AWS re:Invent

ANT206-L

Leadership session: Trends with data lakes and analytics

Andi Gutmans

VP, Analytics & ElastiCache, AWS

AWS re:Invent

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

What to expect

Turn your data into insights



Amazon Redshift
Data warehousing



Amazon EMR
Hadoop + Spark



Amazon Athena
Interactive analytics



Amazon Kinesis
Real-time data analytics



Amazon Elasticsearch Service
Operational Analytics

Modernize your data lake infrastructure



Amazon S3/Glacier



AWS Lake Formation



AWS Glue

The new realities customers are facing



Explosion of data

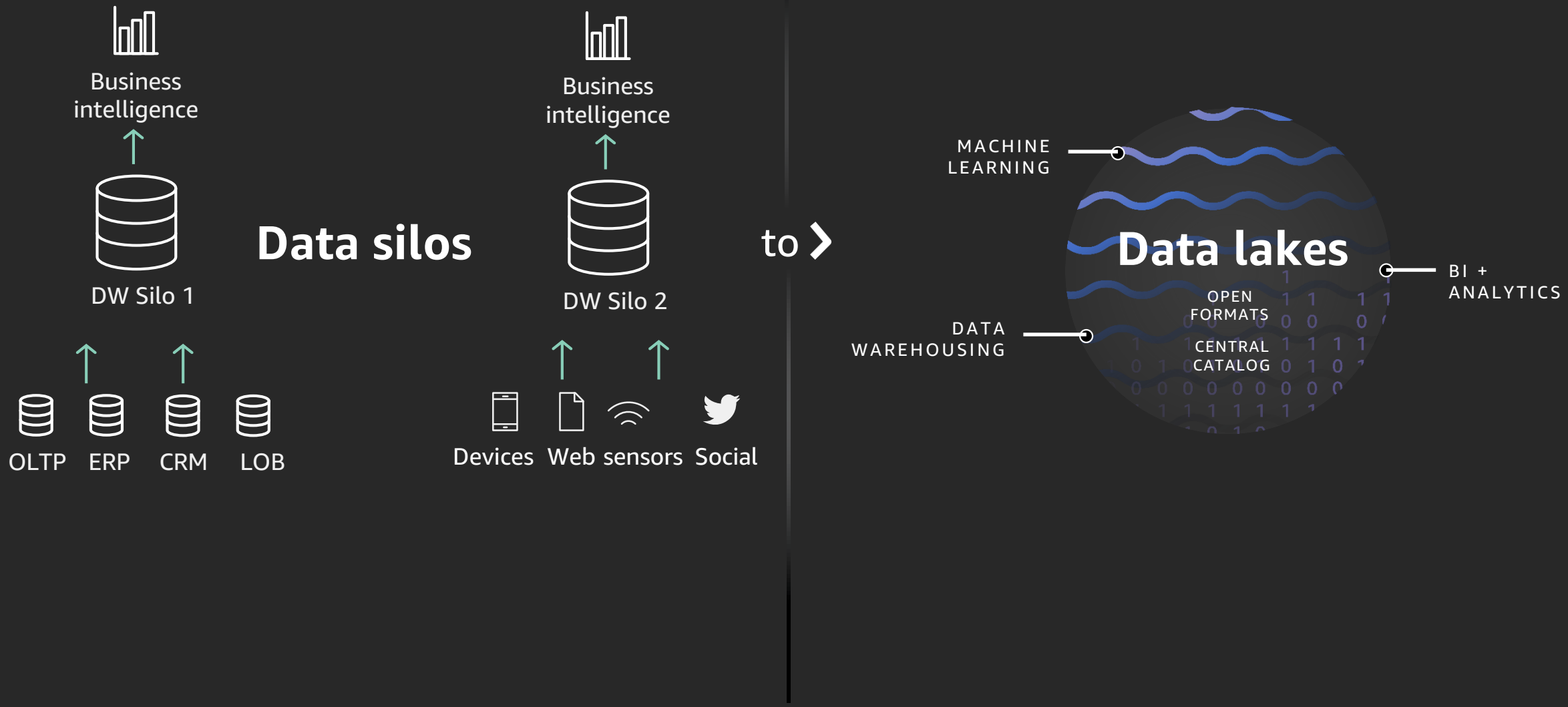


Explosion of personas

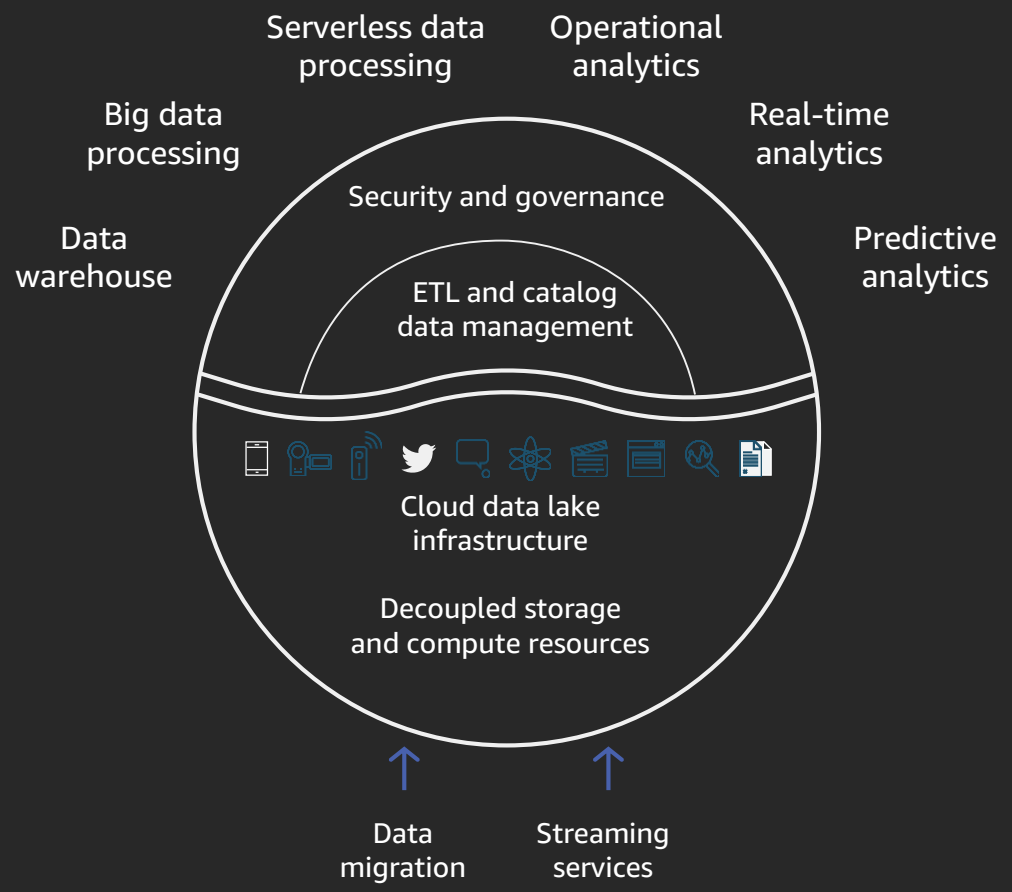


Demand for faster decision-making on real-time data

The convergence of traditional data silos



