aws

SUMMIT ONLINE

8

From raw data to business Insights. What you need to build a modern data lake

Javier Ramirez

Developer Advocate

@supercoco9

aws SUMMIT ONLINE

Data is a strategic asset for every organization

> "The world's most valuable resource is no longer oil, but data."

David Parkins, 2017, The Economist

The Economist
MAY 6TH–12TH 2017
Theresa May v Brussels
Ten years on: banking after the crisis
South Korea's unfinished revolution
Biology, but without the cells
The world's most valuable resource
amazon
f
Data and the new rules of competition

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

Line of business (LOB) knowledge workers

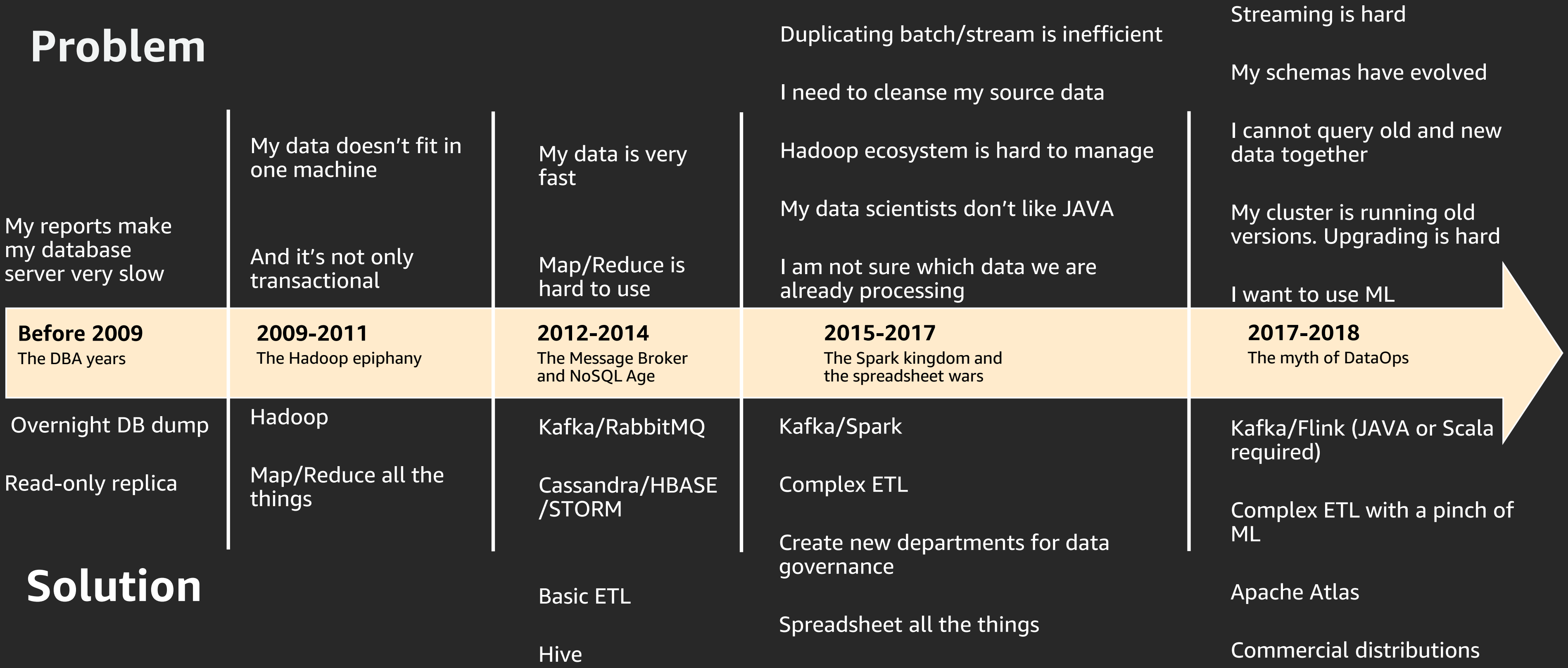
Product managers

IT operations

IT security and governance

VP/director analytics

A brief opinionated history of data analytics



	Before 2009 The DBA years	**2009-2011** The Hadoop epiphany	**2012-2014** The Message Broker and NoSQL Age	**2015-2017** The Spark kingdom and the spreadsheet wars	**2017-2018** The myth of DataOps
Problem	My reports make my database server very slow	My data doesn't fit in one machine And it's not only transactional	My data is very fast Map/Reduce is hard to use	Duplicating batch/stream is inefficient I need to cleanse my source data Hadoop ecosystem is hard to manage My data scientists don't like JAVA I am not sure which data we are already processing	Streaming is hard My schemas have evolved I cannot query old and new data together My cluster is running old versions. Upgrading is hard I want to use ML
Solution	Overnight DB dump Read-only replica	Hadoop Map/Reduce all the things	Kafka/RabbitMQ Cassandra/HBASE /STORM Basic ETL Hive	Kafka/Spark Complex ETL Create new departments for data governance Spreadsheet all the things	Kafka/Flink (JAVA or Scala required) Complex ETL with a pinch of ML Apache Atlas Commercial distributions

