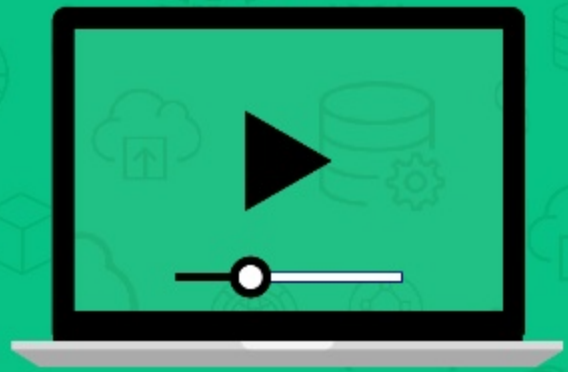
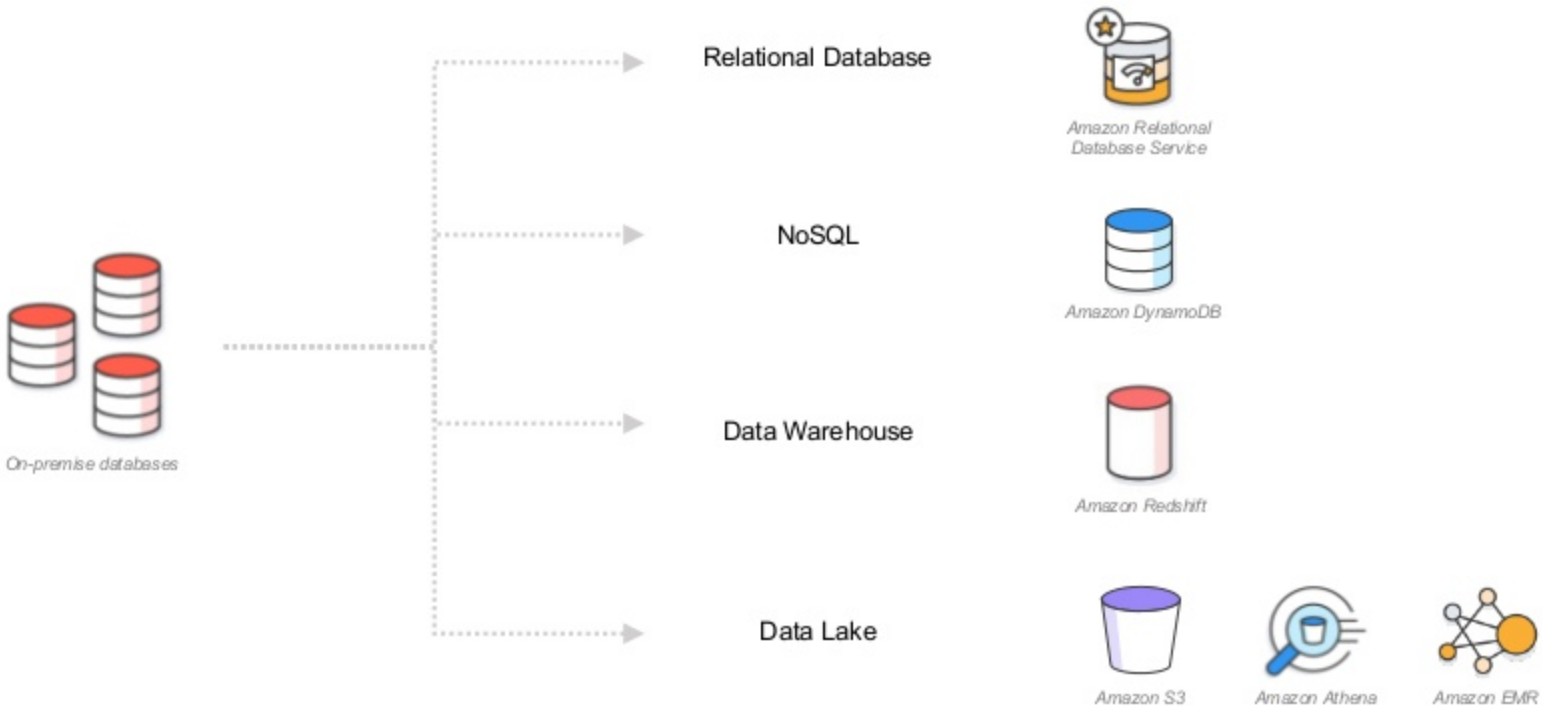
Best Practices for Building Your Data Lake on AWS

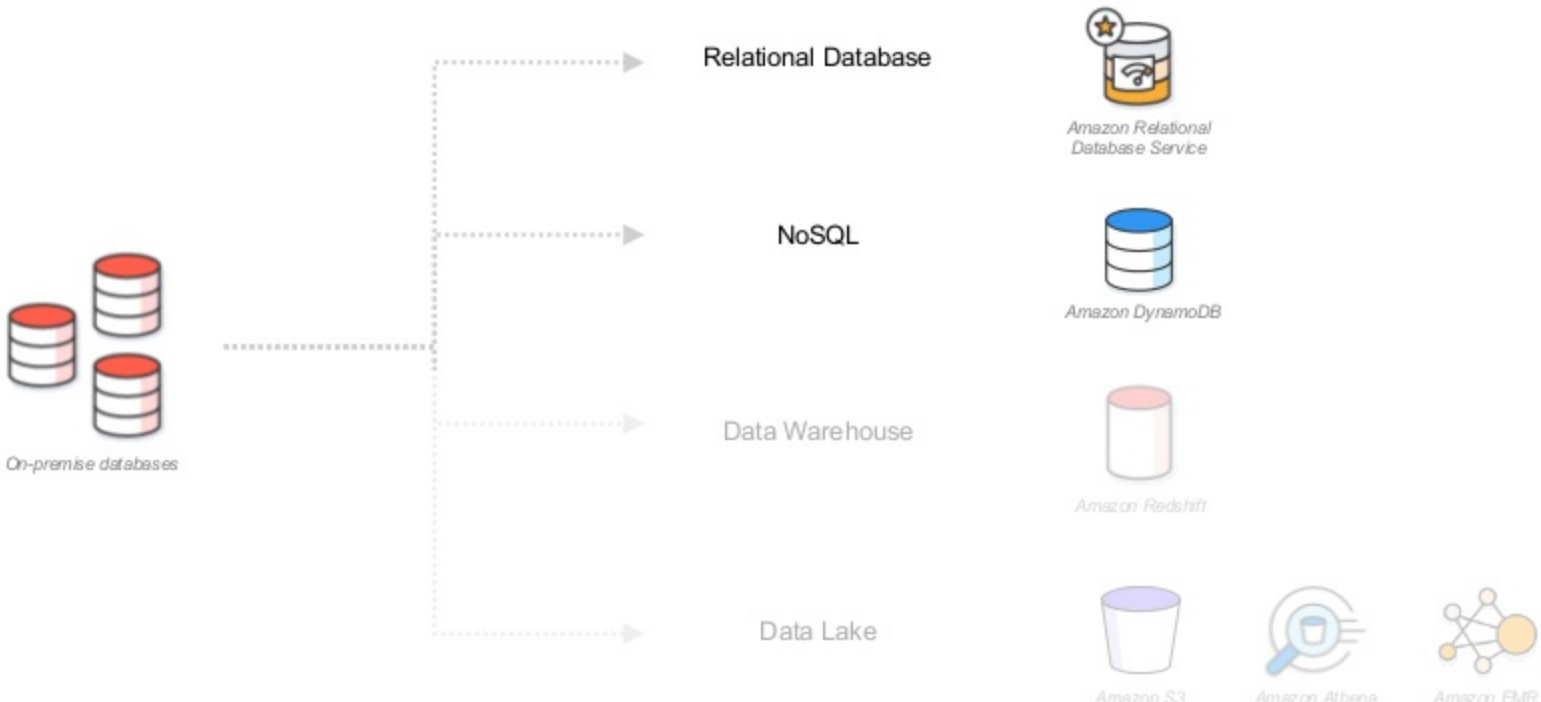


Ian Robinson, Specialist SA, AWS
Kiran Tamana, EMEA Head of Solutions Architecture, Datapipe
Derwin McGeary, Solutions Architect, Cloudwick

