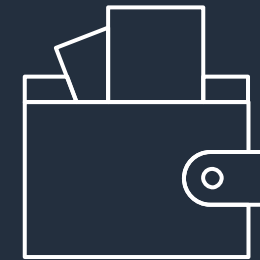
Most secure
infrastructure for
analytics

3



Most
comprehensive
and open

4



Most scalable
and cost
effective

1. Easiest to build data lakes and analytics



A single storage layer (S3) for all analytics and ML

A service to build secure data lakes in days

Deep integration across analytics & infrastructure
(including federated queries)

The fastest way to go from zero to insights,
covering all data for all users

2. Most secure infrastructure for analytics

Services for security and governance



Customers need to have multiple levels of security, identity and access management, encryption, and compliance to secure their data lake



Security



Identity



Encryption



Compliance

Amazon GuardDuty

AWS Shield

AWS WAF

Amazon Macie

VPC

AWS IAM

AWS SSO

Amazon Cloud Directory

AWS Directory Service

AWS Organizations

AWS Certification Manager

AWS Key Management Service

Encryption at rest

Encryption in transit

Bring your own keys,
HSM support

AWS Artifact

Amazon Inspector

Amazon Cloud HSM

Amazon Cognito

AWS CloudTrail

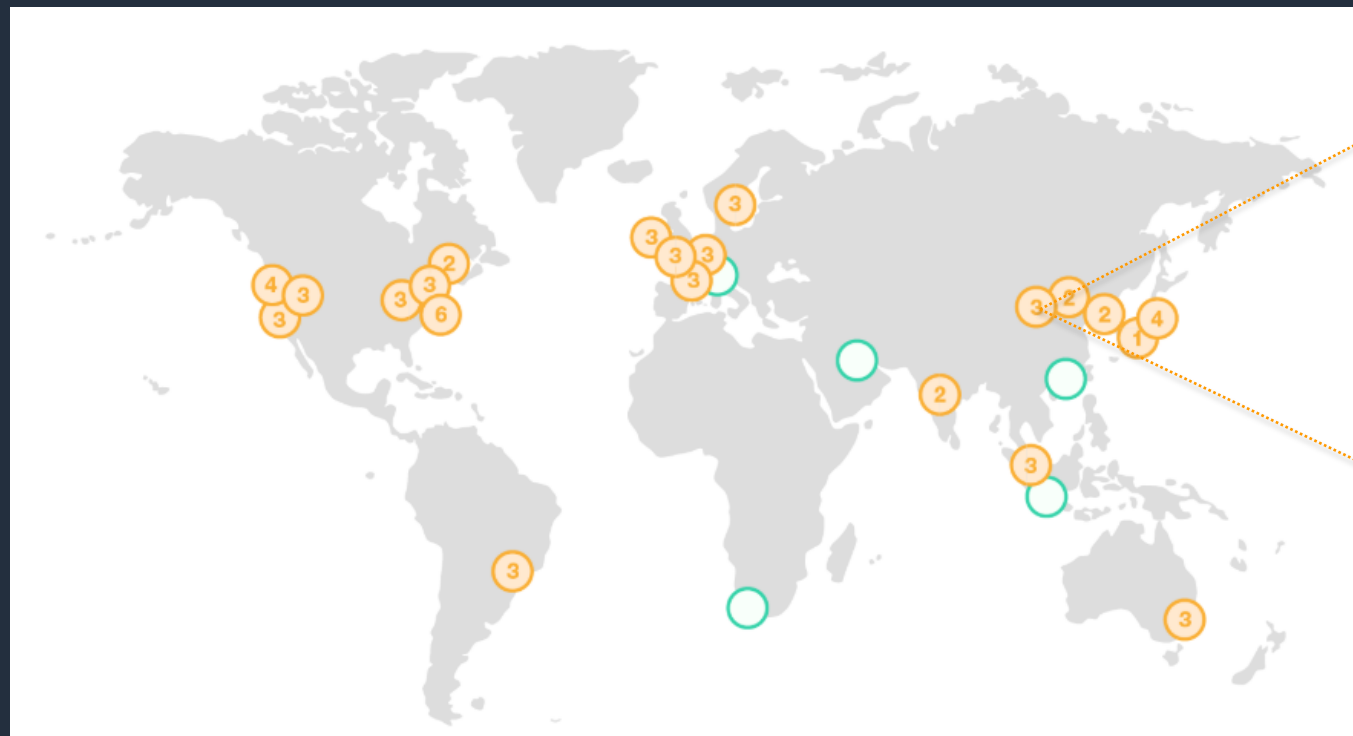
2. Most secure infrastructure: certifications



Global	United States	Asia Pacific	Europe
 CSA Cloud Security Alliance Controls	 CJIS Criminal Justice Information Services	 ITAR International Arms Regulations	 MTCS Tier 3 [Singapore] Multi-Tier Cloud Security Standard
 ISO 9001 Global Quality Standard	 DoD SRG DoD Data Processing	 MPAA Protected Media Content	 My Number Act [Japan] Personal Information Protection
 ISO 27001 Security Management Controls	 FedRAMP Government Data Standards	 NIST National Institute of Standards and Technology	Europe
 ISO 27017 Cloud Specific Controls	 FERPA Educational Privacy Act	 SEC Rule 17a-4(f) Financial Data Standards	 C5 [Germany] Operational Security Attestation
 ISO 27018 Personal Data Protection	 FFIEC Financial Institutions Regulation	 VPAT/Section 508 Accountability Standards	 Cyber Essentials Plus [UK] Cyber Threat Protection
 PCI DSS Level 1 Payment Card Standards	 FIPS Government Security Standards	Asia Pacific	 G-Cloud [UK] UK Government Standards
 SOC 1 Audit Controls Report	 FISMA Federal Information Security Management	 FISC [Japan] Financial Industry Information Systems	 IT-Grundschutz [Germany] Baseline Protection Methodology
 SOC 2 Security, Availability, & Confidentiality Report	 GxP Quality Guidelines and Regulations	 IRAP [Australia] Australian Security Standards	
 SOC 3 General Controls Report	 HIPAA Protected Health Information	 K-ISMS [Korea] Korean Information Security	

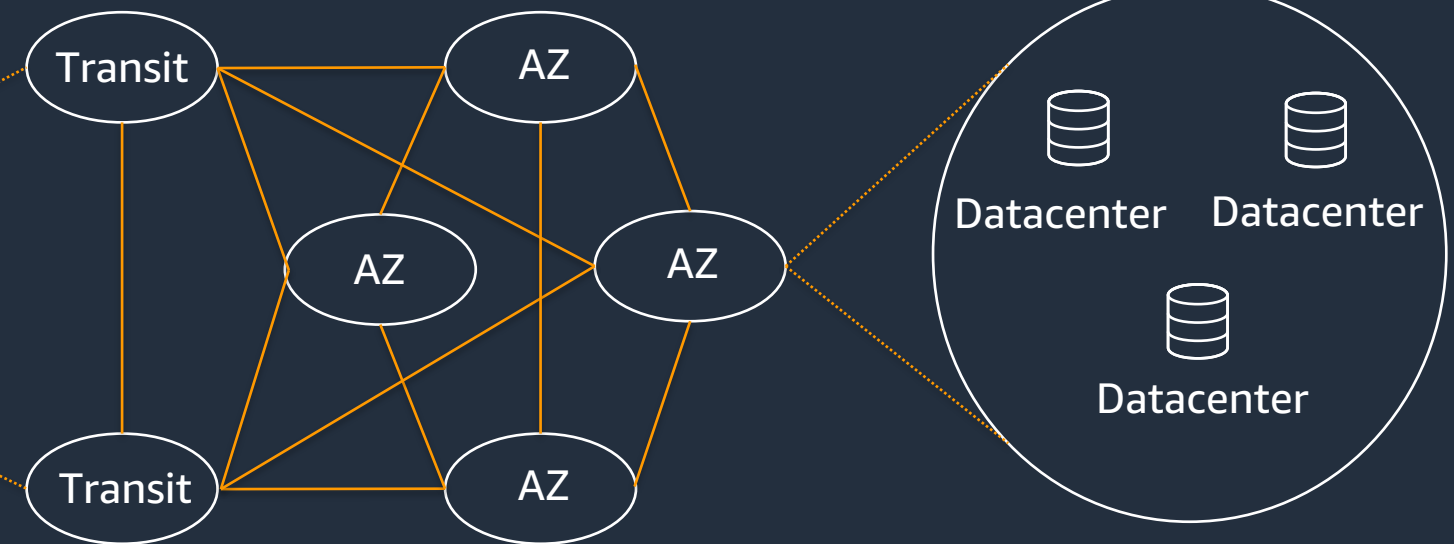
Scale globally with resilience **in every region**

The largest global foot print consistently built with a multi-AZ and multi-datacenter design



AWS Region

AWS Availability Zone (AZ)



A Region is a physical location in the world where we have multiple **Availability Zones**.