The future with cloud data lakes



Customers want:

A single data store that is scalable and cost-effective

To use the standards-based data format of their choice

To analyze their data in a variety of ways

Amazon S3 is the most popular choice for cloud data lakes

Unmatched **durability**, **availability**, and **scalability**

Easiest to use with cost optimization: Intelligent Tiering

Most ways to get data in



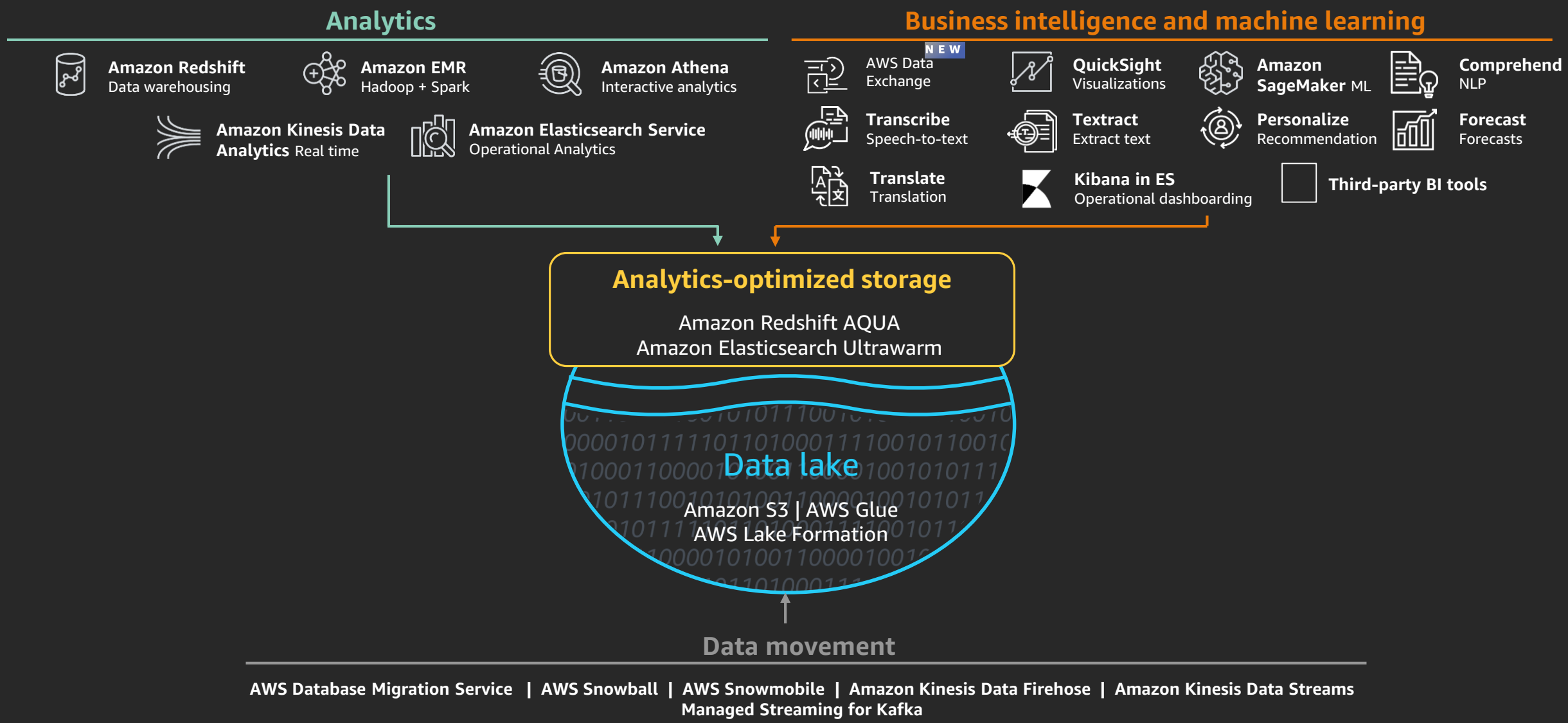
Most **object-level** controls

Broadest portfolio of analytics tools

Best **security**, **compliance**, and **audit** capabilities

Our portfolio

Broadest and deepest portfolio, purpose-built for builders



Designed for the new reality



Cloud-optimized: Architect services ground-up for the cloud and for the explosion of data