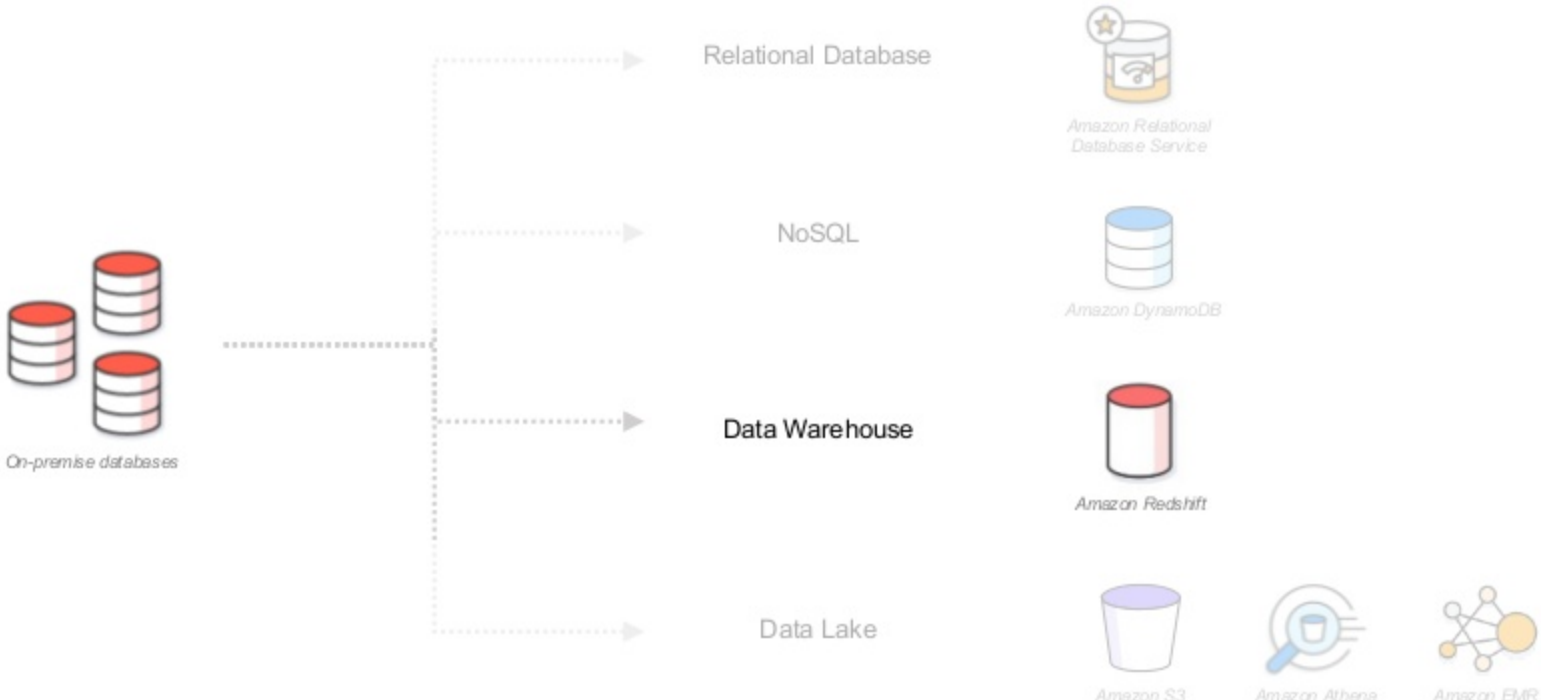
Webinars: Database Freedom



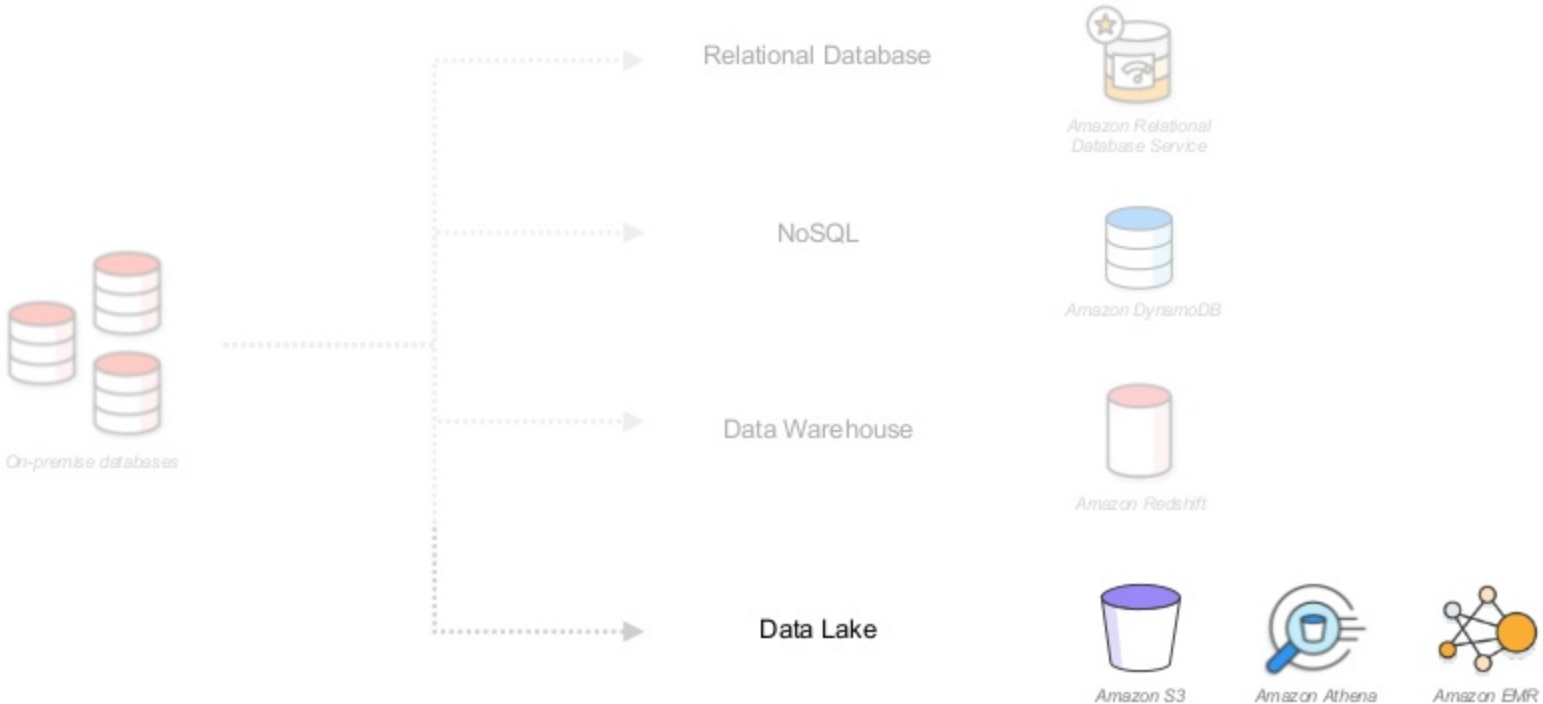
Webinar 1: Migrating Online/Transactional Workloads



Webinar 2: Amazon Redshift and Data Warehousing



Webinar 3: Data Lake Best Practices



Agenda

- What is a Data Lake?
- Key Data Lake Concepts
- Modern Data Architecture
- Putting it All Together
- Partner: Datapipe
- Partner: Cloudwick

What is a Data Lake?

What and Why

Store all your data, forever, at every stage of its lifecycle
Apply it using the right tool for the job

- Agile analytics
- Leverage all data that flows into your organization
- Customer centricity
- Better predictions via Machine Learning
- Competitive advantage

Data Lake versus Enterprise Data Warehouse

EMR

S3

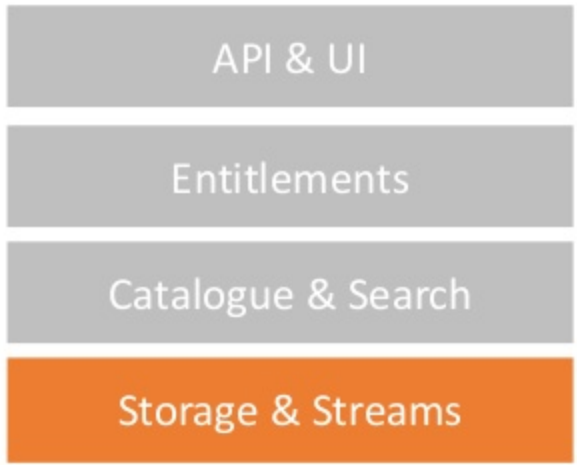
Enterprise DW

Complementary to EDW (not replacement)	**Data lake can be source for EDW**
Schema on read (no predefined schemas)	Schema on write (predefined schemas)
Structured/semi-structured/Unstructured data	Structured data only
Fast ingestion of new data/content	Time consuming to introduce new content
Data Science + Prediction/Advanced Analytics + BI use cases	BI use cases only (no prediction/advanced analytics)
Data at low level of detail/granularity	Data at summary/aggregated level of detail
Loosely defined SLAs	Tight SLAs (production schedules)
Flexibility in tools (open source/tools for advanced analytics)	Limited flexibility in tools (SQL only)

Key Data Lake Concepts


COMPUTE
COMPUTE
COMPUTE
COMPUTE
COMPUTE
COMPUTE
COMPUTE
COMPUTE
STORAGE

Components of a Data Lake



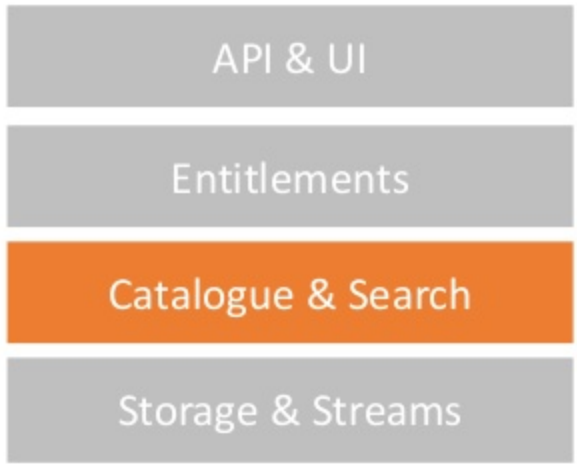
Data Storage

- High durability
- Stores raw data from input sources
- Support for any type of data
- Low cost

Streaming

- Streaming ingest of feed data
- Provides the ability to consume any dataset as a stream
- Facilitates low latency analytics

Components of a Data Lake



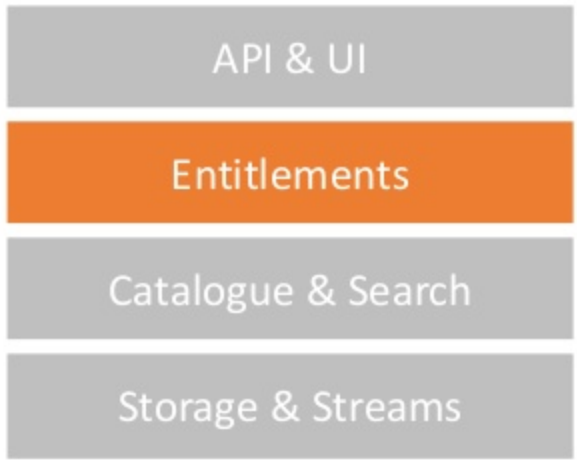
Catalogue

- Metadata lake
- Used for summary statistics and data Classification management

Search

- Simplified access model for data discovery

Components of a Data Lake

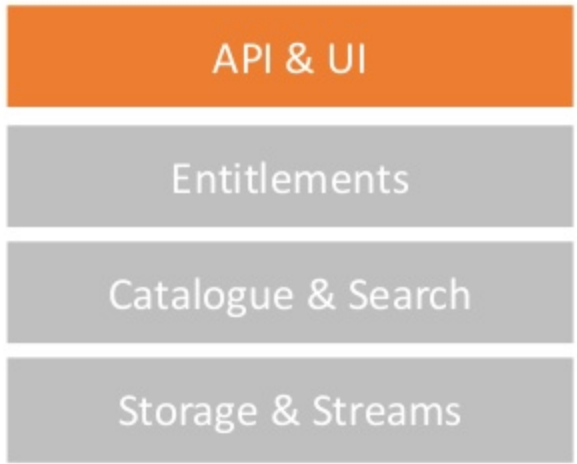


Entitlements system

- Encryption
- Authentication
- Authorisation
- Chargeback
- Quotas
- Data masking
- Regional restrictions