Availability Zones consist of one or more discrete data centers, each with redundant power, networking, and connectivity, housed in separate facilities.

Region & Number of Availability Zones

● Announced Regions
Bahrain, Cape Town, Jakarta, and Milan

3. Most comprehensive and open



Data, visualization, engagement, & machine learning



Data



Dashboards



Digital User Engagement



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

3. Most comprehensive and open



Data, visualization, engagement, & machine learning

NEW



Analytics



Data lake infrastructure & management



Data movement

Database Migration Service | Snowball | Snowmobile | Kinesis Data Firehose | Kinesis Data Streams | Managed Streaming for Apache Kafka

3. Open standards, formats, and Apache open source



Flink

Ganglia

Hbase

HCatalog

HDFS

Hive

Hudi

Java

JupyterHub

Kafka

Livy

Mahout

MapReduce

MxNET

MySQL

Oozie

ORC

Parquet

Phoenix

Pig

Presto

Python

PyTorch

R

Scala

Spark

Sqoop

SQL

TensorFlow

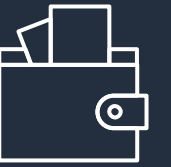
Tez

YARN

Zeppelin

Zookeeper

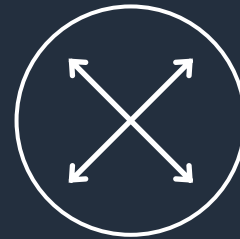
4. Most scalable, cost-effective, high-performance infrastructure for analytics



On-demand, Reserved, and Spot instances to reduce costs



100 Gbps bandwidth network interfaces for performance



Industry leading choice of 200+ instance types to meet workload needs



Five highly available storage tiers and intelligent tiering

4. Most scalable, cost-effective infrastructure for analytics



Some examples of advanced capabilities in analytics services



EMR

Autoscaling

57% less than on-premises per IDC report



Redshift

Less than 1/10th of the cost of traditional, on-premises solutions



Athena & QuickSight

Serverless pay only for what is used

Pricing per session for visualization

More data lakes and analytics than anywhere else

Tens of thousands of data lakes run on AWS across all industries



Consulting partners focus on delivering

Technology solutions

- Data lakes
- Data warehouse analytics
- Real time analytics
- Data governance
- Data platforms
- D&A cloud migration

Business/industry solutions

- Financial services
- Healthcare
- Life sciences
- Retail
- Manufacturing
- Public sector
- Marketing
- Sales
- Human resources
- Customer service
- Finance
- Data exchanges
- Data monetization

Learn analytics with AWS Training and Certification

Resources created by the experts at AWS to help you build and validate data analytics skills



New free digital course: “Data Analytics Fundamentals”



Classroom offerings, including “Big Data on AWS”, feature AWS expert instructors and hands-on labs



Validate expertise with the “AWS Certified Big Data—Specialty” exam or the new “AWS Certified Data Analytics—Specialty” beta exam

Visit aws.amazon.com/training/paths-specialty/

The AWS analytics portfolio

Data, visualization, engagement, & machine learning

NEW



Analytics



Data lake infrastructure & management



Data movement

Database Migration Service | Snowball | Snowmobile | Kinesis Data Firehose | Kinesis Data Streams | Managed Streaming for Apache Kafka



Data movement services

Data movement

Migration & Streaming Services

Most ways to move data to the data lake

Professional services and partners
to help migration



Data movement from
your on-premises
datacenters



Data movement from
real-time sources

Synchronizing data
across environments

Data movement from on-premises datacenters

- Dedicated network connection
- Secure appliances
- Ruggedized shipping containers
- Database migration
- Gateway that lets applications write to the cloud

Data movement from real-time sources

- Connect devices to AWS
- Real-time data streams
- Real-time video streams



Data lake infrastructure & management services

Data lake infrastructure & management



S3/Glacier



Lake Formation



AWS Glue

Customers moving to data lake architectures

Bringing together the best of both worlds



Extends or evolves DW architectures

Store any data in any format

Durable, available, and exabyte scale

Secure, compliant, auditable

Run any type of analytics from DW to Predictive

Build on robust data lake infrastructure with Amazon S3

Data lake infrastructure & management

99.999999999% durability

Global replication capabilities

Management features

Cost-effective storage classes

Most partner integrations



Serverless ETL and data integration with AWS Glue

Data lake infrastructure
& management

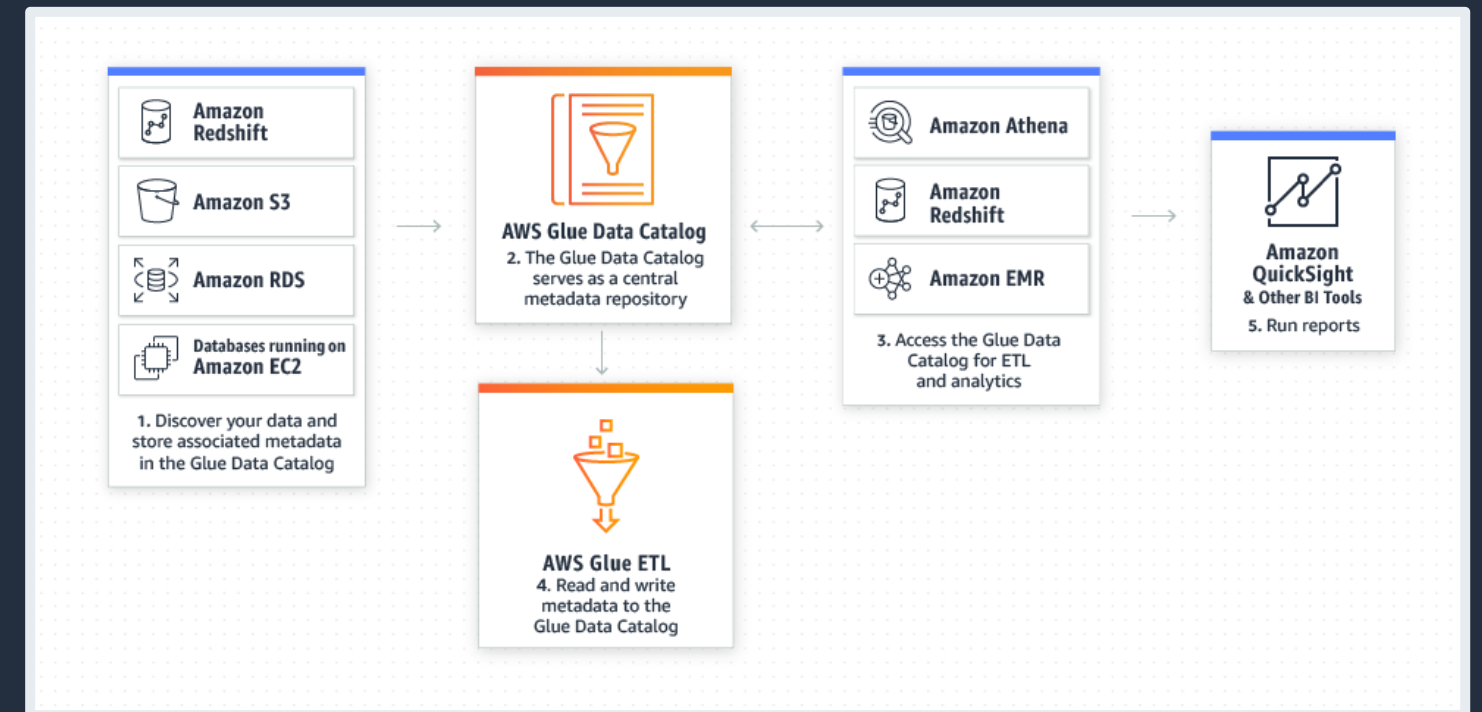
Serverless provisioning, configuration,
and scaling to run your ETL jobs on
Apache Spark and Python

Pay only for the resources used for jobs

Crawl your data sources, identify data
formats and suggest schemas and
transformations

Automates the effort in building,
maintaining and running ETL jobs

Coming soon—faster job start-up times
(under 2 minutes)



AWS Glue

Data lake infrastructure
& management

Simple, flexible, and cost-effective ETL & Data Catalog

Less hassle



Integrated across AWS: supports Aurora, RDS, Redshift, S3, and common database engines in your VPC running on EC2

