aws training and certification

Solutions Training for Partners Data and Analytics (Business)

July 2020

Stephen Hogarth
hogart@amazon.com

Objectives

We aim to...

- Identify business opportunities in analytics
- Provide a high-level overview of key AWS offerings for building a data lake and use cases
- Understand AWS differentiation in analytics

Agenda

- Data in Modern Age
- The Data FlyWheel
- Why AWS for Analytics?
- The Data FlyWheel – DeepDive
- Business Problem to Solution
- Call to Action

aws training and certification

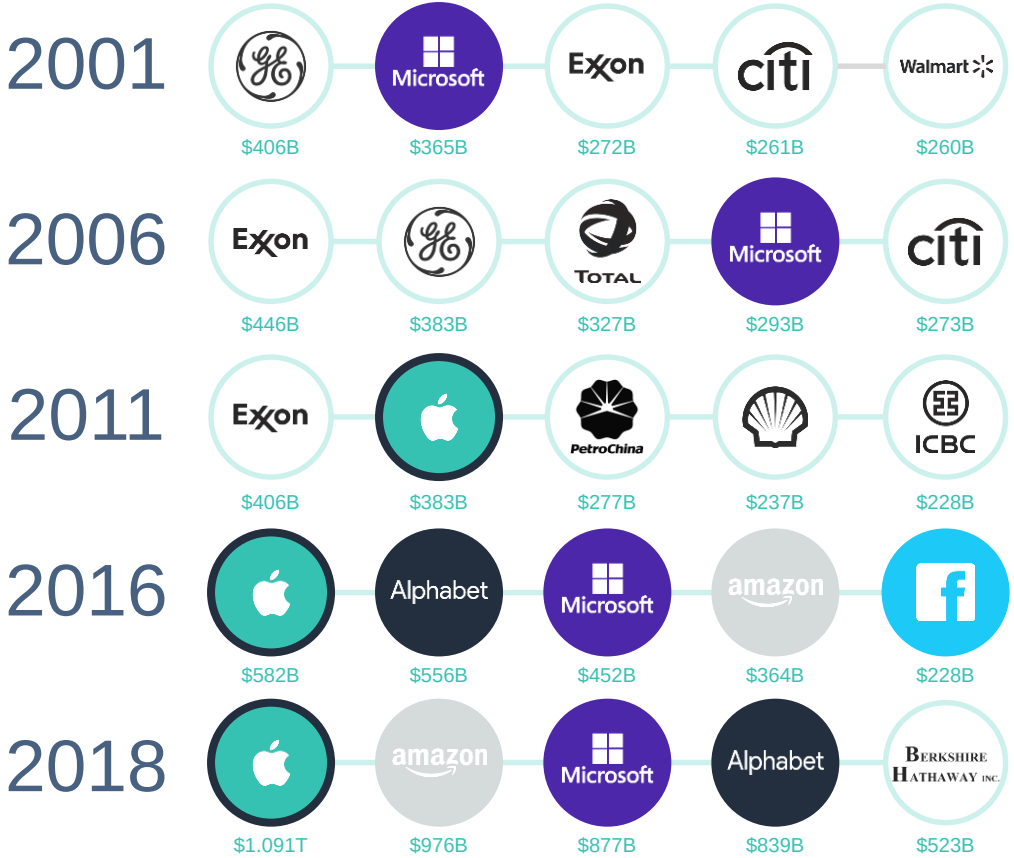
Data in Modern Age

What do these companies have in common?

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



The move toward **data-centric** companies

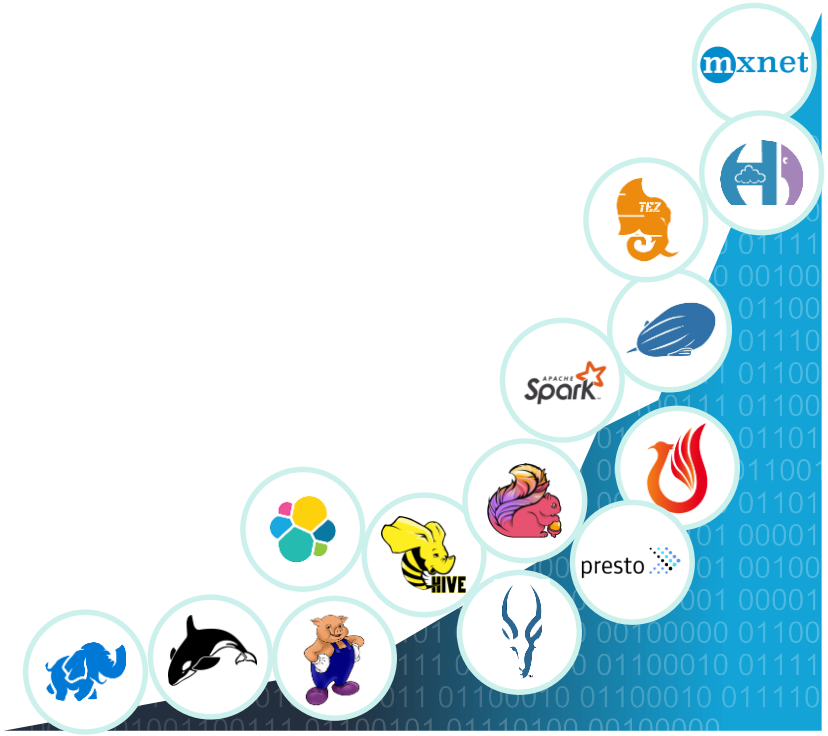
Five largest companies by market cap*

*Copyright: Visual Capitalist, "Chart: The Largest Companies by Market Cap Over 15 Years," August 12, 2016.

There is more data than people think

Data	Data platforms need to	
grows	live for	scale
>10x	**15**	**1,000x**
every 5 years	years	

* IDC, Data Age 20215: The Evolution of Data to Life-Critical Don't Focus on Big Data, Focus on the Data That's Big, April 2017.



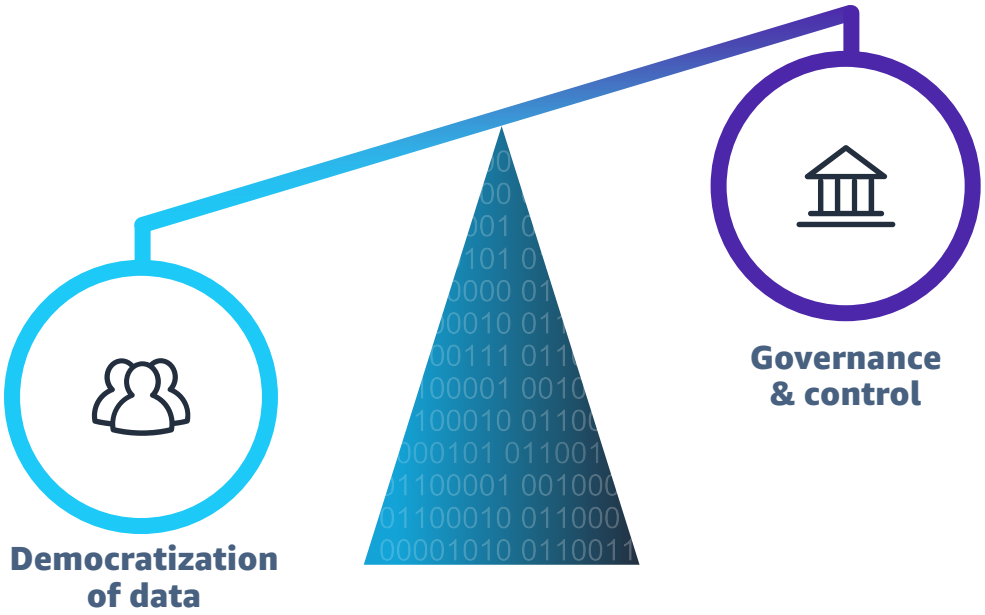
There are more ways to analyze data than ever before

11 **8** **5** **4**

Years ago

Hadoop	Elasticsearch	Presto	Spark

Didn't exist



There are more people working with data than ever before

How do I provide democratized access to data to enable informed decisions while at the same time enforce data governance and prevent mismanagement of the data?

The Data FlyWheel


010010010
01Data01
100010100

1. Move data & workloads to the cloud

2 Store & manage all data

- ✓ Save time
- ✓ Save costs

1 Move data & workloads to the cloud

2 Store & manage all data

✓ Save time
✓ Save costs

✓ Agility
✓ Global distribution
✓ Scale and performance

1 Move data & workloads to the cloud

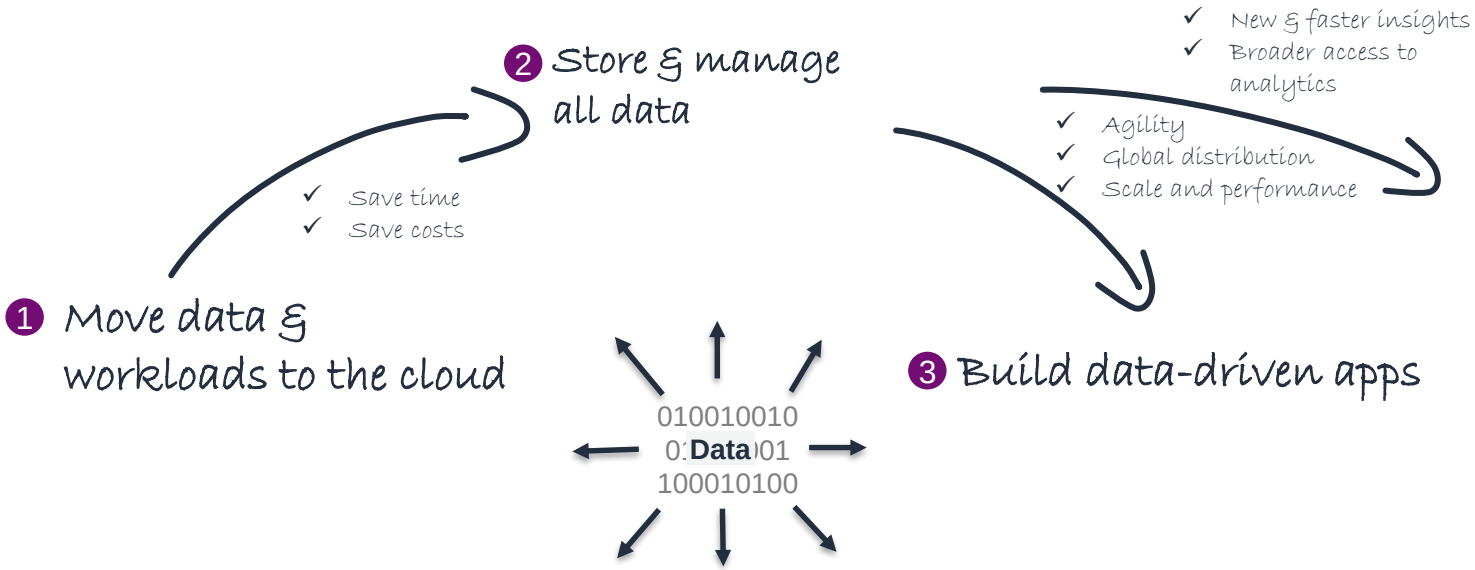
2 Store & manage all data

✓ Save time
✓ Save costs

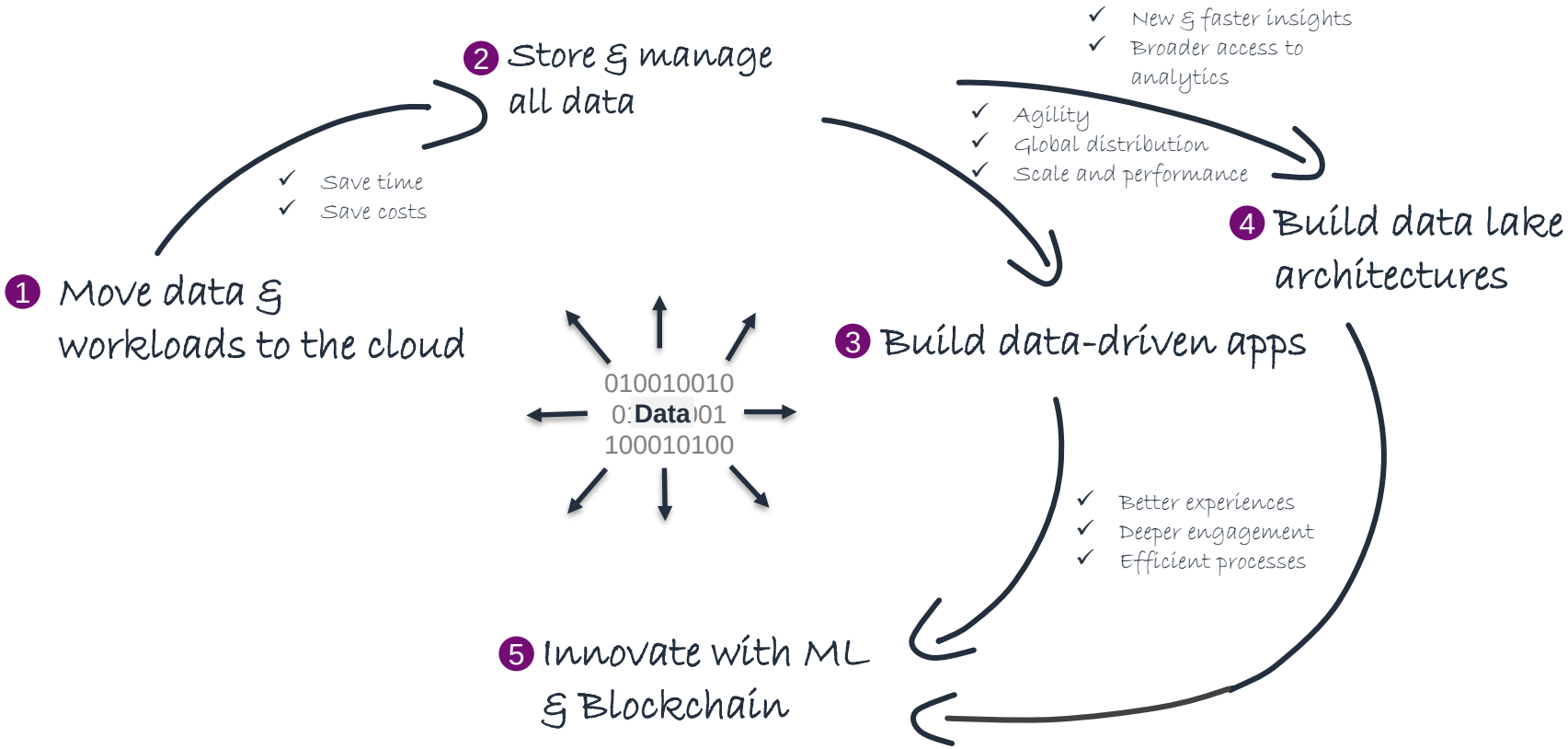
✓ Agility
✓ Global distribution
✓ Scale and performance