aws

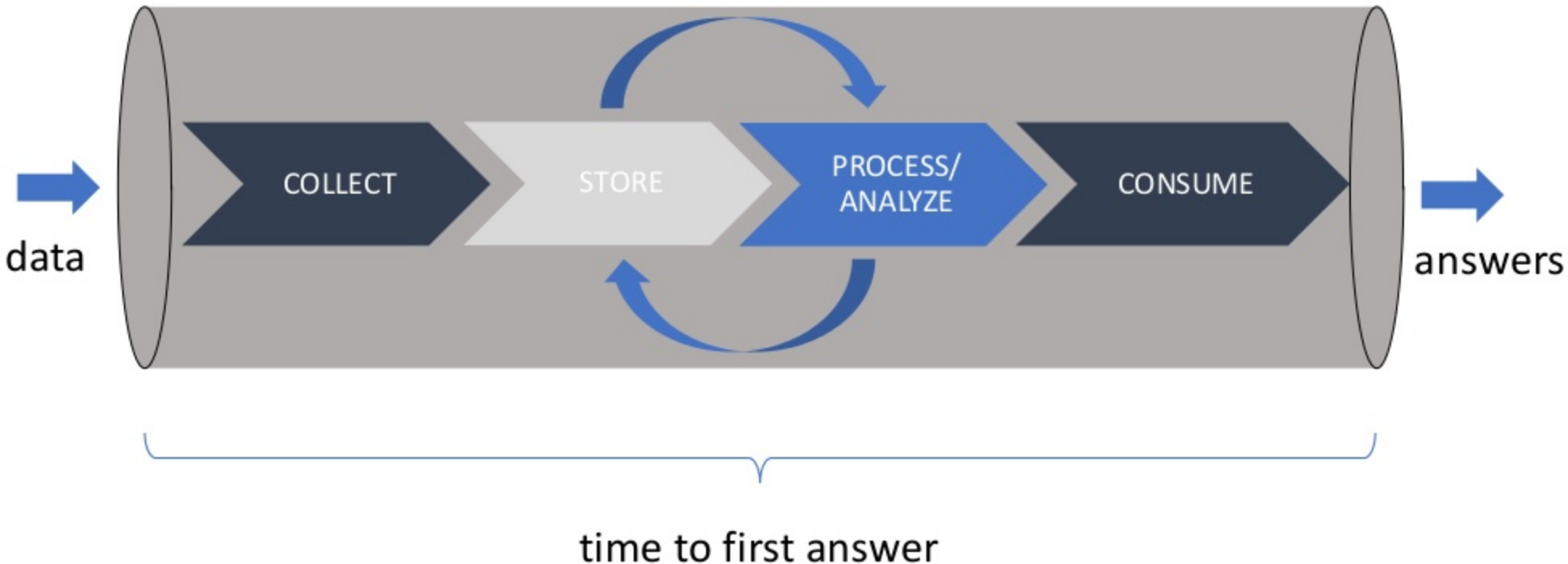
Components of a Data Lake



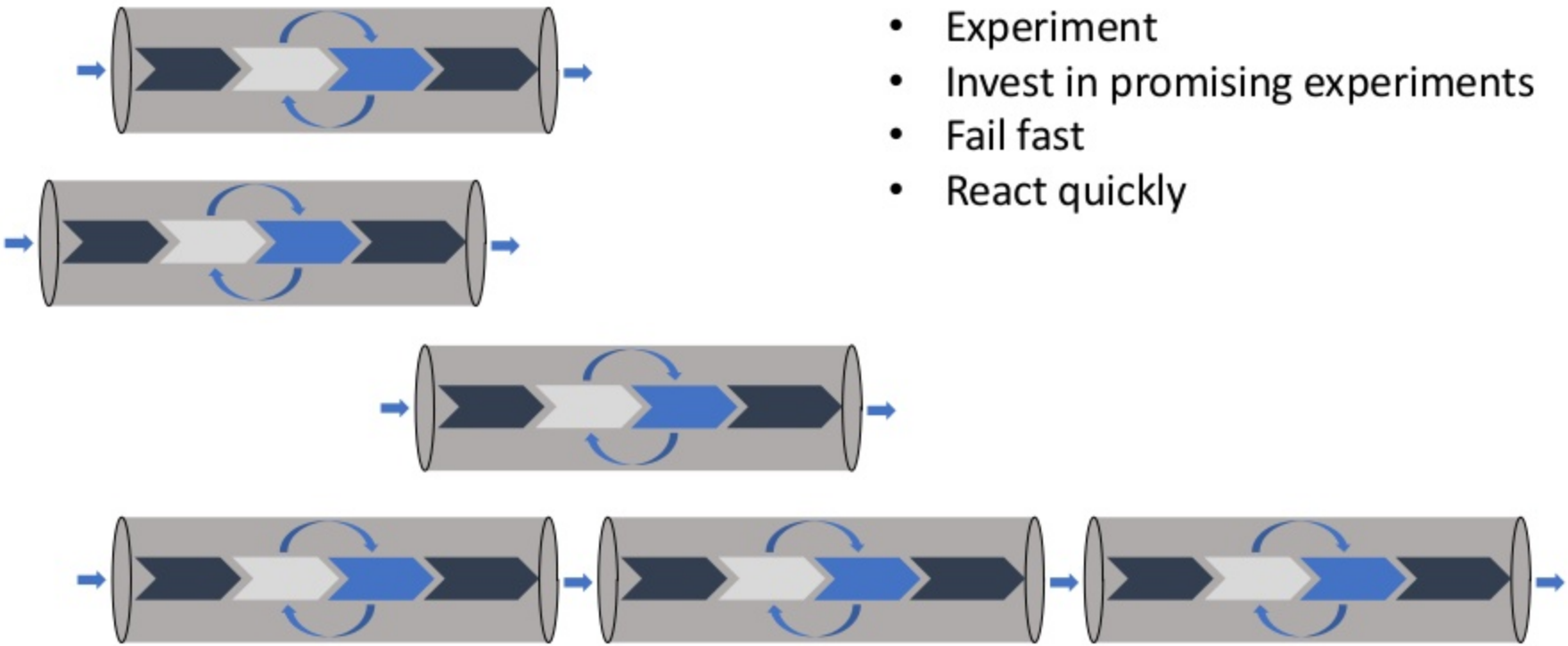
API & User Interface

- Exposes the data lake to customers
- Programmatically query catalogue
- Expose search API
- Ensures that entitlements are respected

The Modern Data Architecture


COLLECT
STORE
PROCESS/
ANALYZE
CONSUME
data
answers
time to first answer

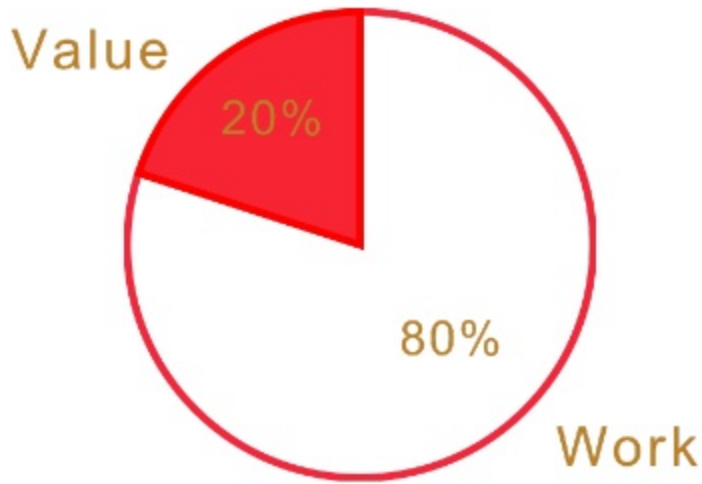
Agile Analytics



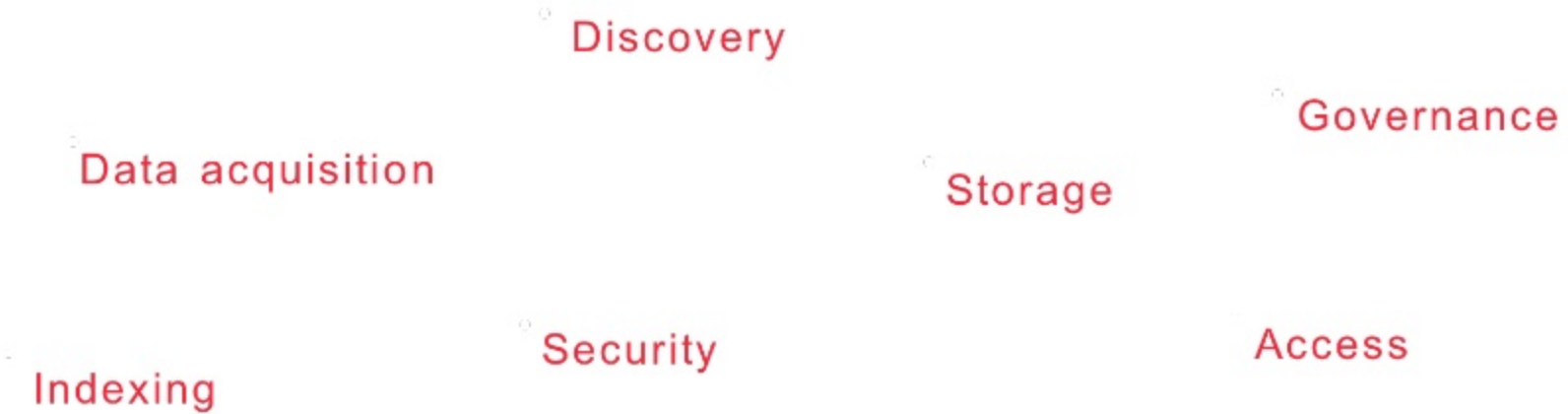
- Experiment
- Invest in promising experiments
- Fail fast
- React quickly

80%

Of What We Consider
Analytics **Is Not Analytics**



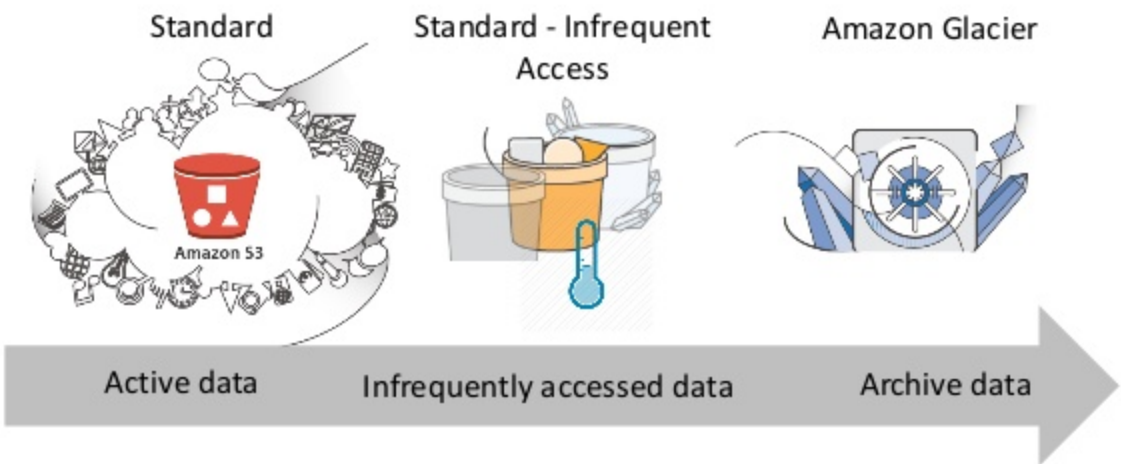
80% Of What We Consider Analytics Is Not Analytics