Serverless



Serverless: no infrastructure to provision or manage

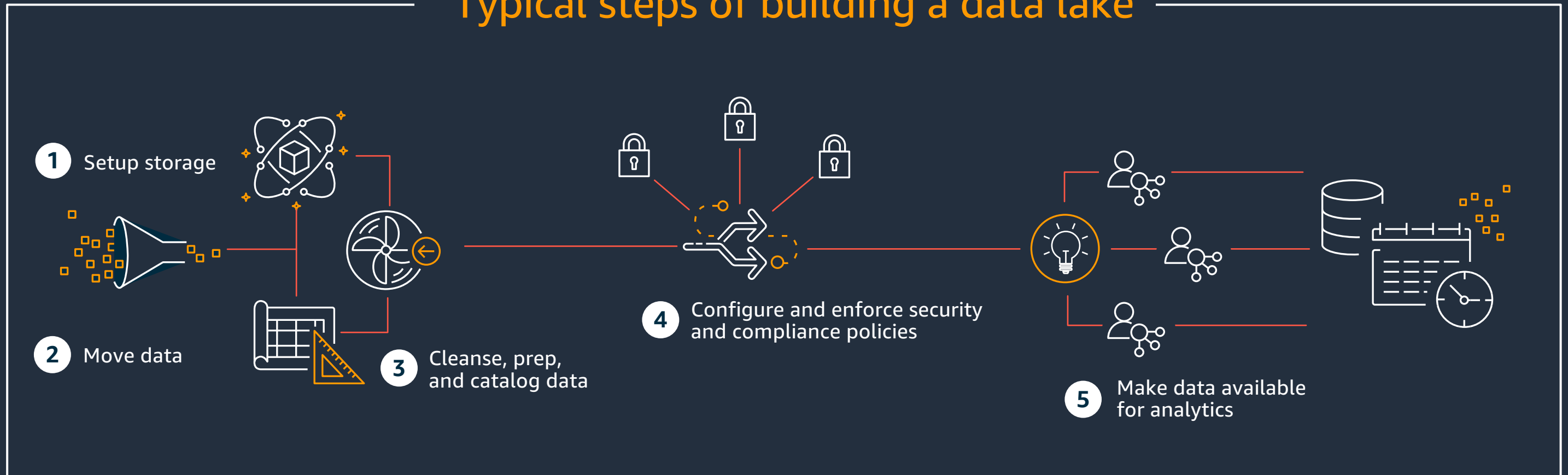
More power



Automatically generates the code to execute your data transformations and loading processes

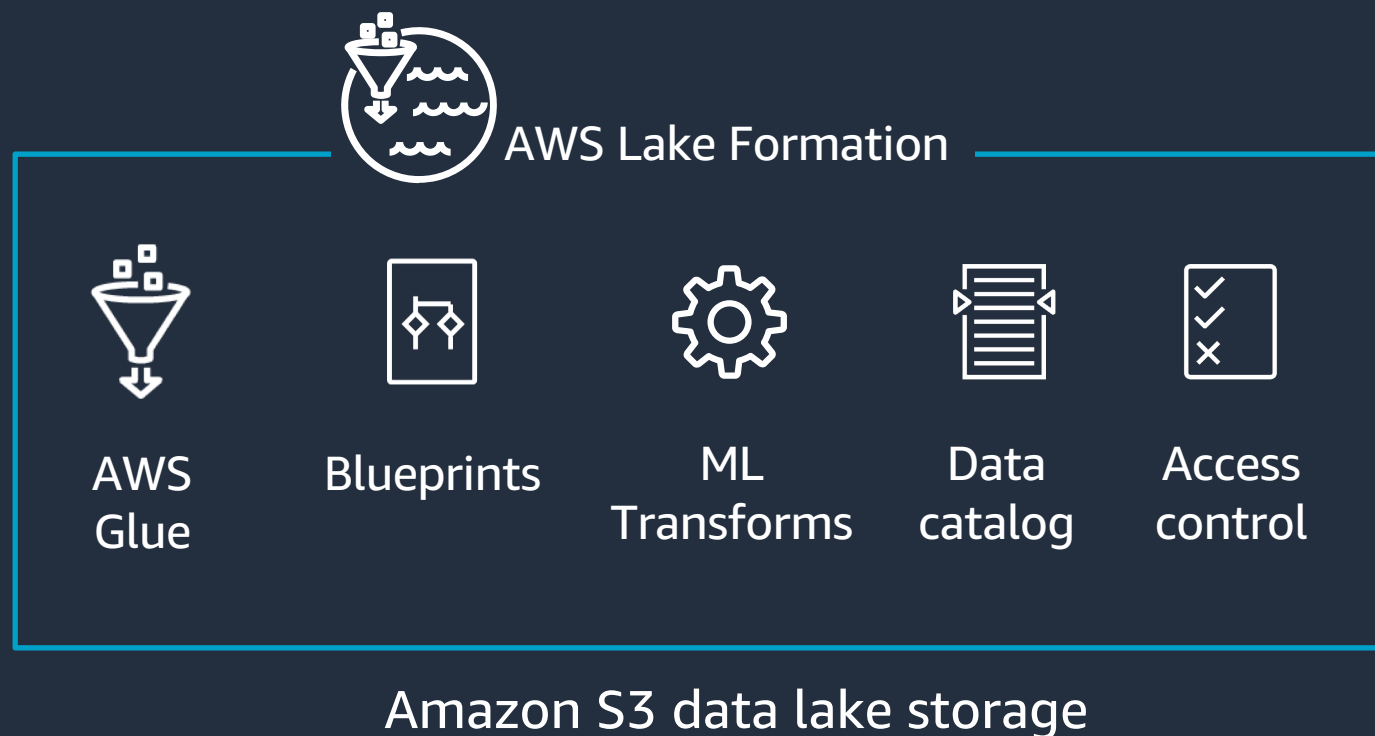
Challenges to making a secure data lake

Typical steps of building a data lake



Build a secure data lake in days with AWS Lake Formation

Data lake infrastructure
& management



Comprehensive set of integrated tools
enable user access consistently

Centralized management of fine-grained
permissions empower security officers


Simplified ingest and cleaning enables
data engineers to build faster




Analytics services


 **Big Data Processing**

 **Data Warehousing**

 **Real-time Analytics**

 **Operational Analytics**

 **Interactive Query**

 **Serverless Data processing**

Amazon EMR

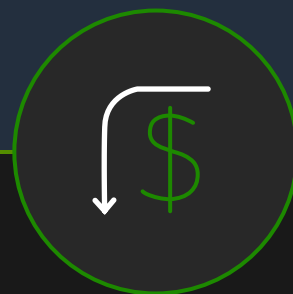
Easily Run Spark, Hadoop, Hive, Presto, HBase, and more big data apps on AWS

Latest versions



Updated with latest open source frameworks within 30 days

Low cost



50–80% reduction in costs with EC2 Spot and Reserved Instances
Per-second billing for flexibility

Use S3 storage



Process data in S3 securely with high performance using the EMRFS connector

Easy



Fully managed no cluster setup, node provisioning, cluster tuning

FINRA increases agility, speed, and cost-savings with an AWS Data Lake



Challenge

FINRA's legacy system did not scale to handle 150 billion events per day. They needed to run complex surveillance queries over 20+ PB of data to detect and analyze illegal market activity.

Solution

FINRA migrated their big data appliance to a S3 data lake and uses Lambda and EMR for data ingestion and EMR and Redshift for data processing.

Benefits

FINRA has been able to increase agility, speed, and cost savings while allowing them to operate at scale. The company estimates it will save \$10 to \$20 million annually.