Purpose-built: Offer a portfolio of purpose-built services, optimized for your workloads

Fully managed: Help you innovate faster through managed services

Now used by a **very large number of customers** for mission-critical applications

More data lakes and analytics than anywhere else



Build a data lake in days with AWS Lake Formation

Fastest to go from data to insights

Move, store, catalog, and clean your data faster



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

ETL and Data Catalog: AWS Glue

Simple, flexible, and cost-effective ETL

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

Streaming data services: Amazon Kinesis



Amazon Kinesis Data Streams

Capture and store data streams



Amazon Kinesis Data Firehose

Analyze data streams in real time



Amazon Kinesis Data Analytics

Load streaming data into streams, data lakes, and data warehouses



Amazon Managed Streaming for Apache Kafka (Amazon MSK)

Capture and store data streams

Streaming data capabilities

NEW!

Access resources within an Amazon VPC using Amazon Kinesis Data Analytics

- Read and write data from resources within your VPCs like Amazon Elasticsearch Service clusters, RDS databases, and Redshift data warehouses

Amazon MSK releases Open Monitoring with Prometheus

- Consume every Apache Kafka metric with low latency
- Enable time-series logging, alarming, and charting through Prometheus



Run Apache Flink and Apache Kafka together using fully managed services on AWS

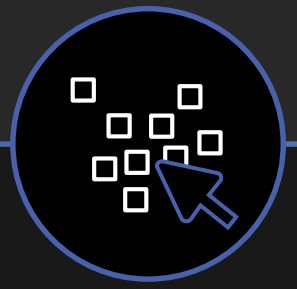
- Use Kinesis Data Analytics to process streaming data stored in Amazon MSK
- Run streaming solutions end-to-end using open source software in fully managed services

Data exchange: AWS Data Exchange (GA)

NEW!

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, Lexis Nexis, and Deloitte

Easily analyze data



Download or copy data to Amazon S3

Combine, analyze, and model with existing data

Analyze data with EMR, Redshift, Athena, and 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

Data warehousing: Amazon Redshift

First and most popular cloud data warehouse

Data lake & AWS integration



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

Query data across various analytics services

Best performance, most scalable



AUQA and RA3: 10x faster than other cloud data warehouses

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, HIPAA

Lowest cost



Cost-optimized workloads by paying compute and storage separately

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

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

Most widely used cloud data warehouse

Tens of thousands of customers use Amazon Redshift



Amazon Redshift has been innovating quickly

Robust result set caching

Large # of tables support ~20000

Copy command support for ORC, Parquet

IAM role chaining

Elastic resize

Groups

Redshift Spectrum: Date formats, scalar json and ION file formats support, region expansion, predicate filtering

Auto analyze

Health and performance monitoring w/Amazon Cloud watch

Automatic table distribution style

Amazon CloudWatch support for WLM queues

Performance enhancements: Hash join, vacuum, window functions, resize ops, aggregations, console, union all, efficient compile code cache

Unload to CSV

Auto WLM

~25 Query Monitoring Rules (QMR) support

AQUA

Concurrency scaling

200+

DC1 migration to DC2

Resiliency of ROLLBACK processing

Manage multi-part query in AWS Management Console

Auto analyze for incremental changes on table

new features in the past 18 months

Spectrum Request Accelerator

Apply new distribution key

Redshift Spectrum: Row group filtering in Parquet and ORC, nested data support, enhanced VPC routing, multiple partitions

Faster Classic resize with optimized data transfer protocol

Performance: Bloom filters in joins, complex queries that create internal table, communication layer

Redshift Spectrum: Concurrency scaling

AWS Lake Formation integration

Auto-vacuum sort, auto-analyze and auto table sort

Auto WLM with query priorities

Snapshot scheduler

Stored procedures

Performance: Join pushdowns to subquery, mixed workloads temporary tables, rank functions, null handling in join, single row insert