Storage
Ingest

Preservation of Original Source Data

Events and Lifecycle Management



S3 as the Data Lake Fabric

- **Unlimited** number of objects and volume
- 99.99% **availability**
- 99.999999999% **durability**
- Versioning
- **Tiered storage** via lifecycle policies
- SSL, client/server-side encryption at rest
- **Low cost** (just over $2700/month for 100TB)
- **Natively supported** by big data frameworks (Spark, Hive, Presto, etc)
- **Decouples** storage and compute
  - Run **transient** compute clusters (with Amazon EC2 Spot Instances)
  - **Multiple, heterogeneous** clusters can use same data

Automated Data Ingestion

S3 Upload

S3 Acceleration

Kinesis

Snowball
Snowball Edge
Snowmobile

Database Migration Service

aws

Storage + Catalog

Scalable (secure, versioned, durable) storage +
Immutable data at every stage of its lifecycle +
Versioned schema and metadata
=
Data discovery, lineage

AWS Glue

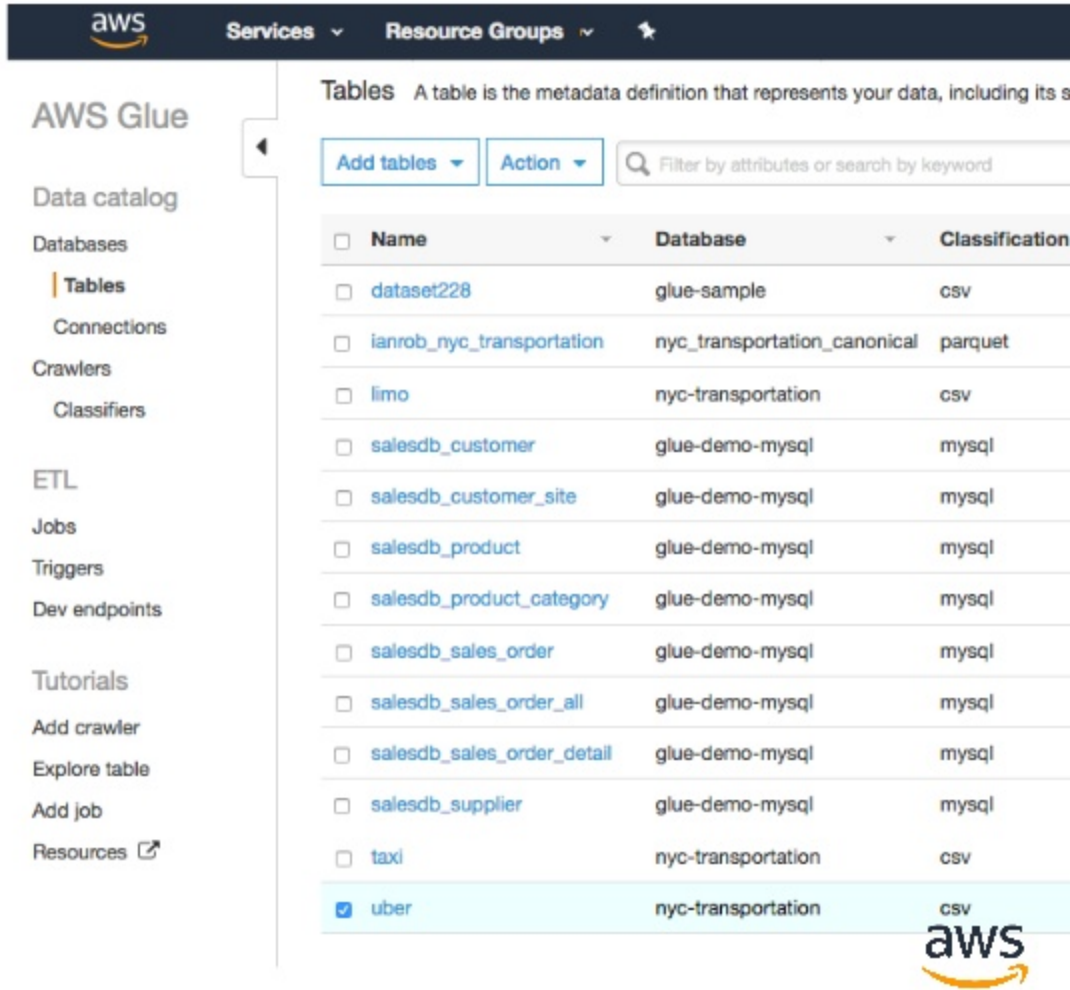
- **Data Catalog** Discover and store metadata
- **Job Authoring** Auto-generated ETL code
- **Job Execution** Serverless scheduling and execution

Glue Data Catalog

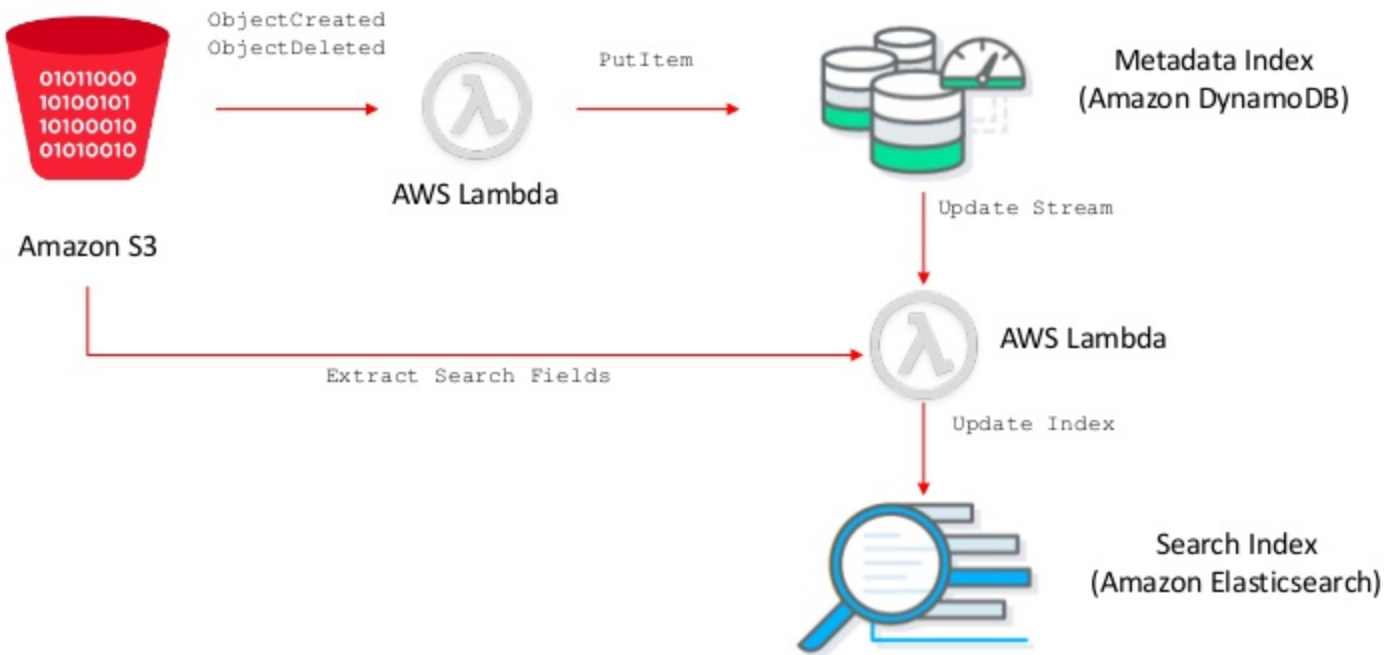
Hive metastore-compatible, highly-available metadata repository:

- **Search** metadata for data discovery
- **Connection info** – JDBC URLs, credentials
- **Classification** for identifying and parsing files
- **Versioning** of table metadata as schemas evolve and other metadata are updated
- **Table definitions** – usable by Redshift, Athena, Glue, EMR

Populate using Hive DDL, bulk import, or automatically through **crawlers**.



