Becoming a Data Driven Organization

Technical Session

Ian Meyers, Director of Product Management, AWS Analytics
Zach Mitchell, Sr. Big Data Architect, AWS Lake Formation
Customers want more value from their data

Growing Exponentially

From new sources

Increasingly diverse

Used by many people

Analyzed by many applications
Customers moving from traditional data warehouse approach to modern data processing and analytics.
Modern Data Architecture on AWS

- Scalable data lakes
- Purpose-built data services
- Seamless data movement
- Unified governance
- Performant and cost-effective
Focusing on business outcomes

**Customer experience**
- Built a customer engagement service using a Modern Data Architecture to serve over eight million developers working with 190k+ businesses in 100+ countries
  - **Twilio**
  - Real-time insights to give tens of millions of users personalized streaming recommendations
  - **Disney+**
  - Increased the use of self-service analytics platform by over 40% for daily active fans—sharing richer information in near real-time
  - **OneFootball**
  - Personalizes searches for better customer experience and gets fewer returns due to improved sizing recommendations
  - **Zappos**

**Agility and innovation**
- Accelerates zero-carbon transition with automated energy predictions and maximized wind farm energy production
  - **ENGIE**
- Helps drive better insights needed to make key race-time decisions, giving a technological edge over competitors
  - **Toyota Racing Development**
- With Amazon Managed Streaming for Apache Kafka, the company is able to experiment with big changes safely with little risk
  - **New Relic**
- Built a sophisticated infectious disease tracker in four months for retirement community residents and employees
  - **Erickson Living**

**Cost optimization**
- Manages over 150 PB of data at $5 per terabyte of data scanned
  - **FINRA**
- Shifting to AWS saves more than $2 million annually in data storage costs
  - **INVISTA**
- AWS Analytics reduced operational costs by over 30% while freeing software engineers of low-value work
  - **Pinterest**
- Amazon EMR as its core ML platform allows for more accurate ML models 80% faster at an 80% lower cost
  - **Eightfold.ai**

**Performance and scale**
- Moved to a Modern Data Architecture to ingest 70 billion records per day, and now runs Amazon Redshift queries 32% faster
  - **Nasdaq**
- Scalability and cost efficiency during a global pandemic with 20x increase in ventilator production while reducing first-pass inspection failures by 60%
  - **Vyaire Medical**
- Scaled ingestion to six billion documents per day using Amazon OpenSearch Service (successor to Amazon Elasticsearch Service)
  - **Pearson**
- Had the tools to support a 101% increase in language learners
  - **Duolingo**
Scalable data lakes
Amazon S3 is the most popular choice for data lakes

- Unmatched durability, availability, and scalability
- Most object-level controls
- Easiest to use with cost optimization: Intelligent tiering
- Brodest portfolio of analytics tools
- Most ways to get data in
- Best security, compliance, and audit capabilities
- Cold storage and archive capabilities
More data lakes run on AWS than anywhere else
Tens of thousands of data lakes run on AWS across all industries
Purpose-built data services
Purpose-built data services

Optimize performance, cost, and scale for your use cases

- **Amazon Athena**: Interactive query
- **Amazon EMR**: Big data processing
- **Amazon OpenSearch Service** (successor to Amazon Elasticsearch Service): Log and search analytics
- **Amazon Kinesis and Amazon MSK**: Real-time analytics
- **Amazon Redshift**: Data warehousing
No compromises on performance, scale, and cost

**3x better price performance** than other cloud data warehouses
Automated performance tuning and near-linear scaling

**Optimized runtimes that provide the best price-performance**
**1.7x faster** than standard Apache Spark; **2.6x faster** than standard Presto

**UltraWarm storage tier reduces costs by 90%**
Store **6x more log data** without increasing costs

Amazon S3 Select retrieves a subset of data leading to queries that run **up to 400% faster**
Amazon S3 intelligent tiering **saves up to 70% on storage costs** for data lakes

**With Graviton2 instance, customers save 25.7% for typical workloads**
Amazon Redshift

Analyze all your data with the fastest and most widely used cloud data warehouse

- Analyze all your data
  Deepest integration with your data lake

- Performance at any scale
  Up to 3x better price performance than other cloud DW

- Lower your costs
  At least 50% less expensive than other cloud DW
Amazon Redshift innovates to meet your needs

**Analyze all your data**
- Modern Data with AWS integration
- Amazon Redshift Spectrum + Lake Formation
- Data lake export
- Federated query