Advisor recommendations for distribution keys

AZ64 compression encoding

Console redesign

Spatial processing

Column level access control with AWS lake formation

RA3

Performance of inter-Region snapshot transfers

Federated query

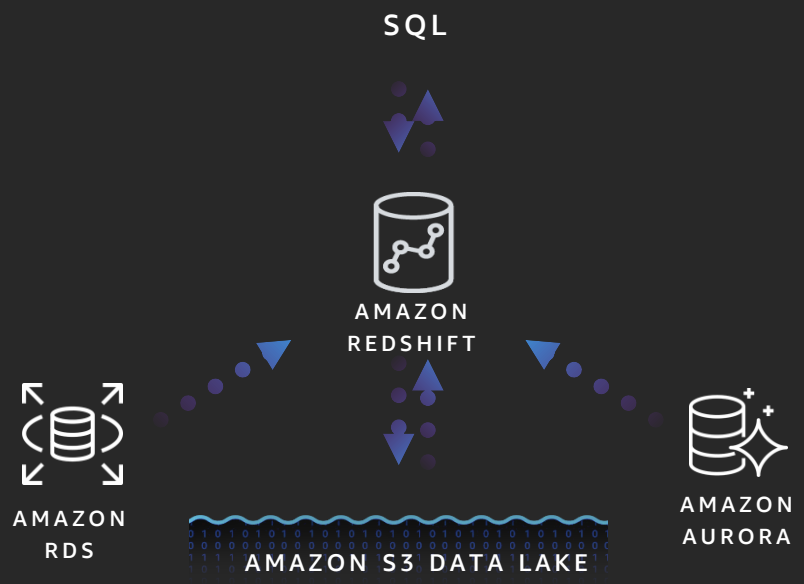
Materialized views

Manual pause and resume

Amazon Redshift Federated Query (Preview)

NEW!

Analyze data across data warehouse, data lakes, and operational databases



Query across multiple systems from Amazon Redshift

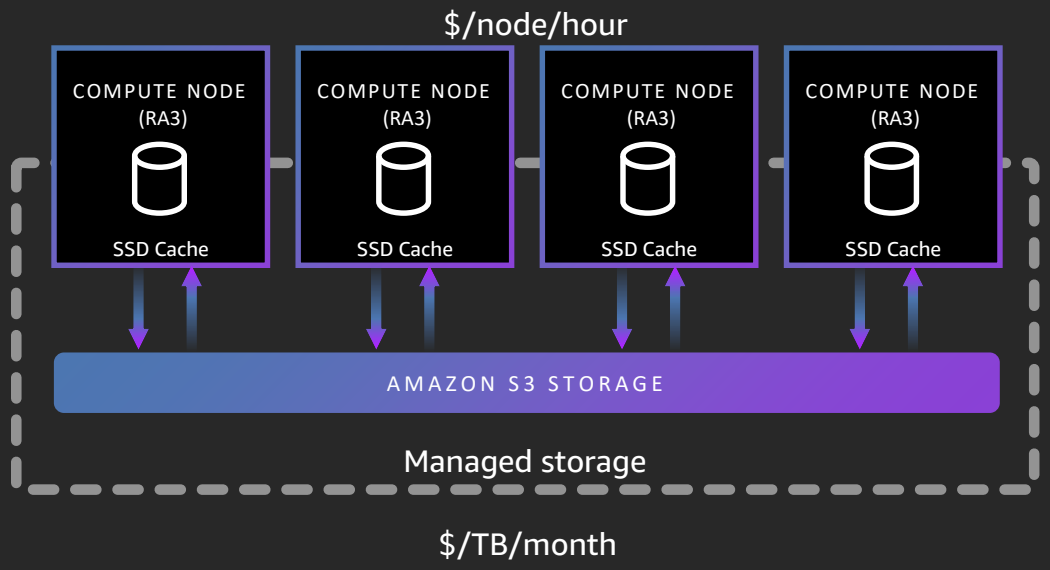
Combine data warehouse and transactional data

Compatible with Amazon RDS and Amazon Aurora (PostgreSQL)

NEW!

Amazon Redshift on RA3 instances (GA)

Optimize your data warehouse by paying for compute and storage separately



Delivers 3x the performance of existing cloud DWs

DS2 customers can migrate and get 2x performance and 2x storage for the same cost

Automatically scales your DW storage capacity

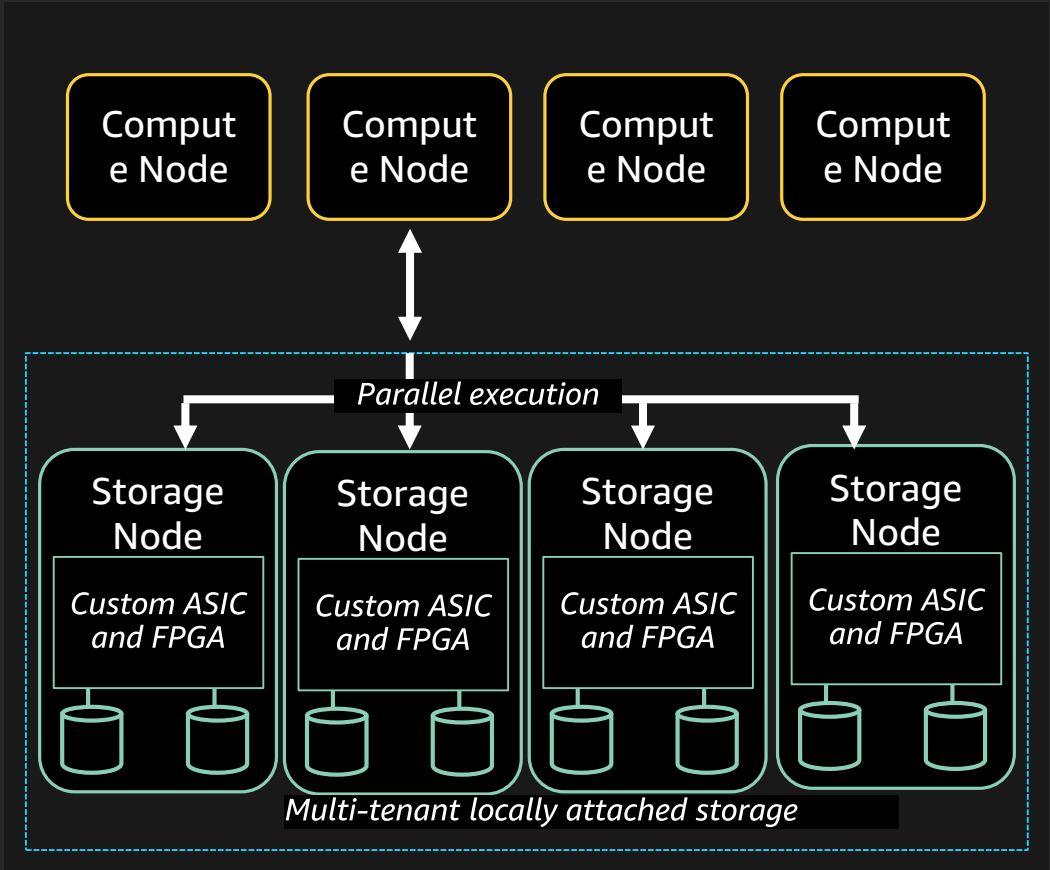
Supports workloads up to 8 PB (compressed)