Customers want more value from their data

Growing exponentially

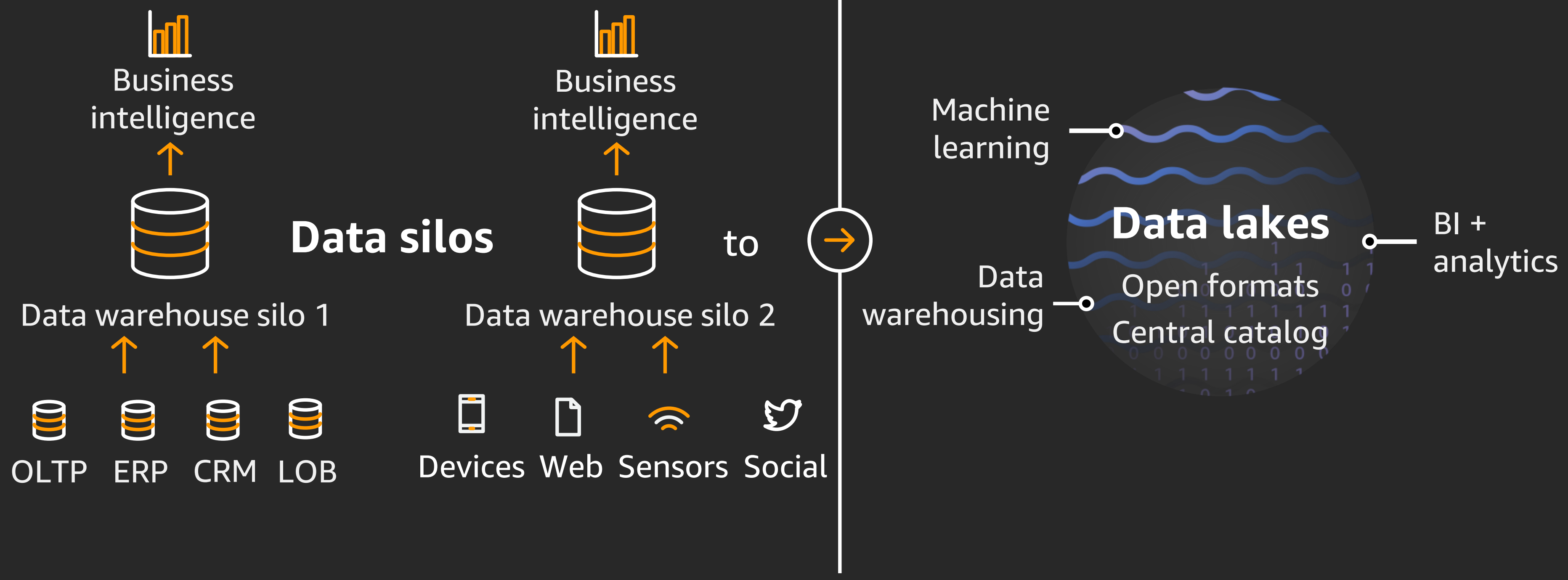
From new sources

Increasingly diverse

Used by many people

Analyzed by many applications

Traditional data warehousing approaches don't scale



Business intelligence

Business intelligence

Data silos

Data warehouse silo 1

Data warehouse silo 2

OLTP ERP CRM LOB

Devices Web Sensors Social

to

Machine learning

Data warehousing

Data lakes

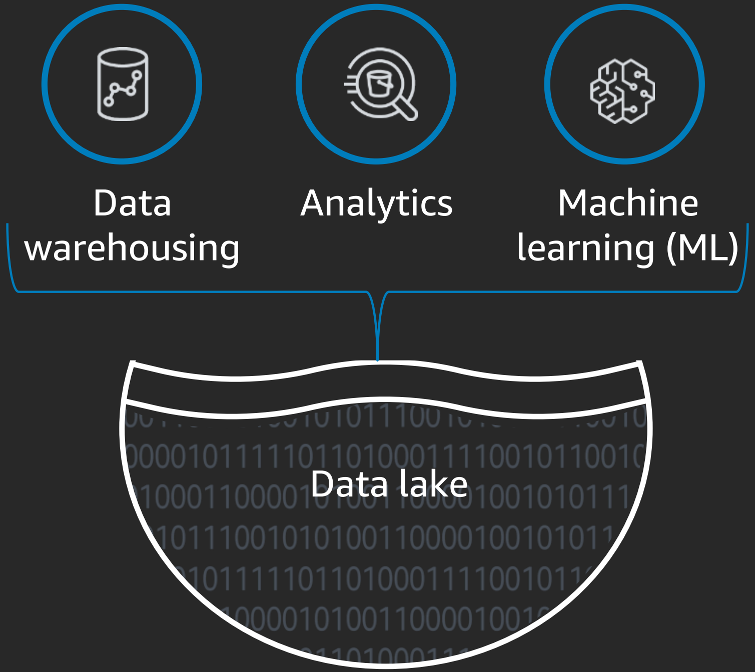
Open formats

Central catalog

BI + analytics

Customers moving to data lake architectures

Bringing together the best of both worlds



Data warehousing

Analytics

Machine learning (ML)

Data lake

Extends or evolves data warehouse architectures

Store any data in any format

Durable, available, and exabyte-scale

Secure, compliant, and auditable

Run any type of analytics from data warehouse to predictive

Modern data analytics 101 – Data Lake Basics

A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale.

Modern data analytics 101 – Data Lake Basics

A good data lake allows self-service and can easily plug-in new analytical engines.

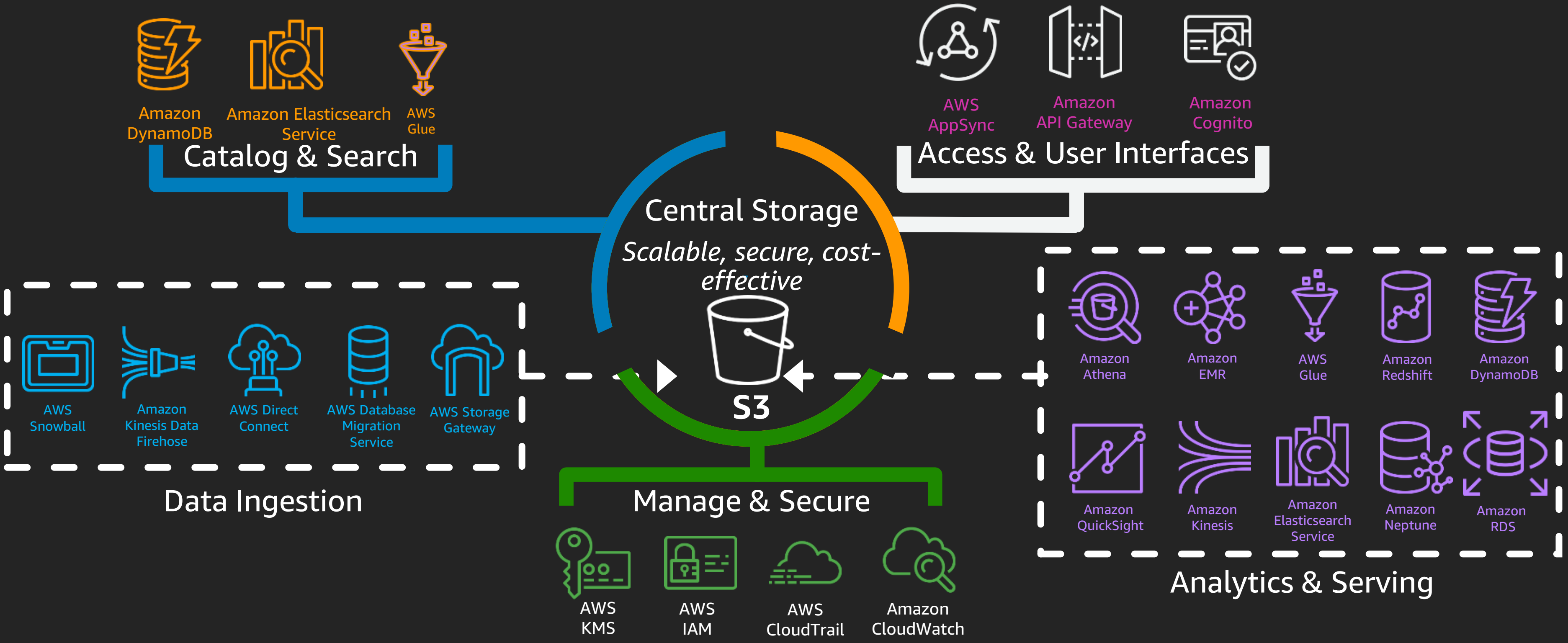
A possible open source solution

- Hadoop Cluster (static/multi tenant)
- Apache NiFi for ingestion workflows
- Sqoop to ingest data from RDBMS
- HDFS to store the data (tied to the Hadoop cluster)
- Hive/HCatalog for data Catalog
- Apache Atlas for a more human data catalog and governance
- Apache Spark for complex ETL –with Apache Livy for REST
- Hive for batch workloads with SQL
- Presto for interactive queries with SQL
- Kafka for streaming ingest
- Apache Spark/Apache Flink for streaming analytics
- Apache Hbase (or maybe Cassandra) to store streaming data
- Apache Phoenix to run SQL queries on top of Hbase
- Prometheus (or fluentd/collectd/ganglia/Nagios…) for logs and monitoring. Maybe with Elastic Search/Kibana
- Airflow/Oozie to schedule workflows
- Superset for business dashboards
- Jupyter/JupyterHub/Zeppelin for data science
- Security (Apache Sentry for Roles, Ranger for configuration, Knox as a firewall)
- YARN to coordinate resources
- Ambari for cluster administration
- Terraform/chef/puppet for provisioning

Some problems you will find

- My team spends more time maintaining the cluster than adding functionality
- Security and monitoring are hard
- Most of my time my cluster is sitting idle; Then it’s a bottleneck
- I don’t have the time to experiment
- Highly specialized profiles: Niches of knowledge and talent problem

Or a cloud native solution on AWS



More data lakes and analytics than anywhere else

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

3M Science. Applied to Life.™
3VICTORS TRAVEL AI
ACQUIA
A+E
amazon
Adobe
AMGEN
ATLASSIAN

Agero
airbnb
BARCLAYS
Capital One
Chick-fil-A
CIVIS
CIVITAS LEARNING
CoffeeMeetsBagel

dataxu
CROWDSTRIKE
DOMO
EA
EPIC GAMES
FAST RETAILING
FINANCIAL TIMES
FOURSQUARE
FUNNY OR DIE

GP Georgia-Pacific
gogo
HEARST corporation
hudl
HYATT
instacart
JustGiving
lyft

MONSTER
mparticle
Nasdaq
NETFLIX
docomo
NU SKIN
Olive
OLX GROUP
Panasonic
Pfizer

Pinterest
pitney bowes
realnetworks
REDFIN
ROVIO
SCHOLASTIC
slack
SONOS
reddit

T··Mobile
Sysco
THOMSON REUTERS
ticketmaster
TUNE
verizon
yelp
Zillow
ZUMBA

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 (Amazon S3) for all analytics and ML
- A service to build secure data lakes in days
- Deep integration across analytics and infrastructure (including federated queries)

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

S3 in action at zalando

As Europe’s leading online fashion platform we deliver to customers in 17 countries. In our fashion store, they can find a wide assortment from more than 2,500 brands

400 teams, with over 8000 S3 buckets, 15PB volume.

They use an event bus whose major purpose is to service communication among distributed microservices, and wanted to save a copy of all published messages in the data lake.

Their fully serverless solution uploads event batches to S3 several million times per day. The data lake also contains web tracking data, and data from their previously existing data warehouse.

Why did zalando choose S3?

Durability, Availability, Scalability, and breadth and depth of AWS services

By using Amazon S3 Intelligent Tiering, they are saving 37% annually in storage. S3 automatically moves objects that have not been touched within 30 days to S3 Standard IA, only moving them back to S3 Standard when they get accessed

“We evaluated multiple cloud providers, and AWS was chosen as the cloud provider of choice due to its durability, availability, and scalability. We also considered the expansive ecosystem of services that AWS offers that we could leverage in the future.”

