aws

INNOVATE

ONLINE CONFERENCE

AWS INNOVATE
ONLINE CONFERENCE

Building next-generation data lakes and analytics on AWS

Ninad Phatak
Big Data Solutions Architect, AISPL

Agenda

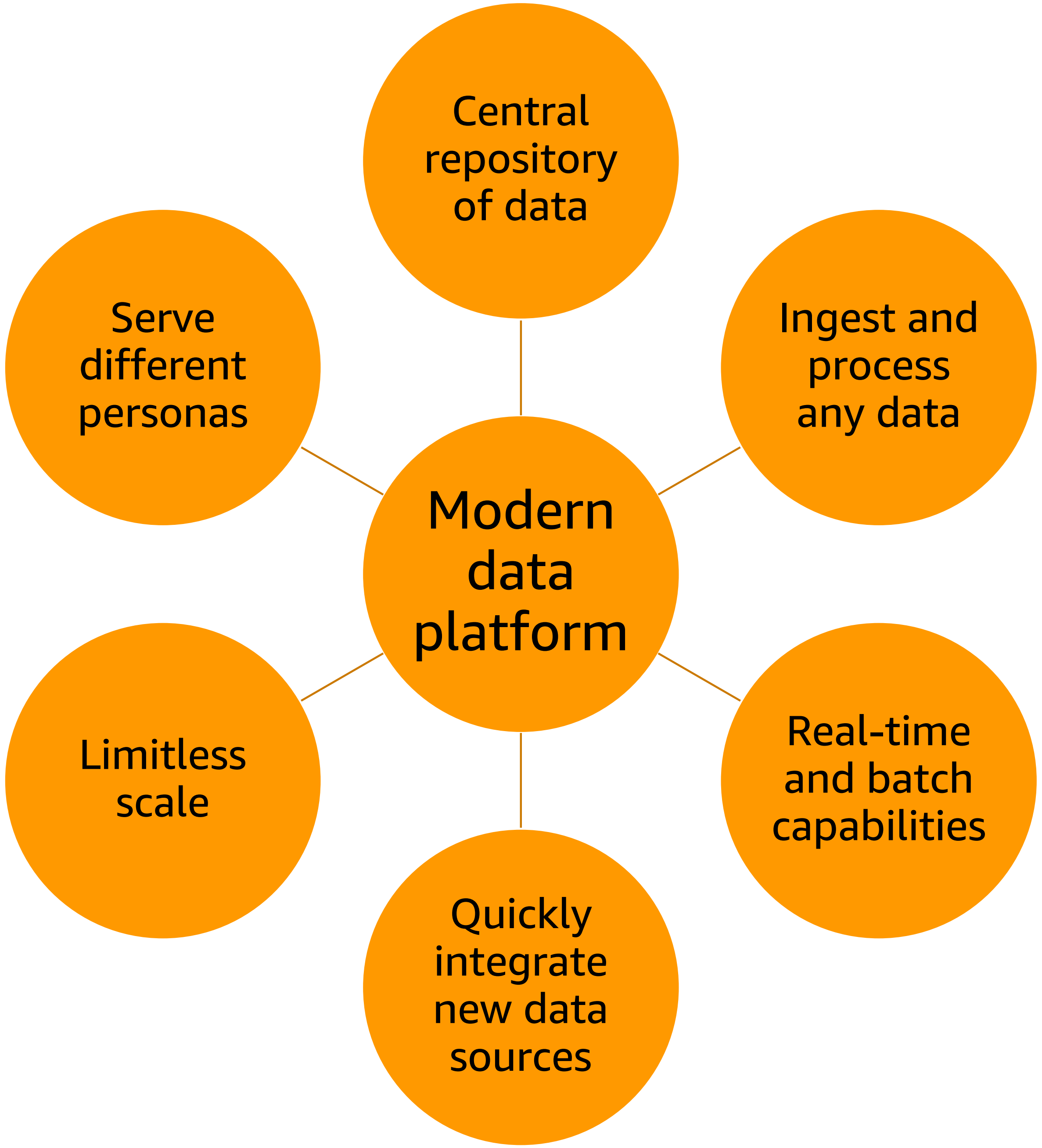
Characteristics of a modern data platform

Components of a data lake on AWS

- Core data lake
- Data ingestion
- Data processing
- Consumption of data and insights