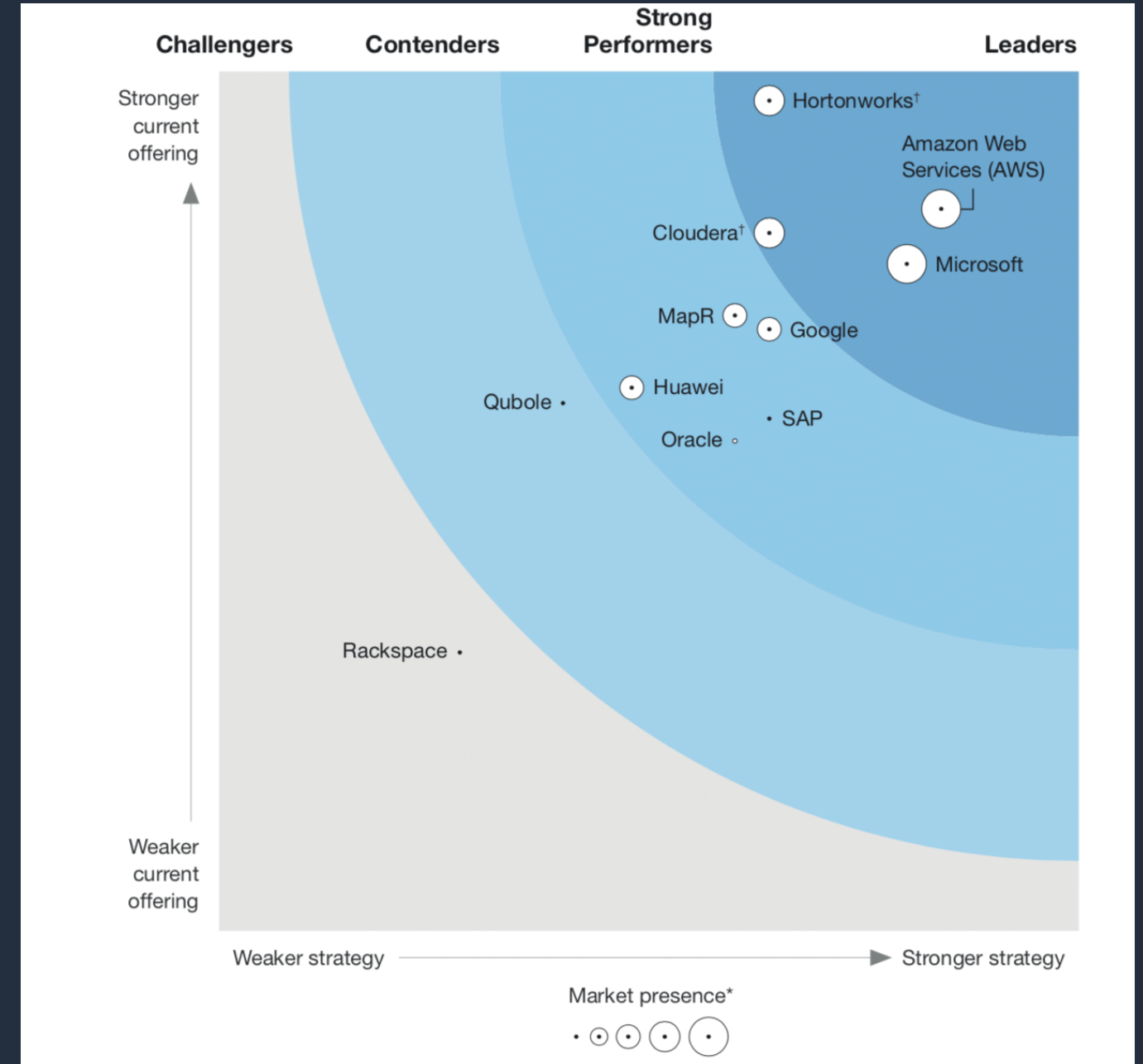
The Forrester Wave

Cloud Hadoop/Spark Platforms Q1 2019

The 11 Providers That Matter Most
and How They Stack Up

by Noel Yuhanna and Mike Gualtieri
February 13, 2019

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.



Amazon Redshift

The most popular and fastest cloud data warehouse

Analytics



Data lake integration

Query exabytes of data directly in open formats with no loading required



Faster performance

2x faster than other cloud DWHs



Secure

Security out of the box, at no extra cost



Cost-effective

Up to 75% less than other cloud DWHs



Input

Clickstream, finance, social and operations data



Amazon S3

Load or stream all data into your S3 data lake



Amazon Redshift

Redshift can query from high performance local disks or directly from Amazon S3 in open data formats



Output

Connect SQL clients and BI tools to give you insights that power business decisions, machine learning algorithms or personalized experiences



“ Amazon Redshift enables us to provide scientists with near real-time analysis of millions of rows of manufacturing data generated by continuous manufacturing equipment, with 1,600 data points per row. Redshift enables us to query our high-volume data at 10 times the performance of our prior data warehousing solution. Because of the performance and scale Redshift provides, we have increased our manufacturing efficiency by optimizing future manufacturing runs. In addition, we have reduced the time needed to gather and prepare data for regulatory submissions by a factor of five and now avoid repeated experimentation, which would otherwise have taken an extra three weeks of scientists' time. ”

—Jim Silva
Director Business Partner

Data warehousing: Amazon Redshift

First and most popular cloud data warehouse

Data lake & AWS integration



Analyze exabytes of data across data warehouse, data lakes, and operational database

Query data across various analytics services

Best performance, most scalable



3x faster with RA3*

10x faster with AQUA*
*vs other cloud DWs

Adds unlimited compute capacity on-demand to meet unlimited concurrent access

Most secure & compliant



AWS-grade security (eg. VPC, encryption with KMS, CloudTrail)

All major certifications such as SOC, PCI, DSS, ISO, FedRAMP, HIPPA

Lowest cost



Cost-optimized workloads by paying compute and storage separately

1/10th cost of Traditional DW at \$1000/TB/year

Up to 75% less than other cloud data warehouses & predictable costs



“ We migrated to Amazon Redshift in 2014 because it was 10 times faster than our prior on-premises system. Today, it is the center of our analytics environment. Since we first started using Amazon Redshift, we have added thousands of analysts and data scientists to analyze tens of petabytes daily. Redshift provides our users with consistently faster performance, even as its usage within the company has grown. ”

—Masayuki Tsuda
General Manger of Service Innovation Department

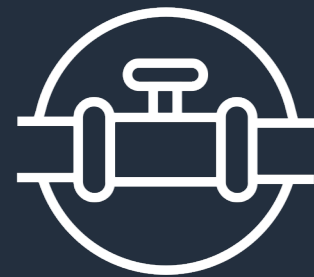
Real-time: Amazon Kinesis

Easily collect, process, and analyze video and data streams in real time



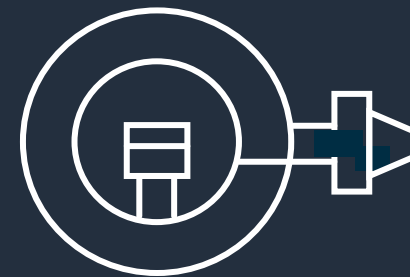
Kinesis Video Streams

Capture, process, and store video streams for analytics



Kinesis Data Streams

Build custom applications that analyze data streams



Kinesis Data Firehose

Load data streams into AWS data stores



Kinesis Data Analytics

Analyze data streams with SQL

Operational Analytics

Fully managed, scalable, secure, Elasticsearch service

Open source Elasticsearch
APIs, Kibana, and
Logstash



Open-source Elasticsearch APIs
Managed Kibana
Integration with Logstash

Fully managed



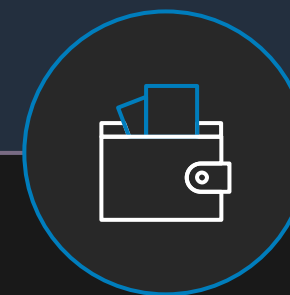
Deploy Elasticsearch clusters
in minutes: simplified hardware
provisioning, software
installation/patching, failure
recovery, backups, and monitoring

Scalable, secure,
and compliant



Scale clusters up/down via a
single API call or a few clicks
Secured network isolation
with VPC, encrypt data
at-rest and in-transit
Compliant: HIPPA, PCI DSS,
and ISO

Pay only for
what you use



Cost-optimized workloads
No upfront fee or
usage requirement
Critical features built-in:
encryption, VPC support,
24x7 monitoring

Open Distro for Elasticsearch



Analytics

An Apache 2.0-licensed distribution for Elasticsearch
Enhanced with enterprise security, alerting, SQL, and more



100% open source

Providing you the freedoms, so you can freely view, use, change, and distribute the code



Enterprise-grade

Delivering security and advanced capabilities such as alerting, SQL, and cluster diagnostics



Community-driven

Providing individuals and organizations the freedom to easily contribute changes to the distro

Amazon Athena

Serverless, interactive query service

Query instantly



Zero setup cost

Point to S3 and start querying

Pay per query



Pay only for queries run

Save 30–90% on per-query costs through compression

Use S3 storage



ANSI SQL

JDBC/ODBC drivers

Multiple formats, compression types, and complex joins and data types

Easy



Serverless: zero infrastructure, zero administration

Integrated with QuickSight

Movable Ink

“ One of the big attractions of Amazon Athena is that it’s serverless and purely consumption-based. We only pay when we’re actually querying the data, and we don’t have to keep a cluster running all the time. Using Amazon Athena, we’re able to query seven years’ worth of data—adding up to hundreds of terabytes—get results at least 50 percent faster, and save nearly \$15,000 per month. ”