Max Schultze, Lead Data Engineer

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 lakes

Security

- Amazon GuardDuty
- AWS Shield
- AWS WAF
- Amazon Macie
- Amazon VPC

Identity

- IAM
- AWS SSO
- Amazon Cloud Directory
- AWS Directory Service
- AWS Organizations

Encryption

- AWS Certificate Manager
- AWS Key Management Service
- Encryption at rest
- Encryption in transit
- Bring your own keys, HSM support

Compliance

- AWS Artifact
- Amazon Inspector
- AWS CloudHSM
- Amazon Cognito
- AWS CloudTrail

2. Most secure infrastructure: Certifications

Global

CSA
Cloud Security Alliance controls

ISO 9001
Global quality standard

ISO 27001
Security management controls

ISO 27017
Cloud-specific controls

ISO 27018
Personal data protection

PCI DSS Level 1
Payment card standards

SOC 1
Audit controls report

SOC 2
Security, availability & confidentiality report

SOC 3
General controls report

United States

CJIS
Criminal Justice Information Services

DoD SRG
DoD data processing

FedRAMP
Government data standards

FERPA
Educational privacy act

ISO FFIEC
Financial institutions regulation

FIPS
Government security standards

FISMA
Federal information security management

GxP
Quality guidelines and regulations

HIPAA
Protected health information

ITAR
International arms regulations

MPAA
Protected media content

NIST
National Institute of Standards and Technology

SEC Rule 17a-4(f)
Financial data standards

VPAT/Section 508
Accountability standards

Asia Pacific

FISC [Japan]
Financial Industry Information Systems

IRAP [Australia]
Australian security standards

K-ISMS [Korea]
Korean information security

MTCS Tier 3 [Singapore]
Multi-Tier Cloud Security Standard

My Number Act [Japan]
Personal information protection

Europe

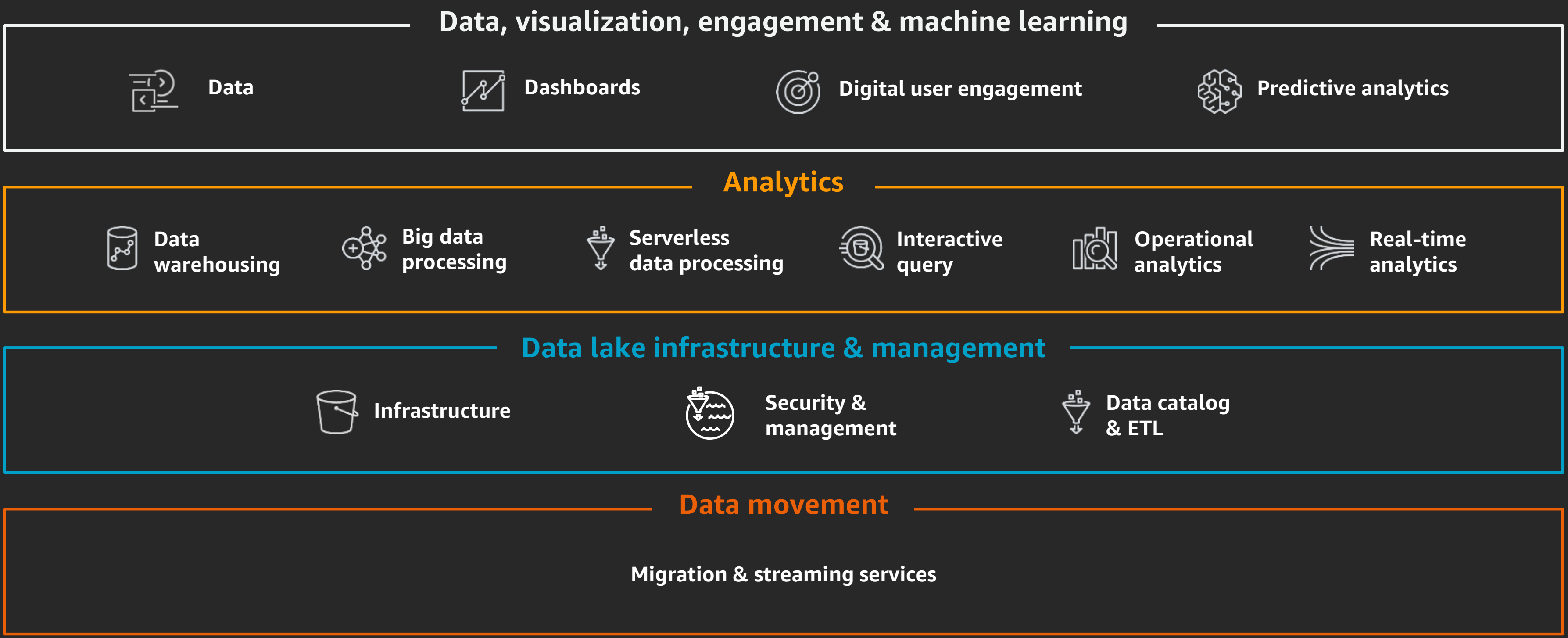
C5 [Germany]
Operational security attestation

Cyber Essentials Plus [UK]
Cyber threat protection

G-Cloud [UK]
UK government standards

IT-Grundschutz [Germany]
Baseline protection methodology

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

- AWS Data Exchange
- Amazon QuickSight
- Amazon Pinpoint
- Amazon SageMaker
- Amazon Comprehend
- Amazon Lex
- Amazon Polly
- Amazon Rekognition
- Amazon Translate

+ Many more

Analytics

- Amazon Redshift
- Amazon EMR (Spark & Hadoop)
- AWS Glue (Spark & Python)
- Amazon Athena
- Amazon Elasticsearch Service
- Amazon Kinesis Data Analytics

Data lake infrastructure & management

- Amazon S3/ Amazon S3 Glacier
- AWS Lake Formation
- AWS Glue

Data movement

AWS Database Migration Service | AWS Snowball | AWS Snowmobile | Amazon Kinesis Data Firehose | Amazon Kinesis Data Streams | Amazon 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

Amazon EMR

Automatic scaling

57% less than on-premises per IDC report

Amazon Redshift

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

Amazon Athena & Amazon QuickSight

Serverless; pay only for what is used

Pricing per session for visualization

Amazon Redshift at CAF

Spanish mobility solutions company CAF builds subway, suburban, inter-city and high-speed trains. LeadMind is their digital solution for fleet management and smart maintenance through advanced analytics.

1000+ trains from 26 fleets. Every train collects 10-20K variables at millisecond frequency. About 1.687M data points per 16 hours of daily train operation.

Their solution combines Redshift, Athena, and EMR for different types of analytics.

The analyzed data highlights the performance of train components—from the operation of the air conditioning to the effectiveness of the braking systems.

Amazon Redshift at CAF

Reduced CAPEX, scalability, IoT security, easy to manage

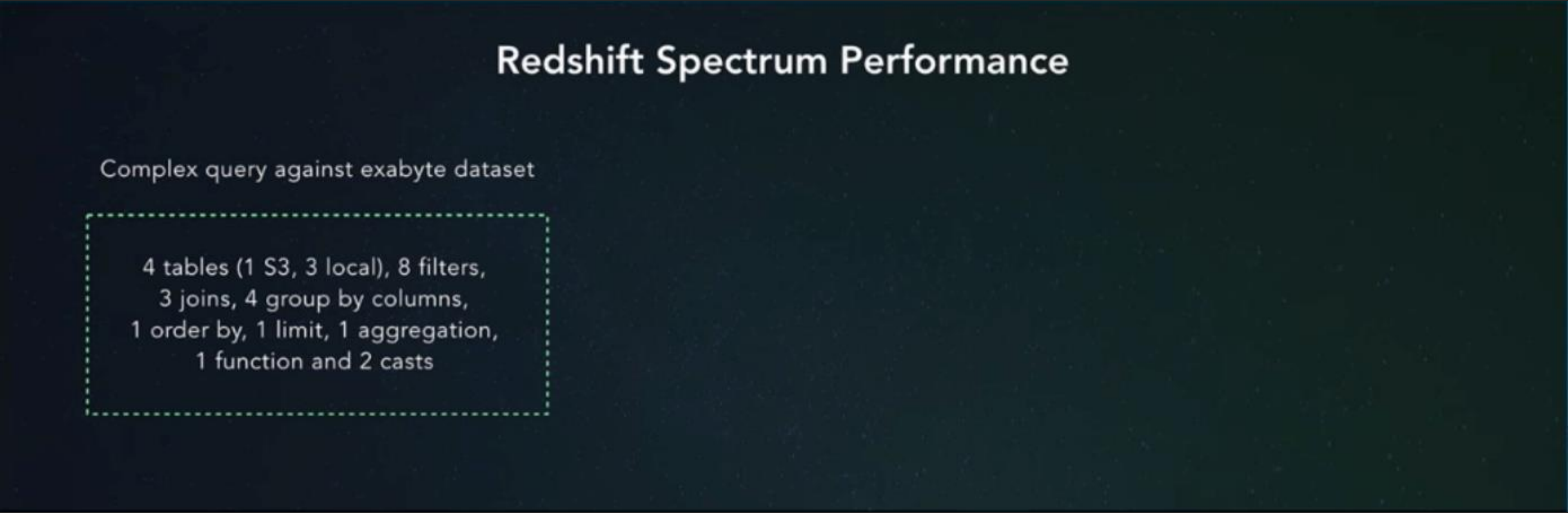
- 10 to 40 times reduction of breakdown rates
- 5% improvement in diagnostic reliability
- Reduction of the maintenance cost, by optimizing the maintenance taks and extending components and equipment life.

