



Best Practices for Building Your Data Lake on AWS

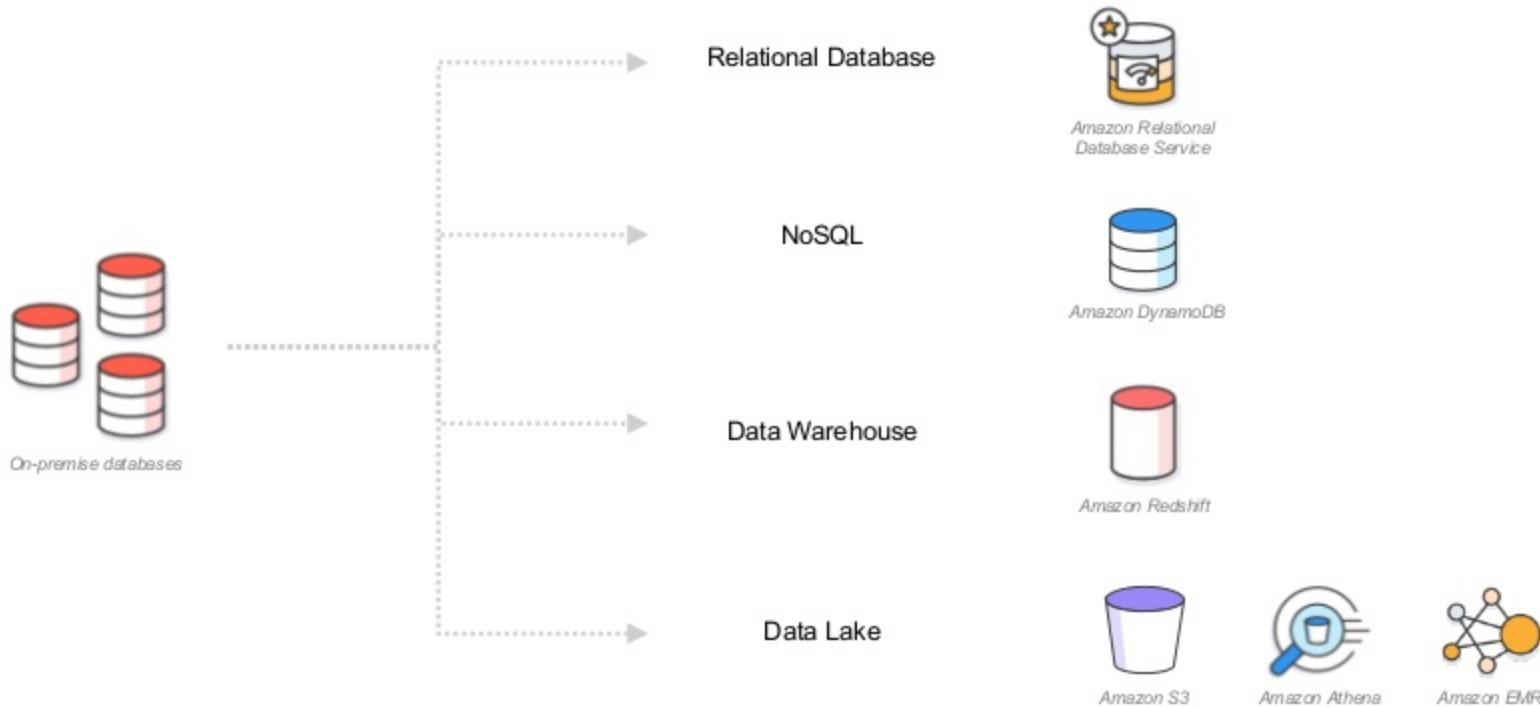


Ian Robinson, Specialist SA, AWS