—Matt Chesler
Director of DevOps

Serverless analytics

Deliver on-demand analytics on the data lake




Serverless
Zero infrastructure
Zero administration



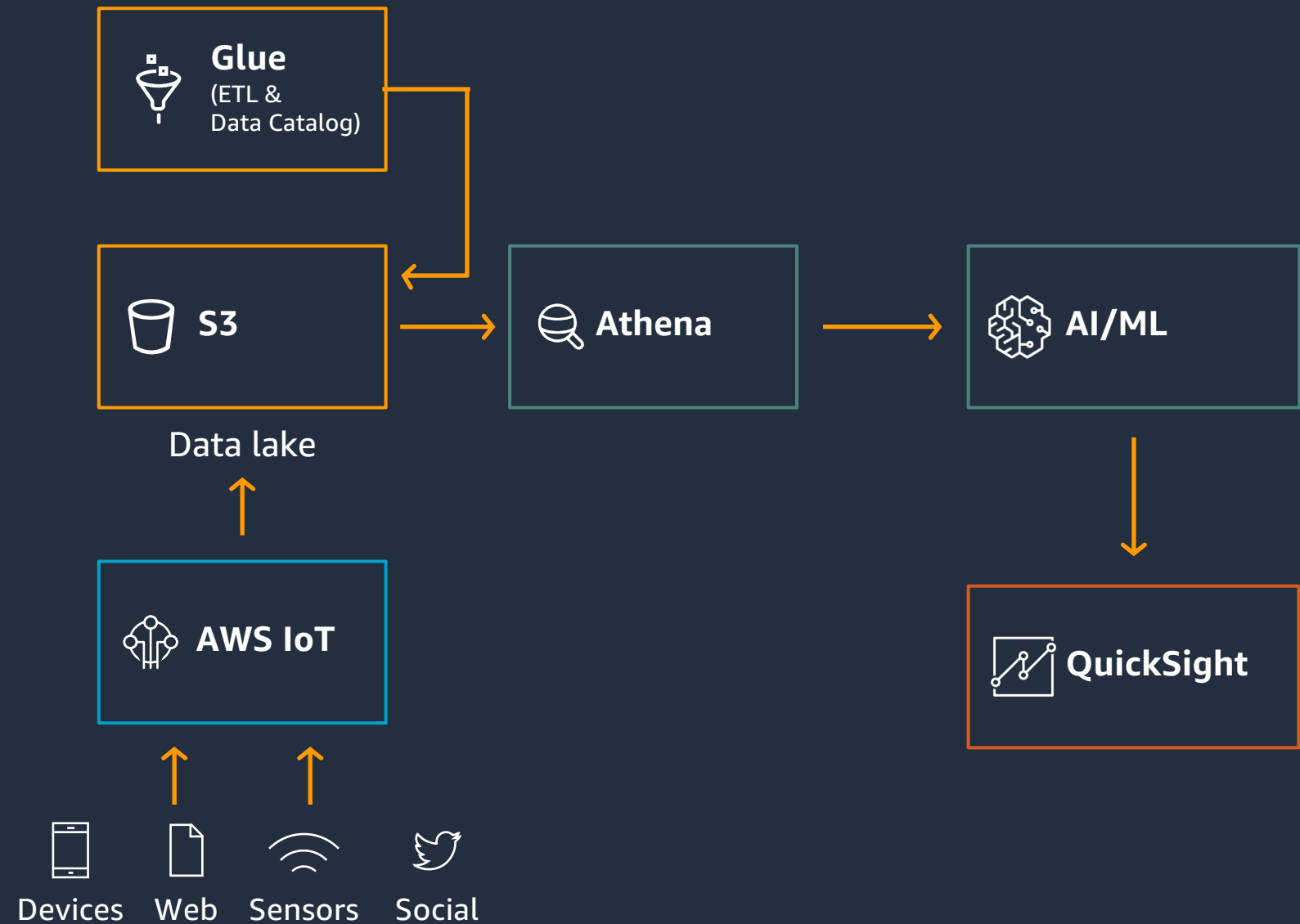
Never pay for
idle resources



Automatically scales
resources with usage



Availability and fault
tolerance built in





Data, visualization, engagement, & machine learning services

Data, visualization, engagement, & machine learning



Data



Dashboards



Digital User Engagement



Predictive Analytics

Data lakes for machine learning

Data, visualization,
engagement, & ML

Easier to discover relevant data

More data makes more accurate and complete models

More data sources provide more context and nuance

More compute resources available when needed

More specialized compute resources when needed

Granular control over what kinds of data is seen

Costs reduced by separating storage from compute

AWS Data Exchange

Easily find and subscribe to 3rd-party data in the cloud

Data, visualization,
engagement, & ML

**Quickly find diverse
data in one place**



>1,000 data products
>80 data providers including
include Dow Jones, Change
Healthcare, Foursquare, Dun
& Bradstreet, Thomson
Reuters, Pitney Bowes, Lexis
Nexis, and Deloitte

Easily analyze data



Download or copy data to S3
Combine, analyze, and model
with existing data

Analyze data with EMR,
Redshift, Athena, and AWS
Glue

**Efficiently access
3rd party data**



Simplifies access to data: No
need to receive physical media,
manage FTP credentials, or
integrate with different APIs

Minimize legal reviews and
negotiations

Amazon QuickSight

First BI service built for the cloud with pay-per-session pricing & ML insights

Data, visualization,
engagement, & ML

Elastic Scaling



Auto-scale 10 to 10K+
users in minutes

Pay-as-you-go

Serverless



Create dashboards in
minutes

Deploy globally
without provisioning a
single server

Deeply integrated with AWS services



Secure, Private access to
AWS data

Integrated S3 data lake
permissions through AWS IAM

API Support

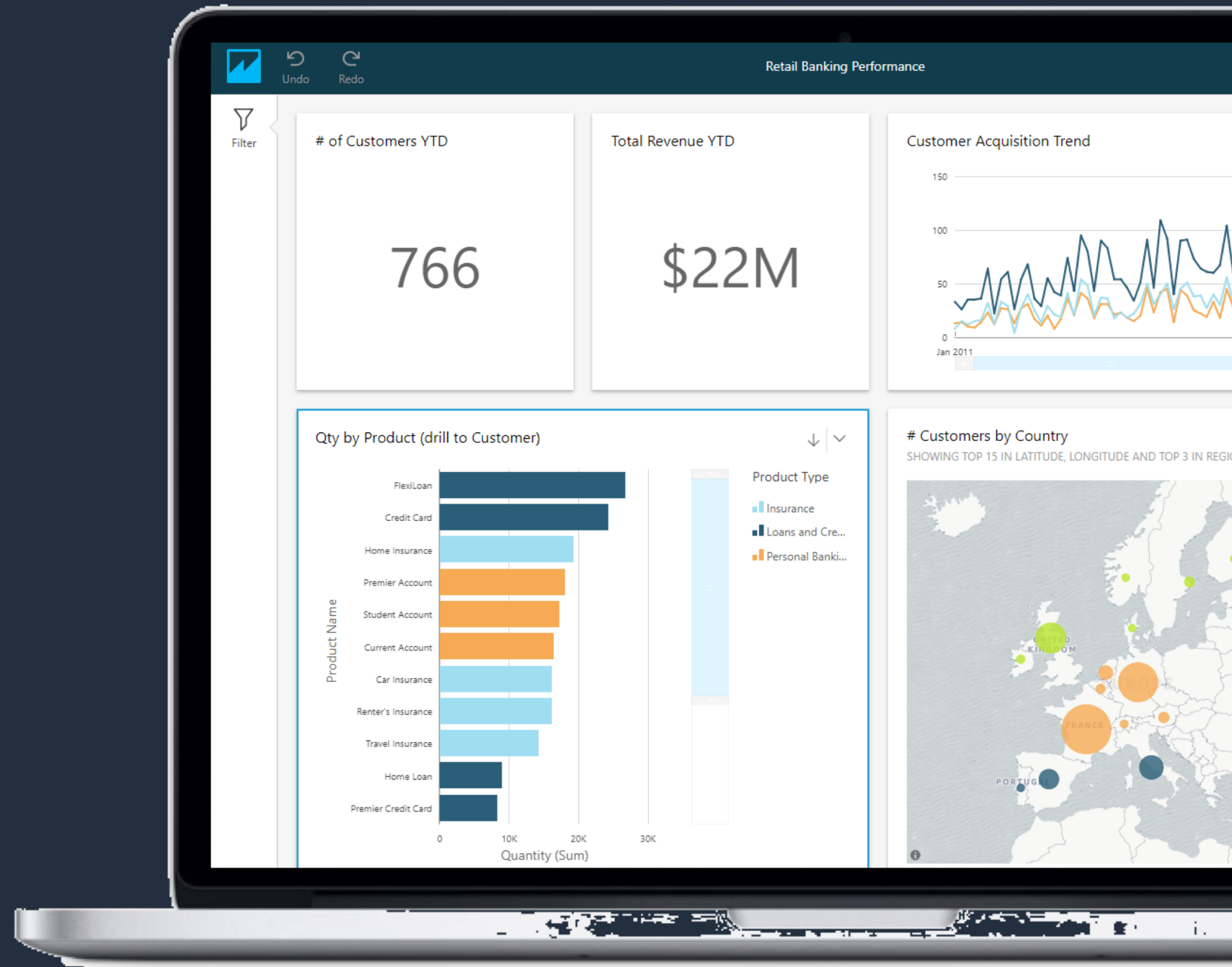


Programmatically onboard users
and manage content

Easily embed in your apps

Create beautiful, interactive dashboards.