Indexing and Searching Using Metadata

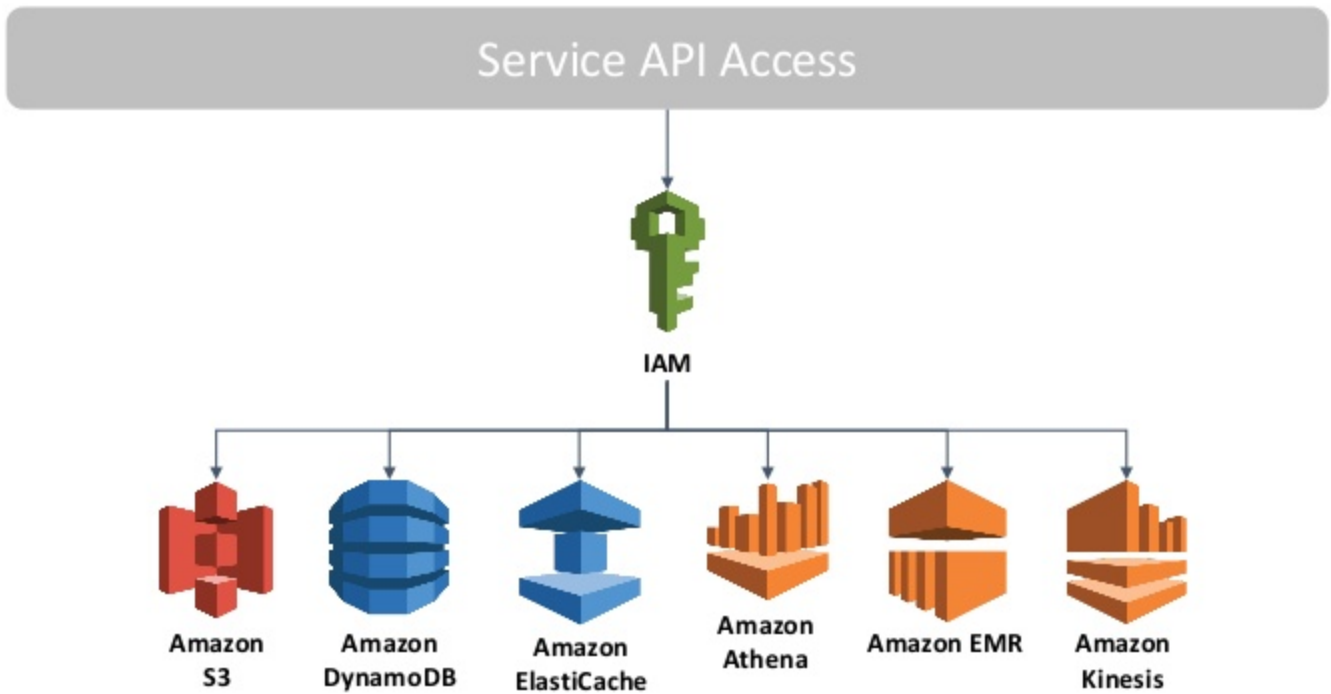


Governance
Security
Privacy

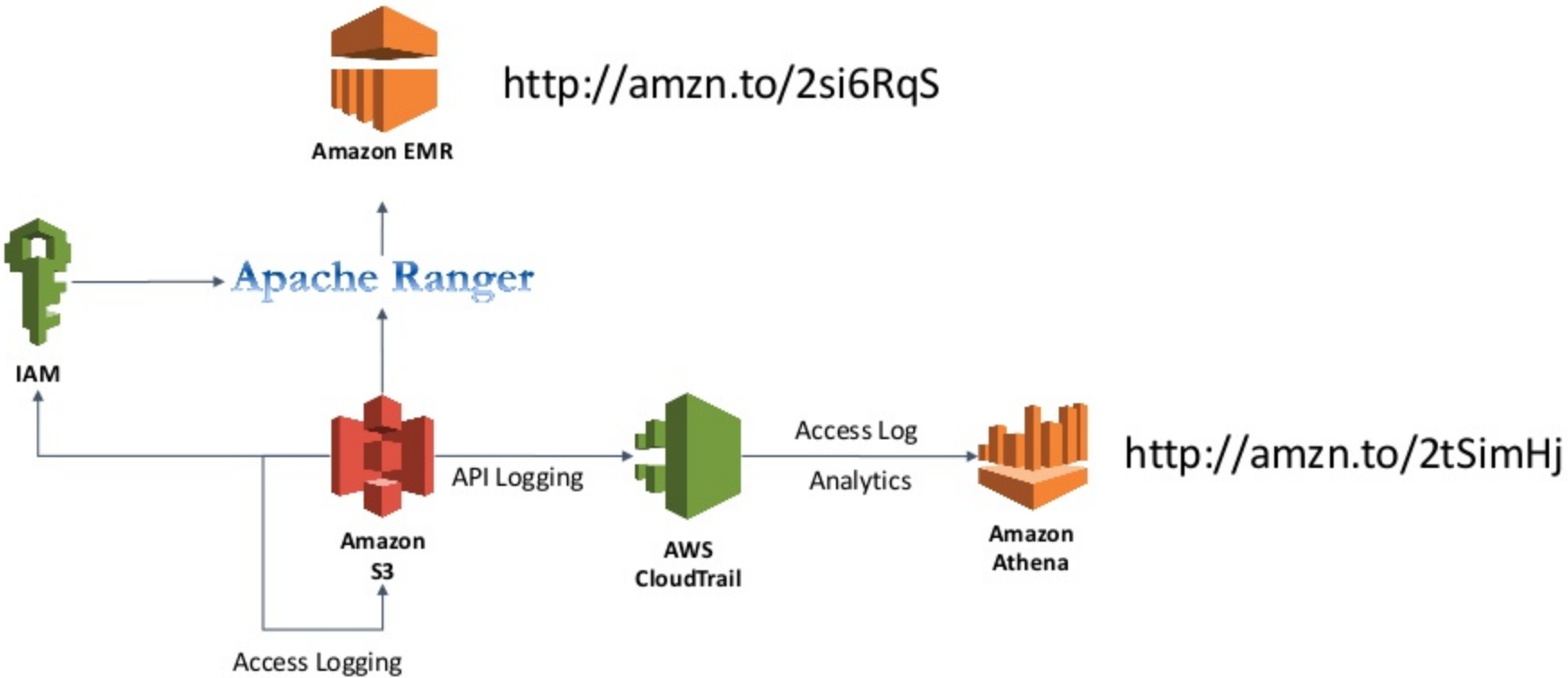
Identity and Access Management

- **Manage** users, groups, and roles
- Identity **federation** with Open ID
- Temporary credentials with Amazon Security Token Service (Amazon **STS**)
- Stored **policy templates**
- Powerful **policy language**
- Amazon **S3 bucket policies**

Security at the Data Level



Third Party Ecosystem Security Tools



Storage Level Support for Access Logging and Audit

aws

Encryption Options

AWS Server-Side encryption

- AWS managed key infrastructure

AWS Key Management Service

- Automated key rotation & auditing
- Integration with other AWS services

AWS CloudHSM

- Dedicated Tenancy SafeNet Luna SA HSM Device
- Common Criteria EAL4+, NIST FIPS 140-2

aws

Discovery
Search
Access

Data Lake API and UI

- **Exposes** the metadata API, search, and Amazon S3 storage services to customers
- Can be based on **TVM/STS Temporary Access** for many services, and a bespoke API for metadata
- Drive all **UI operations** from API?

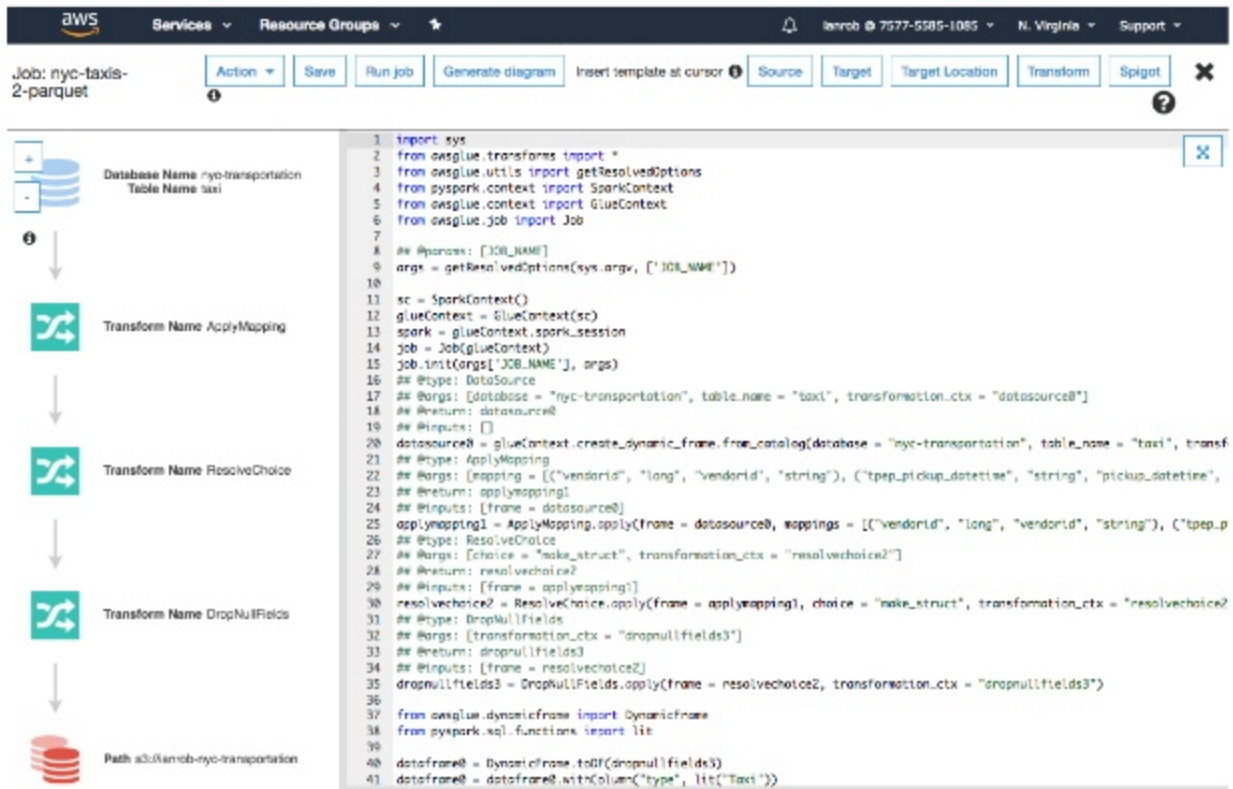
Amazon API Gateway

- Host multiple **versions** and stages of APIs
- Create and **distribute** API keys to developers
- Leverage AWS Sigv4 to **authorize access** to APIs
- **Throttle** and **monitor** requests to protect the backend
- Leverages **AWS Lambda**

Data Quality
Data Preparation
Orchestration and Job Scheduling

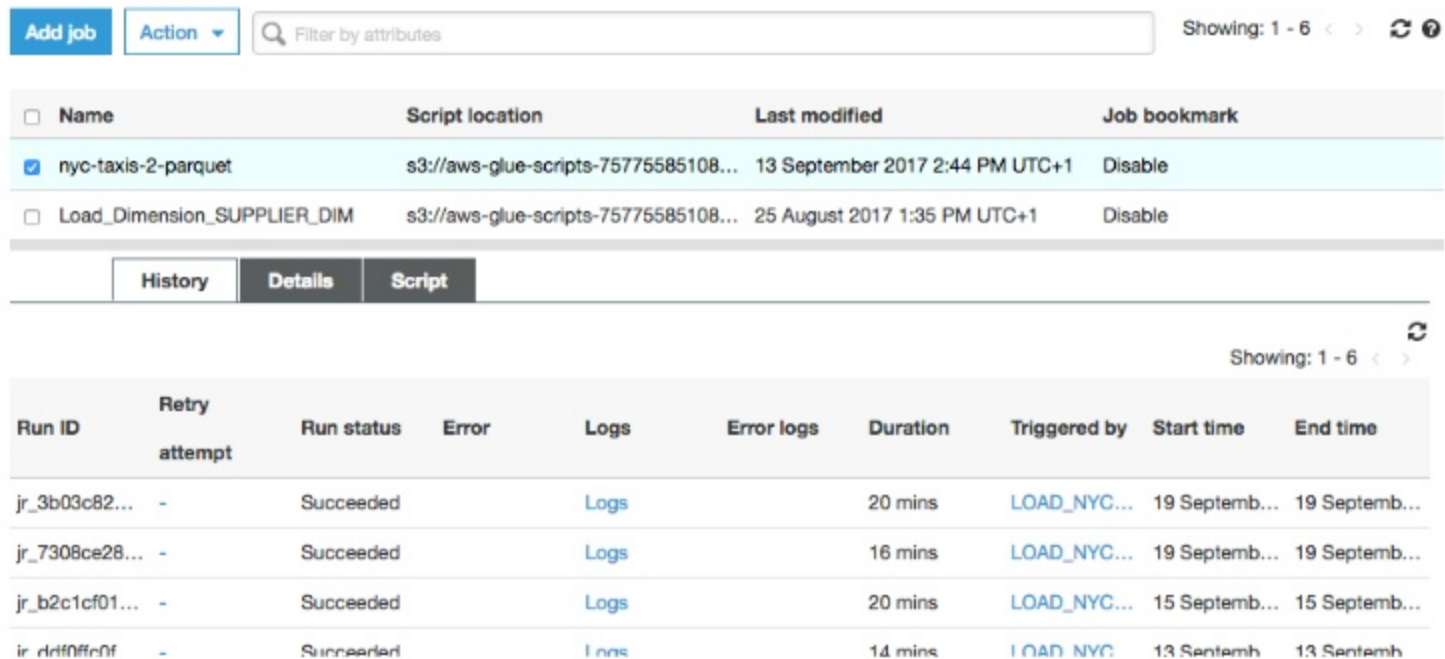
Job Authoring with AWS Glue

- Python code **generated** by AWS Glue
- Connect a **notebook** or IDE to AWS Glue
- **Existing** code brought into AWS Glue