“Thanks to the managed services in AWS and LeadMind ,our data scientists have additional time to create more-effective predictive maintenance models to help customers quickly identify potential issues with trains and maximize standards of safety”

Managing Director of CAF Digital Services

Let’s play a game

Amazon Redshift Spectrum extends your data warehouse queries across your S3 data lake



Redshift Spectrum Performance

Complex query against exabyte dataset

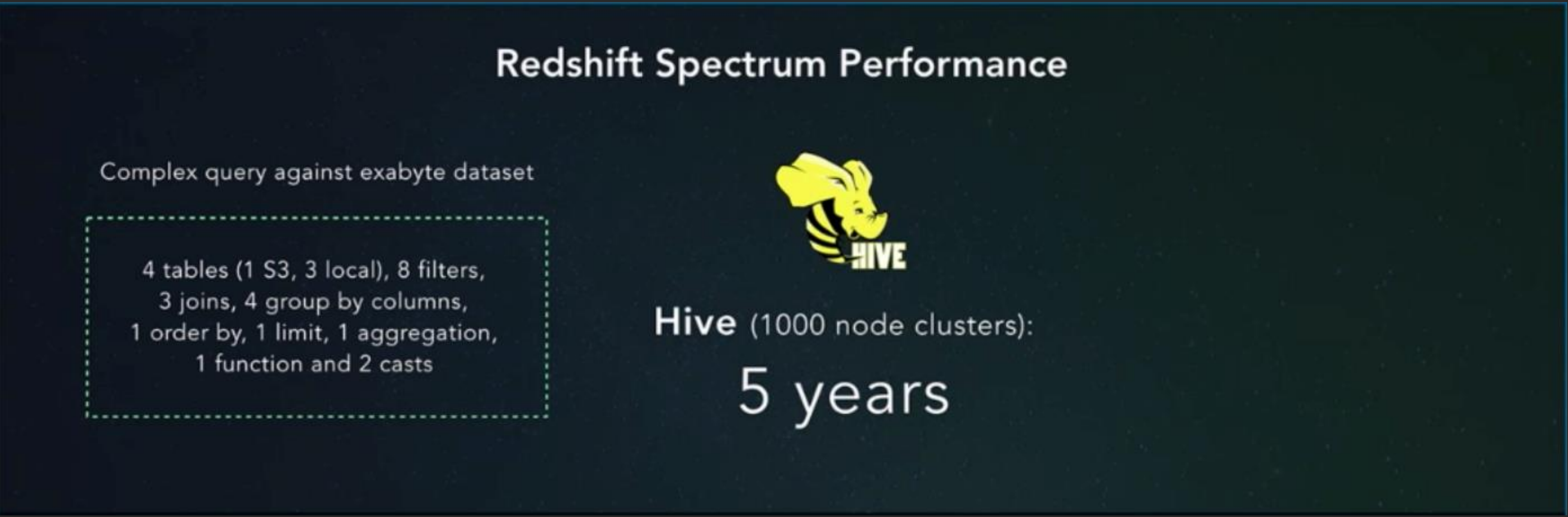
4 tables (1 S3, 3 local), 8 filters,
3 joins, 4 group by columns,
1 order by, 1 limit, 1 aggregation,
1 function and 2 casts

Werner Vogels, Amazon’s CTO, AWS Summit San Francisco 2017

https://youtu.be/RpPf38L0HHU?t=3963

Numbers are fun

Even if some big data technologies are awesome, they struggle when querying huge data sets



Redshift Spectrum Performance

Complex query against exabyte dataset

4 tables (1 S3, 3 local), 8 filters, 3 joins, 4 group by columns, 1 order by, 1 limit, 1 aggregation, 1 function and 2 casts

Hive (1000 node clusters):

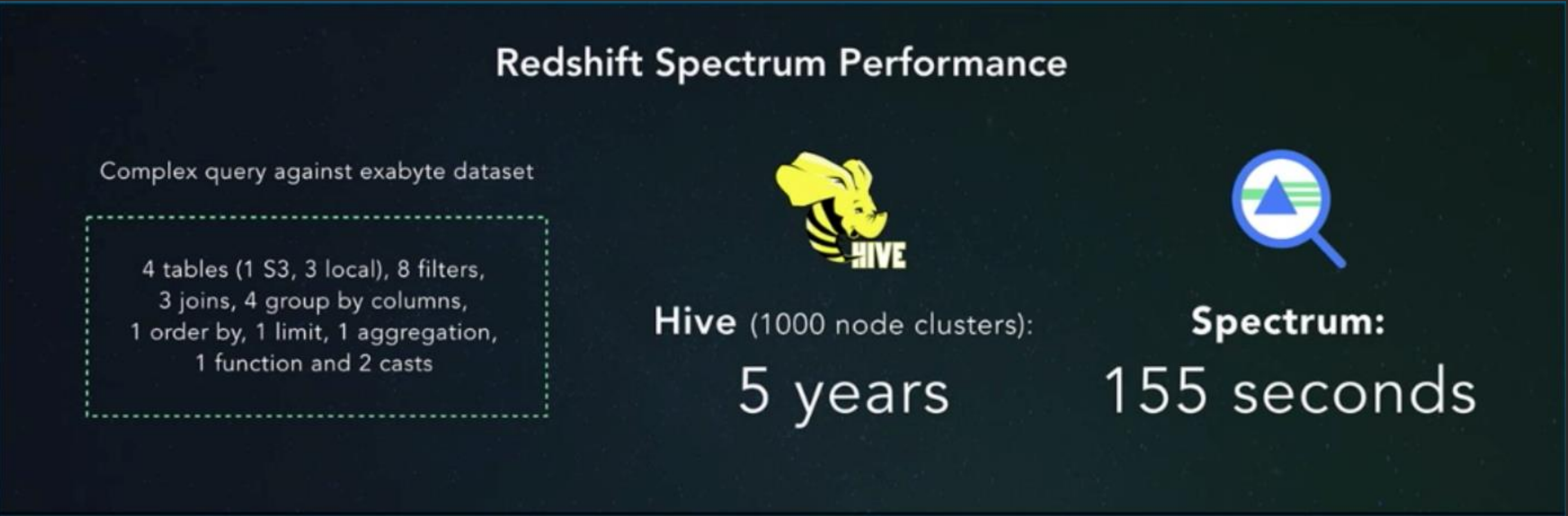
5 years

Werner Vogels, Amazon's CTO, AWS Summit San Francisco 2017

https://youtu.be/RpPf38L0HHU?t=3963

Numbers are fun

Redshift Spectrum takes advantage of the cloud and performs exceptionally well



Redshift Spectrum Performance

Complex query against exabyte dataset

4 tables (1 S3, 3 local), 8 filters,
3 joins, 4 group by columns,
1 order by, 1 limit, 1 aggregation,
1 function and 2 casts

Hive (1000 node clusters):

5 years

Spectrum:

155 seconds

Werner Vogels, Amazon's CTO, AWS Summit San Francisco 2017

https://youtu.be/RpPf38L0HHU?t=3963

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/

APN Data & Analytics and Machine Learning Competency Partners

accenture

Deloitte.

devoteam | Revolve

DoiT INTERNATIONAL

REPLY

Visit the Partner Discovery Zone to meet these partners and view the full list of APN Competency Partners

Thank you!