AWS lake formation

A modern data platform/data lake



Central repository of data

Ingest and process any data

Real-time and batch capabilities

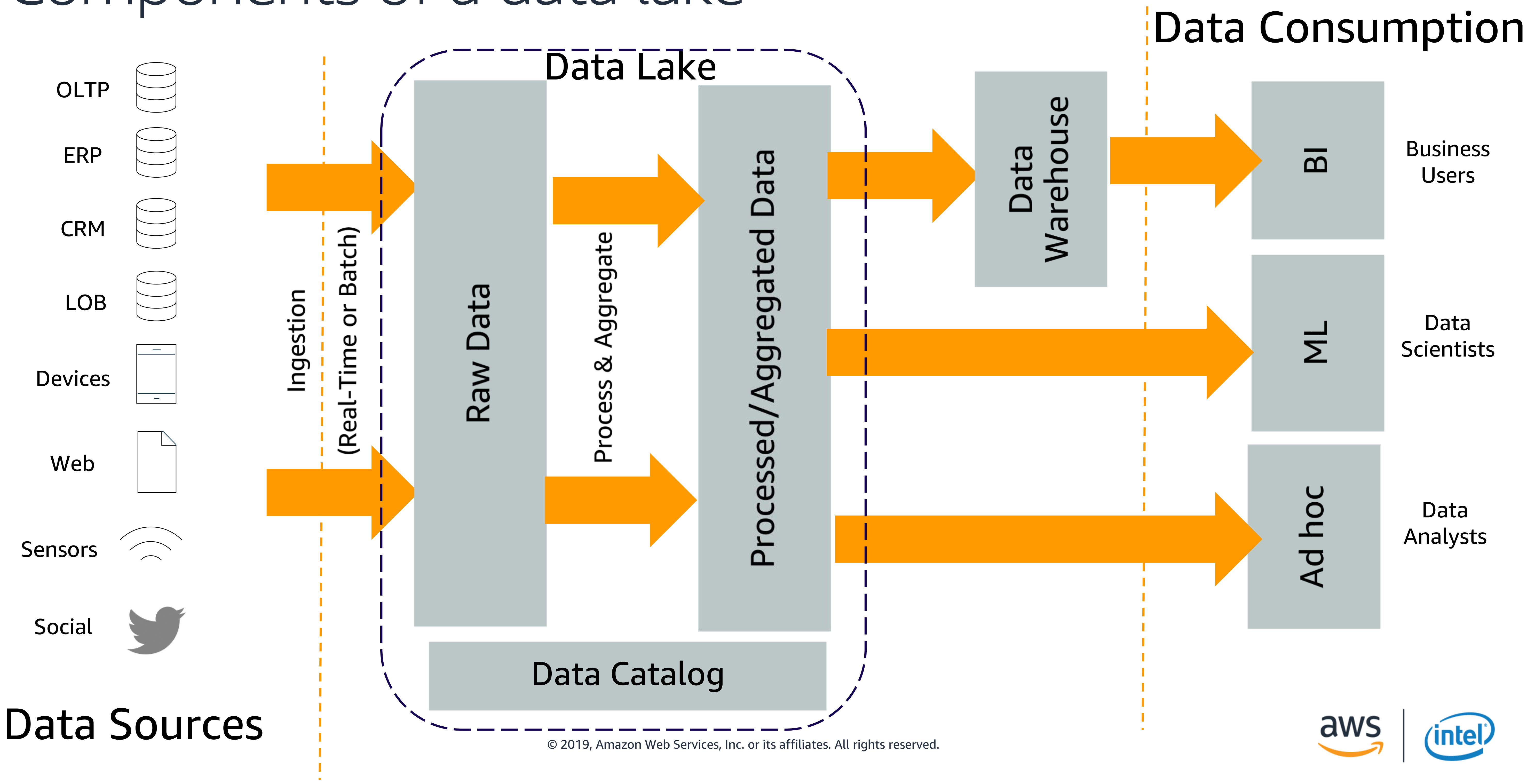
Quickly integrate new data sources

Limitless scale

Serve different personas

Modern data platform

Components of a data lake



The core data lake

Amazon S3: Object storage

Secure, highly scalable, durable object storage with millisecond latency for data access

Store any type of data — web sites, mobile apps, corporate applications, and IoT sensors

Durability, availability, and scalability

Built for eleven 9s of durability; data distributed across 3 physical facilities in an AWS region; automatically replicated to any other AWS region

Security and compliance

Three different forms of encryption; encrypts data in transit when replicating across regions; log and monitor with AWS CloudTrail, use ML to discover and protect sensitive data with Amazon Macie

Query in place

Run analytics and ML on data lake without data movement; Amazon S3 Select can retrieve subset of data, improving analytics performance by 400%

Flexible management

Classify, report, and visualize data usage trends; objects can be tagged to see storage consumption, cost, and security; build lifecycle policies to automate tiering and retention

Amazon S3 Glacier: Backup and archive

Secure, durable, and extremely low-cost storage for data archiving and long-term backup

Store data at $0.004/GB/month

Durability, availability, and scalability

Built for eleven 9s of durability; data distributed across 3 physical facilities in an AWS region; automatically replicated to any other AWS region

Retrieves data in minutes

Three retrieval options to fit your use case; expedited retrievals with Amazon S3 Glacier Select can return data in minutes

Secure

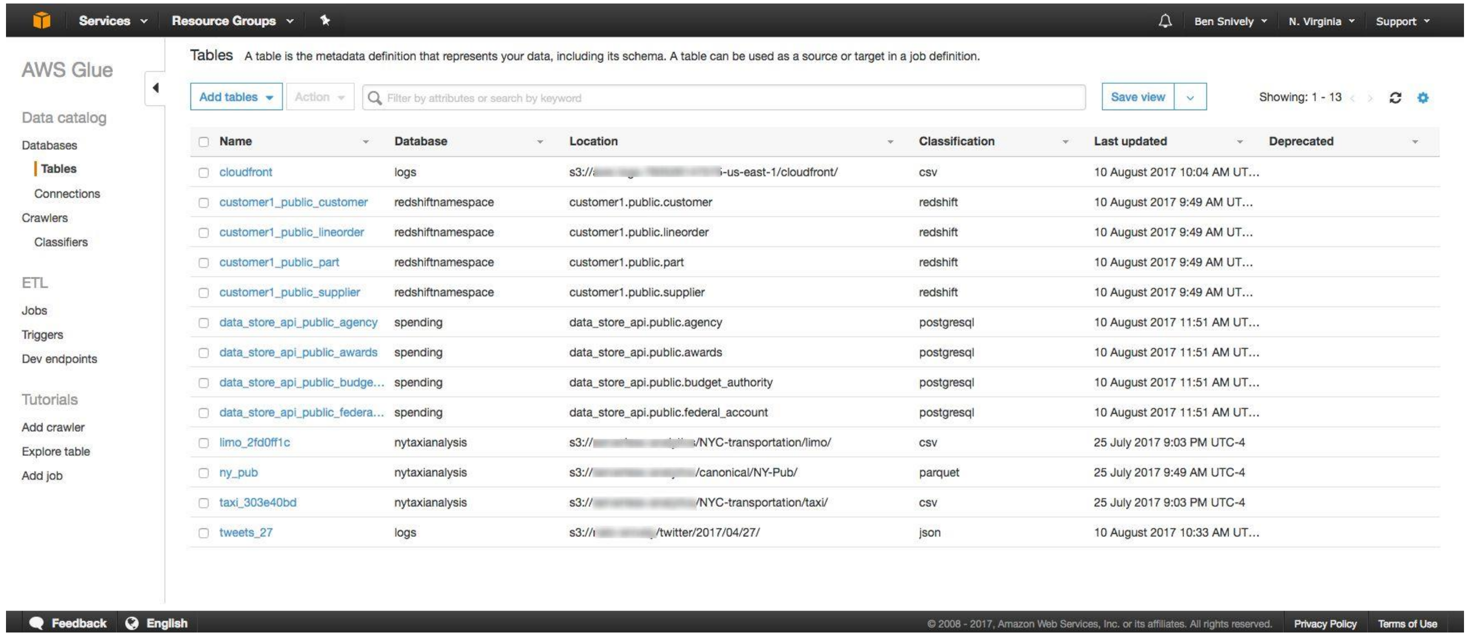
Log and monitor with AWS CloudTrail; Amazon S3 Glacier Vault Lock enables WORM storage capabilities, helping satisfy compliance requirements

Inexpensive

Lowest-cost AWS object storage class, allowing you to archive large amounts of data at a very low cost

AWS Glue Data Catalog

Bring in metadata from a variety of data sources into a single categorized list that is searchable