AQUA – Advanced Query Accelerator

NEW!

COMING IN 2020

Redshift runs **10x faster** than any other cloud data warehouse without increasing cost



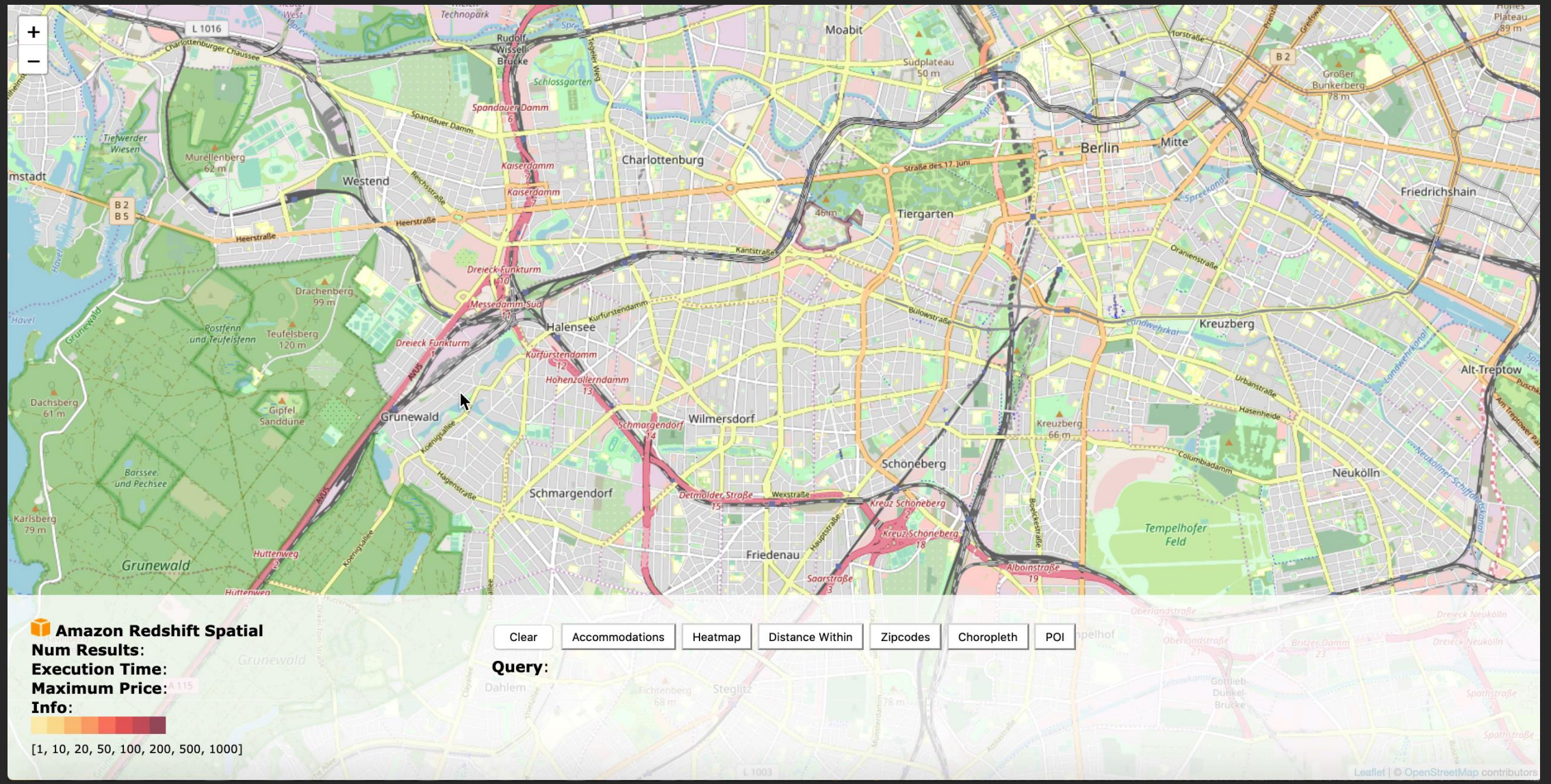
AQUA brings compute to the storage layer so data doesn't have to move back and forth

High-speed cache on top of S3 scales out to process data in parallel across many nodes

AWS custom-designed analytics processors accelerate data compression, encryption, and data processing

100% compatible with the current version of Redshift

Redshift spatial demo



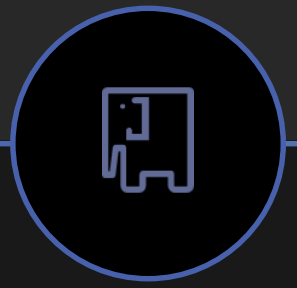
Amazon EMR

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

Lowest cost, fully integrated with auto scaling, Spot, and Amazon S3

Enterprise-grade security, latest versions of OSS frameworks

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 Amazon S3 securely with high performance using the EMRFS connector

Easy



Fully managed no cluster setup, node provisioning, cluster tuning

Apache Hudi for Amazon EMR (GA)

Record-level insert, update, and delete on Amazon S3

NEW!