Javier Ramirez

Developer Advocate

@supercoco9

aws SUMMIT ONLINE

Appendix

The AWS analytics portfolio

Data, visualization, engagement & machine learning

- NEW AWS Data Exchange
- Amazon QuickSight
- Amazon Pinpoint
- Amazon SageMaker
- Amazon Comprehend
- Amazon Lex
- Amazon Polly
- Amazon Rekognition
- Amazon Translate

+ Many more

Analytics

- Amazon Redshift
- Amazon EMR (Spark & Hadoop)
- AWS Glue (Spark & Python)
- Amazon Athena
- Amazon Elasticsearch Service
- Amazon Kinesis Data Analytics

Data lake infrastructure & management

- Amazon S3/ Amazon S3 Glacier
- AWS Lake Formation
- AWS Glue

Data movement

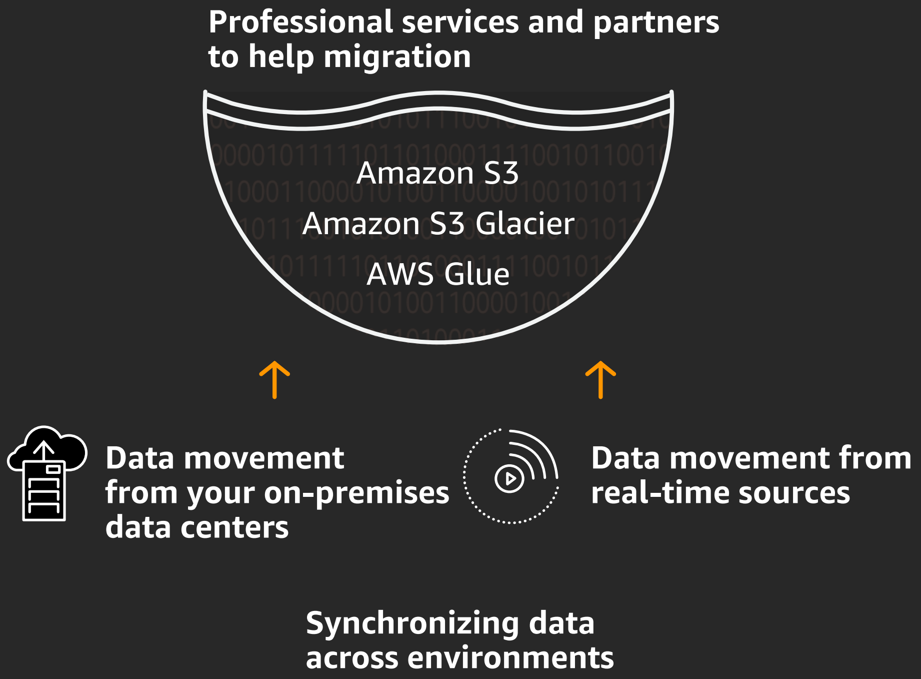
AWS Database Migration Service | AWS Snowball | AWS Snowmobile | Amazon Kinesis Data Firehose | Amazon Kinesis Data Streams | Amazon Managed Streaming for Apache Kafka

Data movement services

Data movement

Migration & streaming services

The most ways to move data to the data lake



Professional services and partners to help migration

Amazon S3

Amazon S3 Glacier

AWS Glue

Data movement from your on-premises data centers

Data movement from real-time sources

Synchronizing data across environments

Data movement from on-premises data centers

Dedicated network connection

Secure appliances

Ruggedized shipping containers

Database migration

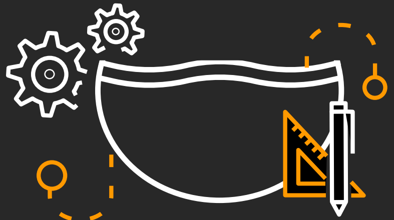
Gateway that lets applications write to the cloud

Data movement from real-time sources

Connect devices to AWS

Real-time data streams

Real-time video streams



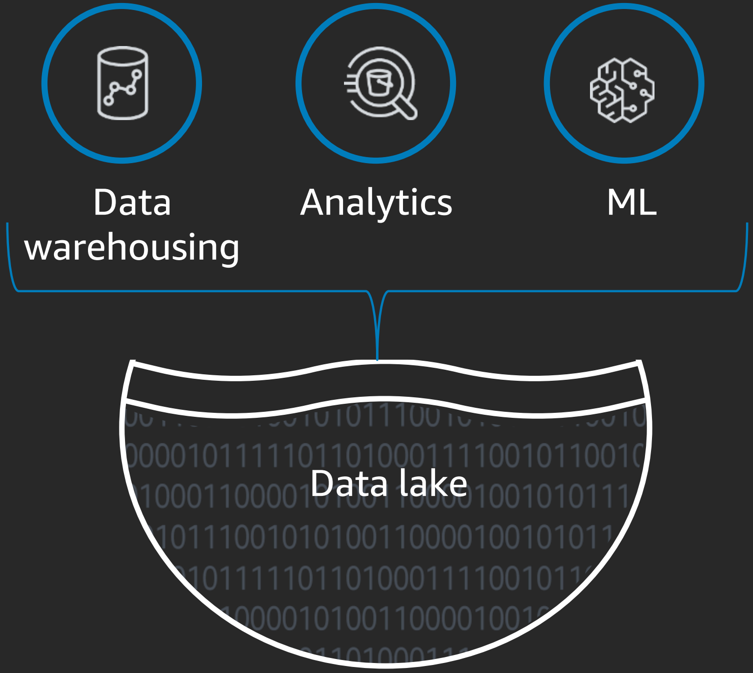
Data lake infrastructure & management services

Data lake infrastructure & management

- Amazon S3/ Amazon S3 Glacier
- AWS Lake Formation
- AWS Glue

Customers moving to data lake architectures

Bringing together the best of both worlds



Extends or evolves data warehouse architectures

Store any data in any format

Durable, available, and exabyte-scale

Secure, compliant, and auditable

Run any type of analytics from data warehouse to predictive

Build on robust data lake infrastructure with Amazon S3

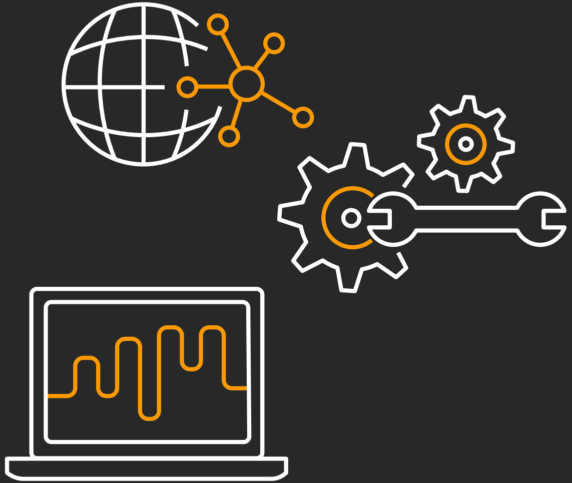
99.999999999% durability

Global replication capabilities

Management features

Cost-effective storage classes

Most partner integrations



Georgia-Pacific uses AWS to save millions of dollars annually

Georgia-Pacific

Challenge

Georgia-Pacific wanted to gain new insights from manufacturing data collected at paper production plants, but it relied on disparate sources to analyze data on material quality, moisture, temperature, and other features

Solution

Georgia-Pacific uses an AWS advanced analytics solution, featuring an Amazon S3 data lake, Amazon Kinesis, and Amazon SageMaker, to collect and analyze data from equipment at manufacturing facilities across North America

Benefits

- Boosts profits by millions of dollars
- Predicts equipment failure 60–90 days in advance
- Runs more production lines in a predictable manner
- Ensures highest quality products

> “We are using AWS data analysis technologies to predict . . . precisely how fast converting lines should run to avoid tearing. **By reducing paper tears, we have increased profits by millions of dollars for one production line.**”

Steve Bakalar
VP, IT & Digital Transformation
Georgia-Pacific

Serverless ETL and data integration with AWS Glue

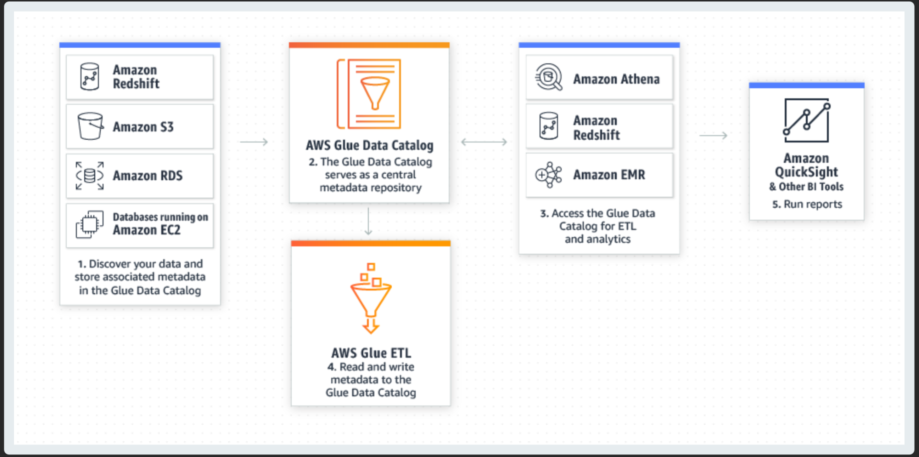
Serverless provisioning, configuration, and scaling to run your ETL jobs on Apache Spark and Python

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)


Amazon Redshift
Amazon S3
Amazon RDS
Databases running on Amazon EC2
1. Discover your data and store associated metadata in the Glue Data Catalog
AWS Glue Data Catalog
2. The Glue Data Catalog serves as a central metadata repository
AWS Glue ETL
4. Read and write metadata to the Glue Data Catalog
Amazon Athena
Amazon Redshift
Amazon EMR
3. Access the Glue Data Catalog for ETL and analytics
Amazon QuickSight & Other BI Tools
5. Run reports

Data lake infrastructure
& management

AWS Glue

Simple, flexible, and cost-effective ETL and data catalog

Less hassle

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

Serverless

Serverless –
no infrastructure
to provision or manage

More power

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

ALICE uses AWS Glue to solve complex data migration challenges

ALICE

Challenge

ALICE acquired a large competitor, GoConcierge, with a global customer base of over 1,000 hotels. Its challenge was to upgrade its customers without getting in the way of its operation. It needed a technology that was highly versatile and flexible in transforming one data structure into another.