Tables A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

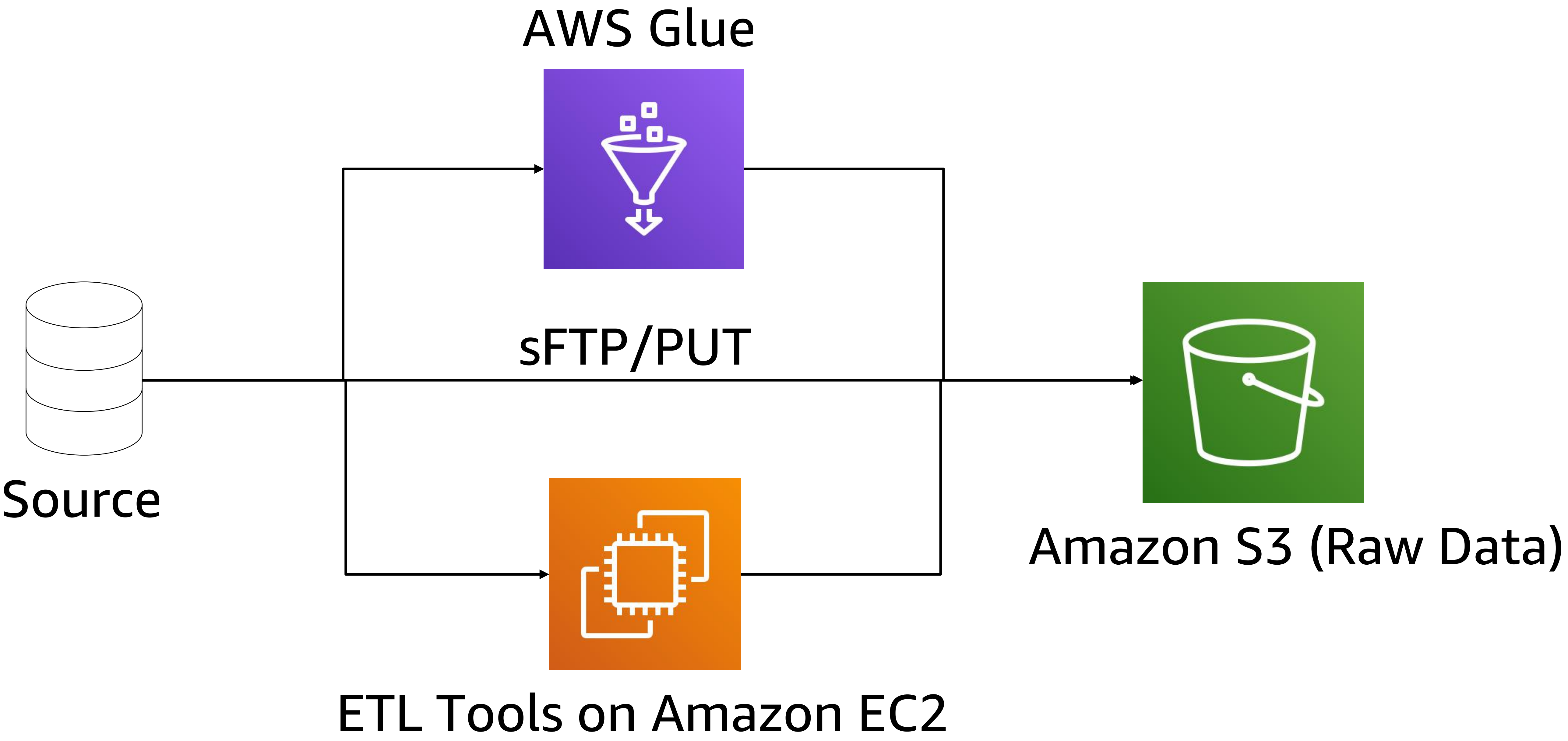
□	Name	Database	Location	Classification	Last updated	Deprecated
□	cloudfront	logs	s3://[redacted]-us-east-1/cloudfront/	csv	10 August 2017 10:04 AM UT...	
□	customer1_public_customer	redshiftnamespace	customer1.public.customer	redshift	10 August 2017 9:49 AM UT...	
□	customer1_public_lineorder	redshiftnamespace	customer1.public.lineorder	redshift	10 August 2017 9:49 AM UT...	
□	customer1_public_part	redshiftnamespace	customer1.public.part	redshift	10 August 2017 9:49 AM UT...	
□	customer1_public_supplier	redshiftnamespace	customer1.public.supplier	redshift	10 August 2017 9:49 AM UT...	
□	data_store_api_public_agency	spending	data_store_api.public.agency	postgresql	10 August 2017 11:51 AM UT...	
□	data_store_api_public_awards	spending	data_store_api.public.awards	postgresql	10 August 2017 11:51 AM UT...	
□	data_store_api_public_budge...	spending	data_store_api.public.budget_authority	postgresql	10 August 2017 11:51 AM UT...	
□	data_store_api_public_federa...	spending	data_store_api.public.federal_account	postgresql	10 August 2017 11:51 AM UT...	
□	limo_2fd0ff1c	nytaxianalysis	s3://[redacted]/NYC-transportation/limo/	csv	25 July 2017 9:03 PM UTC-4	
□	ny_pub	nytaxianalysis	s3://[redacted]/canonical/NY-Pub/	parquet	25 July 2017 9:49 AM UTC-4	
□	taxi_303e40bd	nytaxianalysis	s3://[redacted]/NYC-transportation/taxi/	csv	25 July 2017 9:03 PM UTC-4	
□	tweets_27	logs	s3://[redacted]/twitter/2017/04/27/	json	10 August 2017 10:33 AM UT...	

- Hive metastore-compatible metadata repository
- Used across multiple services
- Automatically discover data using crawlers
- Track data evolution using schema versioning