Apache Hudi is open source, and uses open data formats, enabling data lakes to:

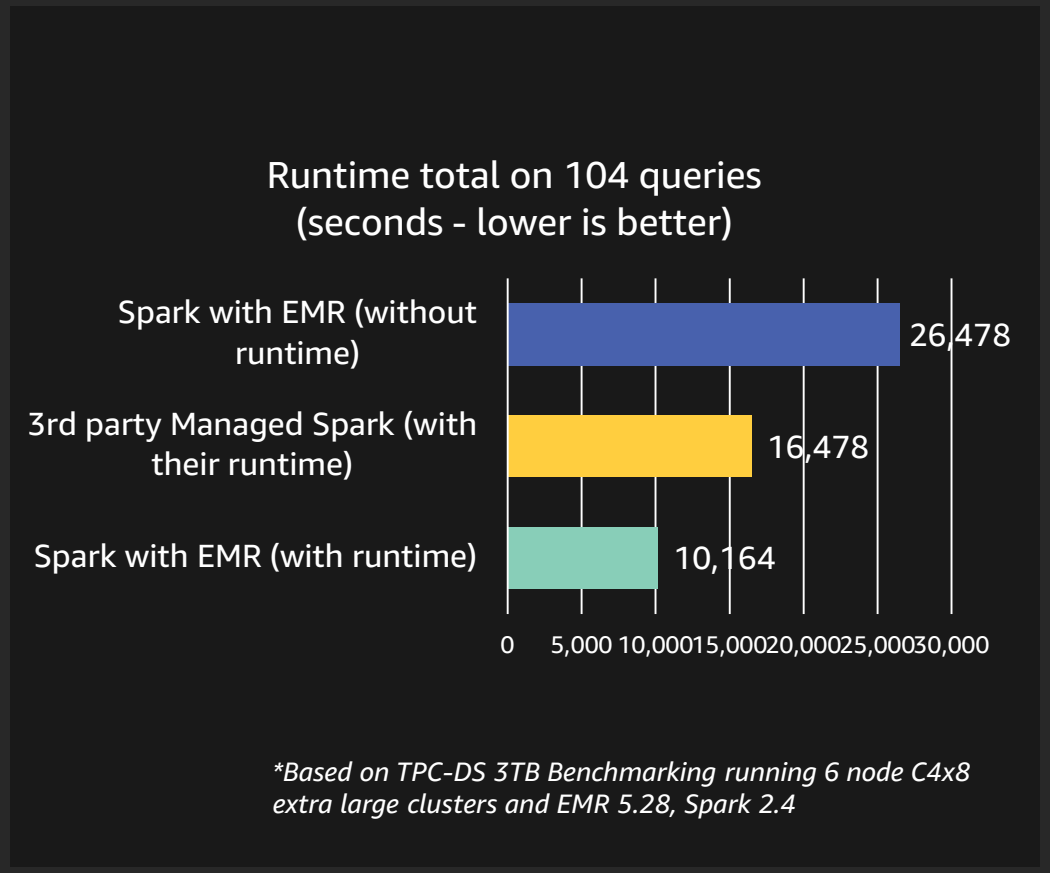
- Comply with data privacy laws
- Consume real-time streams and change data capture
- Reinstate late-arriving data
- Track change history and rollback

Includes support for Spark, Hive, and Presto

Performance improvements in Spark for Amazon EMR

Performance-optimized runtime for Apache Spark, 2.6x faster performance at 1/10th the cost

NEW!



**Based on TPC-DS 3TB Benchmarking running 6 node C4x8 extra large clusters and EMR 5.28, Spark 2.4*

Runtime optimized for Apache Spark performance

Best performance

- **2.6x faster** than Spark with Amazon EMR without runtime
- **1.6x faster** than third-party Managed Spark (with their runtime)

Lowest price

- **1/10th** the cost of third-party Managed Spark (with their runtime)

100% compliant with Apache Spark APIs

Amazon Elasticsearch Service (Amazon ES)