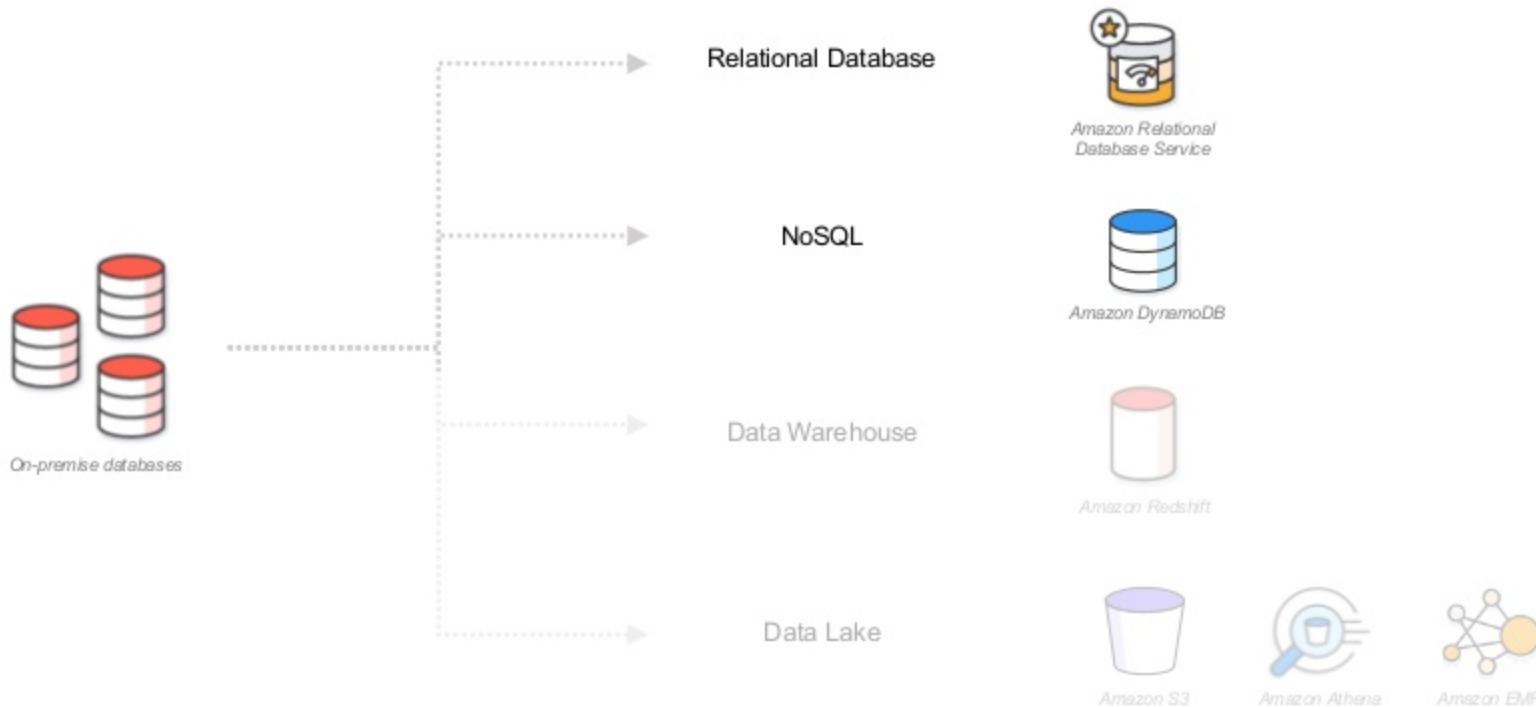
Kiran Tamana, EMEA Head of Solutions Architecture,
Datapipe