Data ingestion

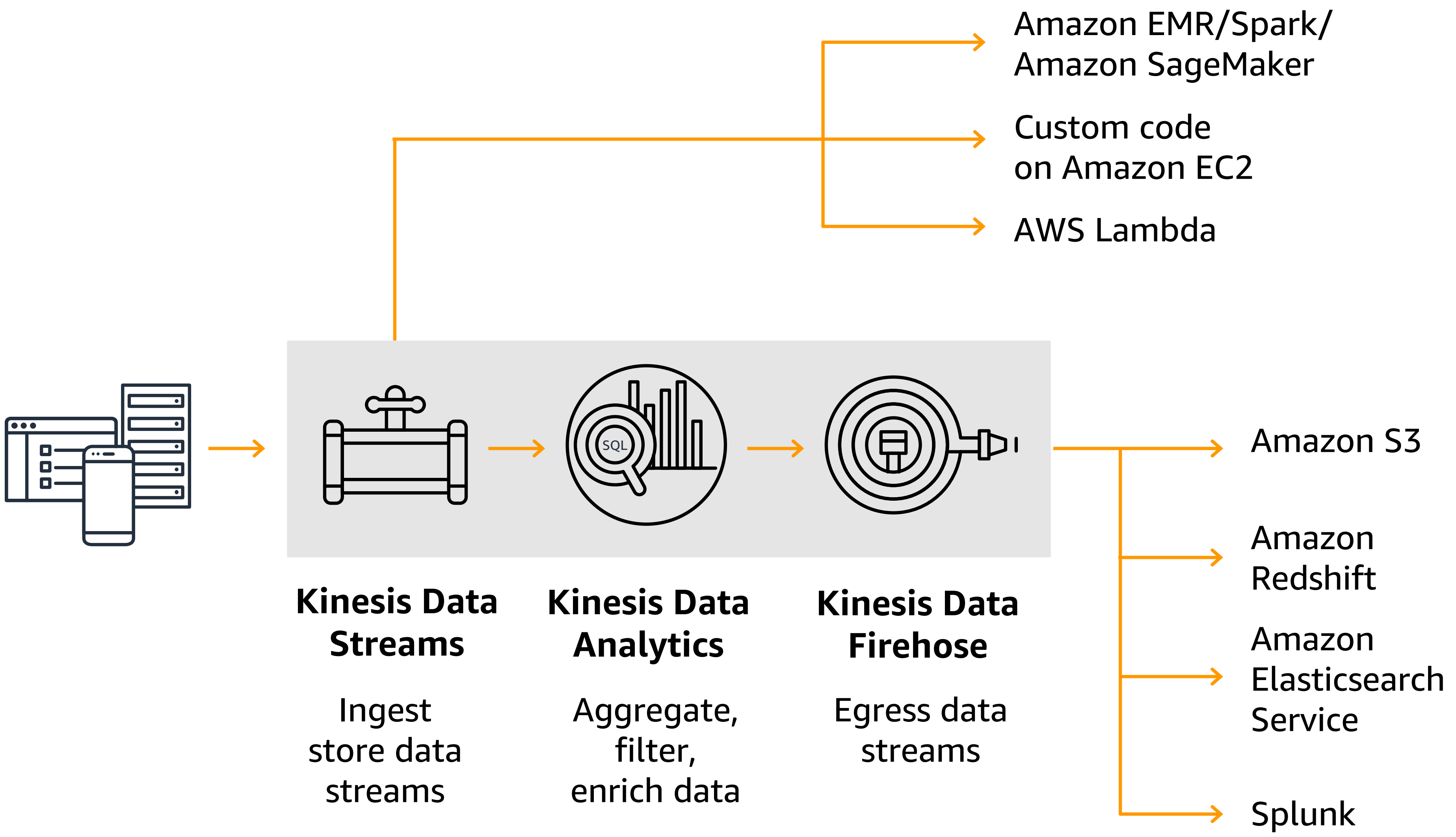
aws INNOVATE
ONLINE CONFERENCE

Batch data ingestion



Real-time data ingestion using Amazon Kinesis services

Collect, process, and analyze data streams in real time



Amazon EMR/Spark/
Amazon SageMaker

Custom code
on Amazon EC2

AWS Lambda

Kinesis Data Streams

Ingest
store data
streams

Kinesis Data Analytics

Aggregate,
filter,
enrich data

Kinesis Data Firehose

Egress data
streams

Amazon S3

Amazon
Redshift

Amazon
Elasticsearch
Service

Splunk

- Real-time
- Fully managed
- Scalable
- Secure
- Cost-effective

Amazon Managed Streaming for Kafka (Amazon MSK)

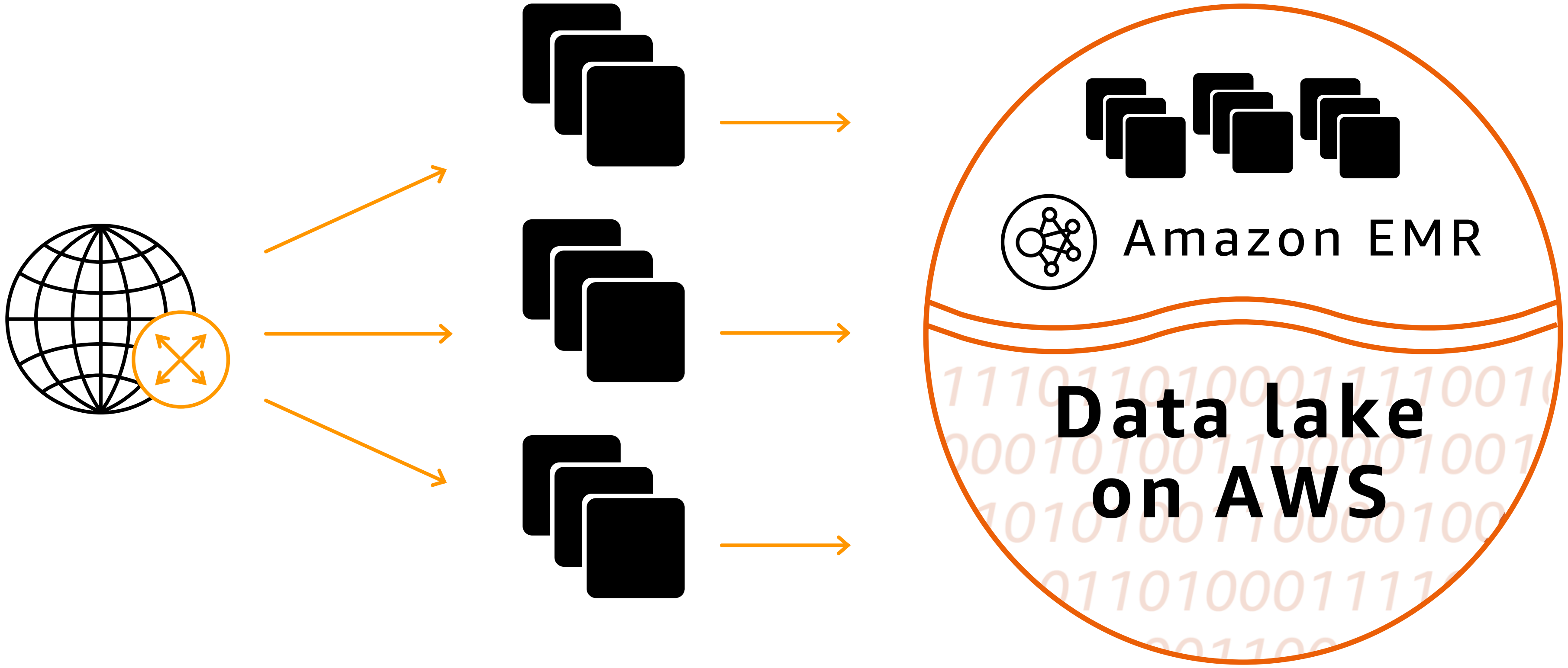
- Fully compatible with Apache Kafka v1.1.1 and 2.1.0
- Clusters are set up automatically
- Provision Apache Kafka brokers and storage
- Create and tear down clusters on demand
- Apache ZooKeeper is under the hood, highly available, and included with each cluster at no additional cost