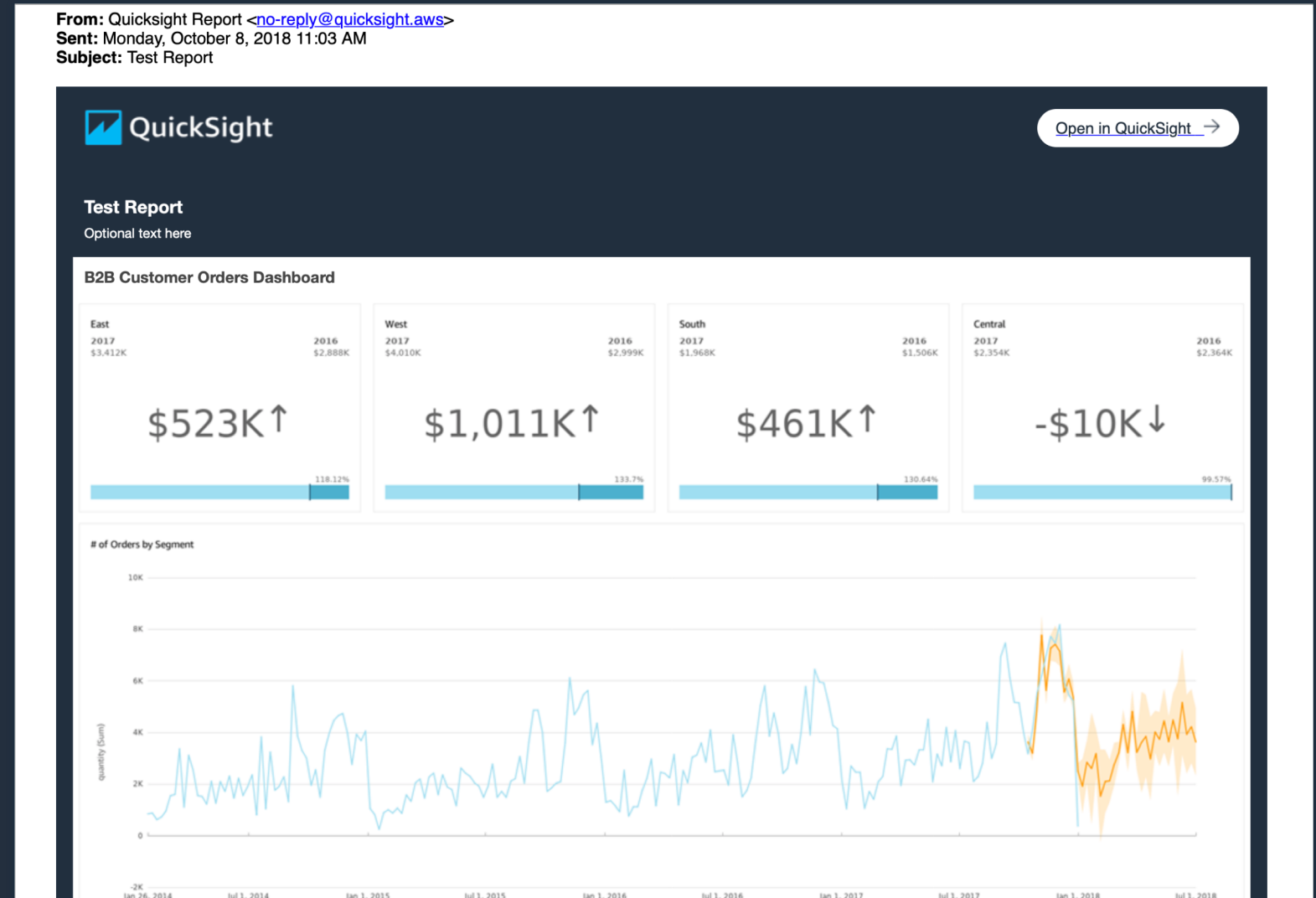
- Add rich interactivity like filters, drill downs, zooming, and more
- Blazing fast navigation
- Accessible on any device
- Data Refresh
- Publish to everyone with a click



Insights Delivered to Your Inbox

QuickSight lets you send report snapshots directly to your users inbox.

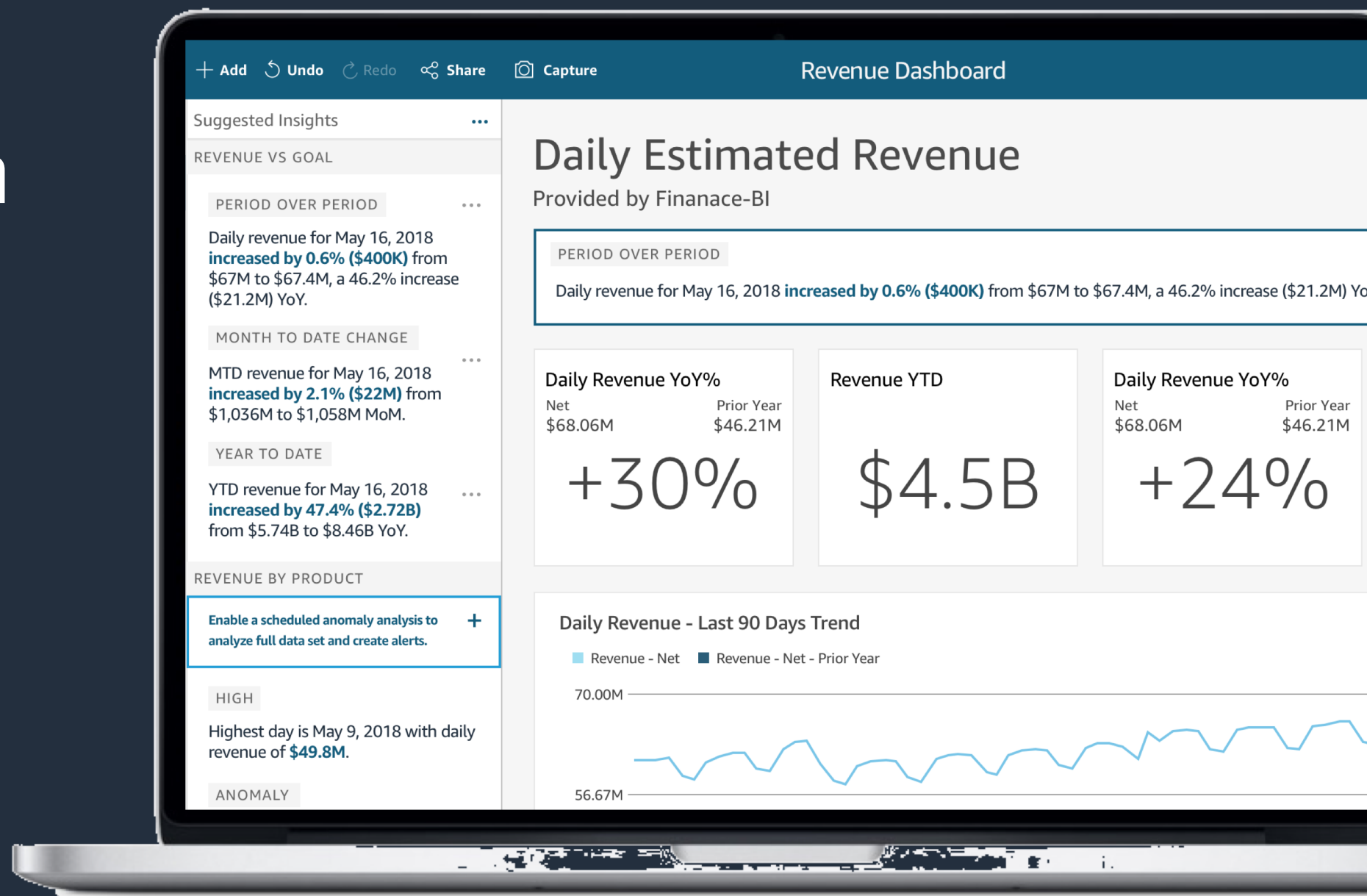
- Schedule email reports on a daily, weekly, or monthly basis
- Sent to individual users or groups
- Users can opt out of any report so they can focus on what's important.
- Uses Pay-per-Session Pricing



Introducing ML Insights

Cutting edge ML tools that automatically discover powerful insights for your users.