Derwin McGeary, Solutions Architect, Cloudwick

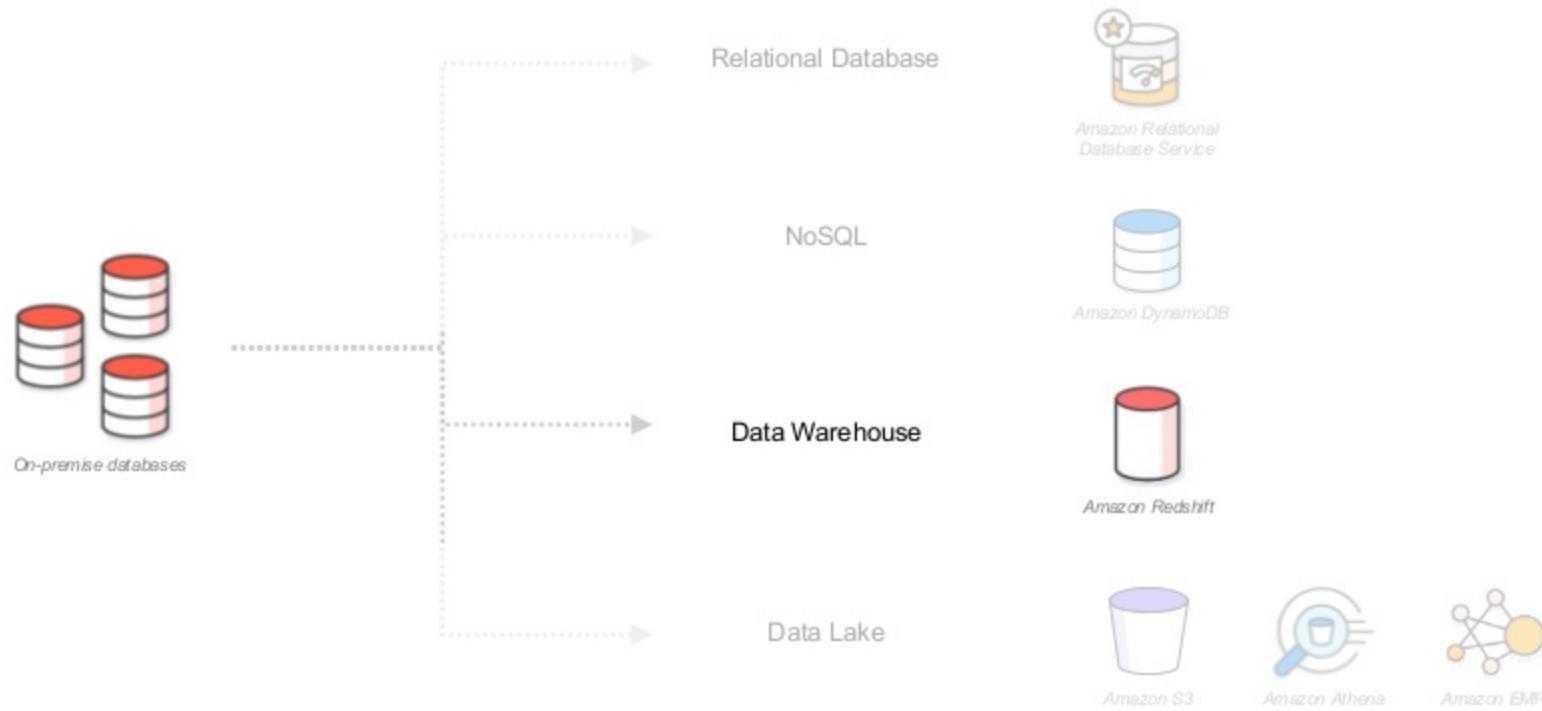
Webinars: Database Freedom