Fully managed, scalable, secure

Open source Amazon ES APIs, Kibana and Logstash

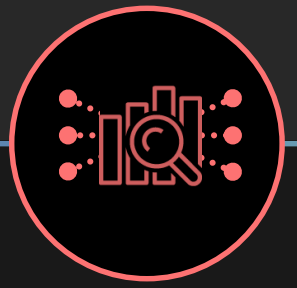


Open-source Amazon ES APIs

Managed Kibana

Integration with Logstash

Fully managed



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

Scalable, secure, and compliant



Scale clusters up/down with a single API call or a few clicks

Secured network isolation with VPC, encrypt data at rest and in transit

Compliant: HIPAA, PCI DSS, and ISO

Pay only for what you use



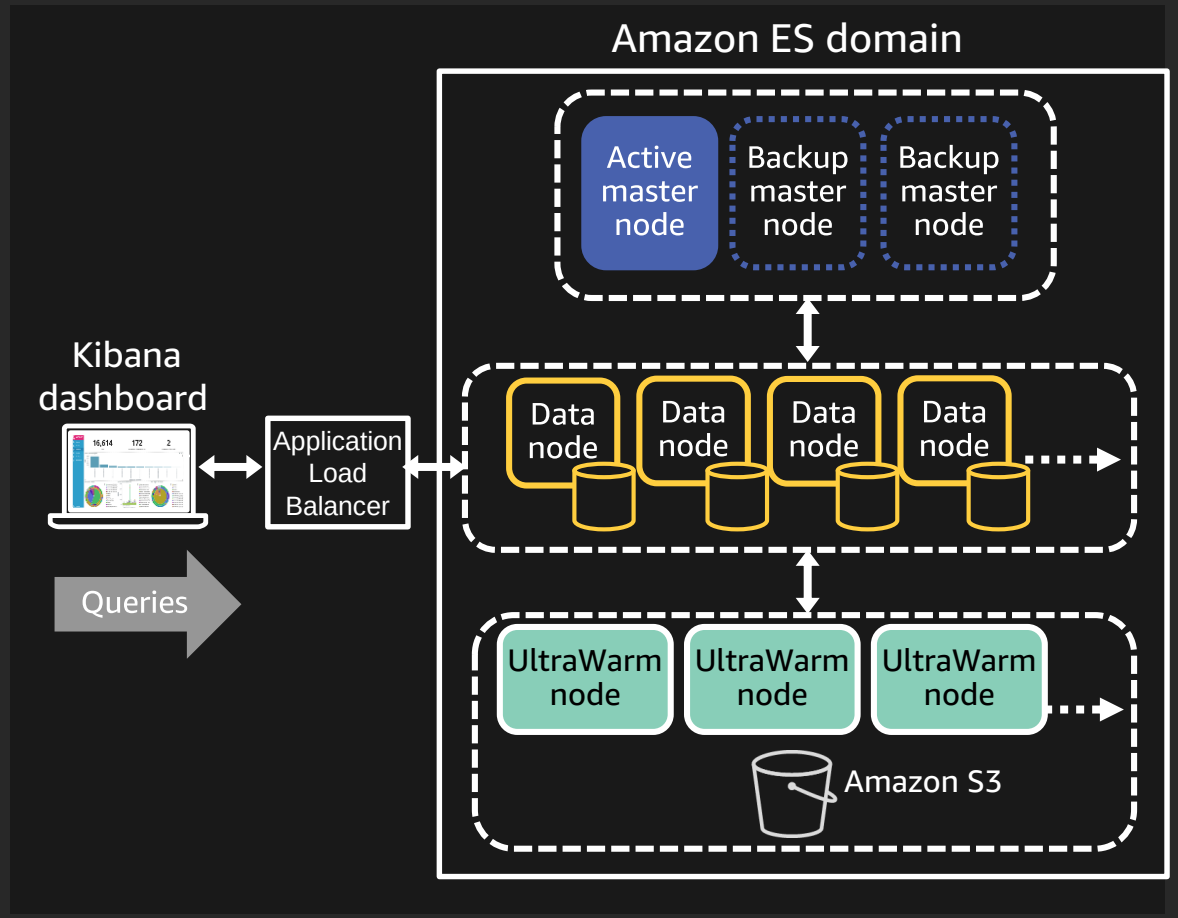
No upfront fee or usage requirement

Critical features built-in: encryption, VPC support, 24x7 monitoring

UltraWarm for Amazon ES (Preview)

A new warm storage tier for Amazon ES

NEW!



90% lower cost

Scale up to 3 PB per domain

Analyze years of operational data

Extends Amazon ES

Amazon Athena

Run SQL queries on data in Amazon S3

No infrastructure to manage

Pay per query

Query instantly



Zero setup cost

Point to Amazon S3 and start querying

Pay per query



Pay only for queries run

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

Open



ANSI SQL

JDBC/ODBC drivers

Multiple formats, compression types, complex joins & data types

Easy



Serverless: Zero infrastructure, zero administration

Integrated with QuickSight

Federated Query for Amazon Athena (Preview)

Run SQL queries on data spanning multiple data stores

NEW!



Run SQL queries on relational, non-relational, object, or custom data sources; in the cloud or on-premises