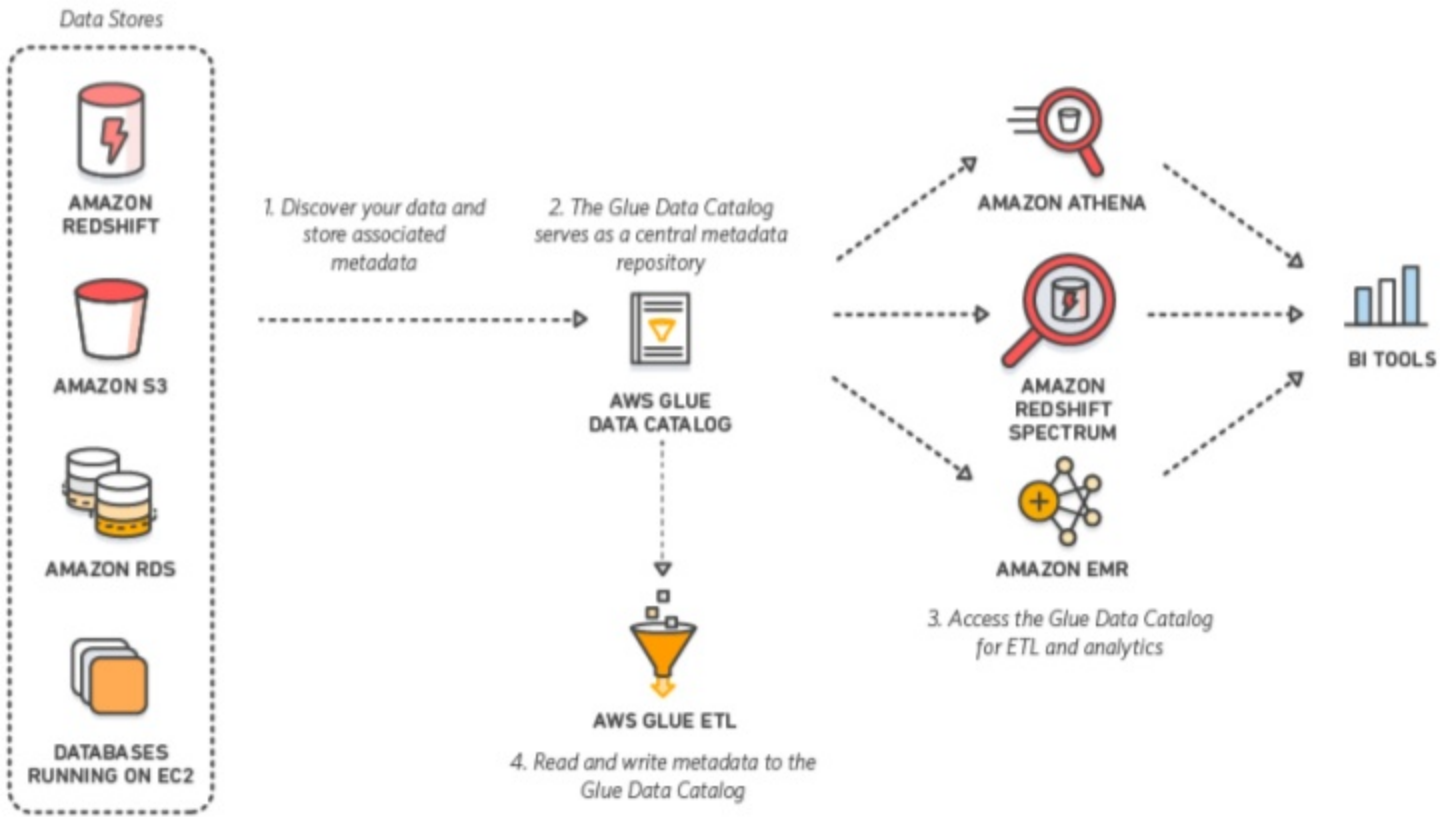


Job Execution with AWS Glue

- Schedule-based
- Event-based
- On demand



Instantly Query Your Data Lake on S3



AWS Batch

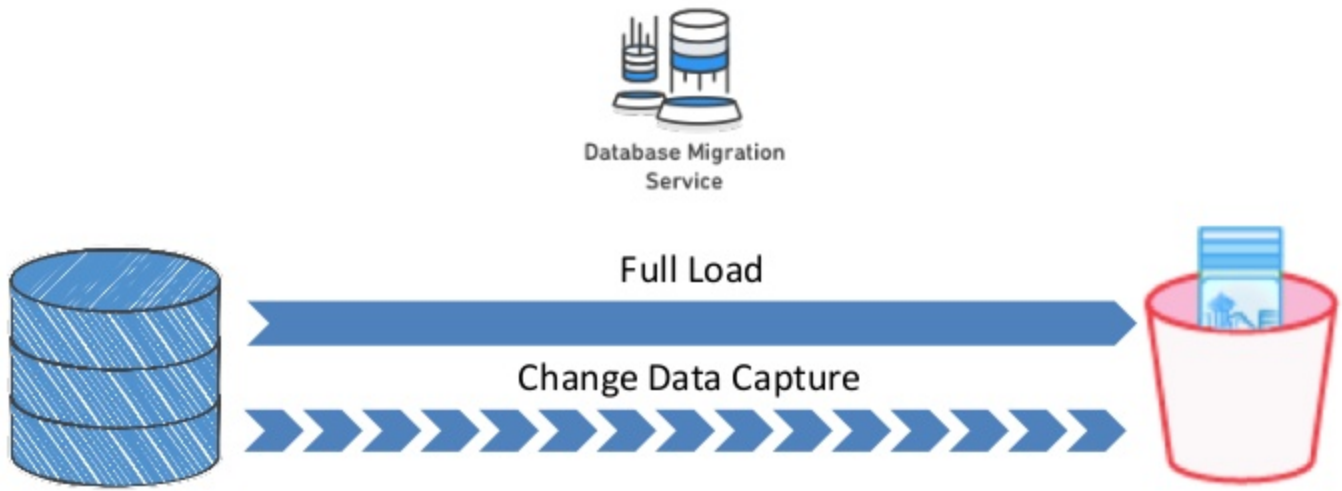
Fully managed

Dynamic provisioning and scaling

Cost optimization through EC2 Spot fleet

Priority-based queues and scheduling

Write Database Changes to S3 with DMS



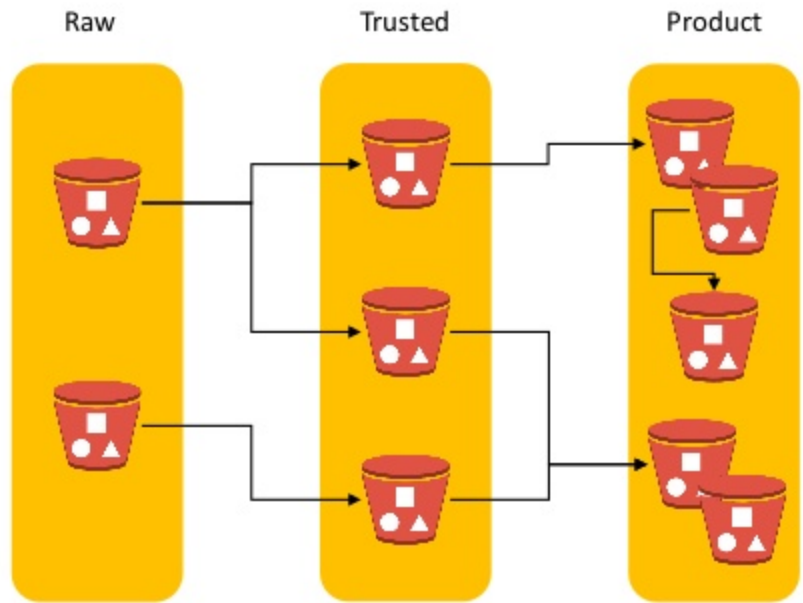

```
<schema_name>/<table_name>/LOAD001.csv
<schema_name>/<table_name>/LOAD002.csv
<schema_name>/<table_name>/<time-stamp>.csv
```

Putting it All Together

Organizing and Managing Your Data Lake

Work backwards from product:

- Define application and analytics datasets
- Identify and land raw sources of data
- Cleanse, enrich, standardize to create trusted sources
- Transform to create product