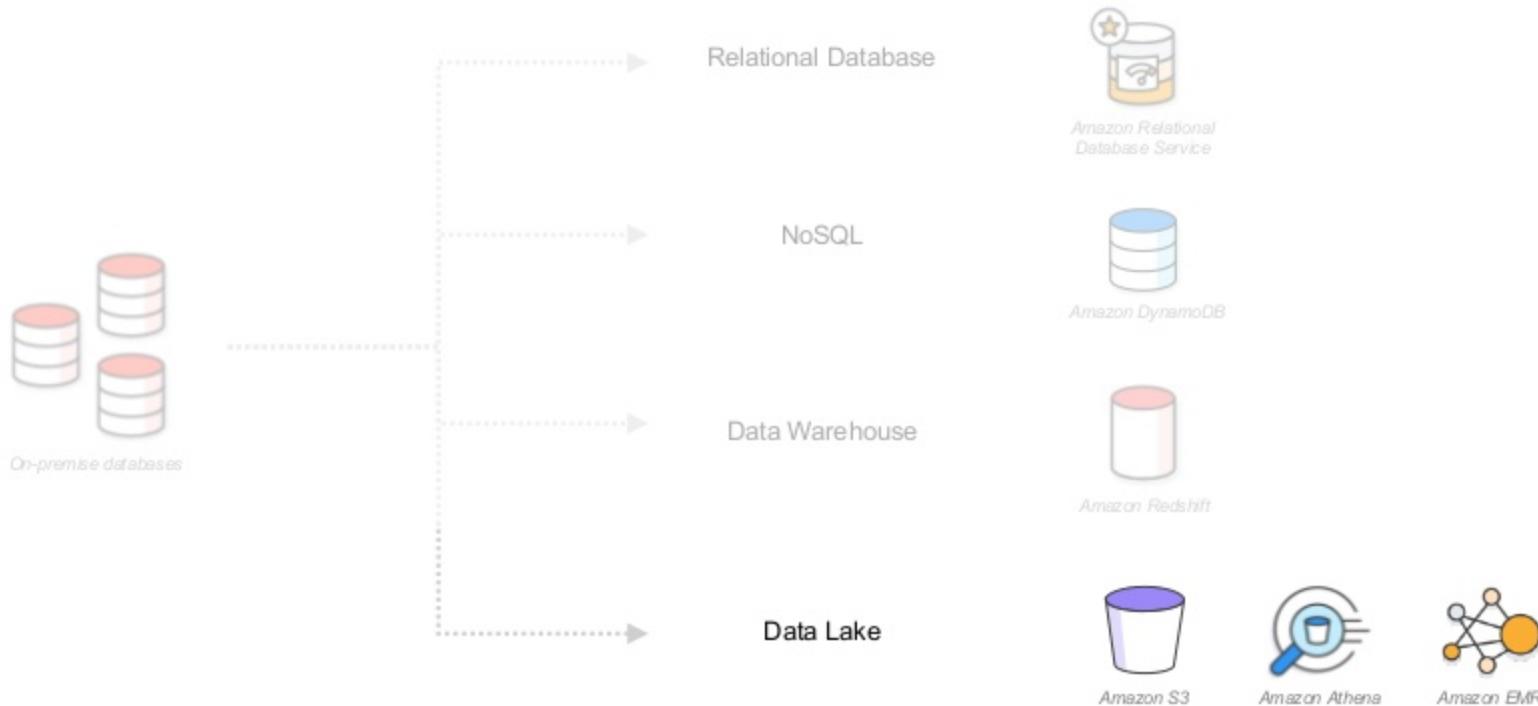
Webinar 1: Migrating Online/Transactional Workloads