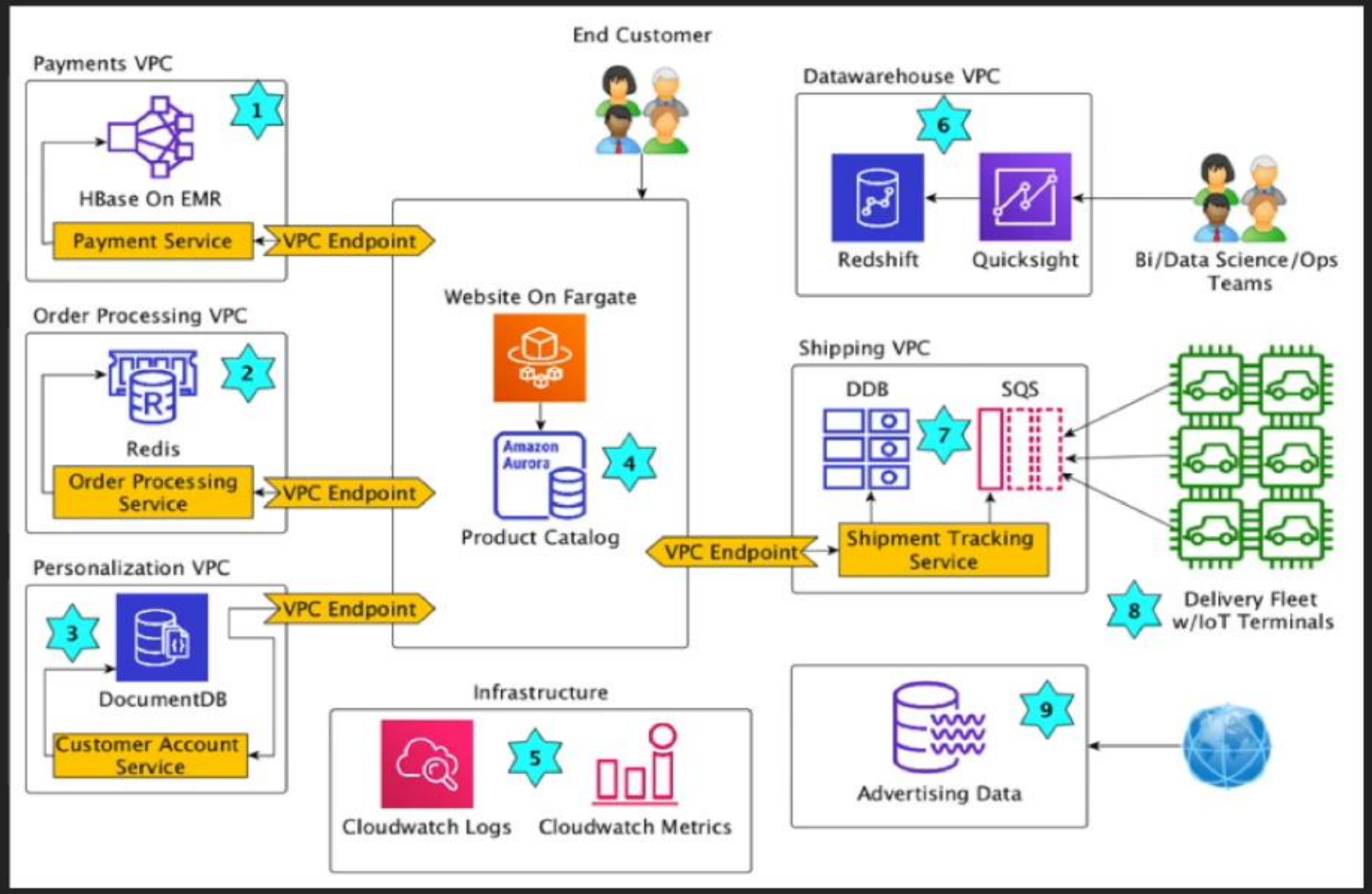
Open Source Connectors for common data sources

Build connectors to custom data sources

Run connectors in AWS Lambda: no servers to manage

Demo: Federated Query for Amazon Athena

Example Architecture



Amazon QuickSight

First BI service with pay-per-session pricing and ML insights



Serverless, cloud-powered BI service (no servers to manage)

Scale from 10s of users to 100s of thousands of users

Pay only for what you use

- Readers: $0.30/30 min session with a $5/user/month max
- Authors: $18/month/Author

Integrates with Amazon S3, Amazon Athena, Amazon Redshift, Amazon RDS, Amazon Aurora, and Amazon EMR

QuickSight used by customers at the largest scale



Capital One has chosen QuickSight to distribute interactive dashboards and reports at scale to tens of thousands of users in their organization.



Best Western is deploying QuickSight across 40,000 users across all their hotel franchises, replacing their legacy reporting solution, which is slow, expensive, hard to scale.



Amazon standardized on QuickSight company-wide, across thousands of teams including AWS, HR, Amazon Business, Finance, Alexa and across hundreds of thousands of employees—enabling fast, easy access to data.

APIs and customizations in QuickSight

NEW!



Deploy and manage dashboards/data via APIs

Match your application UI with QuickSight Themes

Embed dashboards in apps without servers

- Fast, consistent performance
- Pay-per-session

Automatically scale to 10s of 1000s of users

- No server management
- No scripting

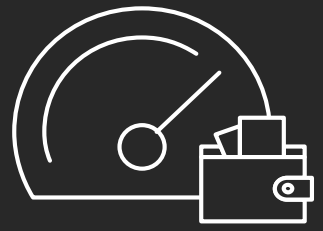
Why choose AWS to turn data to insights

Fastest way to get answers from all your data to all your users



Easiest to build data lakes at scale

- AWS Lake Formation
- Redshift Data Lake Export
- Redshift Federated Query
- Query Federation for Amazon Athena
- Data Streaming for AWS Glue



Best performance at lowest cost

- AQUA for Redshift
- RA3 for Redshift
- Redshift Materialized Views
- UltraWarm Storage Tier for Amazon ES
- Performance improvements for Spark in Amazon EMR



Most comprehensive and open

- Amazon AWS Data Exchange
- Amazon EMR on AWS Outposts
- Record-level insert/updates for Amazon EMR
- ML in Amazon Athena
- ML in Amazon QuickSight



Most secure

- Amazon Westeros
- Amazon Macie
- AWS Lake Formation

Thank you!

Andi Gutmans

gutmans@amazon.com

AWS re:Invent

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws



Please complete the session survey in the mobile app.

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.