Data processing

aws INNOVATE
ONLINE CONFERENCE

Amazon EMR: Enterprise-grade Hadoop & Spark

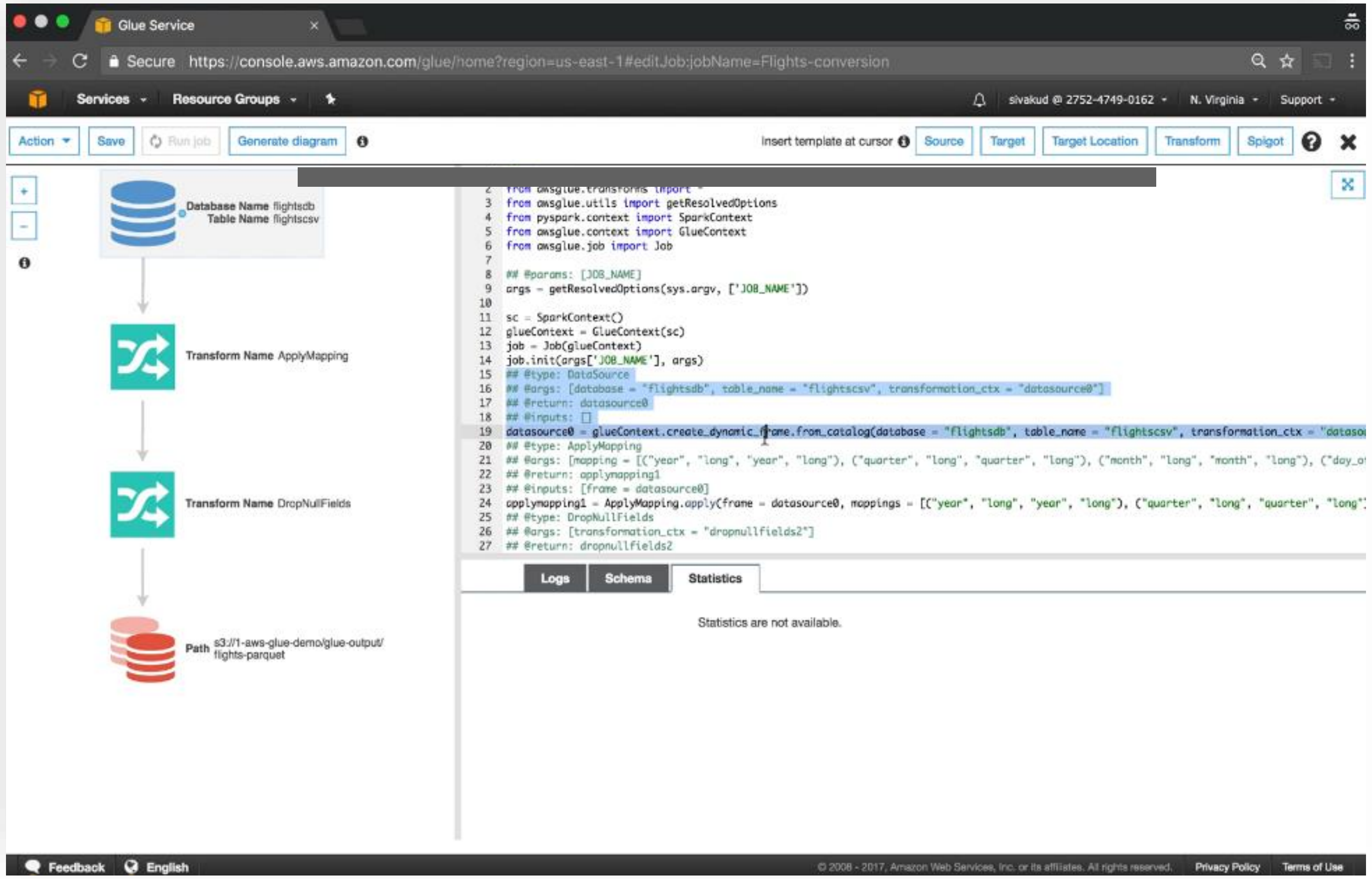
Scale to any size



- Scale compute (Amazon EMR) & storage (Amazon S3) independently
- Store and process any amount of data—PB to EBs
- Provision one, hundreds, or thousands of nodes
- Automatic scaling
- Integrated with Spot Instances

AWS Glue: ETL service

Make ETL scripting and deployment easy


Glue Service
Secure https://console.aws.amazon.com/glue/home?region=us-east-1#editJob:jobName=Flights-conversion
Services
Resource Groups
sivakud @ 2752-4749-0162
N. Virginia
Support
Action
Save
Run job
Generate diagram
Insert template at cursor
Source
Target
Target Location
Transform
Spigot
Database Name flightsdb
Table Name flightscsv
Transform Name ApplyMapping
Transform Name DropNullFields
Path s3://1-aws-glue-demo/glue-output/flights-parquet
Logs
Schema
Statistics
Statistics are not available.
Feedback
English
© 2008 - 2017, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

```
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job

## @params: [JOB_NAME]
args = getResolvedOptions(sys.argv, ['JOB_NAME'])

sc = SparkContext()
glueContext = GlueContext(sc)
job = Job(glueContext)
job.init(args['JOB_NAME'], args)
## @type: DataSource
## @args: [database = "flightsdb", table_name = "flightscsv", transformation_ctx = "datasource0"]
## @return: datasource0
## @inputs: []
datasource0 = glueContext.create_dynamic_frame.from_catalog(database = "flightsdb", table_name = "flightscsv", transformation_ctx = "dataso
## @type: ApplyMapping
## @args: [mapping = [("year", "long", "year", "long"), ("quarter", "long", "quarter", "long"), ("month", "long", "month", "long"), ("day_o
## @return: applymapping1
## @inputs: [frame = datasource0]
applymapping1 = ApplyMapping.apply(frame = datasource0, mappings = [("year", "long", "year", "long"), ("quarter", "long", "quarter", "long")
## @type: DropNullFields
## @args: [transformation_ctx = "dropnullfields2"]
## @return: dropnullfields2
```

- Run your Spark code in a serverless manner
- Automatically generate ETL code
- Code is customizable with Python and Scala
- Endpoints provided to edit, debug, and test code
- Jobs are scheduled or event-based