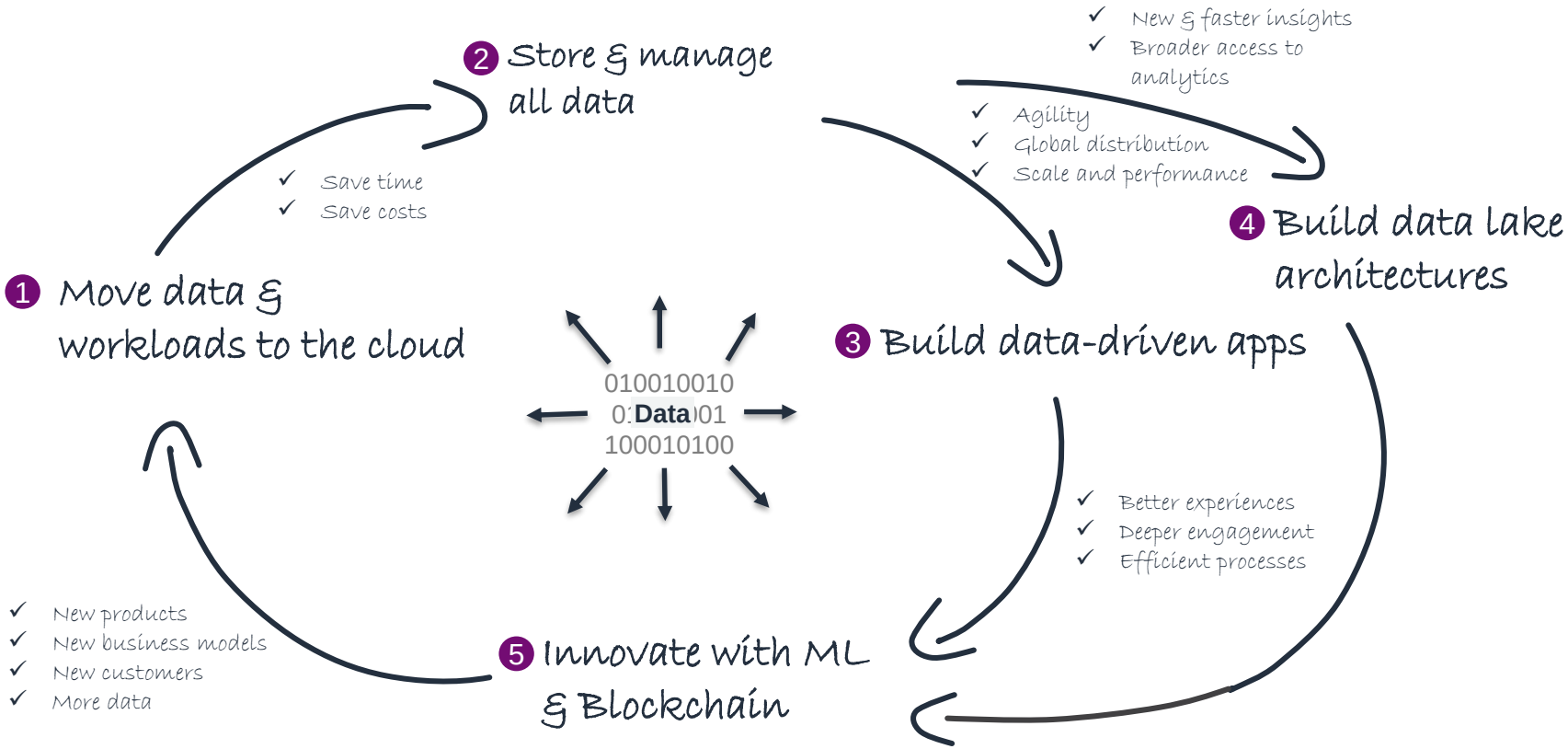
1 Move data & workloads to the cloud

3 Build data-driven apps


2 Store & manage all data
✓ Save time
✓ Save costs
✓ New & faster insights
✓ Broader access to analytics
✓ Agility
✓ Global distribution
✓ Scale and performance
1 Move data & workloads to the cloud
010010010
01Data01
100010100
3 Build data-driven apps

1. Move data & workloads to the cloud
   - ✓ Save time
   - ✓ Save costs
2. Store & manage all data
3. Build data-driven apps
   - ✓ Agility
   - ✓ Global distribution
   - ✓ Scale and performance
4. Build data lake architectures
   - ✓ New & faster insights
   - ✓ Broader access to analytics


1 Move data & workloads to the cloud
✓ Save time
✓ Save costs
2 Store & manage all data
✓ Agility
✓ Global distribution
✓ Scale and performance
3 Build data-driven apps
✓ New & faster insights
✓ Broader access to analytics
4 Build data lake architectures
✓ Better experiences
✓ Deeper engagement
✓ Efficient processes
5 Innovate with ML & Blockchain
010010010
01Data01
100010100


2 Store & manage all data
✓ New & faster insights
✓ Broader access to analytics
✓ Agility
✓ Global distribution
✓ Scale and performance
4 Build data lake architectures
✓ Save time
✓ Save costs
1 Move data & workloads to the cloud
010010010
01Data01
100010100
3 Build data-driven apps
✓ Better experiences
✓ Deeper engagement
✓ Efficient processes
✓ New products
✓ New business models
✓ New customers
✓ More data
5 Innovate with ML & Blockchain

Why AWS for Data Lake?

Why choose AWS for data lakes and analytics?

Most comprehensive

Most secure

Most cost-effective

Easiest to build

Most customers & partners

Most Comprehensive

Broad and deep portfolio, built for builders

Business Intelligence & Machine Learning

- **Amazon QuickSight**
- **Amazon SageMaker**
- **Amazon Comprehend**
- **Amazon Rekognition**
- **Amazon Lex**
- **Amazon Transcribe**
- **AWS DeepLens**

AWS Marketplace
250+ solutions

Databases

- **QLDB** (NEW) — Ledger Database
- **ElastiCache** — Redis, Memcached
- **Aurora** — MySQL, PostgreSQL
- **RDS** — MySQL, PostgreSQL, MariaDB, Oracle, SQL Server
- **Neptune** — Graph
- **DynamoDB** — Key value, Document
- **Timestream** (NEW) — Time Series
- **RDS on VMWare** (NEW)

Analytics

- **Amazon Redshift** — Data warehousing
- **Amazon EMR** — Hadoop + Spark
- **Amazon Elasticsearch service** — Operational Analytics
- **Athena** — Interactive analytics
- **Kinesis Analytics** — Real-time

Blockchain

- **Managed Blockchain** (NEW)
- **Blockchain Templates**

730+ Database solutions

600+ Analytics solutions

25+ Blockchain solutions

Data Lake

- **S3/Amazon Glacier**
- **Lake Formation** (NEW) — Data Lakes
- **AWS Glue** — ETL & Data Catalog

20+ Data lake solutions

Data Movement

Database Migration Service | Snowball | Snowmobile | Kinesis Data Firehose | Kinesis Data Streams | Data Pipeline | Direct Connect

30+ solutions

aws training and certification

Most Secure

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

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

Identity

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

Encryption

- AWS Certification 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

aws training and certification

Most Secure

Compliance and 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

aws training and certification

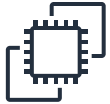
Most Cost Effective

Decouple compute and storage, choice of PAYG analytics services

Storage

Amazon S3 tiers & intelligent tiering

From $0.023 per GB/mo. to as low as $0.004 per GB/mo.



Compute

Spot & Reserved Instances

Save up to 90% off On-Demand prices

Amazon EMR

Automatic scaling

57% less than on-premises per IDC report

Amazon Redshift

Less than a tenth of the cost of traditional solutions

Amazon Athena & QuickSight

Serverless pay only for what you use

Easiest to Build: Serverless Analytics

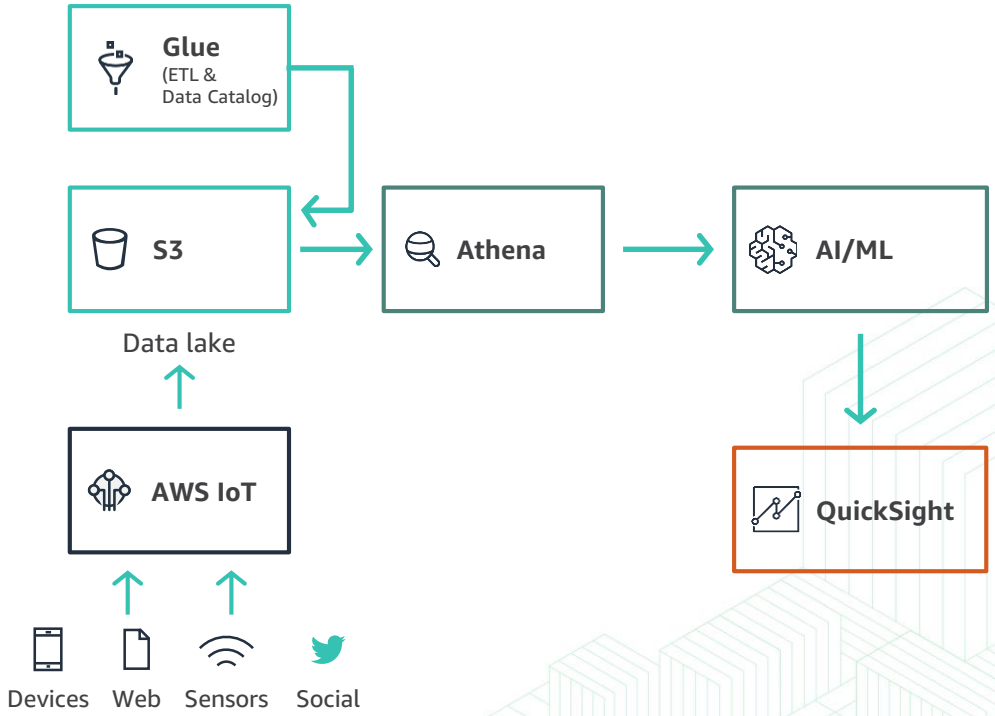
Deliver on-demand analytics on the data lake

Serverless. Zero infrastructure. Zero administration

Never pay for idle resources

Automatically scales resources with usage

Availability and fault tolerance built in



aws training and certification

More data lakes and analytics than anywhere else

More than 10,000 data lakes on AWS

aws training and certification

Most partners to complement AWS offerings

aws training and certification

Open Discussion

Why do data storage and analytics modernization matter to customers?

- Data has become customers' most valuable asset
- However customers can't unlock the value of data, without changing the way they store and analyze data
- Data is locked up in on-premises silos that don't scale cost effectively, don't communicate well with each other, and can't easily be analyzed for insights

Open Discussion

How do AWS storage and analytics services help customers to modernize their data platform?

- AWS provides the broadest and deepest portfolio of storage, data movement, database, analytics, and machine learning services
- AWS provides multiple layers of security, including identity access management, monitoring, encryption, with a broad set of security certifications
- AWS offers various storage classes and pricing options to optimize cost and performance.

Open Discussion

What are business opportunities we can discuss with customers in analytics?

- To store data in a highly durable, available, secure, and cost effective way
- To store and manage various data sets (e.g. streaming data, document, graph, non-relational)
- To build data-driven applications
- To build a data lake
- To bring insights from customer data using machine learning

Case Study: Epic Games

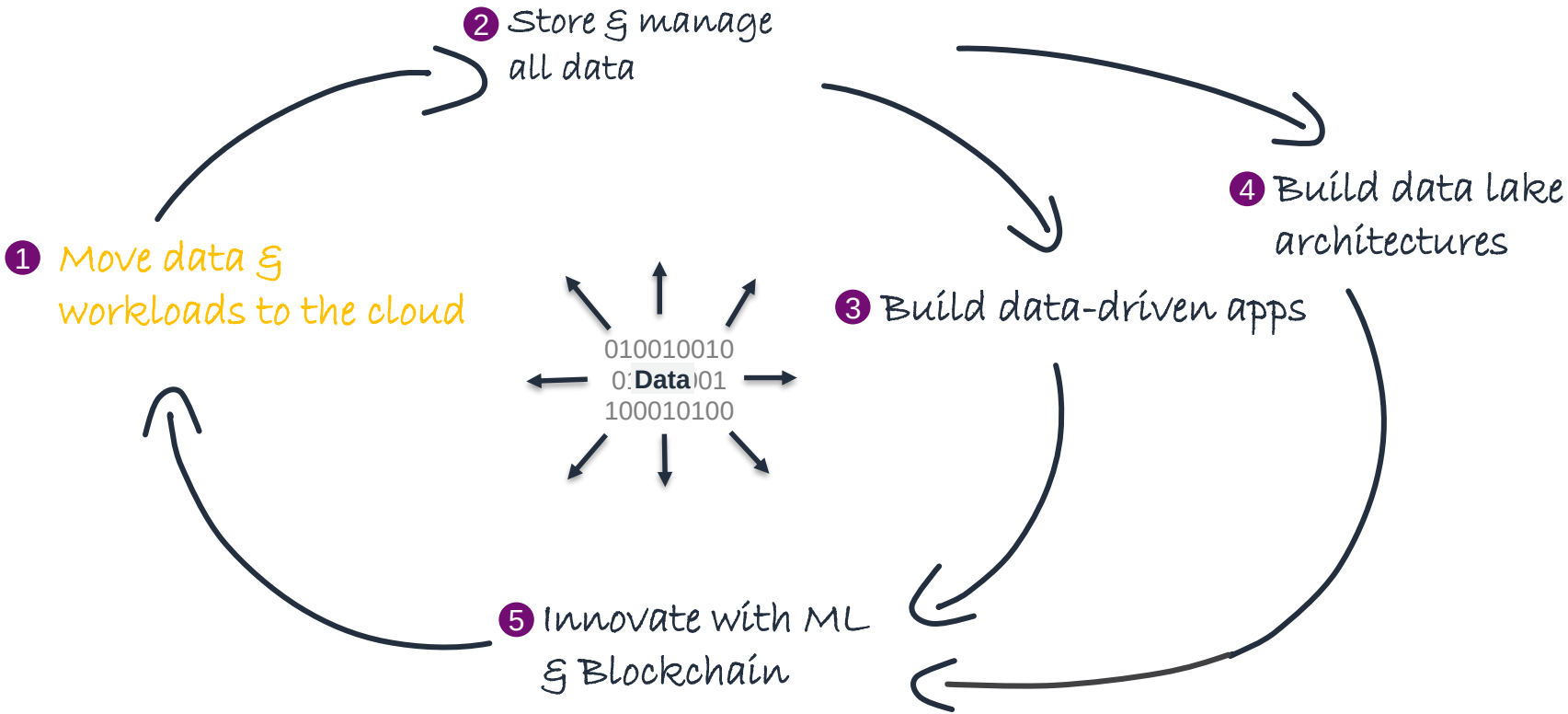
- Fortnite: 125+ million players
- Processes 92mil events a minute
- Data grows 2PB a month
- Over 10m concurrent sessions

- https://www.youtube.com/watch?v=MCLrA401vHw&feature=youtu.be&t=96 7m:16s

- We will reconvene in 10 mins

The Data FlyWheel – Deep Dive

The Data Flywheel



Old-guard commercial databases

Very expensive

Proprietary

Lock-in

Punitive licensing

Unexpected and frequent audits

Customers are moving to open databases

Customers are moving to open databases

+

Commercial-grade performance and reliability?

Amazon Aurora

MySQL and PostgreSQL compatible relational database built for the cloud
Performance and availability of commercial-grade databases at 1/10th the cost

Performance and scalability

5x throughput of standard MySQL and 3x of standard PostgreSQL; scale-out up to 15 read replicas

Availability and durability

Fault-tolerant, self-healing storage; six copies of data across three AZs; continuous backup to S3

Highly secure

Network isolation, encryption at rest/transit

Fully managed

Managed by RDS: no hardware provisioning, software patching, setup, configuration, or backups

Amazon Redshift

Most popular, fastest, most cost effective cloud data warehouse that can extend queries to your data lake

Most popular

More than 15K customers use Redshift and process more than 2 EB of data per day

Fastest

2x faster than the next fastest cloud DW provider; 10x faster than Redshift 2 years ago

Most cost-effective

Redshift costs 1/10 the cost of traditional data warehouses and is up to 75% cheaper than other cloud DW providers

Integrated with the data lake

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

Amazon Athena

Serverless interactive SQL query service to analyze data in Amazon S3

- Interactive query service to analyze data in Amazon S3 using standard SQL
- Serverless
- Pay per Query (only pay for data scanned)
- Built on Presto
- Interactive performance, even for large datasets
- Integrated with Glue Catalog

Amazon Athena & Amazon Redshift

Amazon Athena	Amazon Redshift
Ad-hoc querying	Data warehousing (historical analysis and reporting)
Serverless – no setup or management of cluster	Need to set up cluster
Run interactive queries against data directly in Amazon S3 without worrying about formatting, management etc	Run complex queries that join large numbers of database tables
Scales automatically based on complexity of queries	Can use same Amazon S3 data sources as Amazon Athena

You can also use both services together. If you stage your data on Amazon S3 before loading it into Amazon Redshift, that data can also be registered with and queried by Amazon Athena.