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



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



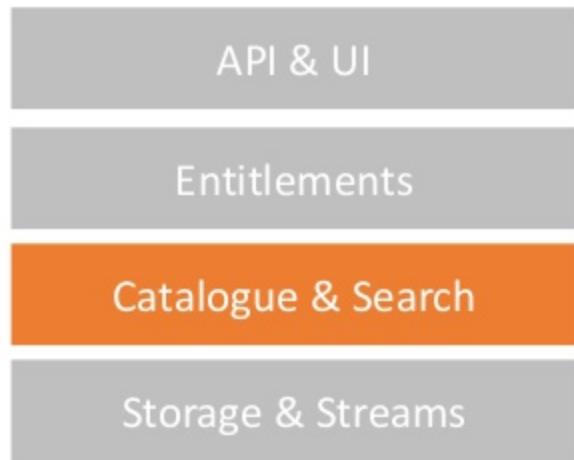
Data Storage

- 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



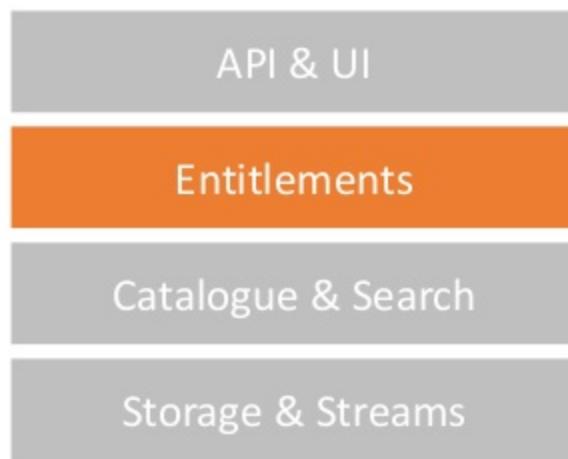
Catalogue

- 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