Consumption of data and insights

aws INNOVATE
ONLINE CONFERENCE

Amazon Redshift: Data warehousing

Fast, powerful, simple, and fully managed data warehouse at 1/10th the cost

Massively parallel, scale from gigabytes to petabytes

Fast at scale

Columnar storage technology to improve I/O efficiency and scale query performance

Open file formats

Analyze optimized data formats on the latest SSD, and all open data formats in Amazon S3

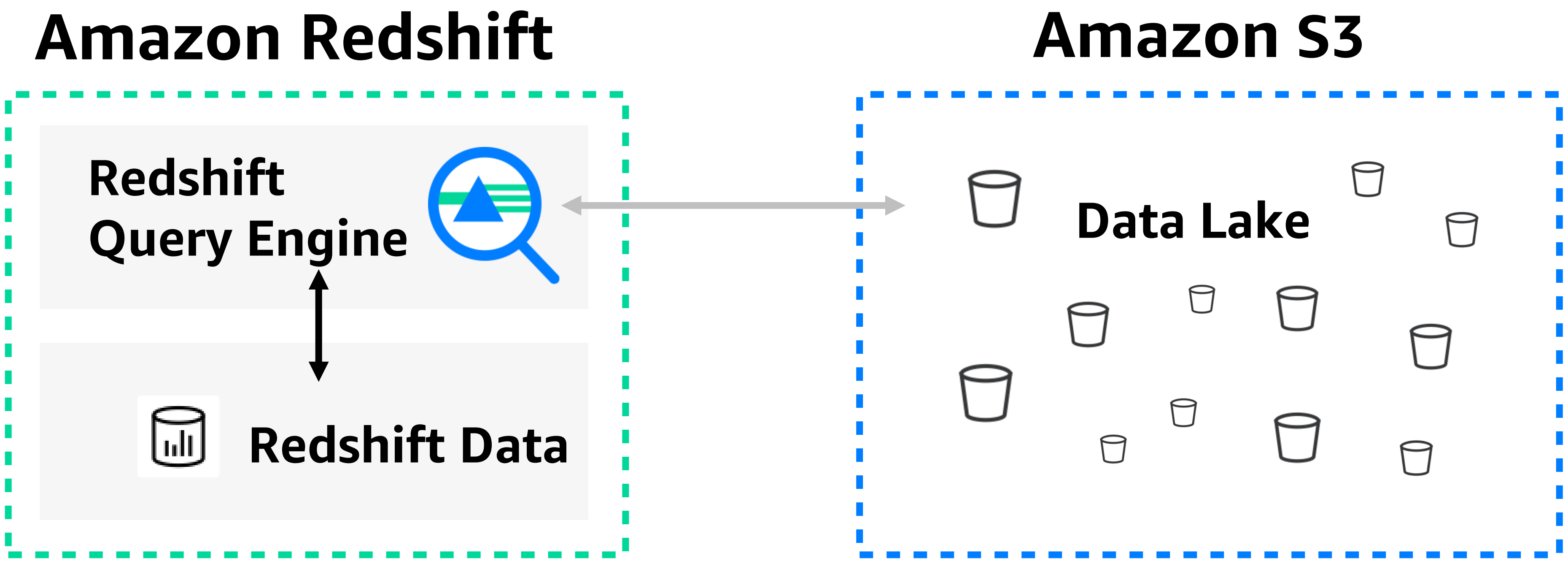
Secure

Audit everything; encrypt data end to end; extensive certification and compliance

Inexpensive

As low as $1,000 per terabyte per year, 1/10th the cost of traditional data warehouse solutions; start at $0.25 per hour

Extend Amazon Redshift queries to Amazon S3



Query S3 directly or join data across Amazon Redshift and S3

Scale Amazon Redshift compute and storage separately

Support for CSV, Parquet, ORC, Grok, Avro, and more formats

Amazon QuickSight

No servers to manage

Amazon QuickSight is a fully managed cloud application, meaning there's no upfront cost, software to deploy, capacity planning, maintenance, upgrades, or migrations.

Pay for what you use

Provide read-only access to interactive dashboards and pay only when your users access them with pay-per-session pricing. There are no upfront costs, no annual commitments, and no charges for inactive users.

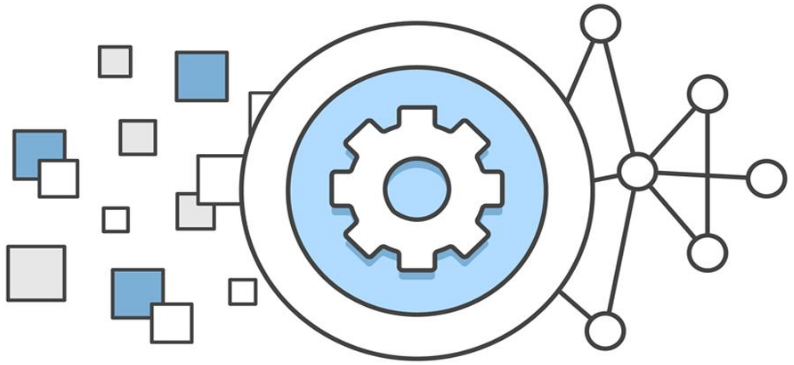
Scalable

Amazon QuickSight automatically scales with your usage and activity, with no need for additional infrastructure. From 10 users to 10,000, Amazon QuickSight seamlessly grows with you.

Fully integrated

Amazon QuickSight integrates with your data sources and other AWS services like Amazon Redshift, S3, Amazon Athena, Amazon Aurora, Amazon RDS, IAM, AWS CloudTrail, Amazon Cloud Directory, and more—providing you with everything you need to build an end-to-end BI solution.