Amazon EMR

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

Enterprise-grade

Easy

Lowest cost

Equinox Fitness uses AWS data lakes and analytics

Faster reports, 80% savings

EQUINOX

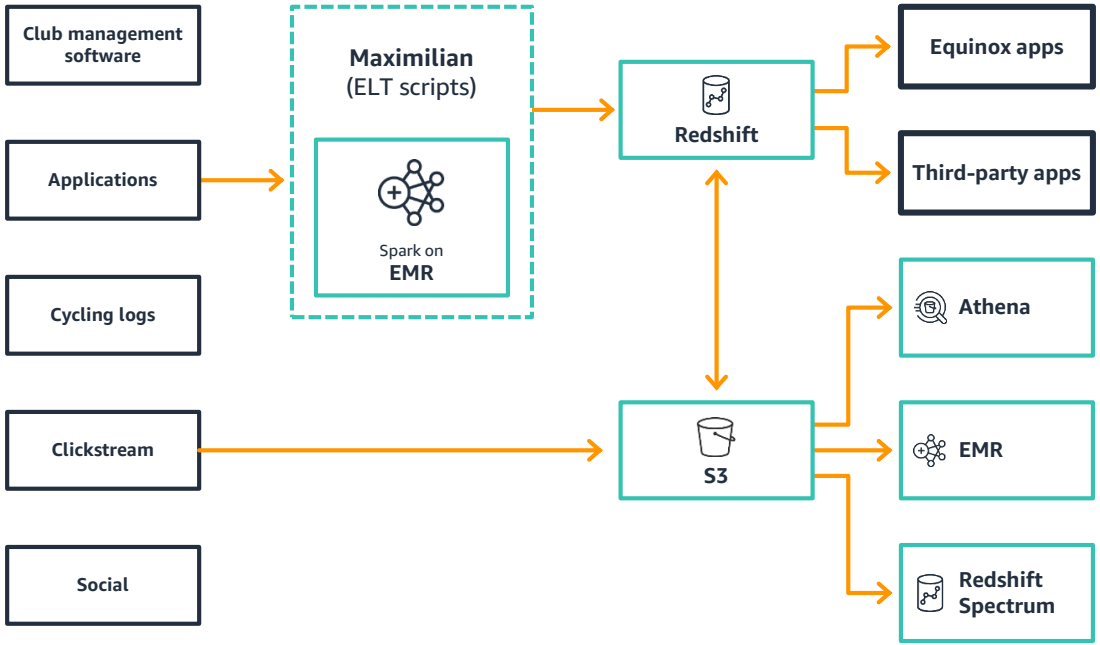
Equinox Fitness integrates luxury and lifestyle offerings centered on movement, nutrition, and regeneration.

They needed to reduce administration and costs or their legacy data warehouse, blend structured and semi-structured data for analytics, and evolve into a data lake strategy.

Equinox migrated from a legacy data warehouse to Amazon Redshift to combine data from disparate sources like clickstream data, cycling log data, club management software, and more. They land data directly in an Amazon S3 Data Lake and perform analytics using Amazon Redshift, Redshift Spectrum, and Amazon EMR.

Their monthly **Amazon Redshift bill is now 20% of prior yearly maintenance of their legacy data warehouse.** AWS Data Lake and Analytics **reduced report delivery time from months to days.**

Equinox Fitness migrated from Teradata to Redshift



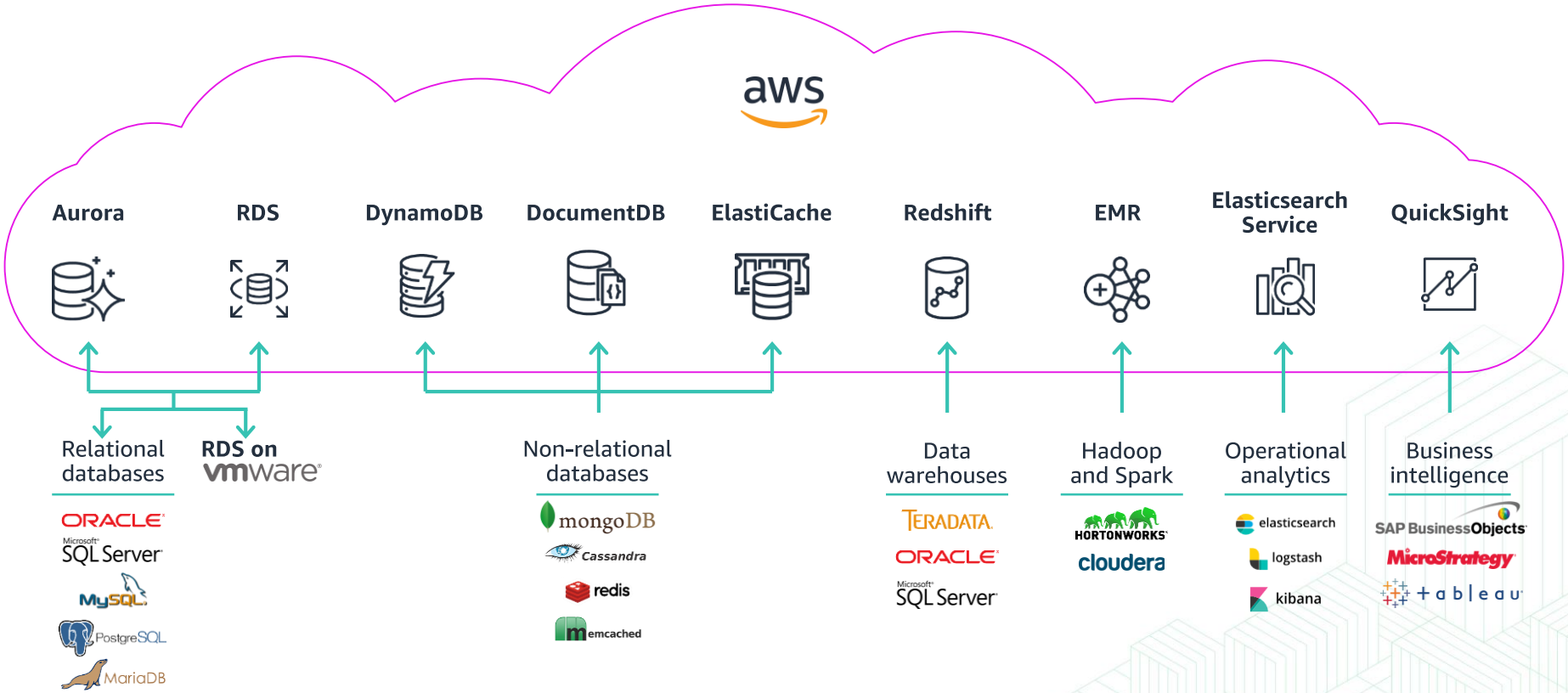
- Migrated from Teradata data warehouse
- Built a DW with Redshift and data lake with S3
- Analytics on data lake with Amazon Athena, Amazon Redshift Spectrum, and Amazon EMR
- Increased user productivity to move faster
- Amazon Redshift costs ~20% of its original Teradata maintenance and support
- Report time reduced from months to days

Re:Invent 2018 video

More Use Cases

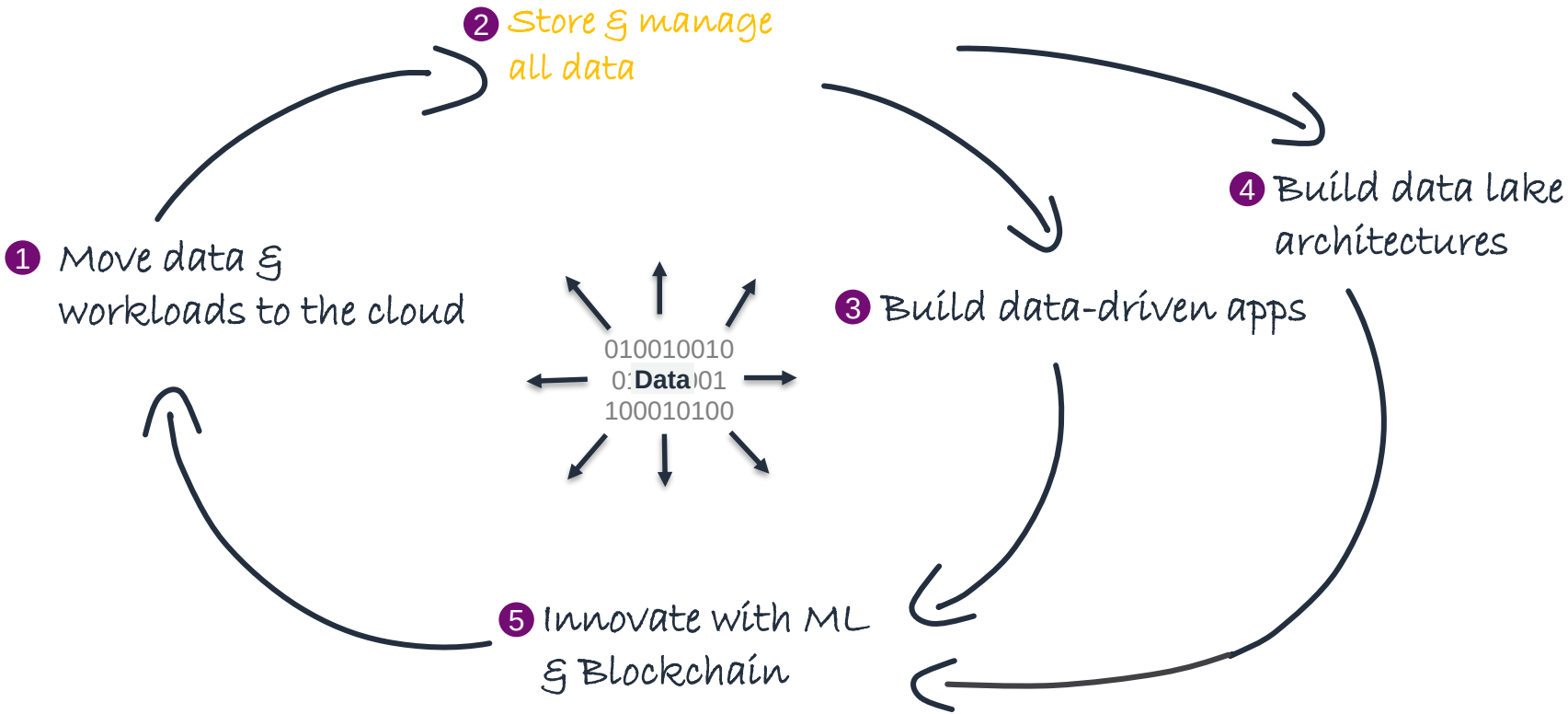
Industry	Use case	Examples
Financial services	Analyze trading and market data, risk analyses, fraud detection	Nasdaq
Healthcare	Analyze clinical records to improve patient outcomes and predict diseases for preventive programs	Pfizer
Advertising	Analyze clickstream and ad impression logs to improve ad targeting	groupm
Gaming	Aggregate data from games and players and analyze in-game behavior	WARNER BROS.
Travel/hospitality	Create personalized experiences and offers for customers	Hilton

Move all types of databases to the cloud

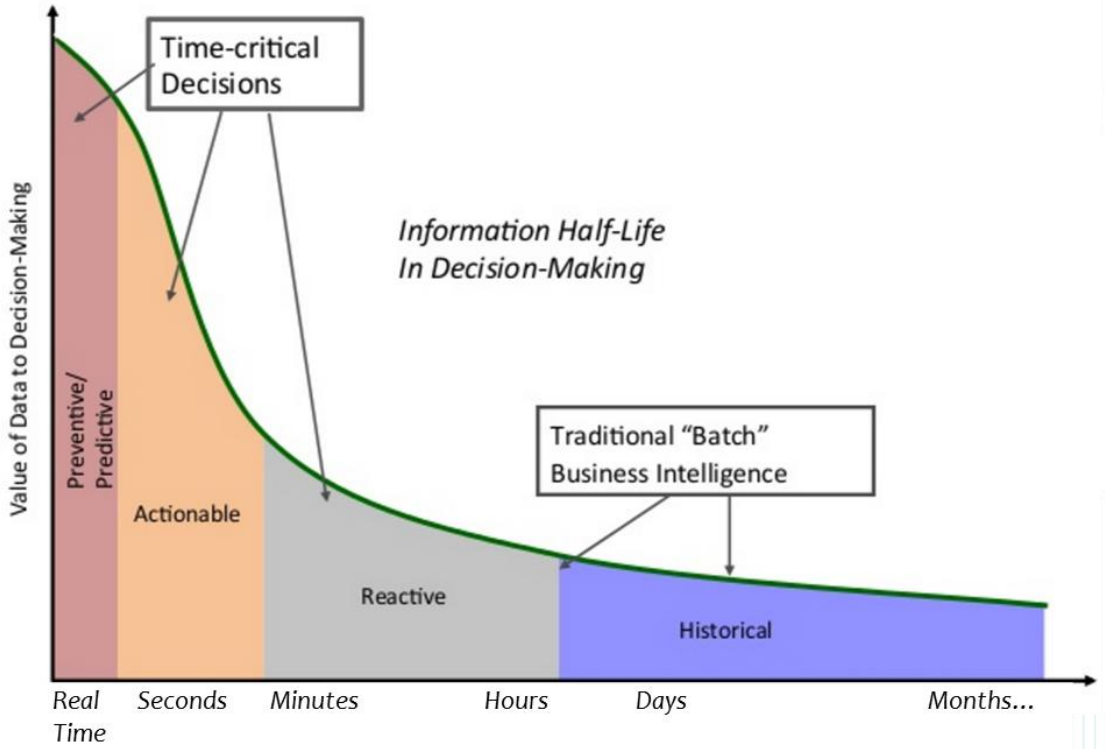


aws training and certification

The Data Flywheel



The value of data diminishes over time



Source: Perishable insights, Mike Gualtieri, Forrester

Common real-time analytics use cases

Messaging between micro-services

Response analytics (Web and mobile app notifications)

Log ingestion

IoT device maintenance

Change Data Capture (CDC)

Streaming ETL into data lakes and data warehouses

Challenges of data streaming

Difficult to setup

Tricky to scale

Hard to achieve high availability

Integration requires development

Error prone and complex to manage

Expensive to maintain

Streaming real-time data with AWS

Easily collect, process and analyze data streams in real time

- Easy to use
- Elastic
- High availability and durability
- Seamless integration with AWS services
- Fully managed
- Pay for what you use

Streaming data with AWS

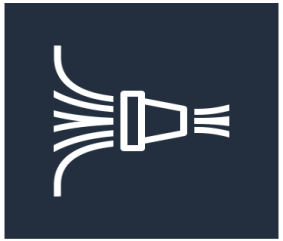
Easily collect, process, and analyze data streams in real time

Amazon Kinesis Data Streams

Capture and store data streams

Amazon Kinesis Data Analytics

Analyze data streams in real time



Amazon Kinesis Data Firehose

Load streaming data into streams, data lakes and warehouses

Thomson Reuters: real-time dashboards

> Using Amazon Kinesis, our solution delivers new events to user dashboards in less than 10 seconds.

—Anders Fritz
Senior Manager, Product Innovation

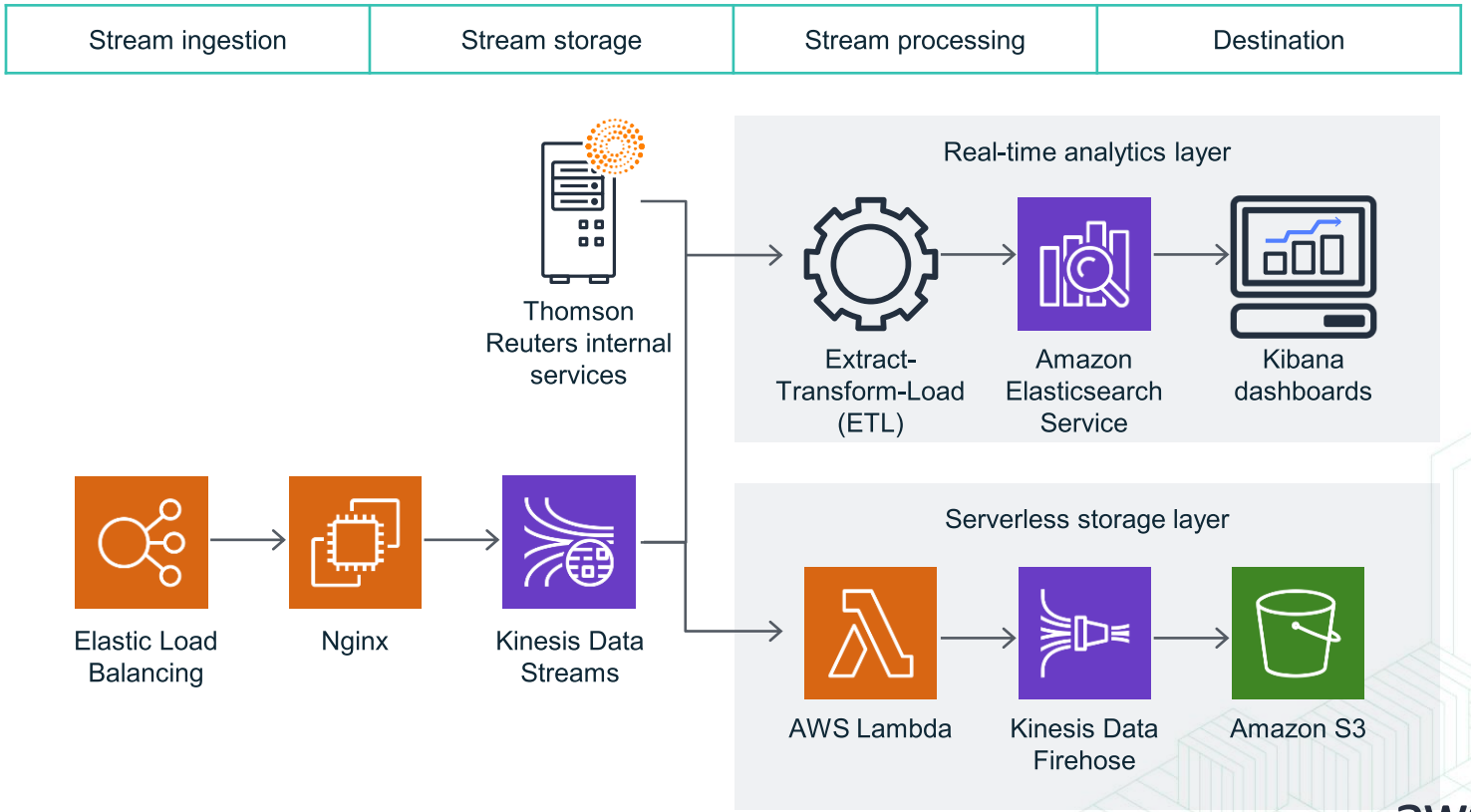
THOMSON REUTERS

Thomson Reuters provides professionals with the intelligence, technology, and human expertise they need to find trusted answers.