Organizing and Managing Your Data Lake

Structuring and Cataloguing

S3 prefixes – Separate and partition data

AWS Glue Data Catalog – Versioned metadata grouped by database/table

Amazon DynamoDB/Amazon Elasticsearch – Indexing

Archiving

S3 lifecycle policies

Ownership, Classification, Access Control

Buckets

Tags

Bucket policies

ACLs

Auditing

AWS CloudTrail and S3 Access Logging

Shifting Analytics
So It's

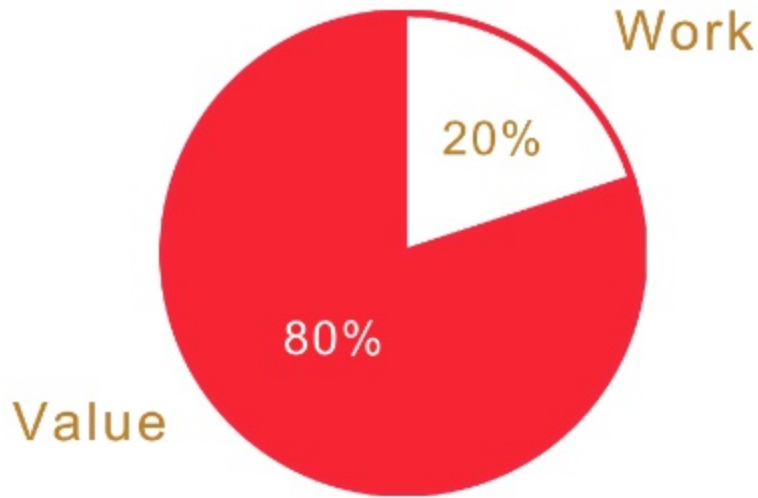
80%

Analytics

And Only

20%

Prep



Building a Data Strategy on AWS

Analytics Capabilities

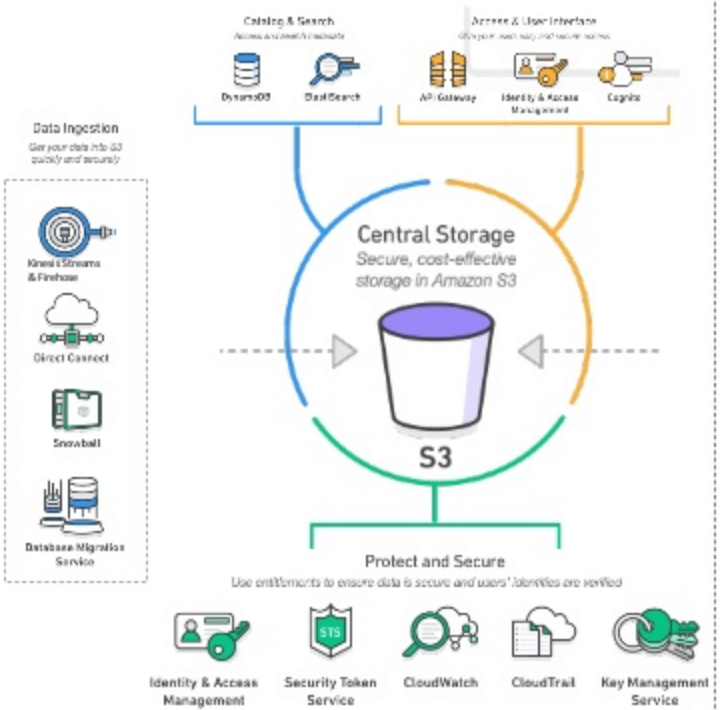
Data processing

Data warehousing

Reporting

Real-time processing

Predictive analytics



Processing & Analytics

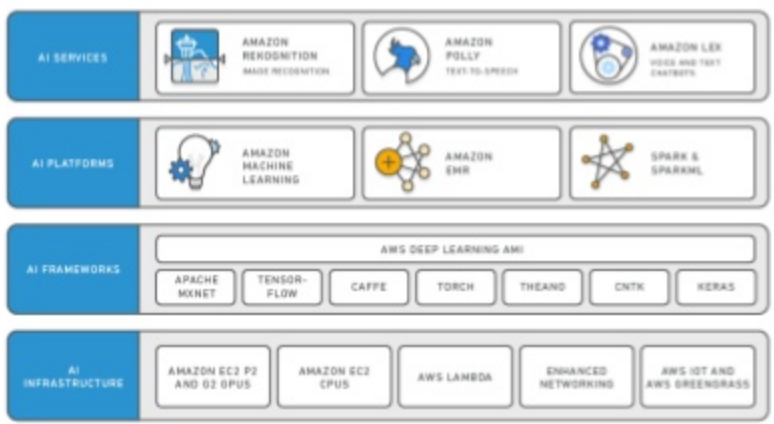
Real-time

- Elasticsearch Service
- Kinesis Analytics, Kinesis Streams
- Spark Streaming on EMR
- Apache Flink on EMR
- AWS Lambda
- Apache Storm on EMR
- EastiCache
- kafka
- DAX

Batch

- EMR Hadoop, Spark, Presto
- Athena Query Service
- Redshift Data Warehouse
- AWS Batch

Predictive



Transactional & RDBMS

- DynamoDB
  NoSQL DB
- Aurora
  Relational Database

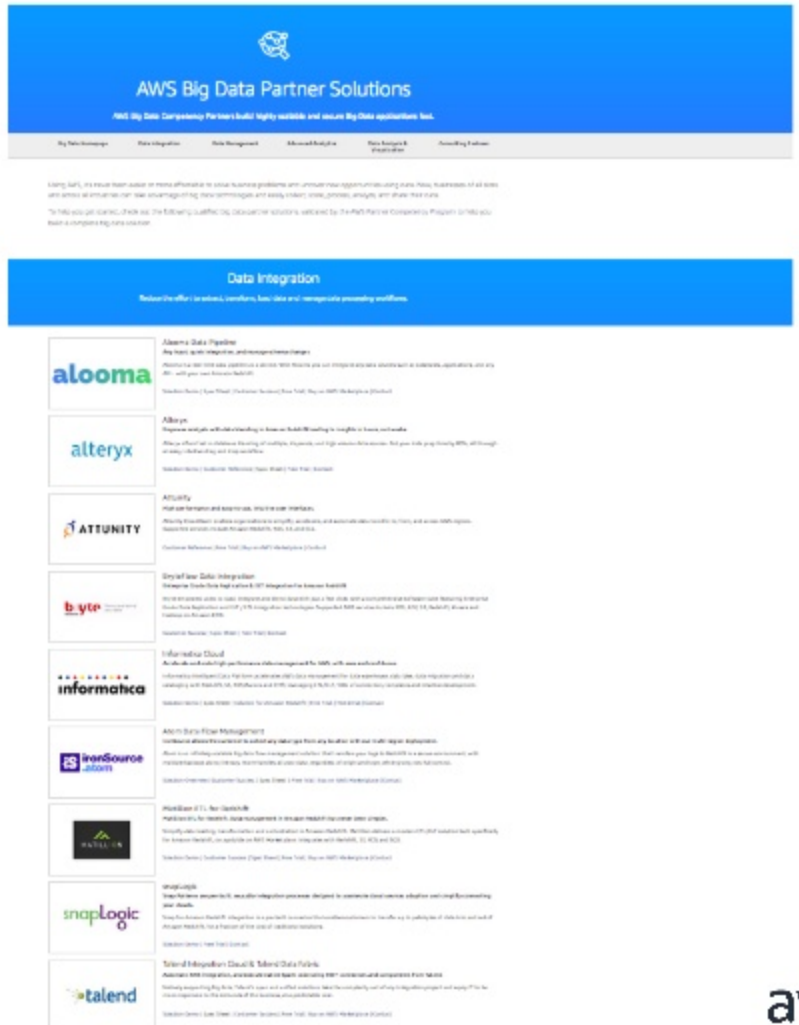
BI & Data Visualization

- Amazon QuickSight
- tableau
- MicroStrategy
- kibana

aws

AWS Partner Ecosystem

Reduce the effort to move, cleanse, synchronize, manage, and automate data related processes



https://aws.amazon.com/big-data/partner-solutions/

AWS DATA LAKES

Kiran Tamana, EMEA Head of Solutions Architecture, Datapipe

DATAPIPE
a rackspace company

PREMIER CONSULTING PARTNER IN EMEA AND NORTH AMERICA RECOGNISED AS A LEADER BY GARTNER WORLDWIDE

aws partner network

Premier

Consulting Partner

- Microsoft Workload Competency
- Migration Competency
- Public Sector Partner
- MSP Partner
- Direct Connect Partner

170+ YEARS — **Combined AWS Professional Services Experience**

2.2k+ DAYS — **Experience Managing AWS Since 2010**

10x WINNER — **Stevie Award for Customer Service**

9+ MILLION — **Monthly Instance Hours Managed**

4 CONTINENTS — **Global Presence and support Datapipe has Data Center locations in 29 datacenters across 9 countries**

Audited AWS Premier Consulting Partner and AWS Direct Connect Partner

Proprietary Keyless Management

DATAPIPE
a rackspace company

DATA LAKES – A MARITIME USE CASE

120,000	163,500	2,800	19,000
Tracking vessels	Shipping companies	Ports	Terminals

A requirement to manage a global network of multiple data types, coming from a wide variety of business critical maritime sources

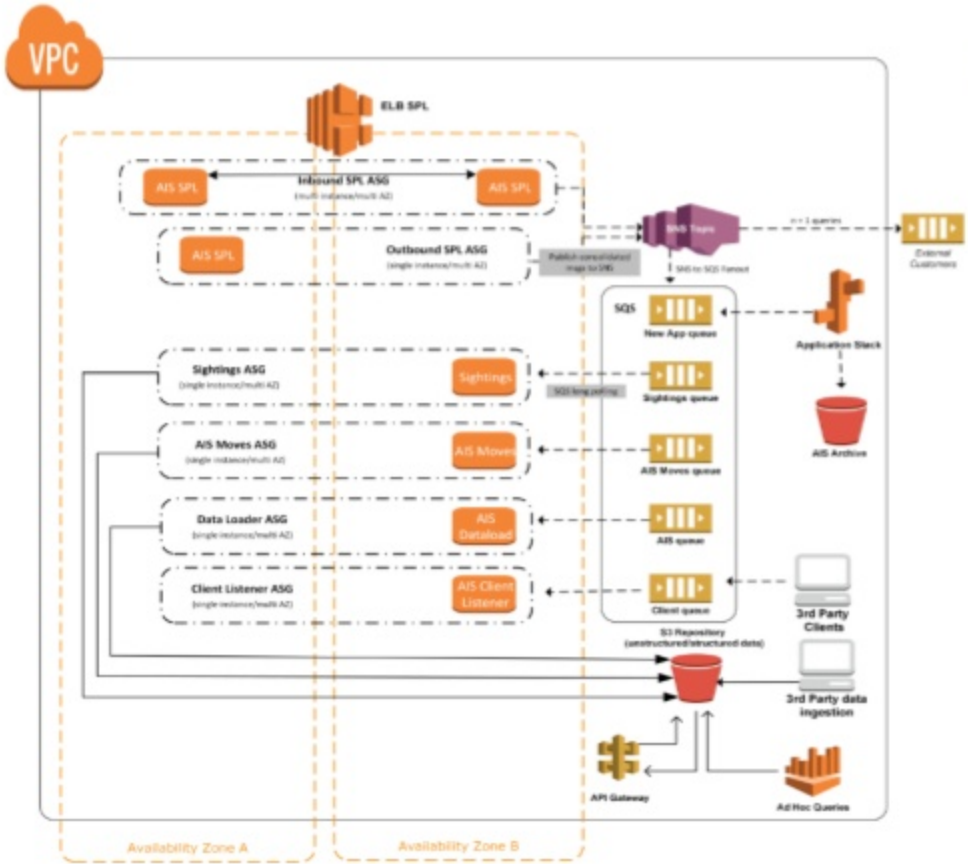
- Vessel tracking via satellite and terrestrial ID data
- Vessel ownership & characteristic data
- Vessel registration, safety and insurance data
- Casualty, arrest and vessel inspection data
- Fleet development data
- Global Ports and Terminal data

DATAPIPE
a redspace company

DEVELOP A MIGRATION STRATEGY

DATAPIPE
a rackspace company

DATA LAKE ARCHITECTURE



CLIENT METHODOLOGY:

INGEST

STORE

DELIVER

DATAPIPE

Best Practices for Building Your Data Lake on AWS

Presented by
Derwin McGeary
Solutions Architect, Cloudwick EMEA

Contact us: services@cloudwick.com

Webinar Agenda

- About Cloudwick
- Reference Data Lake Architecture on AWS
- AWS Data Lake Implementations
- How to Get Started

About Cloudwick

7
Founded
in 2010

3
Global Offices
US | EMEA | APAC

150+
Professionals
US | EMEA | APAC

100+
Data Lake
Engagements

amazon web services | Partner Network
ADVANCED TECHNOLOGY PARTNER
BIG DATA COMPETENCY

400+
Certifications
Big Data, Analytics, Cybersecurity
Visualization & Cloud Partnerships

SOLARFLARE · amazon webservices · MESOSPHERE · databricks · cloudera
PROTECTWISE · H2O.ai · intel · BRICATA · graphistry · sqrrl
DATASTAX · confluent · MAPR · HORTONWORKS · Microsoft Azure