Solution

The ALICE architecture leverages AWS Glue to load the data from the source database, transform it into the target data model, and build new foreign keys for re-establishing the relationship within datasets and between datasets.

Benefits

ALICE migrated over 500 properties, and it continues to be stable. An average hotel takes one hour to run end-to-end. AWS Glue made this a simple process, and it has been the foundation of successfully upgrading customers to the ALICE platform.

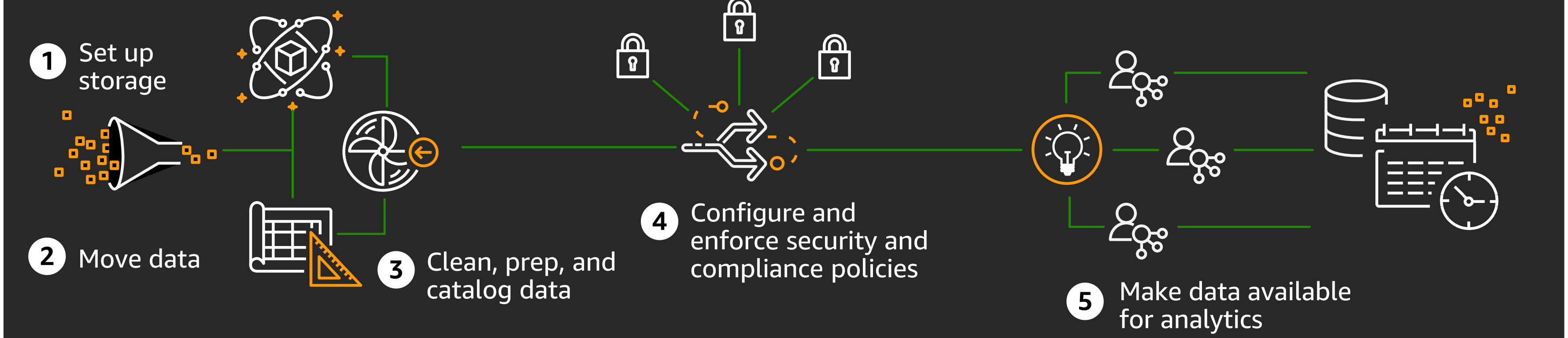
> “It is amazing how seamlessly [AWS] Glue jobs can be integrated into a technology landscape.”

Michael Dreikorn
Tech Lead
ALICE

Challenges to making a secure data lake

Data lake infrastructure & management

Typical steps of building a data lake



1 Set up storage

2 Move data

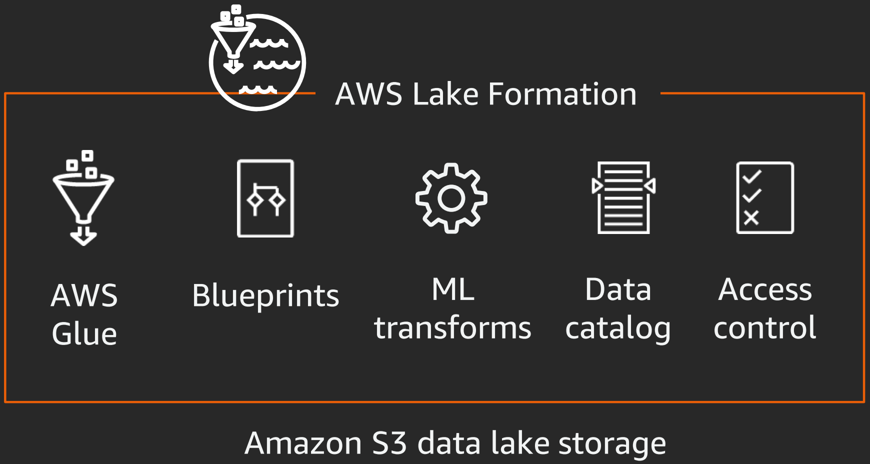
3 Clean, prep, and catalog data

4 Configure and enforce security and compliance policies

5 Make data available for analytics

Data lake infrastructure
& management

Build a secure data lake in days with AWS Lake Formation



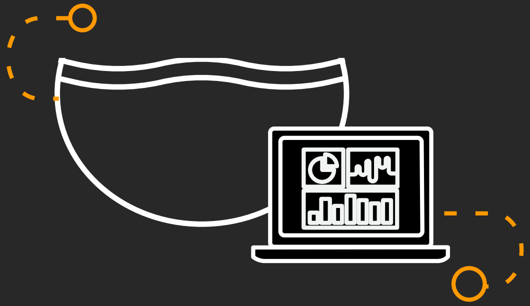
Comprehensive set of integrated tools enables consistent user access

Centralized management of fine-grained permissions empowers security officers

Simplified ingest and cleaning functions enables data engineers to build faster

> “Setting up security and access controls for each AWS account, service, user, and dataset at the level of detail that was required could be cumbersome. AWS Lake Formation streamlines the process with a central point of control while also enabling us to manage who is using our data, and how, with more detail. AWS Lake Formation allows us to manage permissions on Amazon S3 objects like we would manage permissions on data in a database. Our users will be able to find, access, and analyze the data they need with the tools they prefer.”

Kerby Johnson
Enterprise Data Lake Product Owner
Amgen



Analytics services

- Big data processing
- Data warehousing
- Real-time analytics
- Operational analytics
- Interactive query
- Serverless data processing

Amazon EMR

Easily run Apache Spark, Hadoop, Hive, Presto, HBase, and more big data applications 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. The company needed to run complex surveillance queries over 20+ PB of data to detect and analyze illegal market activity.

Solution

FINRA migrated its big data appliance to an Amazon S3 data lake. It uses AWS Lambda and Amazon EMR for data ingestion and Amazon EMR and Amazon Redshift for data processing.

Benefits

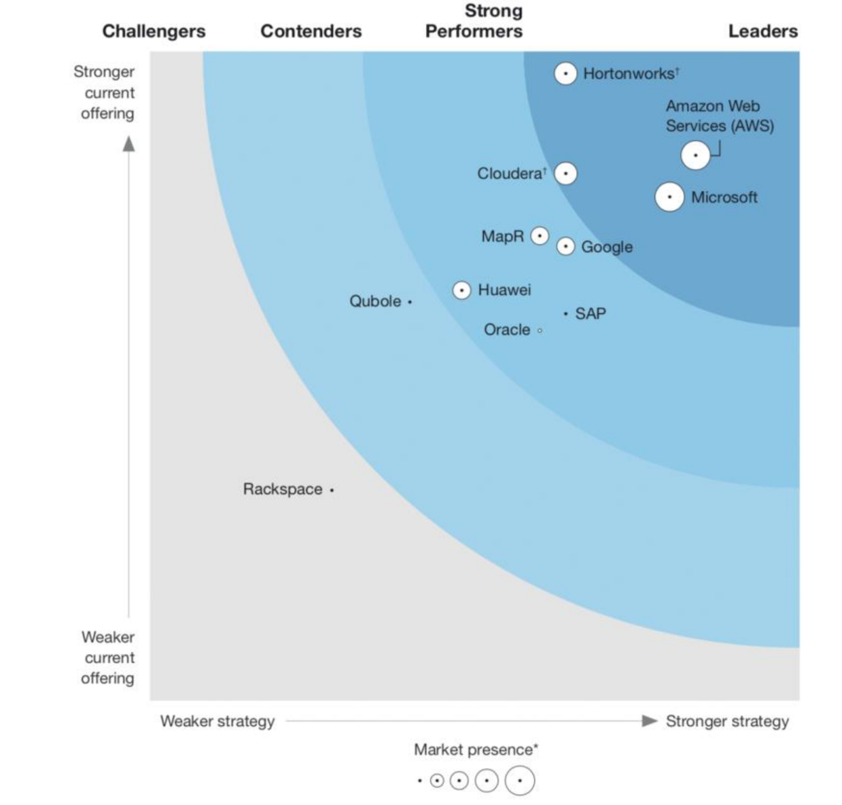
FINRA was able to increase agility, speed, and cost savings while also operating at scale. The company estimates that it will save $10 to $20 million annually.