- Ability to process up to 4,000 events per second, anticipated to scale to 10,000 within one year
- Data pipeline accommodates twofold to threefold traffic increases during breaking news
- No data loss or downtime since launch, thanks to robust failover architecture
- Near-real-time availability of analytics data
- Simultaneous streaming and batching of data in one solution

Web Case Study

Thomson Reuters architecture



Video Case Study

aws training and certification

EPIC GAMES

Fortnite | 200+ million players

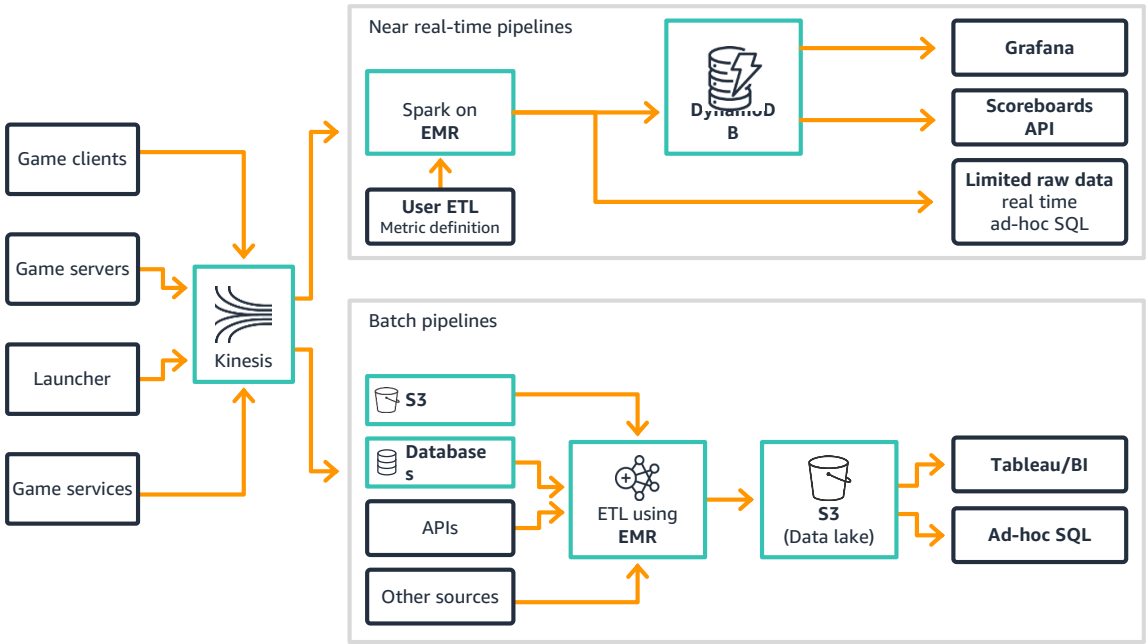
CHALLENGE

Need to create constant feedback loop for designers

Gain up-to-the-minute understanding of gamer satisfaction to guarantee gamers are engaged, thus resulting in the most popular game played in the world

Re:Invent 2018 Video

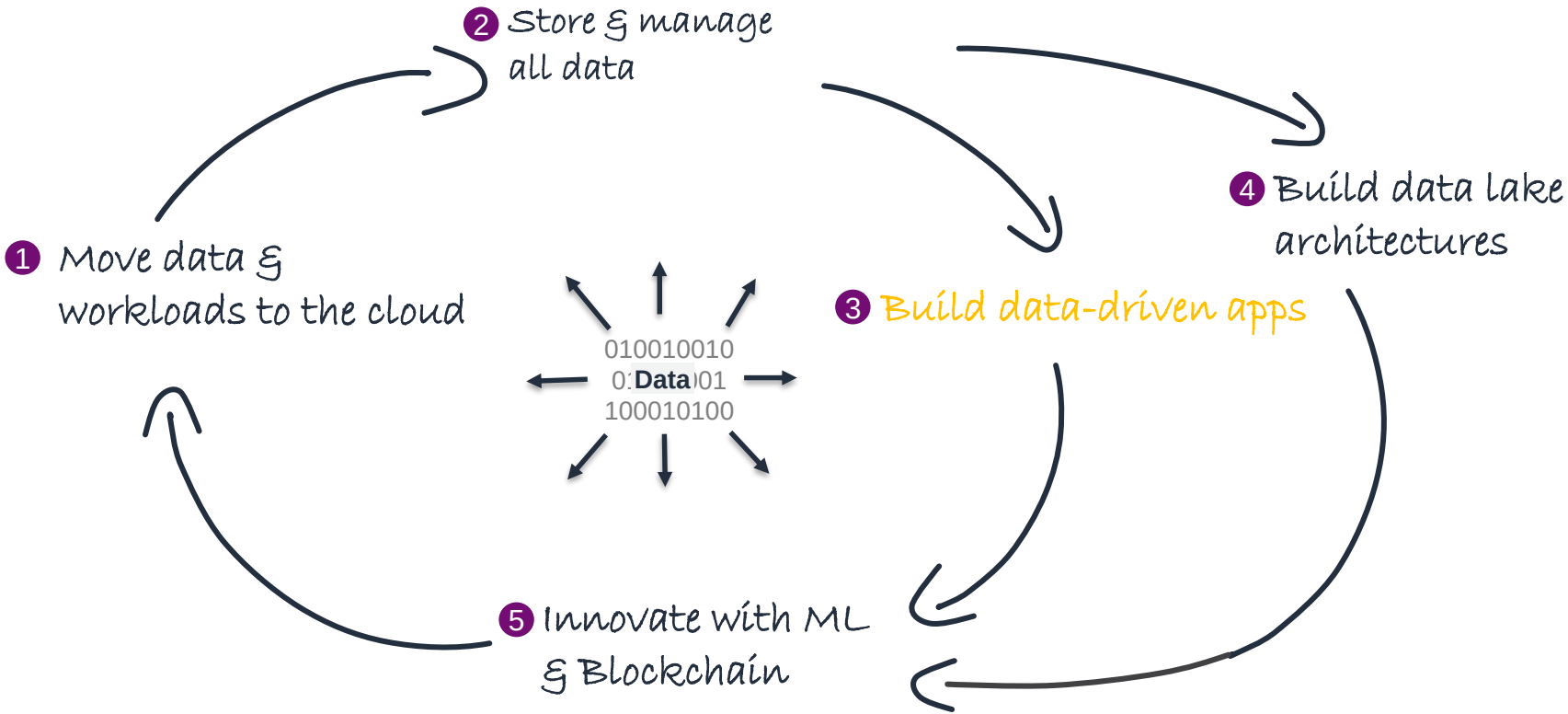
Epic Games uses AWS Data Lakes and Analytics



- Entire analytics platform running on AWS
- Amazon S3 leveraged as a data lake
- All telemetry data is collected with Amazon Kinesis
- Real-time analytics done through Spark on Amazon EMR, DynamoDB to create scoreboards and real-time queries
- Use Amazon EMR for large batch data processing
- Game designers use data to inform their decisions

Re:Invent 2018 Video

The Data Flywheel



Modern apps create new requirements

Ride hailing

Media streaming

Social media

Dating

Users:	1million+
Data volume:	TB–PB–EB
Locality:	Global
Performance:	Milliseconds–microseconds
Request Rate:	Millions
Access:	Web, Mobile, IoT, devices
Scale:	Up-down, Out-in
Economics:	Pay for what you use
Developer access:	Instant API access

Challenge:

Challenges with scaling their internet business to the next level

Solution:

Architected application with purpose-built databases for different needs

- Relational data: **Amazon RDS** for referential integrity and primary transactional database
- User search history: **Amazon DynamoDB** for massive data volume, and quick lookups for personalized search
- Session state: **Amazon ElastiCache** for in-memory store for sub millisecond site rendering

https://aws.amazon.com/blogs/aws/airbnb-reinventing-the-hospitality-industry-on-aws/

Blog post

NFL Next Gen Stats

Challenge

The NFL is America's largest sports organization with 180M fans worldwide. They needed to automate and provide basic game stats in real time (<1 second), and develop new ways of visualizing the action on the field to uncover deeper insights.

Solution

Sensors in the football and player shoulder pads generate real-time data that gets processed by Amazon QuickSight and AI services in real time to make insights available to 100s of users within NFL, the 32 clubs in the league, and broadcast partners through the NFL Next Gen Stats Portal.

Benefits

Their Quicksight dashboard allows users of the portal to quickly derive stats based on a number of different criteria. These statistics were previously generated manually, and delivered to the same audience on a weekly basis. With QuickSight, this data is now available real-time. Using embedded QuickSight dashboards also allows the NFL team to add new metrics and calculations to their application by simply modifying and republishing the dashboard.

> "With the Amazon QuickSight Readers and pay-per-session pricing, we are able to extend these secure, customized and easy to use dashboards for each club without having to provision servers or manage infrastructure—all while only paying for actual usage."
>
> **—Matt Swensson**
> **VP, Emerging Products and Technology**

https://aws.amazon.com/nfl/

ASEAN Case Study - Zalora

ZALORA

- EC2 instances climb from a baseline of 5 to 40, spiking to 1,000 instances at peak time
- S3 for static assets
- RDS for highly scalable primary databases, custom snowflake schema, linked it to well-known ETL tools, and used it to drive core business reporting across all functions.
- Redshift for data warehouse; retail reporting, merchandise planning, monitoring of retail KPIs.