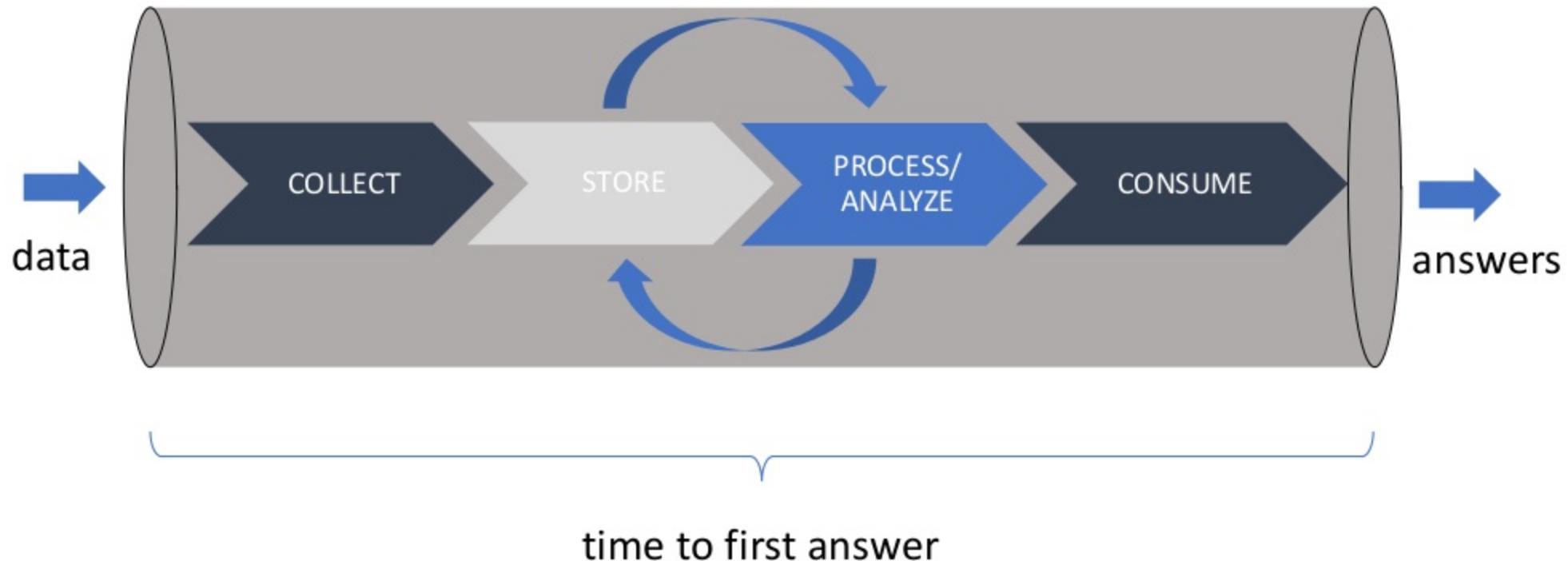
Components of a Data Lake



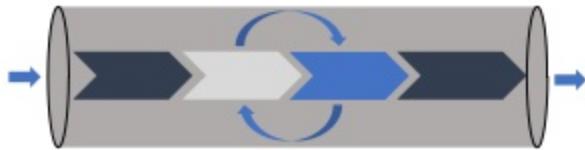
API & User Interface

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

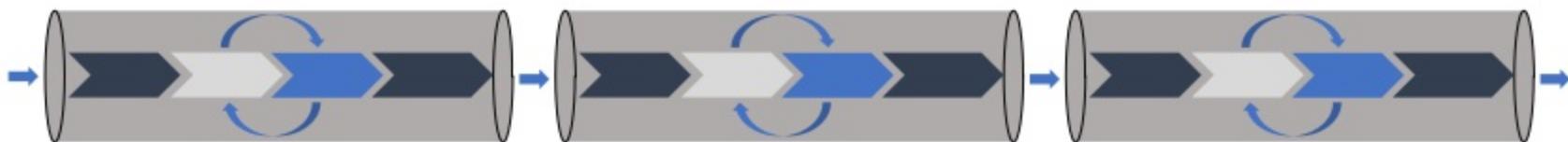
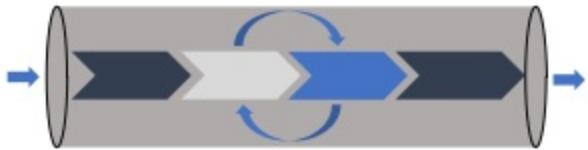
The Modern Data Architecture