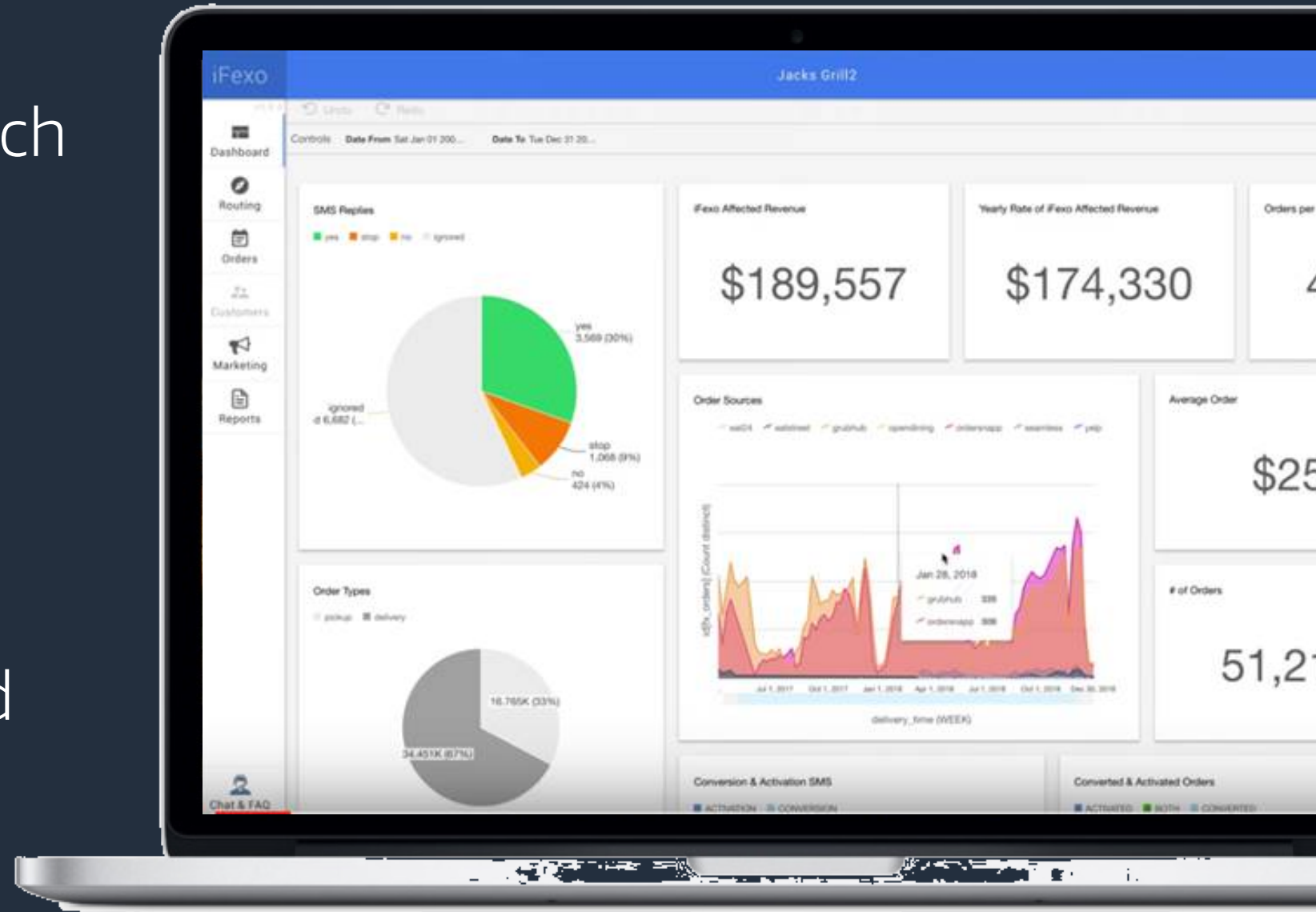
- ML powered Anomaly Detection
- ML Powered Forecasting
- Auto-generated natural language narratives and summaries.



Embedding Dashboards In Your Application

QuickSight allows you to seamlessly integrate interactive dashboards and analytics into your own applications

- Enhance your applications with rich analytics and dashboards
- Easy maintenance, no servers to manage
- Fast! No Custom development or domain expertise needed
- Leverage new features as we add them
- Utilizes Pay-per-Session Pricing.