Global
1000
Proven Success

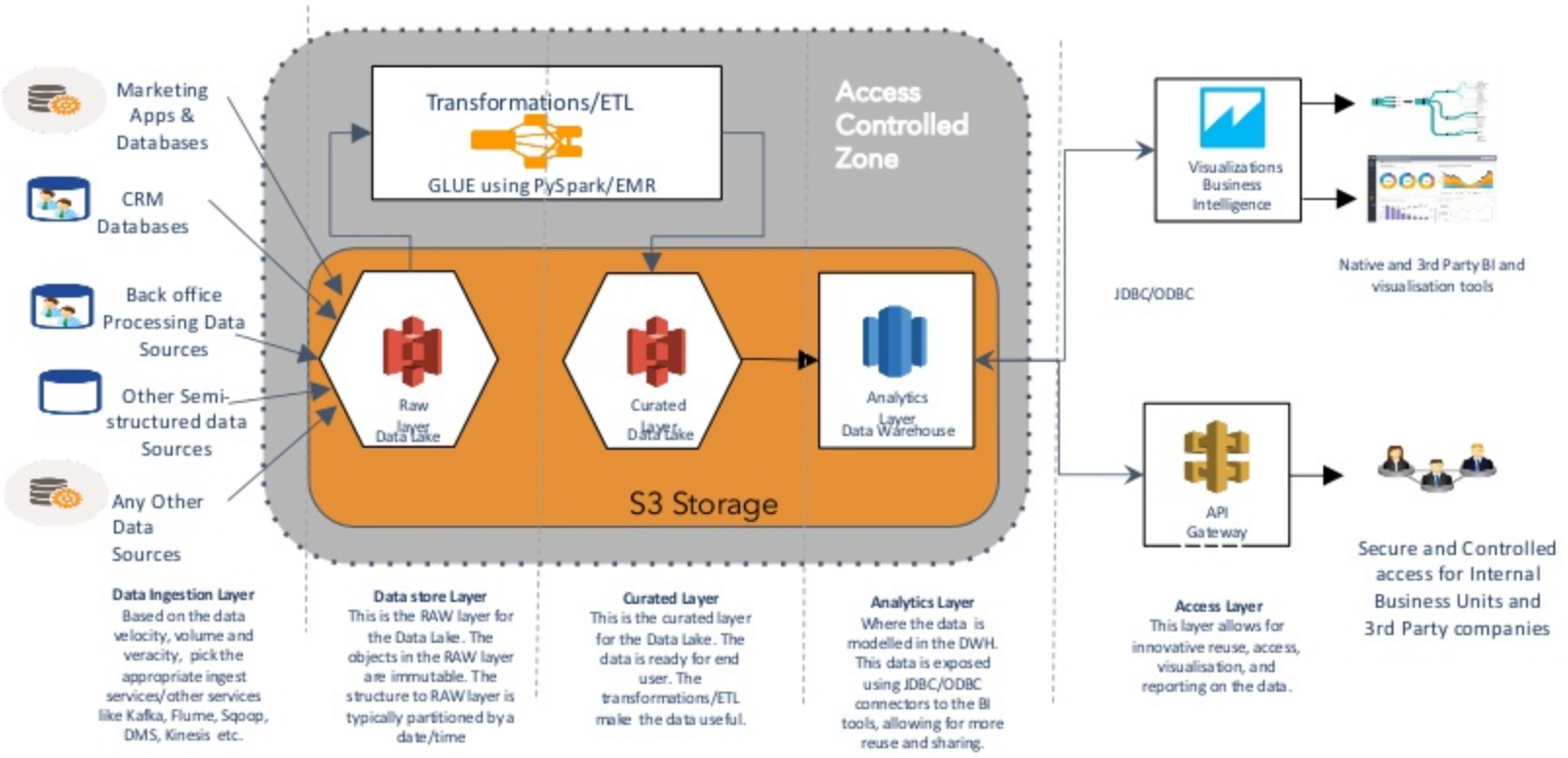
Bank of America · PayPal · VISA · HSBC · LLOYD'S
FINRA · intuit · TARGET · Walmart · SAMSUNG · 3M

aws

Contact us: services@cloudwick.com

Cloudwick

Reference Data Lake Architecture on AWS



AWS Data Lake Implementations

SGN

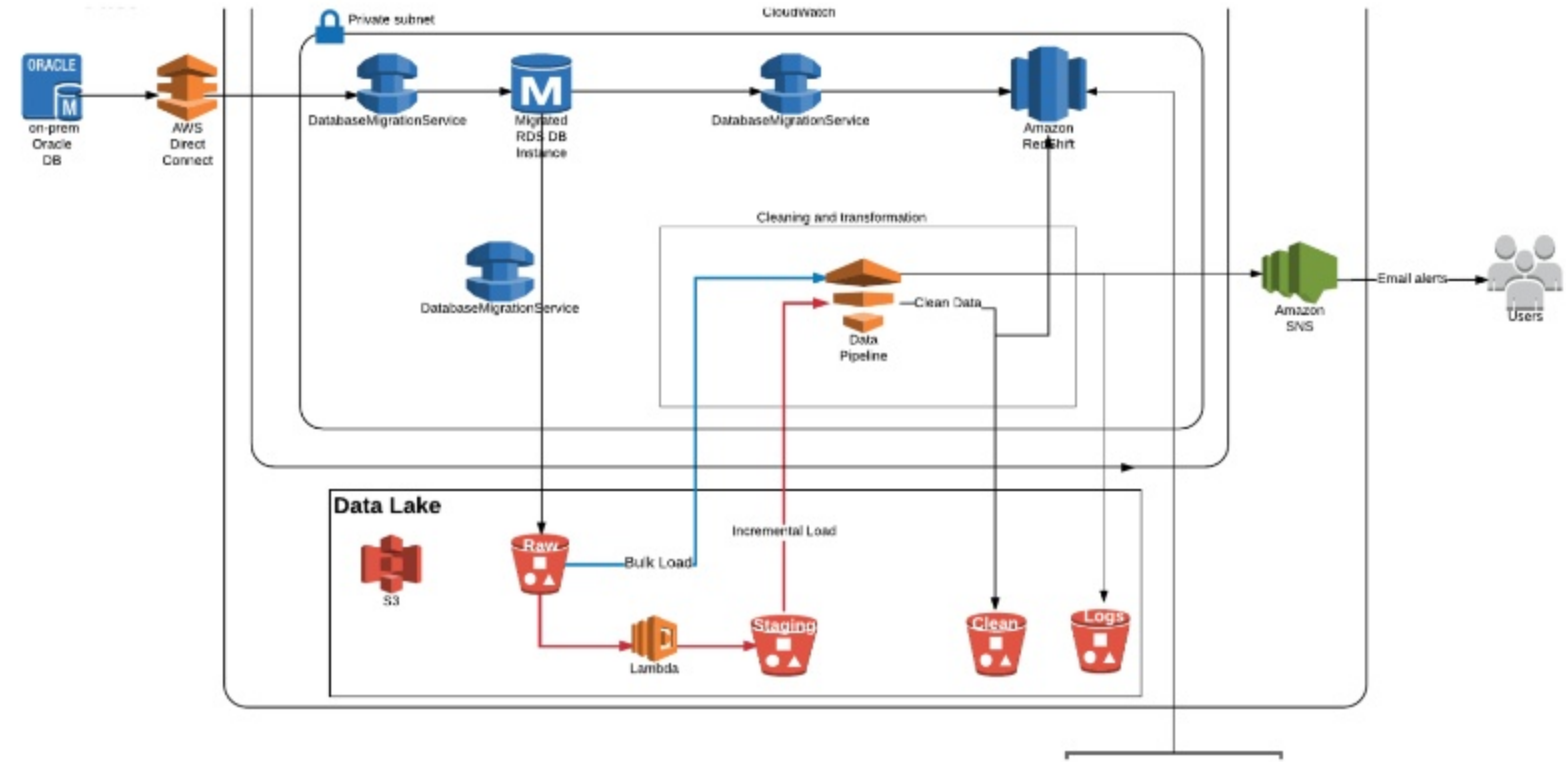
Your gas. Our network.

UK Data Service

SGN Use Case : Background & Challenges

- One of the Largest Gas Distribution Company
- Reporting Responsibilities
- Data Silos
- Challenges
  - Growing Data Sources
  - Existing ETL process
  - Too many silos
  - Data Provision for different use cases
  - Operational and Analytical Reporting

Use Case: Data Lake and Migration for SGN



Where to start?

Data Inventory - what data/metadata do you already have?

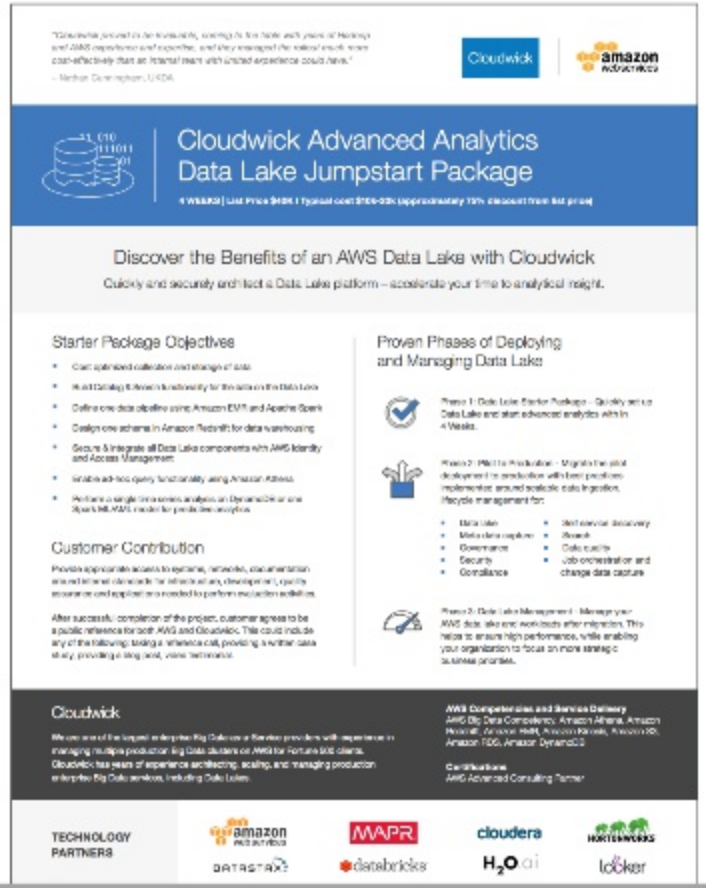
Gap Analysis - what business and technical metadata do you need for your use cases?

Value: Knowing the value of data and getting value from data

Cloudwick can help you get started:

Cloudwick Quickstart
http://tinyurl.com/CloudwickQuickStart

Cloudwick Jumpstart
http://tinyurl.com/CloudwickJumpStart



Thank you!