https://aws.amazon.com/solutions/case-studies/zalora/

This is My Architecture video

ASEAN Case Study - Visenze

https://aws.amazon.com/solutions/case-studies/visenze/

ASEAN Case Study - Kumparan

https://aws.amazon.com/solutions/case-studies/kumparan/

ASEAN Case Study – Warung Pintar

https://aws.amazon.com/solutions/case-studies/warung-pintar/

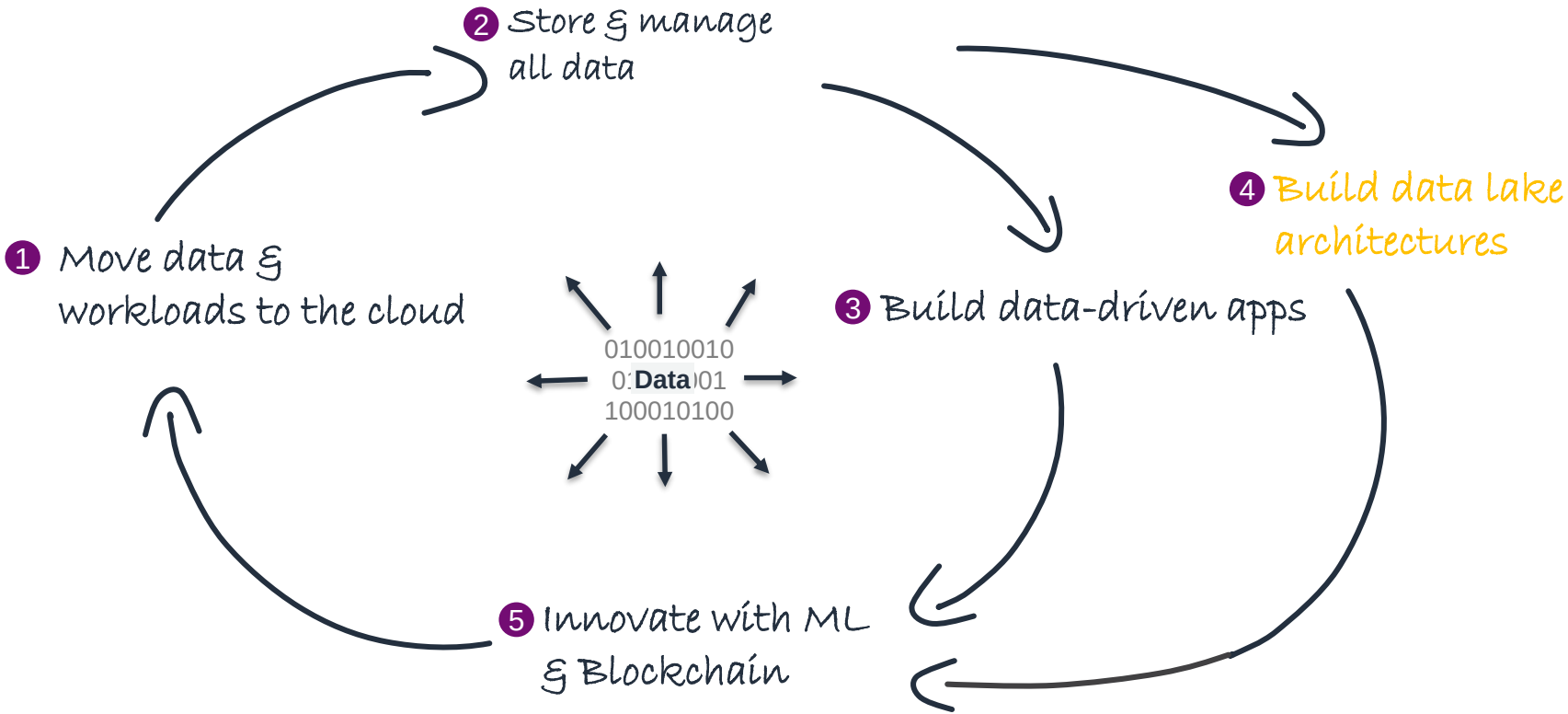
ASEAN Case Study – Sunday Insurance

https://aws.amazon.com/solutions/case-studies/sunday_insurance/

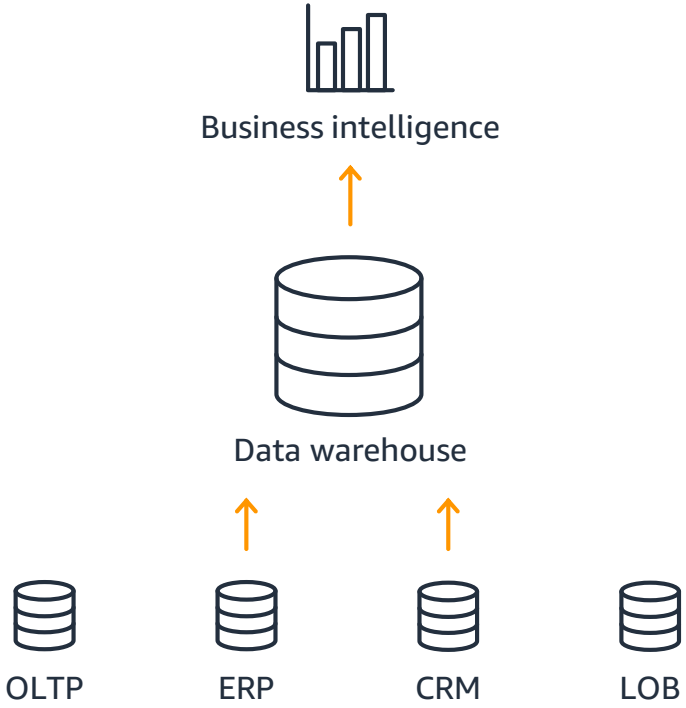
ASEAN Case Study - iFlix

https://aws.amazon.com/solutions/case-studies/iflix/

The Data Flywheel

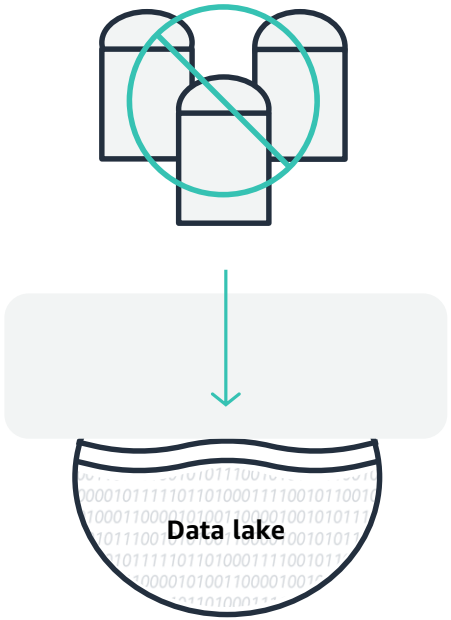


Traditionally, Analytics Used to Look Like This



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

Customers need to…

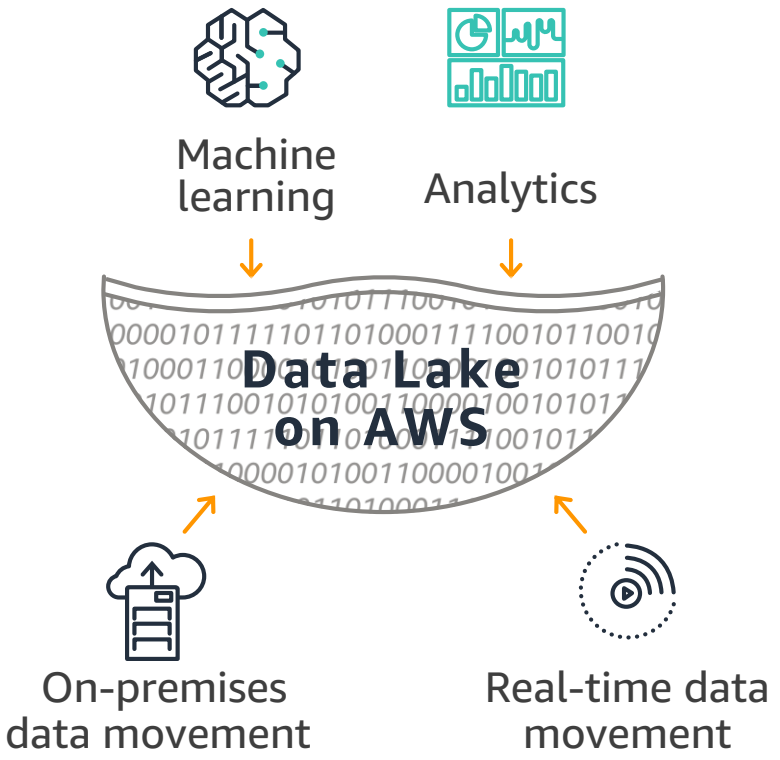


Capture and store new non-relational data at EB scale.

Secure and combine data from new and existing sources.

Do new types of analysis on their data.

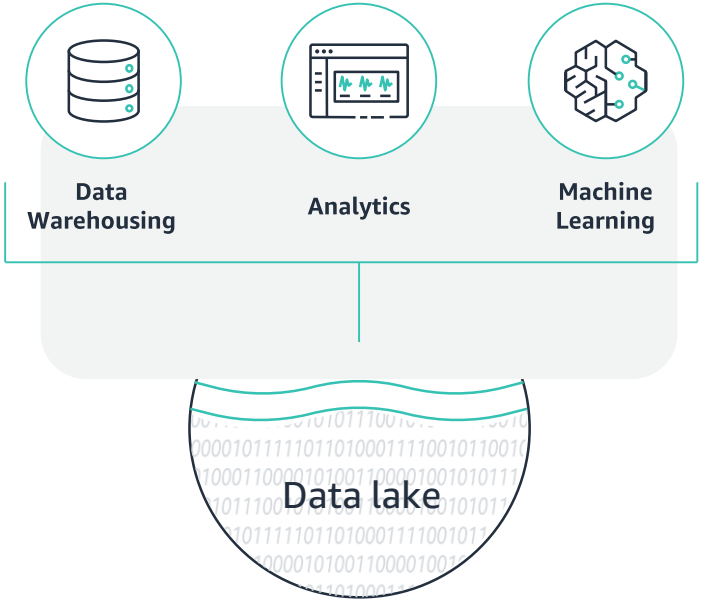
What is a Data Lake?



A **Data Lake** is a **centralized repository** that allows you to store all your **structured and unstructured data** at any scale. You can store your data as-is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions.

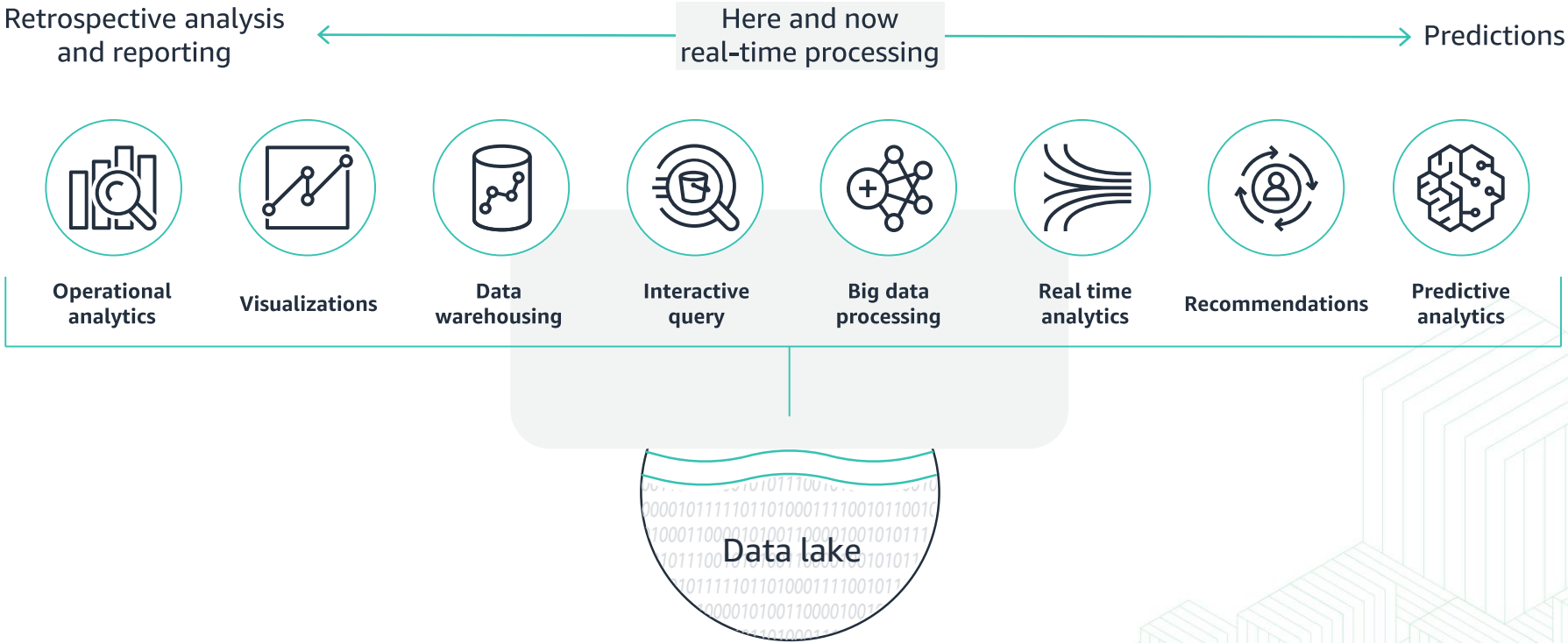
Data lake architectures

Bringing together the best of both worlds

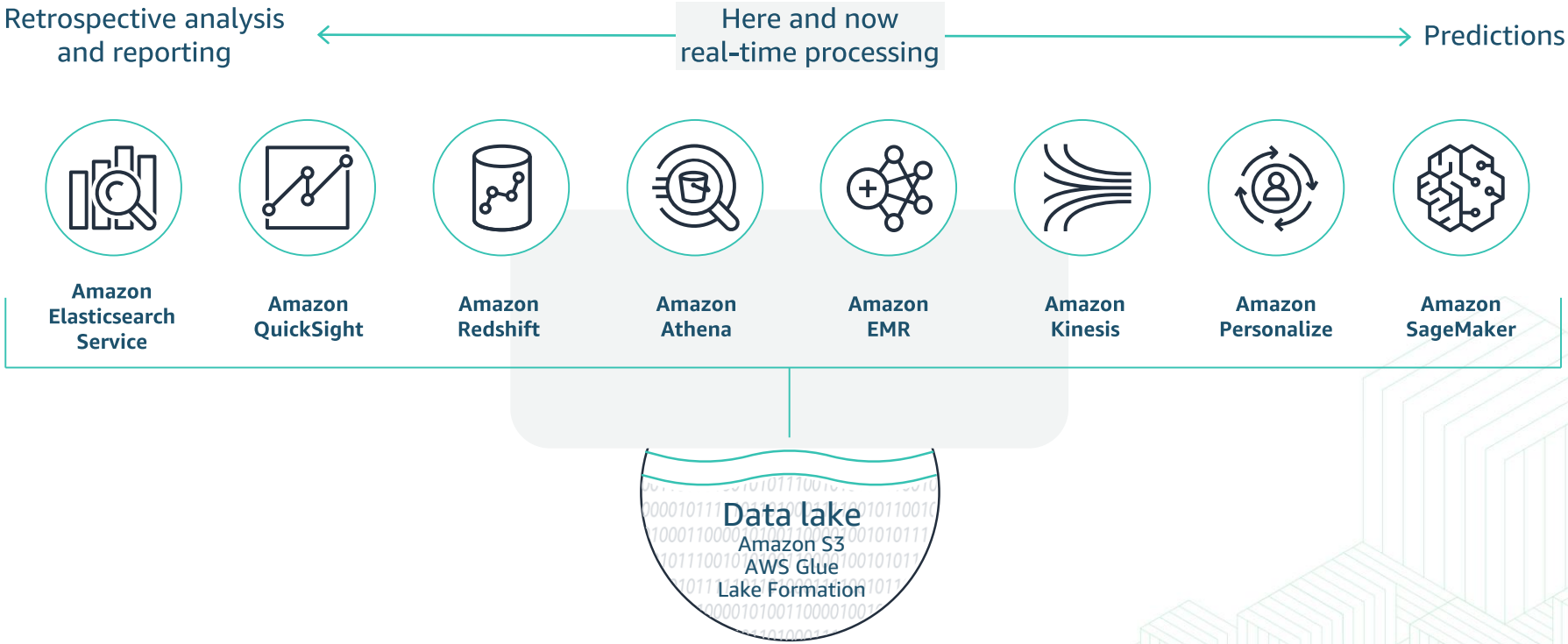


- Extends or evolves their data warehouses
- Durable and available; exabyte scale
- Secure, compliant, auditable
- Run any type of analytics from DW to predictive
- Decoupling of compute and storage
- On-demand resources, tiering, cost choices

Any type of analytics on the data lake



Any type of analytics on the data lake



Amazon S3 (Simple Storage Service)

Durable

Designed for
99.99999999999%
of durability

Available

Designed for
99.99% availability

High performance

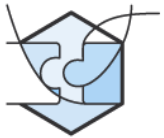
- Multiple upload
- Range GET

Easy to use

- Simple REST API
- AWS SDKs
- Read-after-create consistency
- Event notification
- Lifecycle policies

Scalable

- Store as much as you need
- Scale storage and compute independently
- No minimum usage commitments



Integrated

- Amazon EMR
- Amazon Redshift
- Amazon DynamoDB

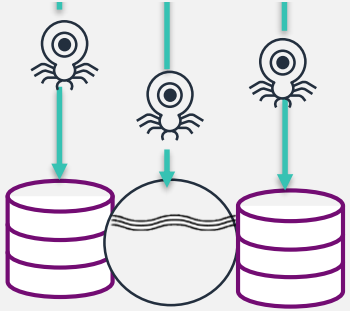
Common Challenges with ETL

- Volume of data grows
- Data sources are added
- 70% of ETL jobs are hand-coded
- Data formats change over time
- Target schemas change

AWS Glue
Serverless Data catalog & ETL service

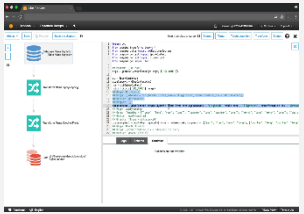
Data Catalog

Discover data and extract schema



ETL Job authoring

Auto-generates customizable ETL code in Python, Scala, and Spark

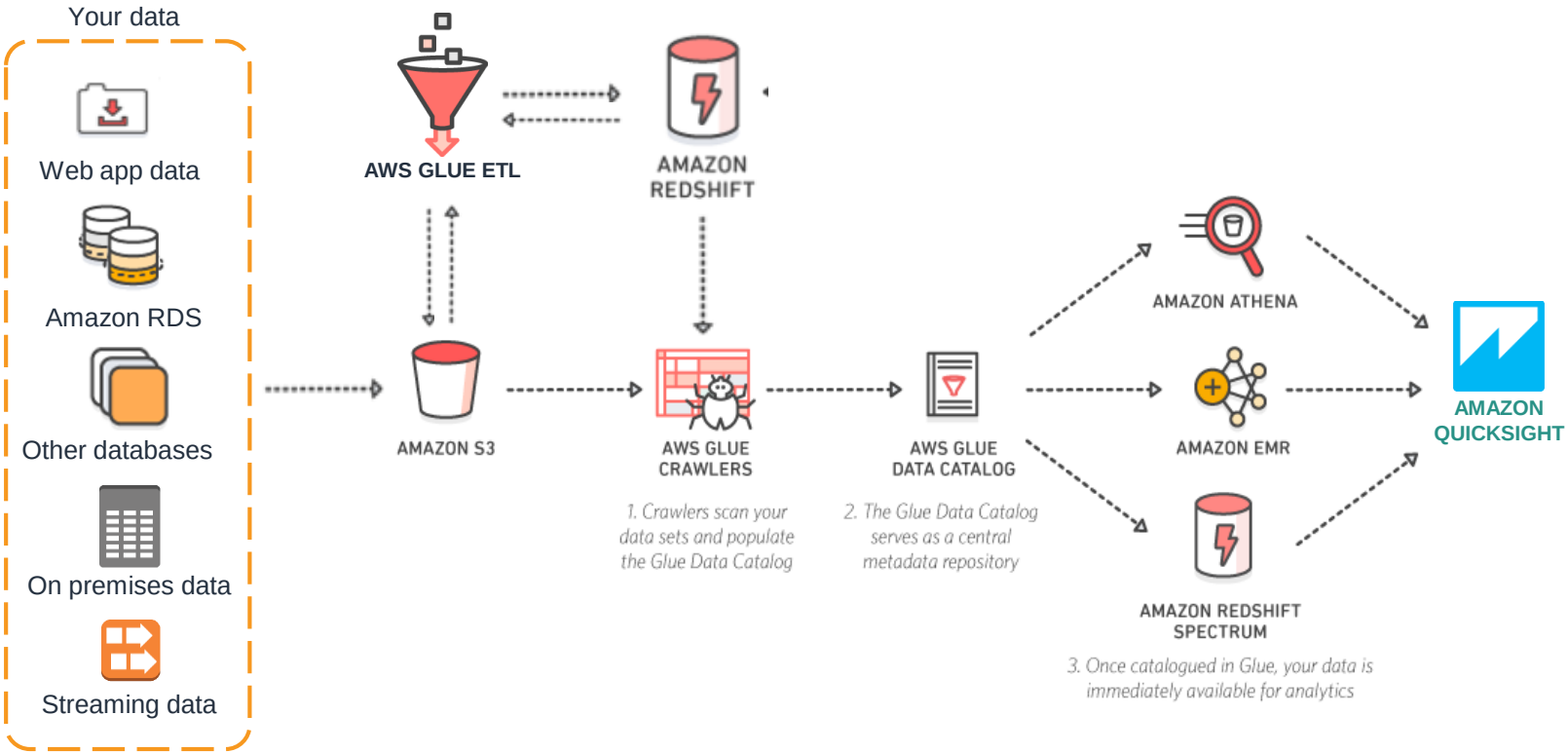


- Automatically discovers data and stores schema
- Data searchable, and available for ETL
- Generates customizable ETL code
- Schedules and runs your ETL jobs
- Serverless, flexible, and built on open standards

[Knowledgent's Intelligent Clinical Trial App Video]

aws training and certification

Use Case:
Interoperate data lake and data warehouse



ning and
ification

Case Study: FinAccel

"AWS Glue allows us to **pay only for computing power** that we need to run the jobs. It is amazing that leveraging AWS Glue has enabled our **small team of data engineers to run the whole data infrastructure** in our company."