Successfully engage your customers with Amazon Pinpoint

Data, visualization,
engagement, & ML

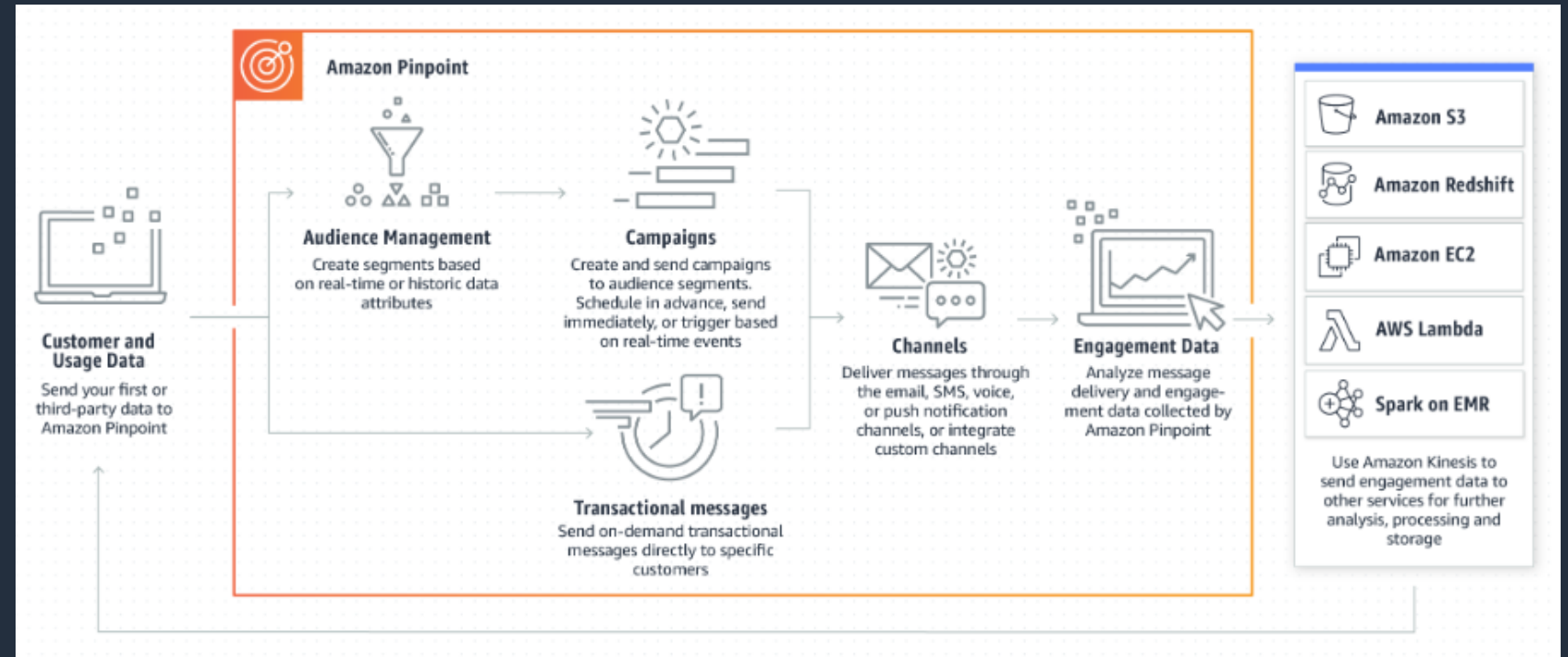
Understand our customers

Segment based
upon understandings

Target in a contextually
relevant way

Communicate in best channel

React to customers responses



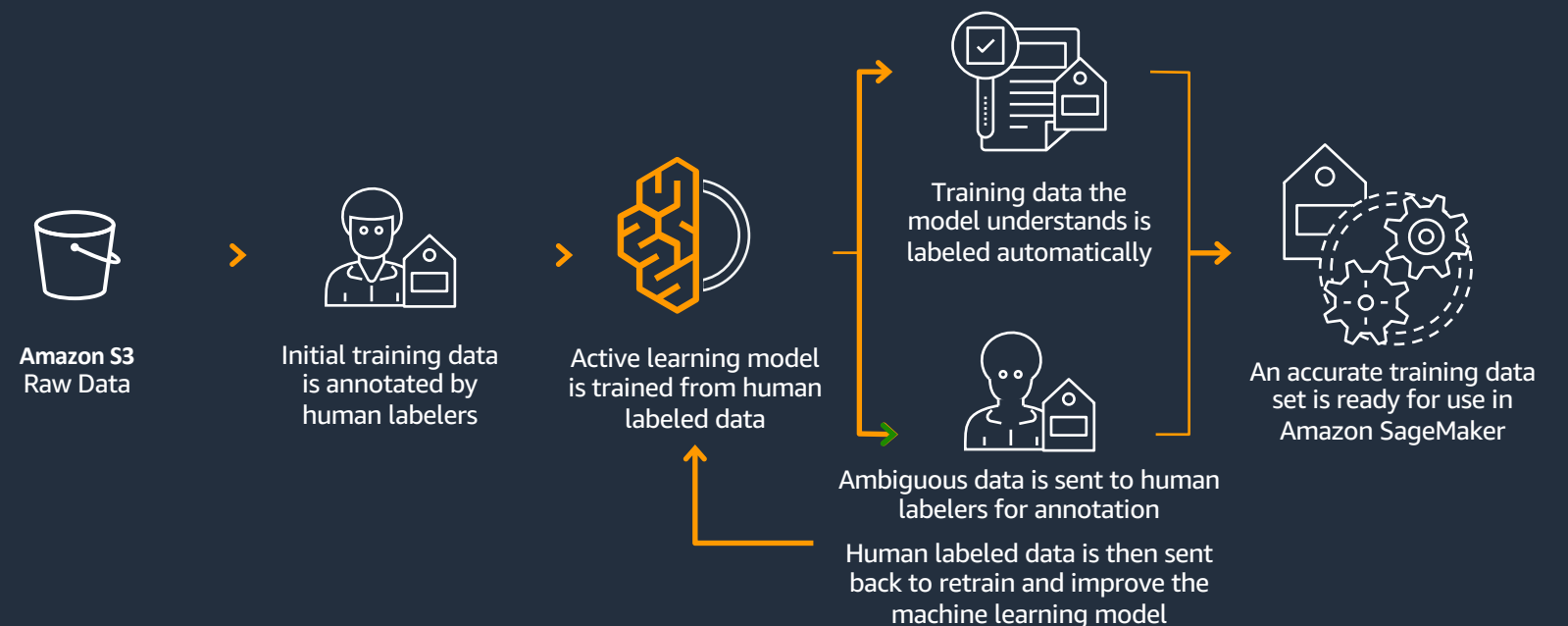
Predictive insights with AWS ML & AI services

Data, visualization,
engagement, & ML

AI services that enable developers
to plug-in pre-built
AI functionality into their apps

ML platform services that make it
easy for any developer to get
started and get deep with ML

ML frameworks and interfaces for
machine learning practitioners



Amazon.com Data Lakes on AWS

Amazon.com lowers cost and gains faster insights with an AWS Data Lake

Challenge

Amazon needed to analyze a massive amount of data to find insights, identify opportunities, and evaluate business performance.

The Oracle DW did not scale, was difficult to maintain, and costly.

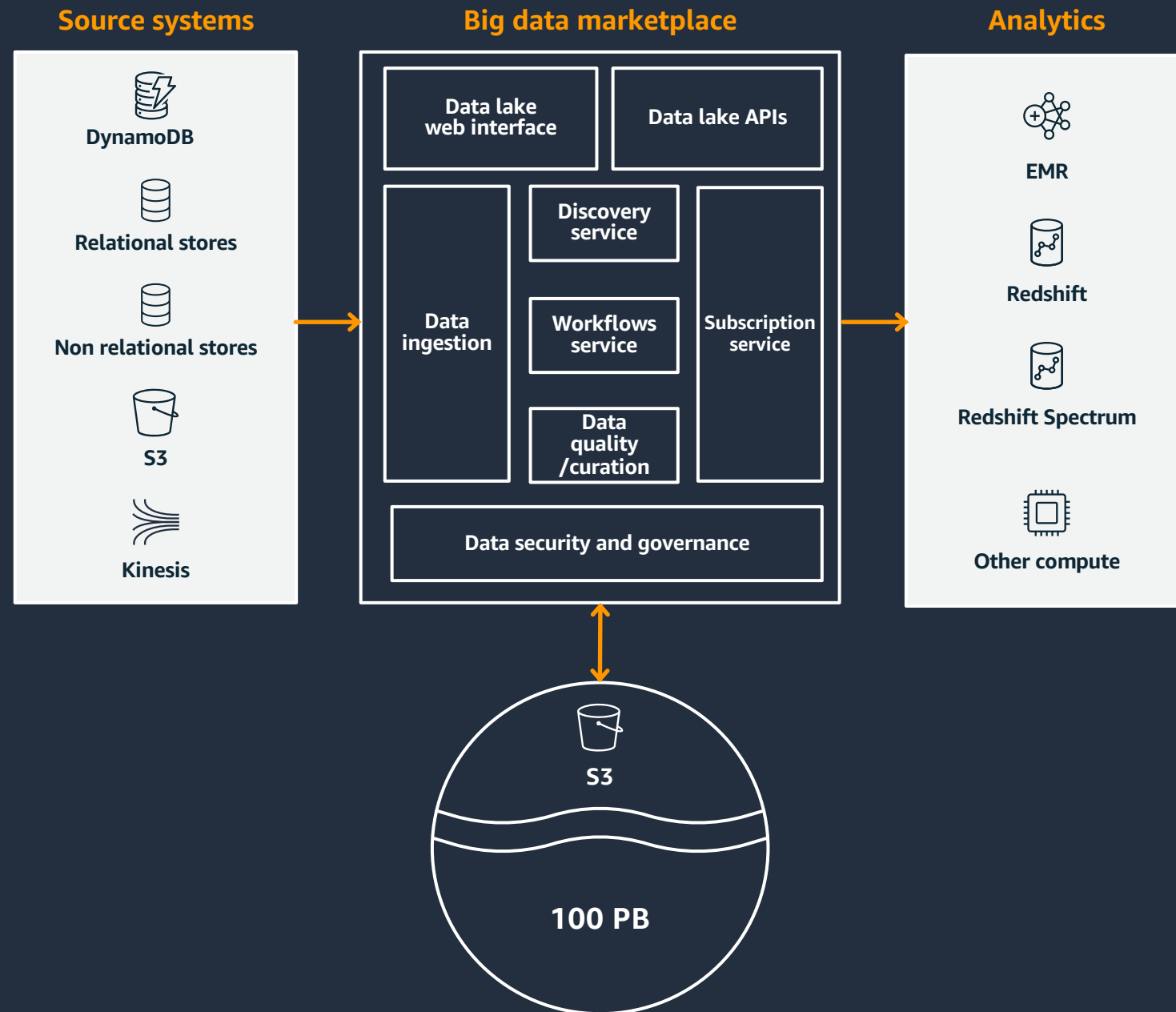
Solution

Amazon deployed a data lake with Amazon S3, and now runs analytics with Amazon Redshift, Redshift Spectrum, and Amazon EMR.

Benefits

They doubled the data stored from 50 PB to 100 PB, lowered costs, and were able to gain insights faster.

Amazon uses an AWS Data Lake



- 50 PB of data
- 600,000 analytics jobs/day

Next steps...

Dive deeper into specific AWS services

Set up a proof-of-concept

Talk about how professional services can help

1

Sign up for an AWS account

Instantly get access to the AWS Free Tier

2

Learn with 10-minute tutorials

Explore and learn with simple tutorials

3

Start building with AWS

Begin building with step-by-step guide to help you launch your AWS project

Thank you!