**Performance & scale**
- Fast and self-tuning
- Concurrency scaling
- RA3 nodes & managed storage
- AQUA

**Low cost & best value**
- Predictable costs
- On-demand and RIs
- Cross-AZ cluster recovery
- Pause and resume

**NEW!**
- Data sharing
- Materialized views
- Automated perf. tuning

**NEW!**
- Amazon Redshift ML
- Materialized views
- Automatic workload manager

© 2021, Amazon Web Services, Inc. or its Affiliates.
Amazon Redshift ML

Create, train, and deploy machine learning (ML) models using familiar SQL commands

- Simple, optimized, and secure integration between Redshift and Amazon SageMaker
- Train and deploy an ML model using a SQL command in your data warehouse
- Embed predictions like fraud detection, risk scoring, and churn in queries and reports
Amazon EMR

Easily run Spark, Hadoop, Hive, Presto, HBase, and other big data frameworks

- Automate provisioning, configuring, and tuning
  Easy setup, management, and monitoring, with latest open-source framework updates within 30 days

- Run workloads faster and more cost-effectively
  1.7x faster than standard Apache Spark 3.0 at 40% of the cost, and 2.6x faster than open-source Presto 0.238 at 80% of the cost

- Automatically scale up and down
  Manage cluster size based on utilization to reduce costs

- Simple and predictable pricing
  Per-second pricing, and save 50%–80% with Amazon EC2 Spot and Reserved Instances
Amazon EMR differentiated performance

1.7x faster performance than standard Apache Spark 3.0 at 40% of the cost

Up to 2.6x faster performance than open-source Presto 0.238 at 80% of the cost

11.5% average performance improvement with Graviton2

25.7% average cost reduction with Graviton2
Amazon OpenSearch Service
(successor to Amazon Elasticsearch Service)
Search, visualize, and analyze up to petabytes of text and unstructured data

- Fully managed
  Operate OpenSearch with the leading contributor of the community-driven, open source software.

- Easily accessible
  Quickly search and analyze your unstructured and semi-structured data to easily find what you need.

- Cost-effective
  Eliminate operational overhead and reduce cost with automated provisioning, software installation, patching, storage tiering, and more.
The OpenSearch Project
An Apache 2.0-licensed search and analytics suite

100% open source
Providing you the freedoms, so you can freely view, use, change, and distribute the code

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

Community-driven
Providing individuals and organizations the freedom to easily contribute changes
Amazon Athena
Query data in S3 using SQL

- **Serverless**
  Quickly query S3 data without managing infrastructure, and pay only for the queries you run

- **Open and standard**
  Use ANSI SQL for querying with support for Parquet, CSV, JSON, Avro and other standard data formats

- **Fast interactive performance**
  Parallel execution to deliver most results within seconds, with no cluster management required

- **Cost effective**
  Pay only for queries run and save 30–90% by compressing, partitioning, and converting your data into columnar formats

© 2021, Amazon Web Services, Inc. or its Affiliates.
Seamless Data Movement
Seamless data movement
Move your data, at scale, to where you need it the most

- Extract, transform, load
- Visual data preparation
- Data replication
- Data warehouse to/from data lake
- Federated query
AWS Glue
Simple, scalable, and serverless data integration

Connect to more sources
Easily ingest data from hundreds of popular data sources

Simplify workflow orchestration
Easily run and manage thousands of data integration jobs

No servers to manage
Pay only for the resources your jobs consume

Simplify development
Visually develop and manage data integration jobs
AWS Glue Elastic Views
Easily combine and replicate data across multiple data stores

Create materialized views across a wide variety of databases and data stores using familiar SQL

Continually monitors source databases for changes and updates targets within seconds

Serverless and automatically scales capacity up and down to accommodate your workloads

Handles the heavy lifting of copying and combining data without requiring custom code
Federated query in Amazon Redshift and Athena

Unified analytics across databases, data warehouse, and data lake

Integrate operational database with data warehouse and Amazon S3 data lake

Analytics on operational data without data movement and ETL delays

Flexible and easy way to ingest data, avoiding complex ETL pipelines

*Other sources available in Amazon Athena: Amazon ElastiCache for Redis, Amazon DocumentDB, Amazon DynamoDB, HBase in Amazon EMR

© 2021, Amazon Web Services, Inc. or its Affiliates.
Unified governance
AWS Lake Formation

Build a secure data lake in days

Build data lakes quickly
Move, store, and catalog your data faster; simplify data management with governed storage

Simplify security management
Centrally define and enforce security, governance, and auditing policies

Provide self-service access to data
Share datasets easily and securely within your organization and with partners

© 2021, Amazon Web Services, Inc. or its Affiliates.
AWS Lake Formation
Common Data Sharing Topologies

Centralized: Hub & Spoke

Hybrid: Distributed Storage

Data Mesh

© 2021, Amazon Web Services, Inc. or its Affiliates.
What is Data Mesh?