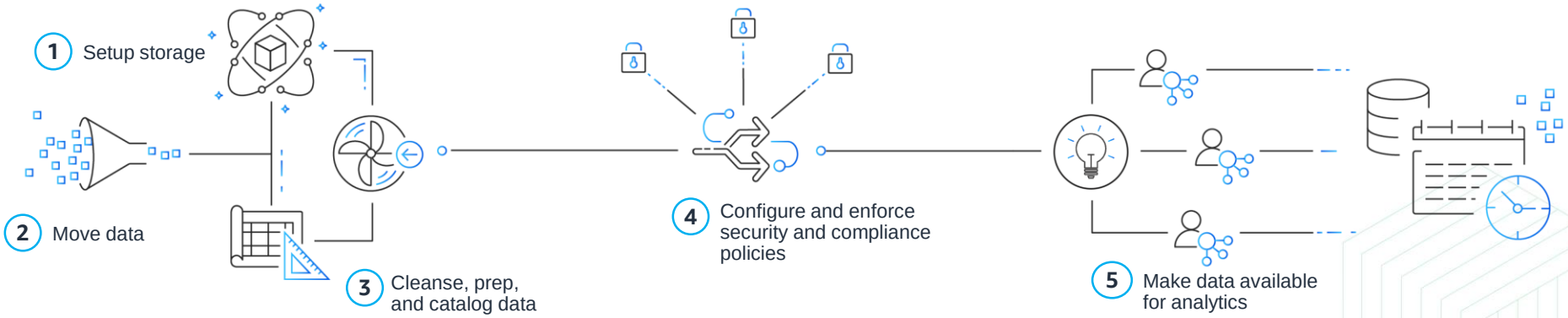
- Umang Rustagi, Co-founder and COO, FinAccel

https://aws.amazon.com/solutions/case-studies/finaccel/ 4:23

Please also see: https://aws.amazon.com/solutions/case-studies/visenze/ 4:01

We will resume in 15 mins

Typical steps of building a data lake



AWS Lake Formation

Identify, ingest, clean, and transform data

Move, store, catalog, and clean your data faster with **machine learning**

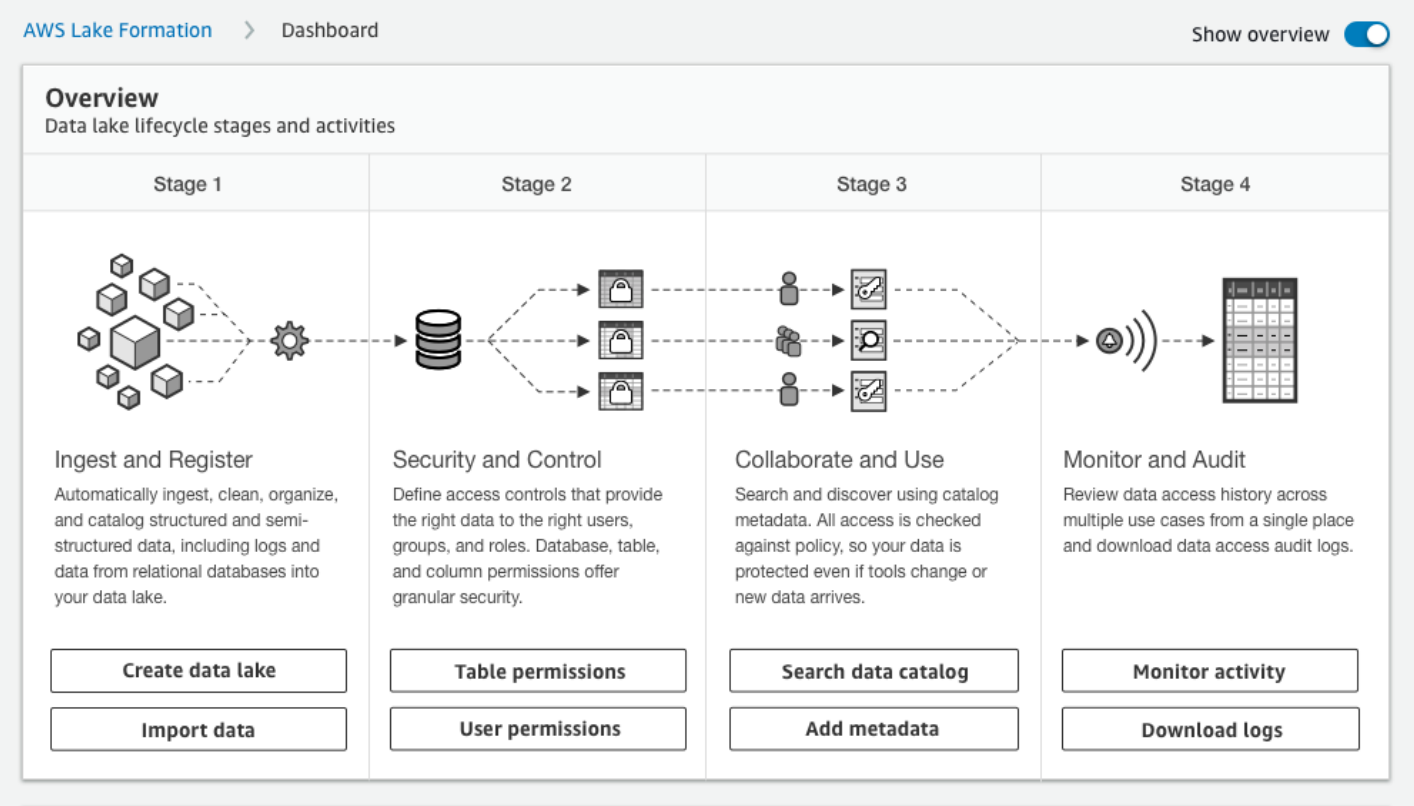
Enforce security policies across multiple services

Enforce security policies across multiple **services**

Gain and manage new insights

Empower analysts and data scientists to gain and manage new **insights**

How it Works



Sysco Foods uses AWS data lakes and analytics

Unlock data for analysis by data scientists to business users

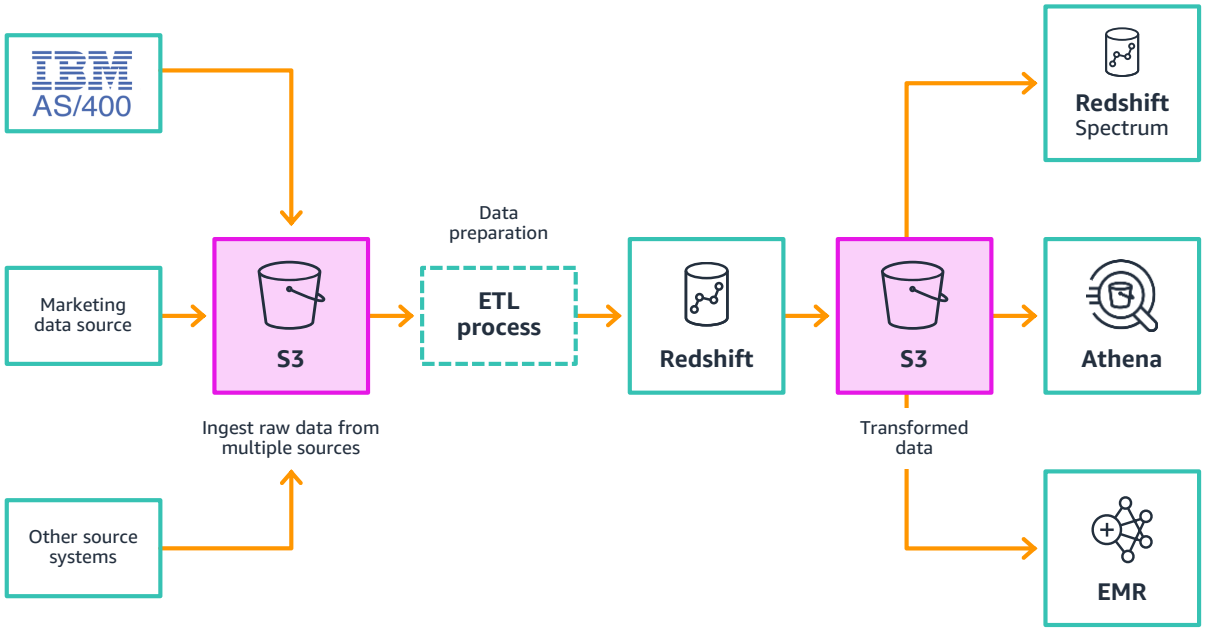
Sysco is the leader in selling, marketing, and distributing food.

They wanted to eliminate data silos by combining large volumes of data distributed across multiple systems, and to reduce the high costs of maintaining their on-premises data warehouse.

Sysco migrated their on-premises data warehouse to AWS where they could use Amazon S3 as a Data Lake and Amazon Redshift, Redshift Spectrum, Amazon EMR, and Amazon Athena for analytics.

“By bringing the data to S3 from various sources, including niche datamarts such as Lexington, and EDW into S3, the true potential of the data was unlocked by allowing different personas to use the ecosystem in new ways.“

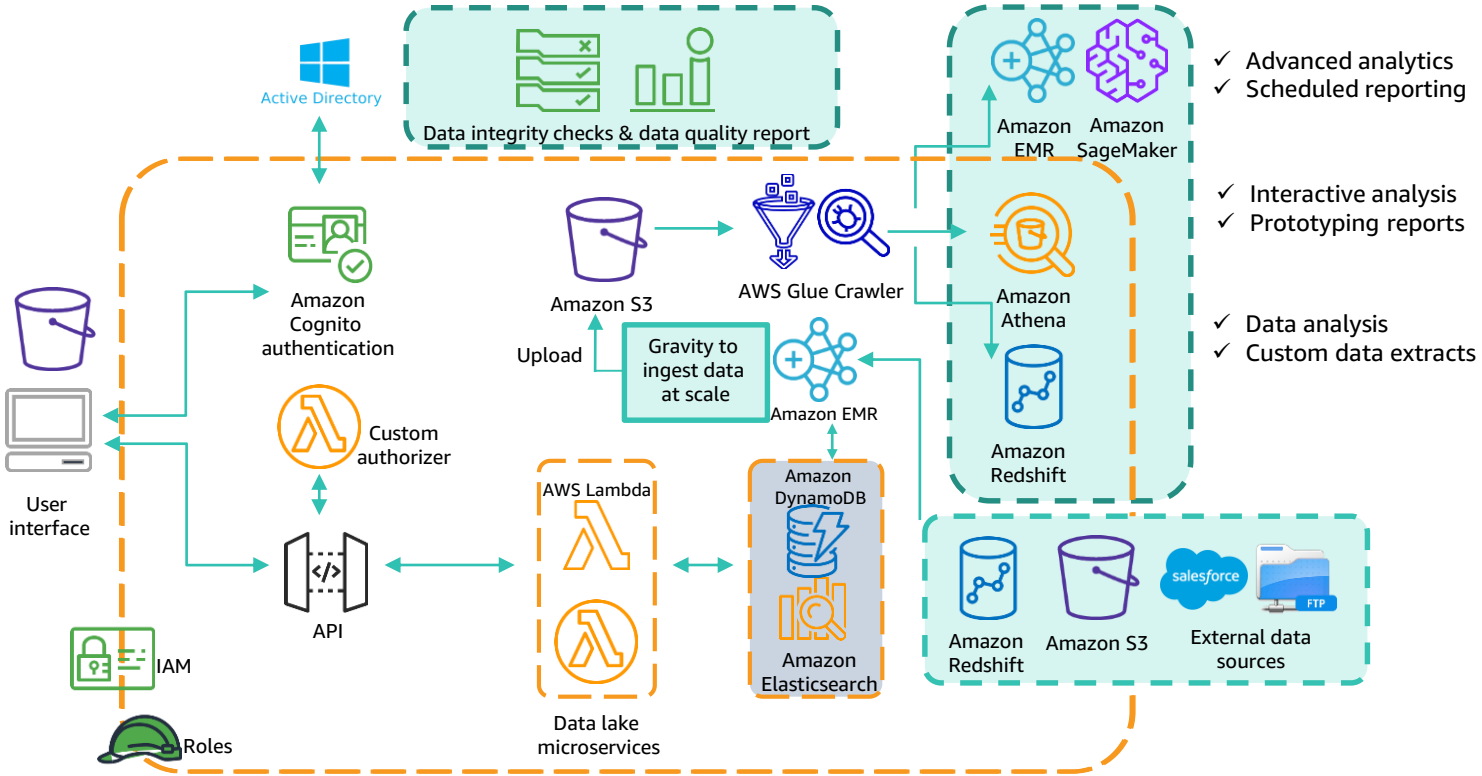
Sysco Foods uses AWS data lakes and analytics



- Sysco is the leader in selling, marketing and distributing food
- Challenge: large volumes of data in multiple systems
- Consolidated data into a single S3 data lake
- Data scientists use EMR notebooks, Athena and Amazon Redshift Spectrum used by business users for reporting

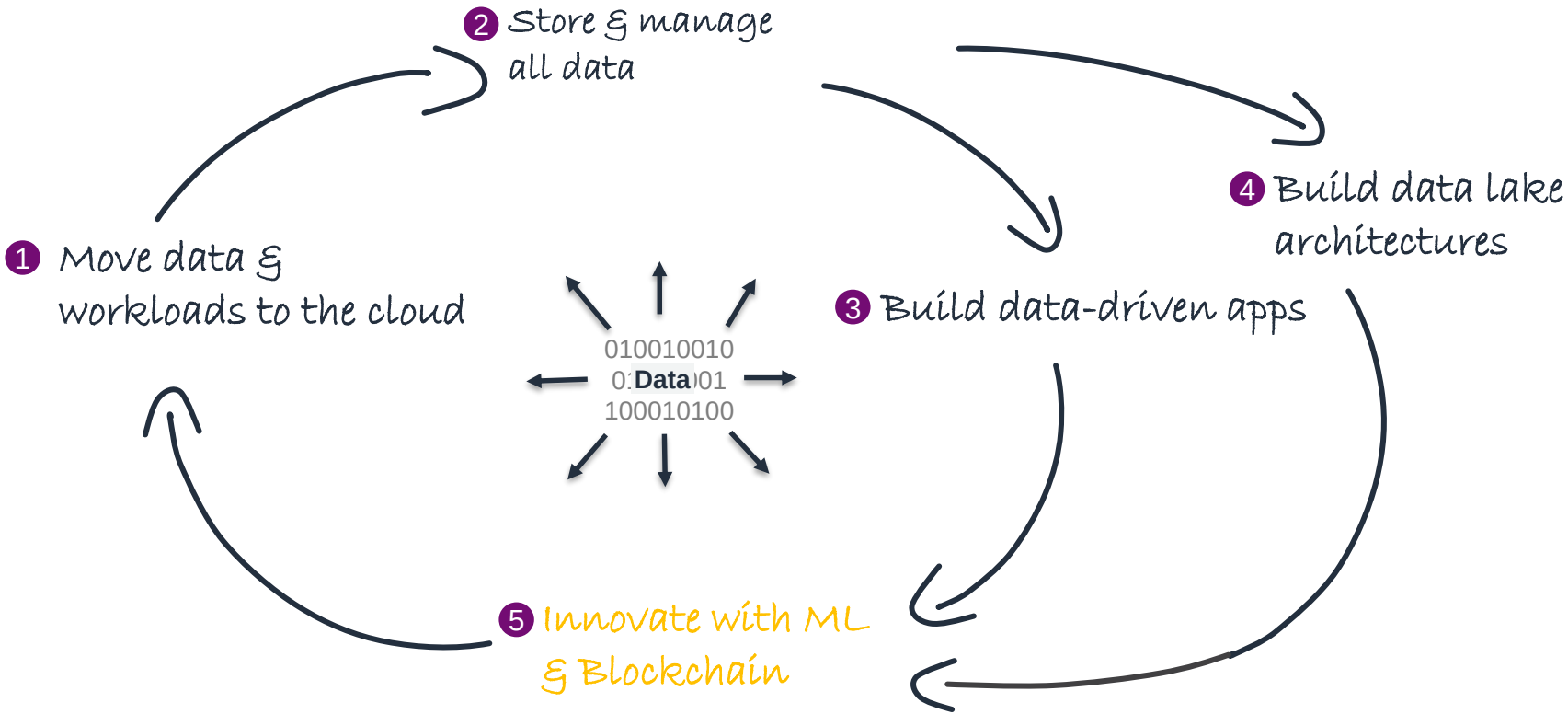
https://youtu.be/ua1dCSeViEo?t=875

aws training and certification



https://youtu.be/ua1dCSeViEo?t=870

The Data Flywheel



AWS for AI and Machine Learning

Broadest and deepest set of AI and ML services

- 200 new features & services launched this last year alone
- Unmatched flexibility

Accelerate your adoption of ML with SageMaker

- 70% cost reduction in data-labeling
- 10x faster performance
- 75% lower inference cost

Built on the most comprehensive cloud platform optimized for ML

- AWS holds the top spots on Stanford's benchmark, for fastest training time, lowest cost, lowest inference latency

aws training and certification

AWS Machine Learning Customers

10,000+ customers | 85% of TensorFlow projects in the cloud happen on AWS

aws training and certification

THE AWS ML STACK

Broadest and deepest set of capabilities

AI Services

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

ML Services

 **Amazon SageMaker**

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

ML Frameworks + Infrastructure

FRAMEWORKS	INTERFACES
TensorFlow, mxnet, PYTORCH	GLUON, Keras

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

aws AI

Accelerating financial analysis

Using TensorFlow on Amazon SageMaker, Siemens Financial Services developed an NLP model to extract critical information to accelerate investment due diligence, reducing time to summarize diligence documents from 12 hours down to 30 seconds.