Amazon Athena

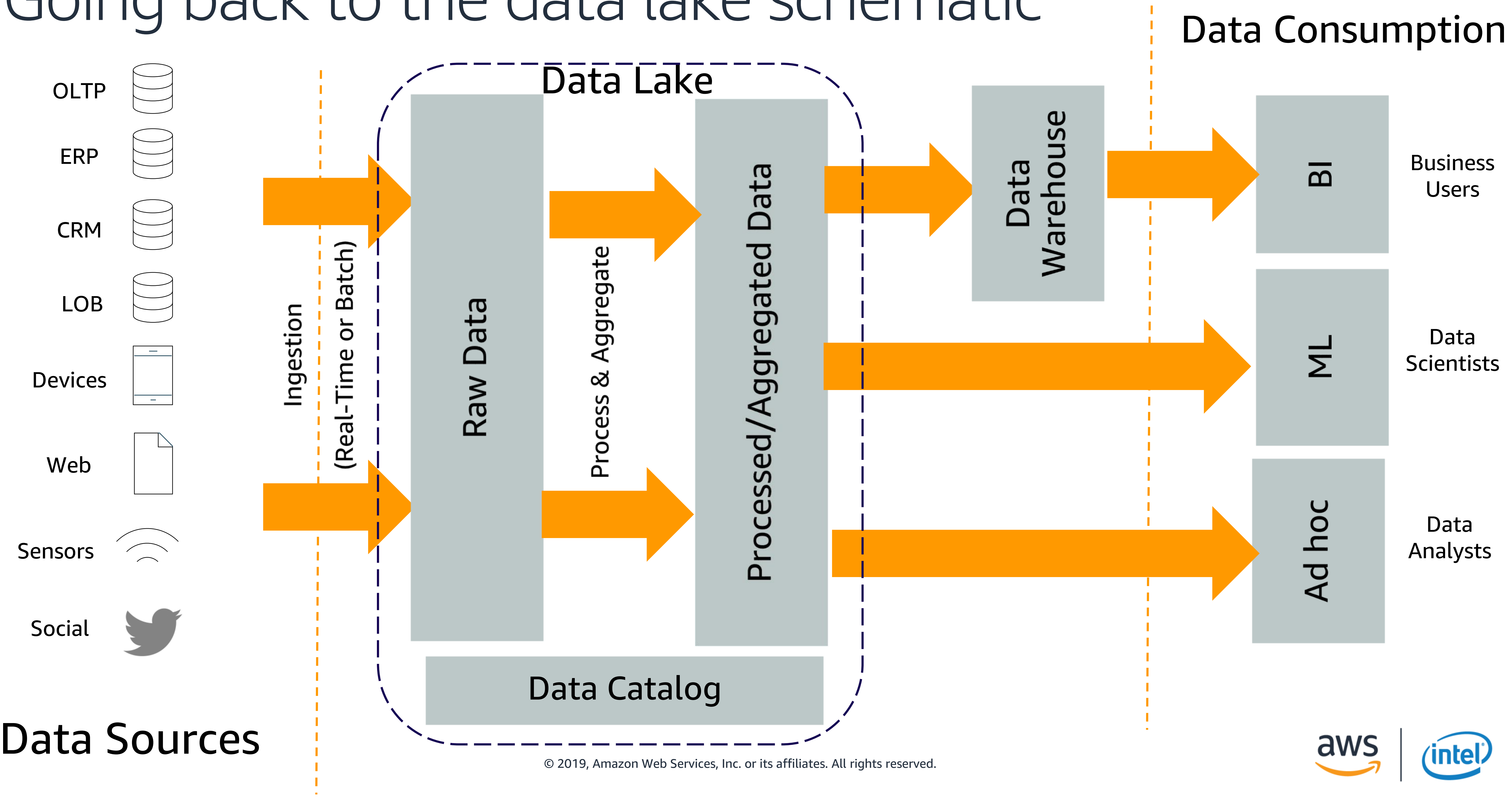


- Query data from Amazon S3 directly
- Use ANSI SQL with support for complex joins, nested queries, and window functions
- Leverages Presto under the covers
- Pay $5/TB scanned by your query
- Workgroups for cost control

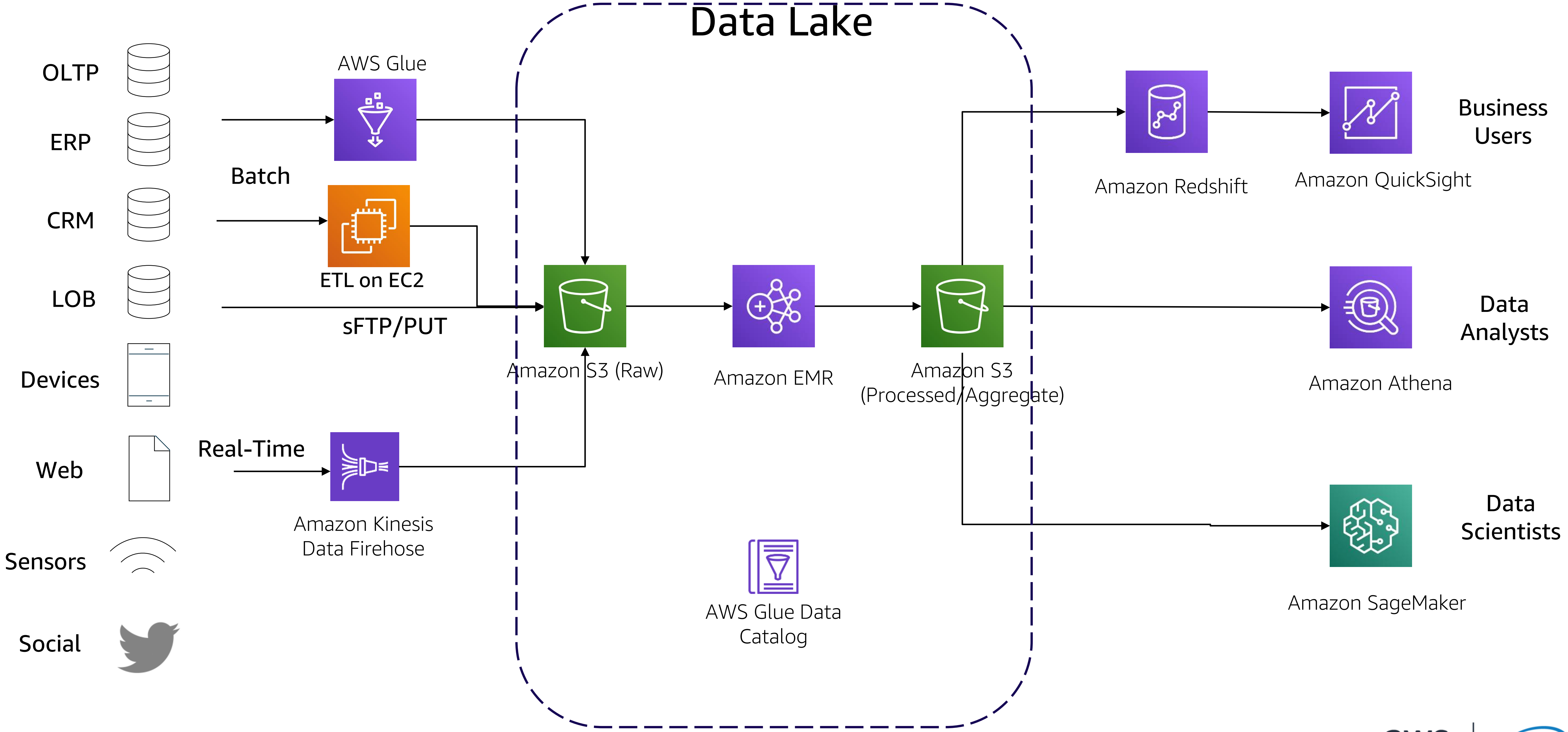
Amazon SageMaker: Build, train, and deploy ML models at scale