The Forrester Wave

Cloud Hadoop/Spark Platforms, Q1 2019 Report

The 11 Providers That Matter Most and How They Stack Up

Noel Yuhanna and Mike Gualtieri, February 13, 2019



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

Data lake integration

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

Faster performance

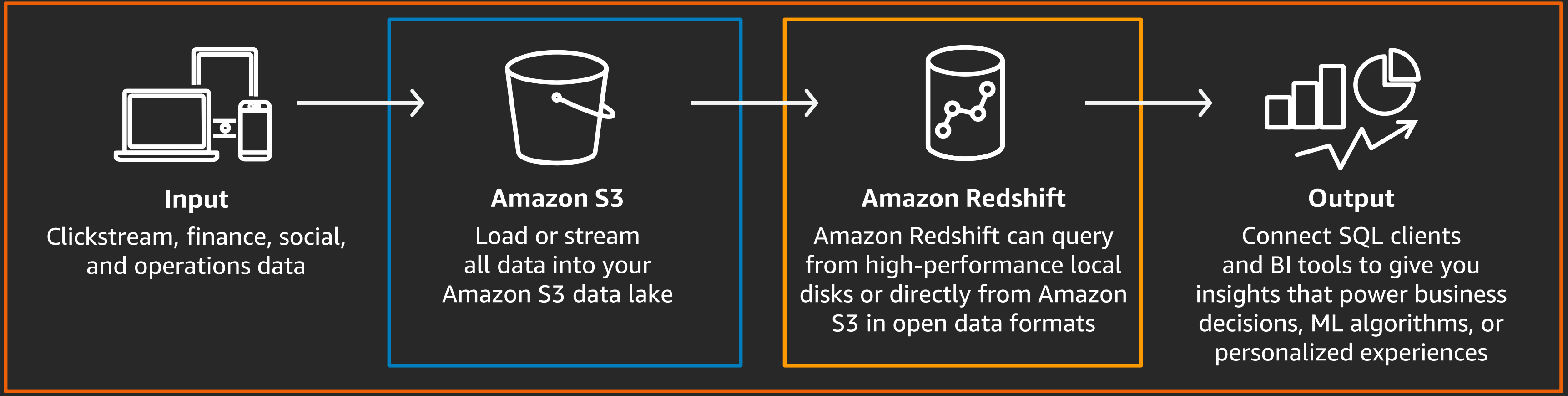
2x faster than other cloud data warehouses

Secure

Security out of the box, at no extra cost

Cost-effective

Up to 75% less than other cloud data warehouses



Input

Clickstream, finance, social, and operations data

Amazon S3

Load or stream all data into your Amazon S3 data lake

Amazon Redshift

Amazon 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, ML algorithms, or personalized experiences

Pfizer

> “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. [Amazon] 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 [Amazon] 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
Pfizer

Data warehousing: Amazon Redshift

First and most popular cloud data warehouse

Data lake & AWS integration

Analyze exabytes of data across data warehouses, data lakes, and operational databases

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 (e.g., VPC, encryption with AWS KMS, AWS CloudTrail)

All major certifications such as SOC, PCI DSS, ISO, FedRAMP, and HIPAA

Lowest cost

Cost-optimize workloads by paying compute and storage separately

1/10 the cost of a traditional data warehouse at $1,000/TB/year

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

NTT docomo

> “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. [Amazon] Redshift provides our users with consistently faster performance, even as its usage within the company has grown.”

Masayuki Tsuda
General Manger of Service Innovation
DOCOMO

Real time: Amazon Kinesis

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

Amazon Kinesis Video Streams

Capture, process, and store video streams for analytics

Amazon Kinesis Data Streams

Build custom applications that analyze data streams

Amazon Kinesis Data Firehose

Load data streams into AWS data stores

Amazon Kinesis Data Analytics

Analyze data streams with SQL

yieldmo

> “Amazon Kinesis makes it simple to scale our solution end-to-end, including the capture, processing, and delivery of actionable insights. This empowers our customers to better understand their user base.”

Indu Narayan
Director of Data
Yieldmo

Operational analytics

Fully managed, scalable, secure Amazon 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 and 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

Compliance – HIPAA, 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, and 24/7 monitoring

AUTODESK.

> “Ultimately, we are improving our software products and offering better service to our customers because of the real-time visibility we’re getting into log data. Amazon Elasticsearch Service enables data forensic activities to take place and help find and fix application problems faster.”

Tommy Li
Senior Solutions Architect
Autodesk

Analytics

Open Distro for Elasticsearch

An Apache 2.0–licensed distribution for Elasticsearch, enhanced with enterprise security, alerting, SQL, and more

100% open source

Providing you the freedom to 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

Get started with flexible deployment options

Visit the website for Open Distro for Elasticsearch

Download the Elasticsearch and Kibana packages

Load and query data

Docker

RPM

Debian

Analytics

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 Amazon 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
Moveable Ink

Serverless analytics

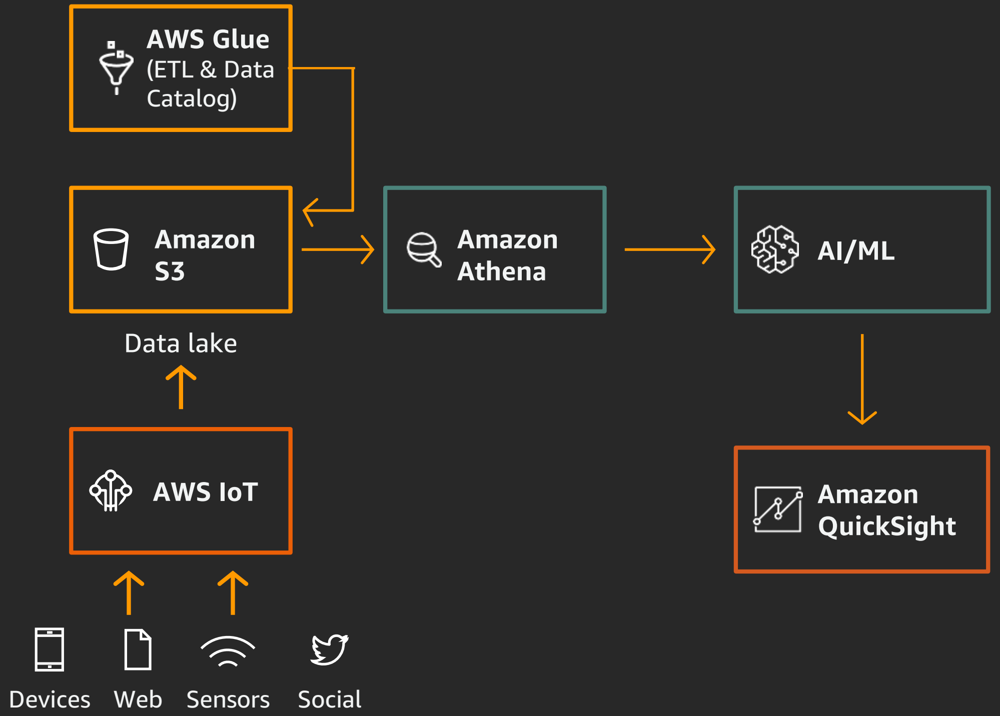
Deliver on-demand analytics on the data lake

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

Epic Games continually improves Fortnite for 250+ million players globally

Challenge

The company needed a way to process and analyze over 100 PB of data (125M events per minute) 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 lakes for machine learning

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 are seen

Costs reduced by separating storage from compute

AWS Data Exchange

Easily find and subscribe to third-party data in the cloud

Quickly find diverse data in one place

>1,000 data products

>80 data providers including Dow Jones, Change Healthcare, Foursquare, Dun & Bradstreet, Thomson Reuters, Pitney Bowes, LexisNexis, and Deloitte

Easily analyze data

Download or copy data to S3

Combine, analyze, and model with existing data

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

Efficiently access third-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

Elastic scaling

Automatically scale 10 to 10,000+ 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 applications

Rio Tinto

> "At Rio Tinto, safety is paramount, and we want to empower everyone to make decisions with the best data available. Amazon QuickSight allows our analysts to create insightful dashboards quickly for our critical risk management program, enabling us to move from static spreadsheets to interactive data. However, rolling out these dashboards at scale to the field was going to be costly and complicated. We asked AWS for a better solution, and they listened. The ability to have 'read' access to dashboards in QuickSight, with usage-based pricing, will help us scale the dashboards to more end users across the world and only pay for what we use."