SIEMENS

aws AI

Personalizing customer experiences

Domino’s uses Amazon Personalize to customize and scale relevant marketing communications to customers based on time, context, and content, thereby improving and enhancing their experience with the Domino’s brand.

Summary – Data Flywheel

From Discovery to Visualization

Discover

AWS Glue (Crawlers)

Ingest

AWS Glue

Kinesis Streams & Data Firehose

Database Migration Service (DMS)

Snowball

Snowmobile

Direct Connect

Managed Streaming for Kafka (MSK)

Store

S3

S3 Glacier

RDS

Aurora

DynamoDB

Secure

Identity and Access Management (IAM)

Key Management Service (KMS)

Macie

Cloudtrail

Cloudwatch

Catalog

AWS Glue (Data Catalog)

EMR (Hive Metastore)

Prepare

AWS Glue ETL (Serverless Apache Spark)

EMR (Apache Spark & Hadoop)

Analyze

Sagemaker

Redshift

Athena

Kinesis Data Analytics

EMR (Apache Spark & Hadoop)

Elasticsearch

AI services

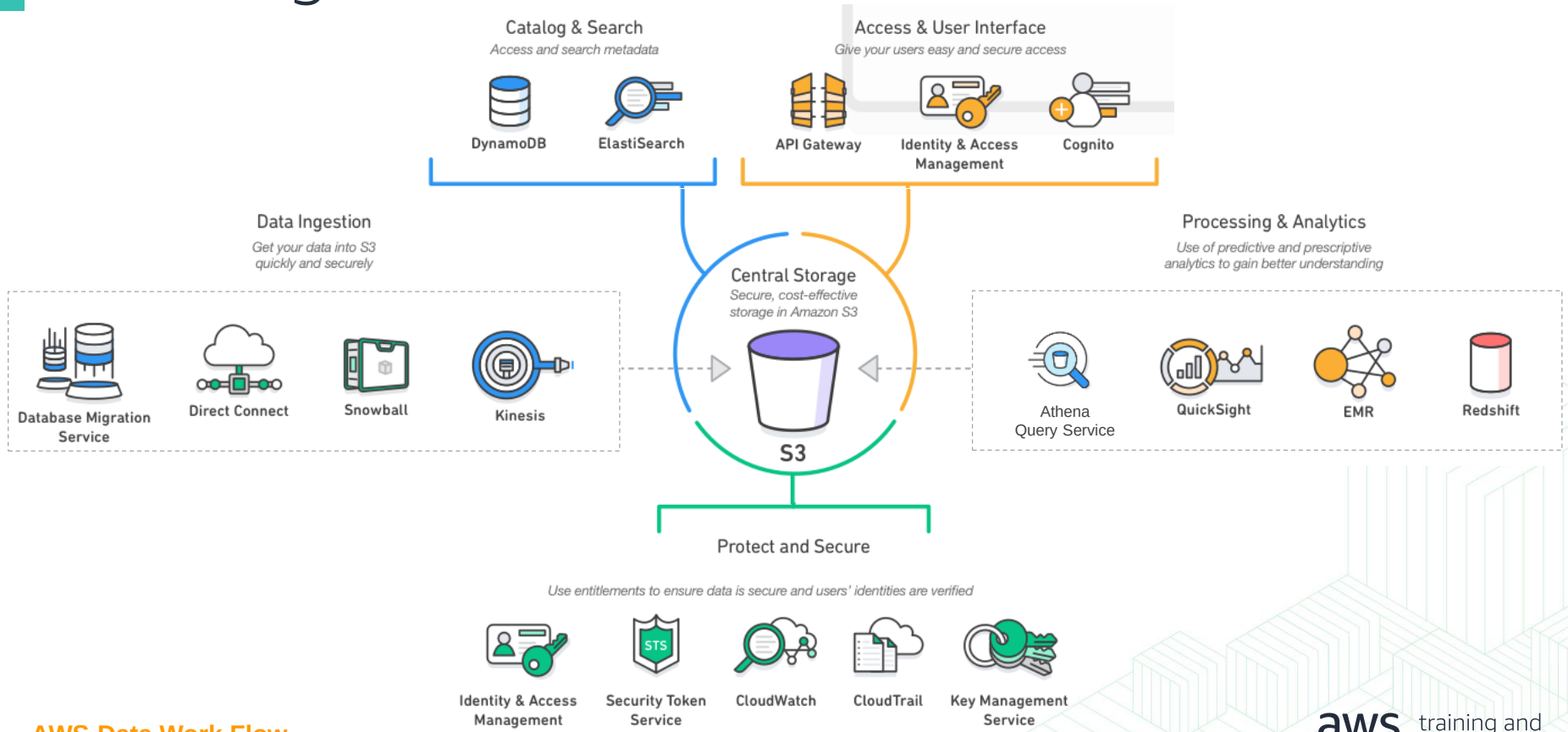
Visualize

Quicksight

EMR Notebooks

aws training and certification

Building a Data Lake on AWS



AWS Data Work Flow

aws training and certification

Best Practices of building a Data Lake

- **Build decoupled systems**
  - ✓ Data → Store → Process → Store → Analyze → Answers
- **Use the right tool for the job**
  - ✓ Data structure, latency, throughput, access patterns
- **Leverage AWS managed and serverless services**
  - ✓ Scalable/elastic, available, reliable, secure, no/low admin
- **Use log-centric design patterns**
  - ✓ Immutable logs, data lake, materialized views
- **Be cost-conscious**
  - ✓ Big data ≠ big cost
- **Talk to your consumer groups before you startup building**

Business Problem to Solution

Open Discussion

- Whom in the organization do you target to win over?
- What are the questions to ask to prospective customers?
- What best practice can you share with your group?

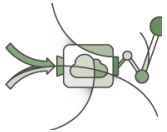
Business Outcomes on a Modern Data Architecture

Outcome 1 : Modernize and consolidate

- Insights to enhance business applications and create new digital services

Outcome 2 : Innovate for new revenues

- Personalisation, demand forecasting, risk analysis



Outcome 3 : Real-time engagement

- Interactive customer experience, event-driven automation, fraud detection

Modernise and Consolidate

Customer Challenges

"I don't want to be **locked into a proprietary data** format or system."
"I don't want to renew **software licensing** for existing and aging data warehouse."
"My hardware has reached it's **peak capacity** and I'll need to invest in more hardware to meet my SLAs."
"Our current infrastructure is going **out of support.**"

"I want to **free my IT staff from** spending majority of their time in **patching and maintaining hardware.**"

"I want to analyze **new data sources.**"
"I want to **monetize insights** from data, making competitive advantage."
"We have a new initiative that will generate historical data that **needs to be analyzed** "

Discovery Questions – to Business

- How easily can you **understand customer value** across your different departments?
- What would it mean for your business to have a **better cost per insight** ratio?
- Have you been able to **quantify the cost of running your operations** as a result of not having a single view of customer value?
- What would it mean for your business to **analyze historical data alongside new data**, like telemetry or social data?
- What would be the impact if you could **analyze more or all of your data**, rather than specific time periods?

Discovery Questions – to Technology

- Are you using a **data warehousing appliance** today? Which one?
- How satisfied are you with your current **licensing/support agreements** with your vendor? Are you nearing **capacity?**
- How effective is your data warehouse in meeting time-sensitive analytics projects that require **additional capacity, data sets and analytics techniques?**
- Are you using data science and machine learning to **gain insights** of your data?

"Have you considered a modern data architecture built natively for the cloud, that gives you the flexibility to use the right tool for the job?"

Re:Invent 2019 Andy Jassy Keynote

Innovate for New Revenues
Real-time engagement

Customer Challenges

"I want to have **better insight** of my business."
"We want to **democratize data**."
"I can't get a **360 degree view** of our customers."
"We want to **consolidate data** from multiple sources."

"We need insights on streaming data, **near real-time.**"

"Complex and costly data collection and processing systems are limiting the **type and amount of data** we can collect and analyze."

"We are growing our data science team and want to give them **easier access to data.**"
"We want to shift our data engineers and infrastructure engineers **towards data science.**"

Discovery Questions – to Business

- What would it mean for your business if you could **extract more value or monetize** your data?
- What would it mean for your business if you could **forecast, predict** and proactively change course?
- How are you managing fraud detection, predictive maintenance, customer 360, IoT, clickstream, operational analytics, root-cause analysis to reduce mean time to detection and mean time to recovery?
- Do you have visibility into your **customer churn/retention?**

Discovery Questions – to Technology

- What does your **big data/analytics landscape** look like today?
- Do you have any **challenges scaling and/or processing your data?**
- How is your team dealing with the **various types of data** being generated by new applications?
- How do you **manage access to data** and ensure quality?
- How **secure** is your data today?
- How siloed is your data today?

Target Personas and Use Cases

Target Personas

Title	CIO / CTO Line of Business Owners	Chief Data Officer	Director of Data Analytics Solutions Architects DB Administrators
Role	lead large scale migrations or re-factoring as part of some overall transformation initiative	Lead analytics and database architecture	lead POCs and product evaluations. Often the primary decision maker for data warehouse transitions
Priorities	Cost, meeting new business requirements, time-to-market	Scale, reduced complexity, innovation	Scale, price to performance, query times, ease of operation.
Concerns	Overall migration risks, skills gap, upfront investment	Vendor lock-in, model transitions, cost control	Satisfying business users, skills gap, migration risks, security

Use Case – Financial Services

Business Problem and Opportunities

- Need to analyze trading and market data
- Need to access historic data set for growing internal groups
- Risk analysis
- Fraud detection
- Need of cost-effective and scalable BI tool
- Fast, interactive queries supporting wide range of users

NASDAQ decreased time-to-market threefold, analyzes more data

Challenge

Each trading day, between 30–50 billion orders, trade and quote messages must be stored and be available for querying. NASDAQ wanted to make their historical data footprint, which previously existed in a large number of disparate systems, available to analyze as a single data lake.

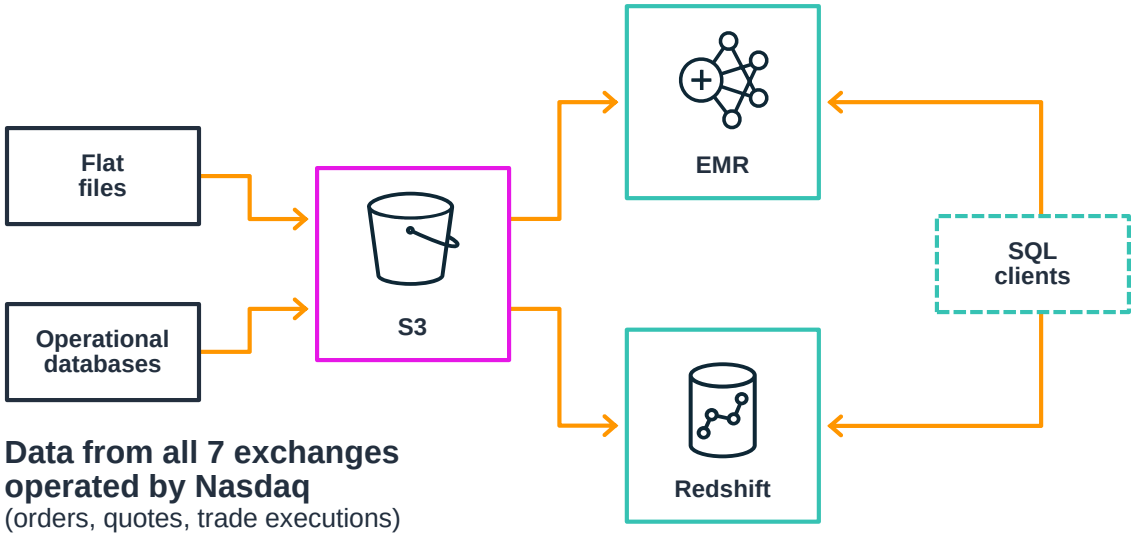
Solution

NASDAQ deployed Amazon S3 as their data lake, Presto on Amazon EMR to process historical data, and Amazon Redshift for fast, interactive queries supporting a wide range of users with different skillsets.

Benefits

They have saved in costs and time-to-market. By leveraging the AWS services, NASDAQ has avoided the need to buy expensive hardware and were able to decrease their time to market threefold. They can now support years of message data at nanosecond scale in their data lake with the option to increase the amount of data stored indefinitely.

Nasdaq uses AWS to build a data lake



Data from all 7 exchanges operated by Nasdaq
(orders, quotes, trade executions)

- Migrate legacy on-premises warehouse to Amazon Redshift
- 4.8B rows inserted per trading day (orders, trades, quotes)
- Ingest data from multiple sources, validates, and stages in S3
- Redshift reads data out of S3 for fast queries
- Presto on EMR and S3 used for analysis of their massive historical data set

Video Case Study

Re:Invent 2019 Video

aws training and certification

Capital One reduces time from data to insights

Challenge

Capital One has thousands of use cases with both operational and analytics needs. They wanted a BI tool that was cost-effective and highly scalable, to grow alongside their business.

Solution

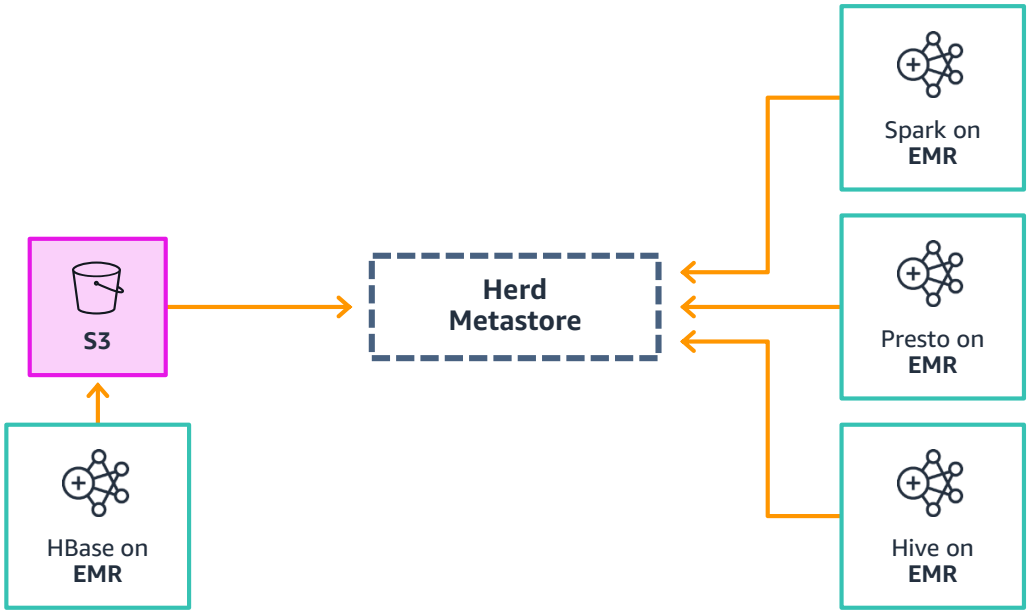
They chose Amazon Quicksight because it was modern, self-service, and easy to onboard. Integrated with their AWS environment, it was easy to set up new environments, removed the pain of maintenance, and allows for pay-per use.