- Pre-built notebooks for common problems
- Built-in, high-performance algorithms
- One-click training on the highest-performing infrastructure
- Model optimization
- One-click deployment
- Fully managed with automatic scaling for 75% less

Going back to the data lake schematic



Data lake on AWS

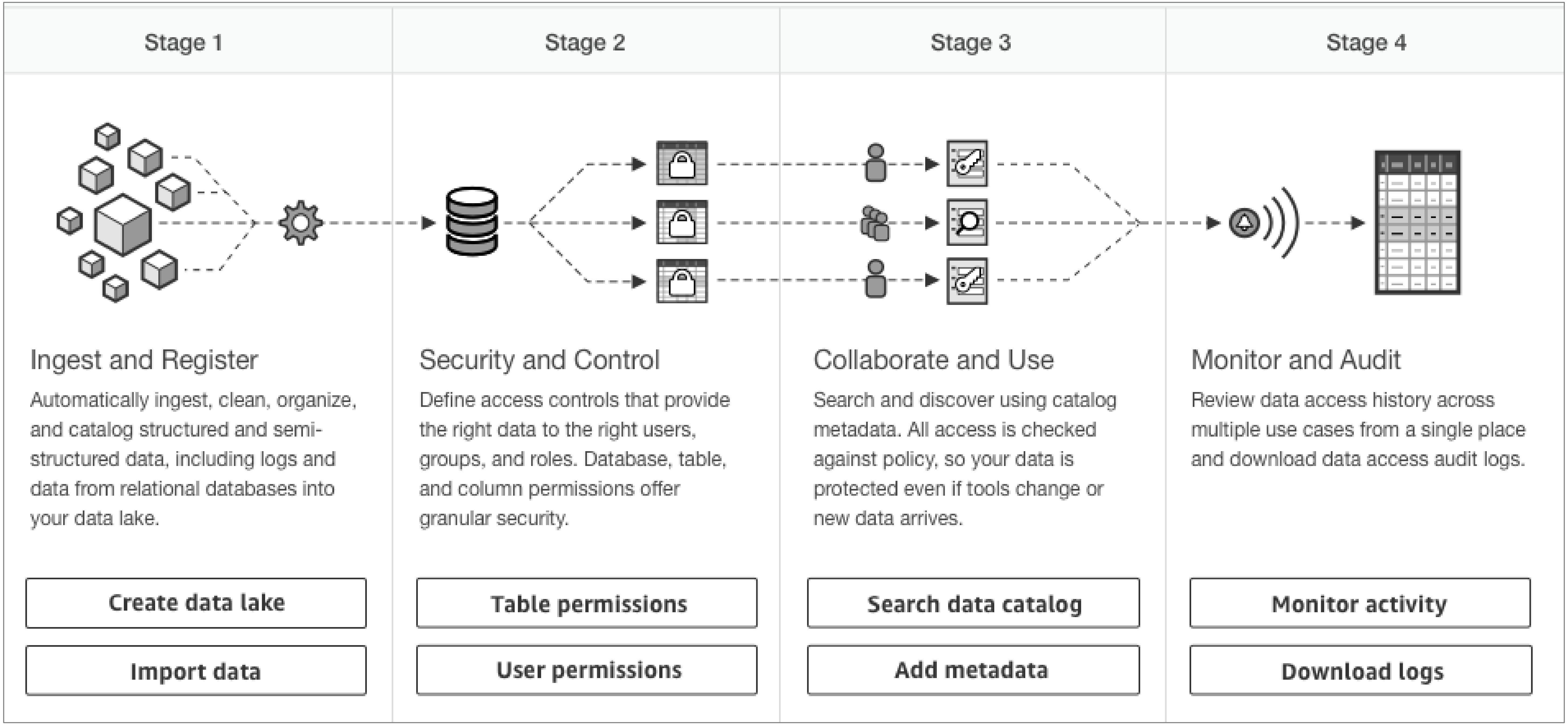


AWS Lake Formation

aws INNOVATE
ONLINE CONFERENCE

AWS Lake Formation

Build a secure data lake in days



Stage 1

Ingest and Register

Automatically ingest, clean, organize, and catalog structured and semi-structured data, including logs and data from relational databases into your data lake.

Stage 2

Security and Control

Define access controls that provide the right data to the right users, groups, and roles. Database, table, and column permissions offer granular security.

Stage 3

Collaborate and Use

Search and discover using catalog metadata. All access is checked against policy, so your data is protected even if tools change or new data arrives.

Stage 4

Monitor and Audit

Review data access history across multiple use cases from a single place and download data access audit logs.

AWS Lake Formation relieves data lake heavy lifting

Loading data

- **Blueprints / Data Importers** – templates for data ingestion, ETL, metadata (schema) and partition management

- **ML Transformations** – ML algorithms that customers can use to create their own ML Transforms (e.g., record de-duplication, match finding)

AWS Lake Formation relieves data lake heavy lifting

Discoverability and governance

- **Enhanced governance layer** – security and governance layer at the Data Catalog level

- **Enhanced Data Catalog** – enable users to record more metadata and tag Data Catalog objects (i.e., databases, tables, columns)

Learn from AWS experts. Advance your skills and knowledge. Build your future in the AWS Cloud.

Digital Training

Free, self-paced online courses built by AWS experts

Classroom Training

Classes taught by accredited AWS instructors

AWS Certification

Exams to validate expertise with an industry-recognized credential

Ready to begin building your cloud skills?
Get started at: https://www.aws.training/

Why work with an APN Partner?

APN Partners are uniquely positioned to help your organization at any stage of your cloud adoption journey, and they:

- Share your goals—focused on your success
- Help you take full advantage of all the business benefits that AWS has to offer
- Provide services and solutions to support any AWS use case across your full customer life cycle

APN Partners with deep expertise in AWS services:

AWS Managed Service Provider (MSP) Partners

APN Partners with cloud infrastructure and application migration expertise

AWS Competency Partners

APN Partners with verified, vetted, and validated specialized offerings

AWS Service Delivery Partners

APN Partners with a track record of delivering specific AWS services to customers

Find the right APN Partner for your needs: https://aws.amazon.com/partners/find/

Thank you for attending AWS Innovate

We hope you found it interesting! A kind reminder to **complete the survey.**
Let us know what you thought of today's event and how we can improve the event experience for you in the future.

- aws-apac-marketing@amazon.com
- twitter.com/AWSCloud
- facebook.com/AmazonWebServices
- youtube.com/user/AmazonWebServices
- slideshare.net/AmazonWebServices
- twitch.tv/aws