Anthony Deakin
Critical Risk Management
Rio Tinto

Data, visualization, engagement & ML

Successfully engage your customers with Amazon Pinpoint

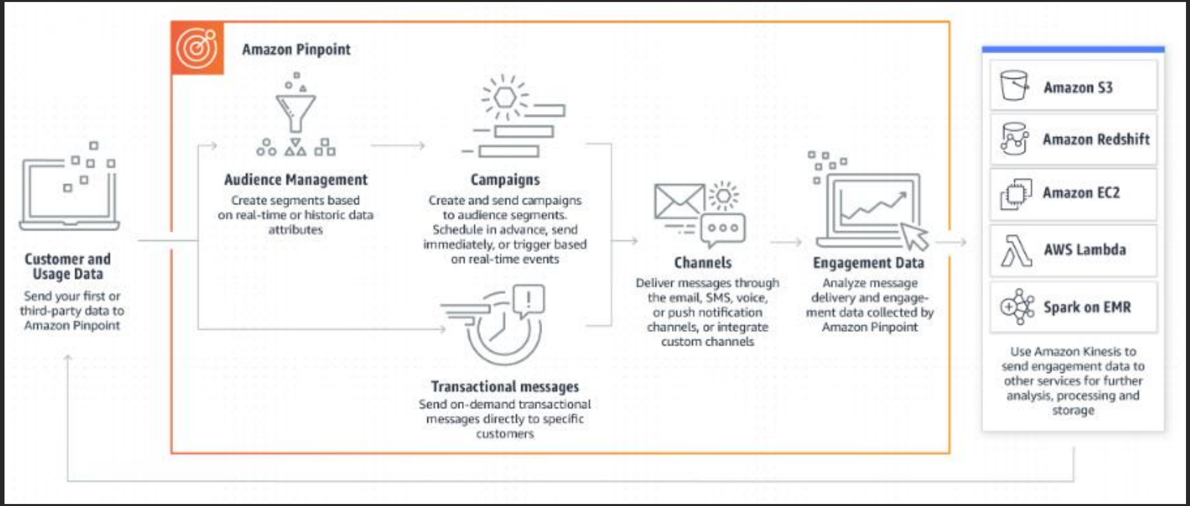
Understand your customers

Segment based on understandings

Target in a contextually relevant way

Communicate in best channel

React to customer responses



Amazon Pinpoint

Customer and Usage Data

Send your first or third-party data to Amazon Pinpoint

Audience Management

Create segments based on real-time or historic data attributes

Campaigns

Create and send campaigns to audience segments. Schedule in advance, send immediately, or trigger based on real-time events

Transactional messages

Send on-demand transactional messages directly to specific customers

Channels

Deliver messages through the email, SMS, voice, or push notification channels, or integrate custom channels

Engagement Data

Analyze message delivery and engagement data collected by Amazon Pinpoint

Use Amazon Kinesis to send engagement data to other services for further analysis, processing and storage

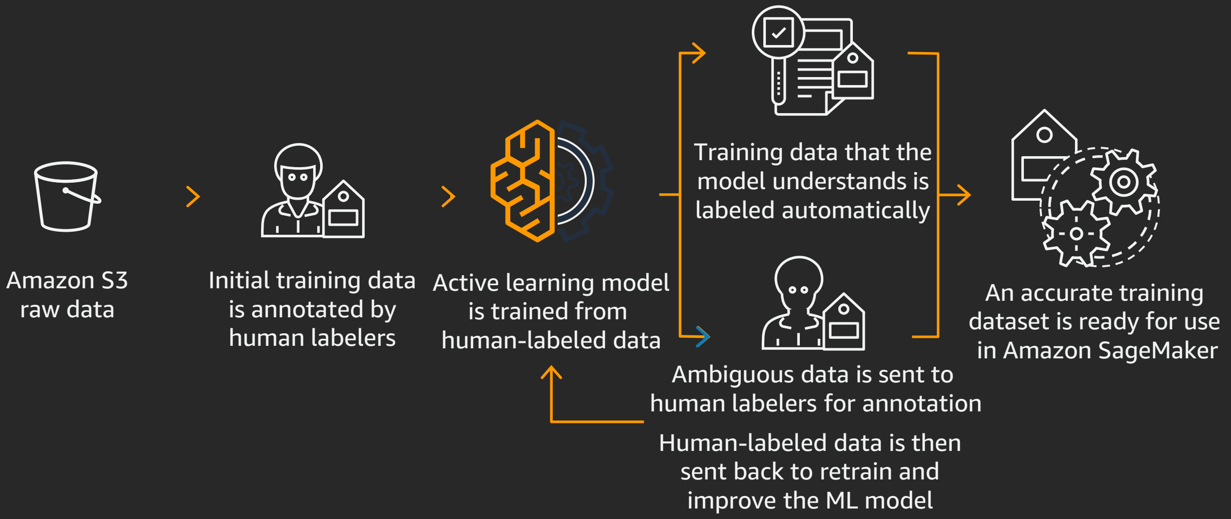
Data, visualization, engagement & ML

Predictive insights with AWS ML & AI services

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

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

ML frameworks and interfaces for ML practitioners



Amazon S3 raw data

Initial training data is annotated by human labelers

Active learning model is trained from human-labeled data

Training data that the model understands is labeled automatically

Ambiguous data is sent to human labelers for annotation

Human-labeled data is then sent back to retrain and improve the ML model

An accurate training dataset is ready for use in Amazon SageMaker

Kinect
ENERGY GROUP

> “We started using [Amazon] SageMaker a little less than a year ago. We’ve grown our own AI team several times over in the past six months to fully exploit the new advantage that AWS’s technologies provide. We’re profiting in new ways by setting prices based on weather that hasn’t yet happened. We’ve gone ‘all in’ with AWS, including storing our data in S3.”

Andrew Stypa
Lead Business Analyst
World Fuel Services

Amazon.com data lakes on AWS

Amazon.com lowers costs 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 data warehouse did not scale, was difficult to maintain, and was costly.

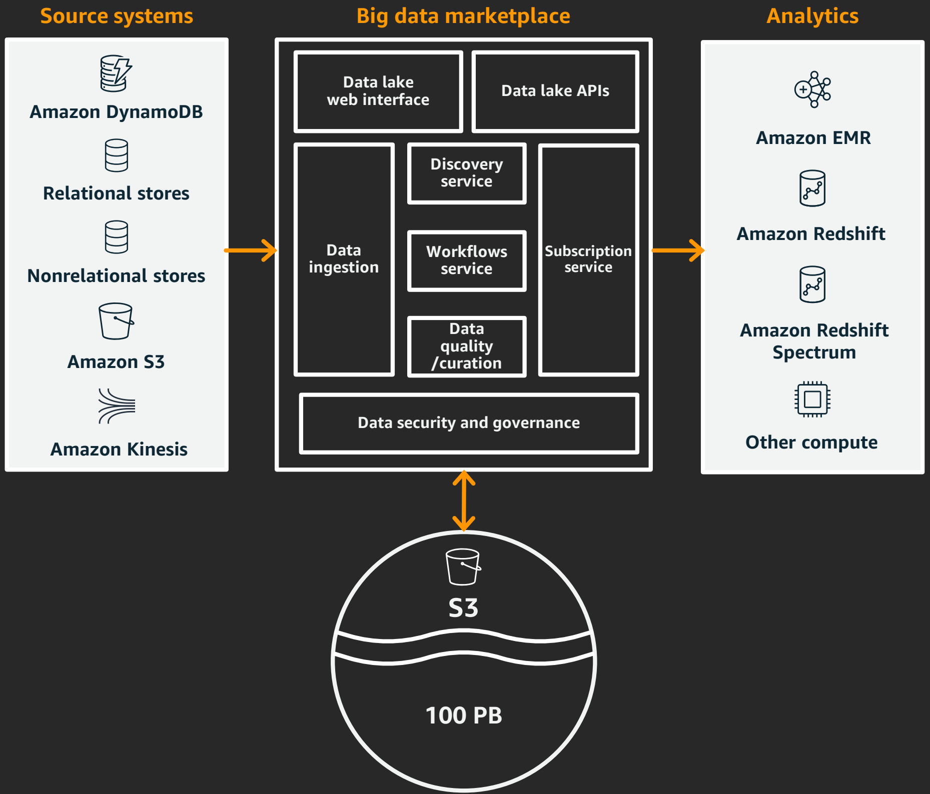
Solution

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

Benefits

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

Amazon uses an AWS data lake



- 50 PB of data
- 600,000 analytics jobs per 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 a step-by-step guide to help you launch your AWS project