Benefits

They can now get insights from data in a day, and set up embedded analytics within a week. Quicksight enabled them to redefine their BI Center for Excellence to focus on things that will move the business forward, instead of only operations.

https://aws.amazon.com/solutions/case-studies/capital-one/
Video case study

FINRA uses AWS data lakes and analytics



- Required fast access across trillions of trade records (20PB+)
- Migrated from on-premises system
- Use Apache HBase on Amazon EMR to store and serve this data
- Use EMR engines—Spark, Presto, and Hive to process data
- Lower costs by 60% over their on-premises system

Use Case – Media and Entertainment

Business Problem and Opportunities

- Massive increase in volume of content produced
- Evolution of how content is consumed
- Ever-changing methodologies to report on data
- Need more granular insights, needing bigger data sets to augment data
- Limitations in storage capacity
- Need to analyze and query live data across relations databases
- Need a high performance data warehouse
- Need to serve large amount of queries concurrently in short amount of time

Nielsen transforms analytics with Amazon Redshift

Challenge

The Nielsen TV measurement products measure everyone, everywhere. With the evolution of how content is consumed, and the massive increase in the total volume of content produced their technology and platforms needed to evolve.

They were running a legacy on premise data warehouse based on Netezza with a monolithic architecture that had limited scale.

Solution

Nielsen built a new data lake-based architecture. The new solution uses Amazon Redshift for daily analysis on data with a tight SLA, and Amazon EMR for transient batch processes.

They also use Amazon Athena on Amazon S3 extracts of Amazon Redshift data for data quality checks.

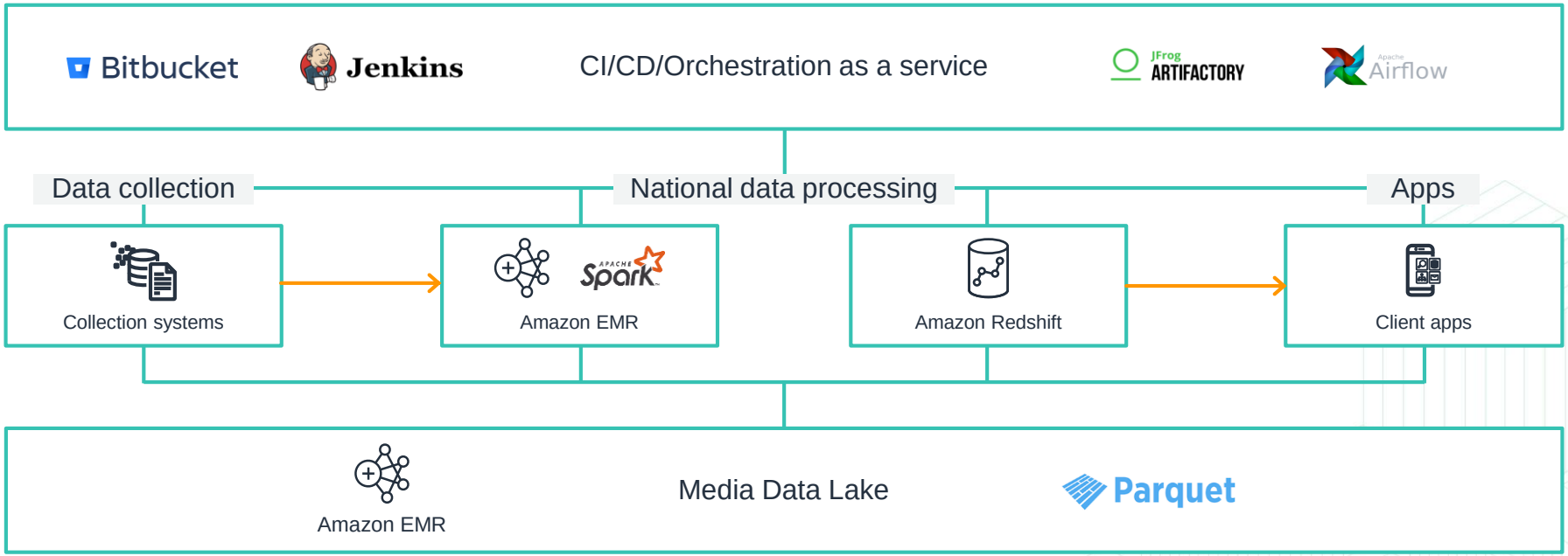
Benefits

The new purpose-built solution provides seamless end to end orchestration for the entire platform. It provides the scale and elasticity needed to provide the ever-changing analytics needs of the team and increased efficiency.

This architecture provides flexibility to continue to optimize costs with offers like spot instances.

Video case study (re:Invent 2019)

nielsen



aws training and certification

> We utilize many AWS and third-party analytics tools, and we are pleased to see Amazon Redshift continue to embrace the same varied data transform patterns that we already do with our own solution. We've harnessed Amazon Redshift's ability to query open data formats across our data lake with Redshift Spectrum since 2017, and now with the new Redshift Data Lake Export feature, we can conveniently write data back to our data lake. This all happens with consistently fast performance, even at our highest query loads. We look forward to leveraging the synergy of an integrated big data stack to drive more data sharing across Amazon Redshift clusters, and derive more value at a lower cost for all our games.

—Kurt Larson
Technical Director of Analytics Marketing Operations

aws

Use Case – Retail

Business Problem and Opportunities

- Drive mission critical insights to support business (customer behavior analytics, personalized recommendations, demand forecasting…)
- Supply chain analytics, inventory management
- Churn prediction
- Price Optimization for products, promotion planning

Amazon.com lowers cost and gains faster insights with an AWS data lake

amazon.com

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

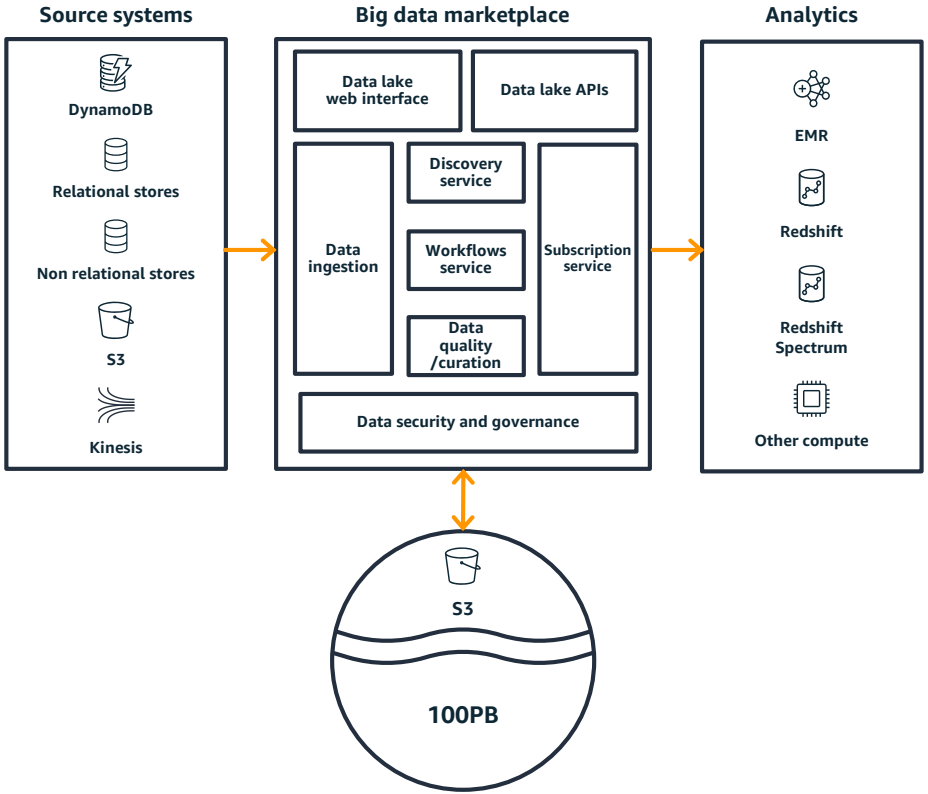
They doubled the data stored from 50PB to 100PB, lowered costs, and were able to gain insights faster.

Video case study

Re:Invent 2019

aws training and certification

Amazon uses an AWS Data Lake



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

aws training and certification

Case Study: NFL Health and Safety

- Using on field tracking and stats to improve player health
- Force of contact and concussion tracking and analysis
- Leading the world in this type of analysis.

- https://blog.aboutamazon.com/amazon-ai/partnering-with-the-nfl-to-transform-player-health-and-safety 1m:10s

- We will resume in 10 mins

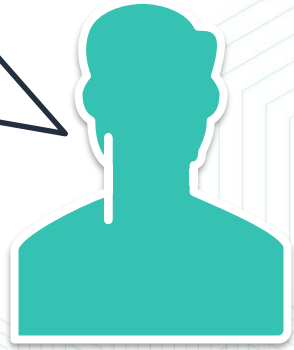
Customer Concerns

How would you handle these questions?

Objection Handling

Our on-prem analytics solution is enough.
We don't see the need to move to cloud.

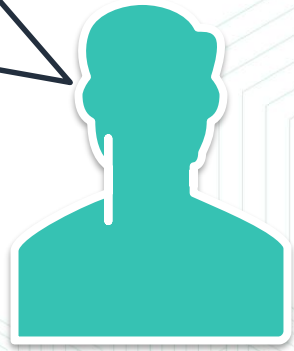
It may look enough for now, however **the volume and variety of your data** will increase continuously.
At that point your current solution will have challenges in cost and scale to manage.
Moving to cloud enables you to leverage a **fully-managed, highly scalable, and durable solution** in a more **cost-efficient** way.



Objection Handling

How is a data lake different from a data warehouse, which I already have?

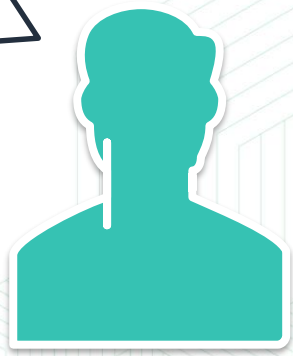
A data lake is a **scalable central repository of large quantities and varieties of data, both structured and unstructured.**
Data lakes complement traditional data warehouses, providing more flexibility, cost-effectiveness, and scalability for ingestion, storage, transformation, and analysis of your data.
The traditional challenges around the construction and maintenance of data warehouses and limitations in the types of analysis can be overcome using data lakes.



Objection Handling

I heard it’s a difficult to migrate to Amazon Redshift.

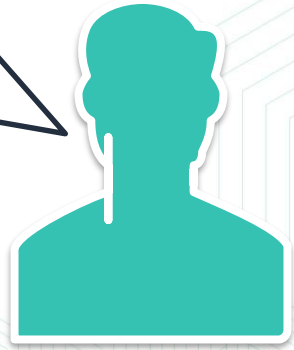
To analyze and automatically convert existing schema, you can use AWS **Schema Conversion Tool (SCT).**
In the migration process, you can use AWS **Data Migration Service (DMS)** that has no charge for 6months for selected source DB engines.
Also**, the source data warehouse remains fully operational during the migration.**
To help with the migration, we can also work with AWS program teams such as **Database Freedom** and **Migration Acceleration Program (MAP).**



Objection Handling

We will lose control of our data once moved to cloud.

Your data will be **encrypted in transit and at rest. Only you, as a customer, has access to your data.** You can leverage identity and access controls, configuration of firewalls, monitoring and logging, DDoS mitigation, management tools of AWS to keep your data secure.
Lastly, you can leverage AWS security partners.



Call to Action

Work with AWS

- Identify customers who would benefit from Data Analytics on AWS
- Register and Submit Opportunities to APN Customer Engagements (ACE) Platform
- Submit for Migration and PoC Funding

Position to AWS Customers

Consulting Partners

- Service Delivery Program (Redshift, DMS, EMR, Kinesis)
- Competency Program (Data and Analytics)

Technology Partners

- AWS Marketplace
- AWS Quickstart

Public Customer References (webpage, logo, AWS events, This is My Architecture videos)

APN Marketing Funding

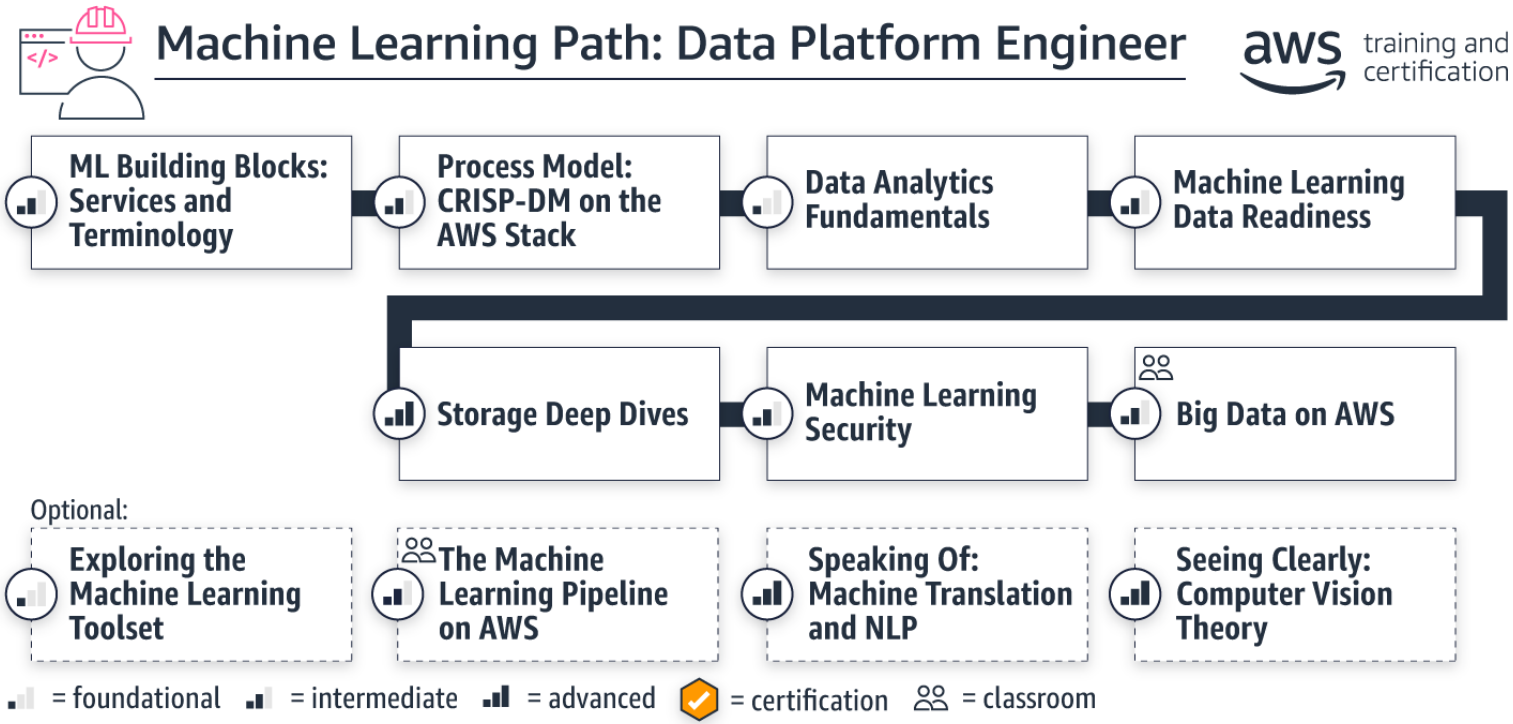
APN Marketing Central

Learn – AWS Certified Data Analytics - Specialty

https://aws.amazon.com/training/path-data-analytics/

aws training and certification

Learn - Data Platform Engineer Learning Path



https://aws.amazon.com/training/learning-paths/machine-learning/data-platform-engineer/

aws training and certification

Resources – Learning AWS Analytics

- Data Lakes and Analytics on AWS
- Data, Analytics, and Machine Learning Resource Hub
- Build Your Data Lake on Amazon S3 Infographic
- AWS Machine Learning Blog
- AWS Big Data Blog
- Data & Analytics Partner Solutions

Training Materials

- Data Analytics Fundamentals (free, digital): https://www.aws.training/Details/eLearning?id=35364
- Big Data on AWS (paid, classroom): https://aws.amazon.com/training/course-descriptions/bigdata/
- Exam Readiness: AWS Certified Big Data – Specialty (paid, classroom): https://www.aws.training/training/schedule?courseId=20370
- AWS Data Lake Solution GitHub: https://github.com/awslabs/aws-data-lake-solution
- Analytics Hands-on Labs
- Overview on Data Lakes (6 webinars)

Before we end…

Core Services:

- Aurora – Managed SQL database
- Redshift – Managed SQL Data Warehouse
- Athena and Glue – Reads flat files via SQL queries
- EMR - Managed Hadoop
- Kinesis – data ingestion at scale
- S3 – storing all your flat files
- Lake Formation

And more:

- Quicksight, Database Migration Service, Lambda, Glacier, Direct Connect, Private Link, Kafta

Questions?

aws training and certification

Thank you!

aws training and certification