A decentralized, domain-oriented data system to drive governed sharing across Lake House Architectures

- Amazon Redshift
- Amazon EMR
- Amazon Athena

Consumer 1

- BI Reports

Consumer 2

- Amazon S3 data lake storage
- Amazon EMR
- Notebook

…. Consumer N

Amazon Redshift
Amazon EMR
SageMaker

Federated Governance

- Unified Policy Management
- Centralized Governance & Audit
- Federated Access Control
- Organization wide Sharing

Amazon S3

LOB 1

Amazon S3

LOB 2

…. Amazon S3

LOB N

© 2021, Amazon Web Services, Inc. or its Affiliates.
Why Data Mesh?

• Encourage data-driven **agility**

• Support domain-local **governance** through lightweight centralized policy

• Isolate data resources with clear **accountability**

• Expose data as **products** which are owned and can be shared
What Is A Data Domain?

Any environment which can be reached over a network and which produces or consumes data. Usually one or more AWS Accounts.

Self Service functionality is provided within each Data Domain to connect the Domain’s data to the Mesh.

Contains a local Lake Formation catalog used to manage metadata. Uses data processing technologies both provided centrally and those which unique to the domain’s requirements.

AWS Lake House Architecture is a best-practice approach to build a data domain.
Central AWS Account where data products are registered

Data Products = Lake Formation Databases, Tables, Columns, and Rows

Create centrally managed Access Control Tags and Tag Access Policies

Support centralized auditing of sharing

Stores data permissions which implement sharing with a Consumer. Permissions can be direct or based on Tags.
What is the Data Mesh?

Applies security & governance policies to Producer & Consumer Accounts and the Data Products they publish, which may include:

- Use of a consistent Identity Model (SSO)
- Use of Lake Formation based Security
- Regional constraints
- AWS service restrictions through AWS Organizations & Service Control Policies
- AWS Service Catalog for reusable patterns
ENGIE builds the Common Data Hub on AWS, accelerates zero-carbon transition

Challenge
ENGIE’s decentralized global customer base had accumulated lots of data, and it required a smarter, unique approach and solution to align its initiatives and to efficiently provide data across its global business units.

Solution
ENGIE built its Common Data Hub data lake on AWS, enabling the company’s business units to collect and analyze data to support a data-driven strategy and to lead the zero-carbon transition.

Result
• Collected 95 TB of data across 351 projects
• Automated energy predictions
• Maximized wind farm energy production
Customer Example - Why JPMorgan Chase built a “data mesh” cloud architecture: 
**Drive significant value to enhance their enterprise data platform**

**Modernize Data Platform**
- Move Beyond Monolithic Data Lake
- Build Loosely Coupled Arch. For Data
- Aggregate fit-for-purpose Data Products
- Distributed Data Pipelines
- Governance and Compliance

**Three Major Business & Technical Principles:**
- Cost Savings
- Business Value – Business (Domain) Use Cases
- Data Reuse (Data Producers/Consumers)

(Source: https://wikibon.com/breaking-analysis-how-jp-morgan-is-implementing-a-data-mesh-on-the-aws-cloud)

© 2021, Amazon Web Services, Inc. or its Affiliates.
Modern Data architecture on AWS

Scalable data lakes

Purpose-built data services

Seamless data access

Unified governance

Performant and cost-effective

AWS LAKE FORMATION

AWS GLUE

Amazon Athena

Amazon EMR

Amazon OpenSearch Service

Amazon Redshift

Amazon Aurora

Amazon DynamoDB

Amazon SageMaker

Amazon S3
Want to build a data vision and strategy?

- Joint engagements with business and technology stakeholder alignment
- Create an organizational vision for innovation with data to drive business outcomes
- Define the first pilot, learn, and build

**Jumpstart the data flywheel**

Have a strategy and need help executing it?

- Joint engineering engagements between customers and AWS technical resources
- Create tangible deliverables to accelerate strategic databases, analytics, and ML initiatives
- Leave with an architecture, working prototype, path to production, and deeper knowledge of AWS services

**Come with an idea, leave with a solution**
Thank you