Agile Analytics

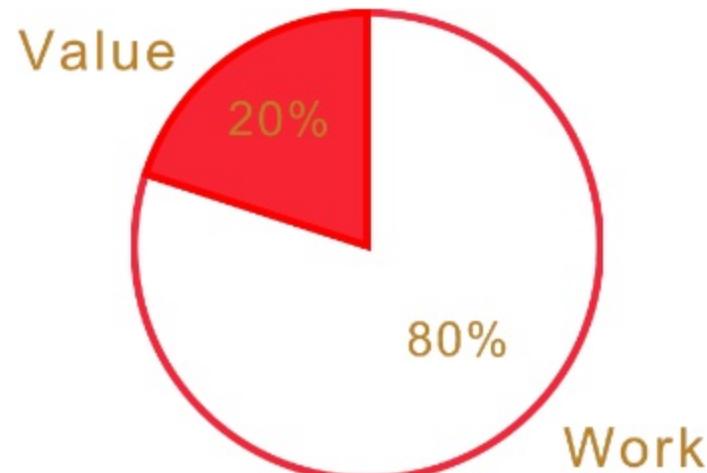


- Experiment
- Invest in promising experiments
- Fail fast
- React quickly



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



S3



EFS



EBS

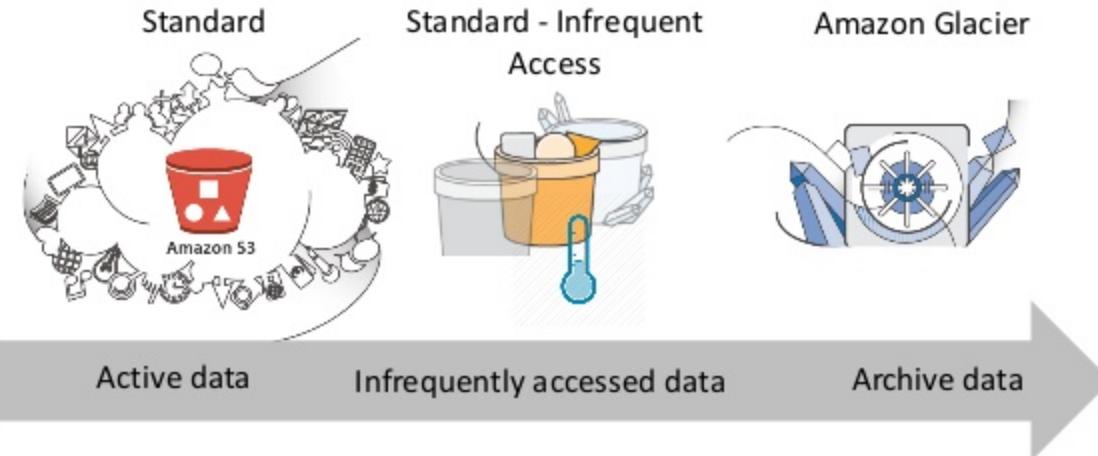
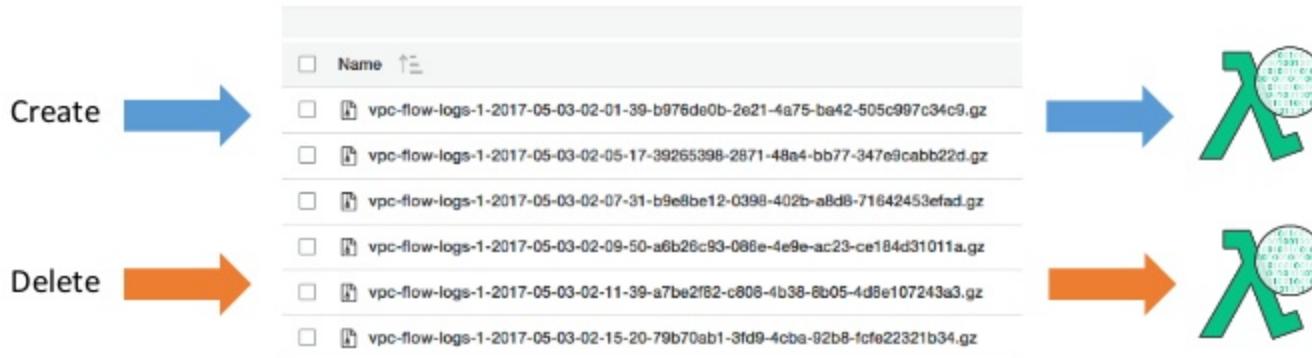


DynamoDB



RDS

Events and Lifecycle Management



S3 as the Data Lake Fabric

- **Unlimited** number of objects and volume
- **99.99% availability**
- **99.99999999% 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

Storage + Catalog

Scalable (secure, versioned, durable) storage +
Immutable data at every stage of its lifecycle +

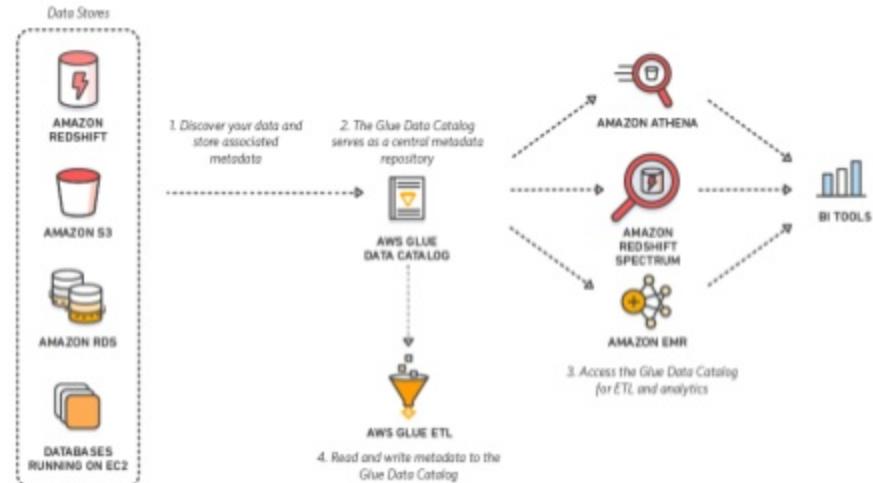
Versioned schema and metadata

=

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:

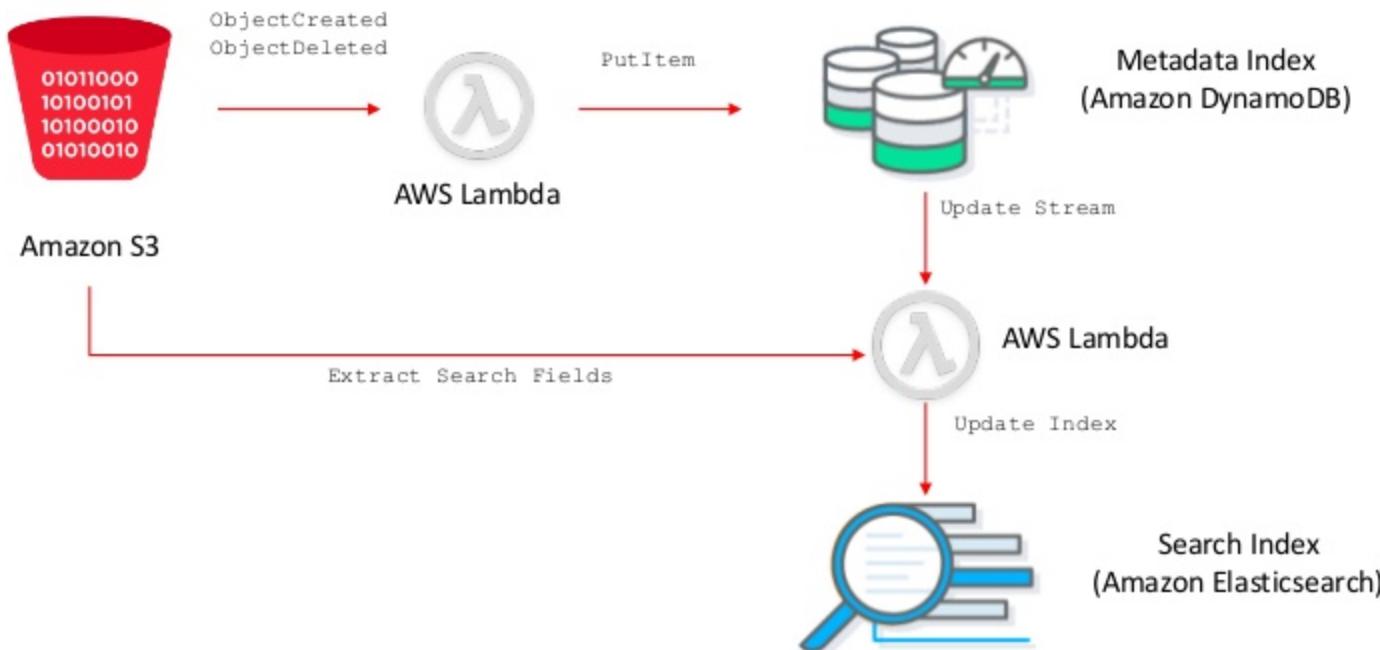
- **Search** metadata for data discovery
- **Connection info** – JDBC URLs, credentials
- **Classification** for identifying and parsing files
- **Versioning** of table metadata as schemas evolve and other metadata are updated
- **Table definitions** – usable by Redshift, Athena, Glue, EMR

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

The screenshot shows the AWS Glue Data Catalog interface. The left sidebar navigation includes: AWS Glue, Data catalog (selected), Databases, Tables (highlighted in orange), Connections, Crawlers, Classifiers, ETL, Jobs, Triggers, Dev endpoints, Tutorials, Add crawler, Explore table, Add job, and Resources. The main content area displays a table of tables with columns: Name, Database, and Classification. A search bar at the top right says "Filter by attributes or search by keyword".

Name	Database	Classification
dataset228	glue-sample	csv
ianrob_nyc_transportation	nyc_transportation_canonical	parquet
limo	nyc-transportation	csv
salesdb_customer	glue-demo-mysql	mysql
salesdb_customer_site	glue-demo-mysql	mysql
salesdb_product	glue-demo-mysql	mysql
salesdb_product_category	glue-demo-mysql	mysql
salesdb_sales_order	glue-demo-mysql	mysql
salesdb_sales_order_all	glue-demo-mysql	mysql
salesdb_sales_order_detail	glue-demo-mysql	mysql
salesdb_supplier	glue-demo-mysql	mysql
taxi	nyc-transportation	csv
uber	nyc-transportation	csv

Indexing and Searching Using Metadata



Governance
Security
Privacy

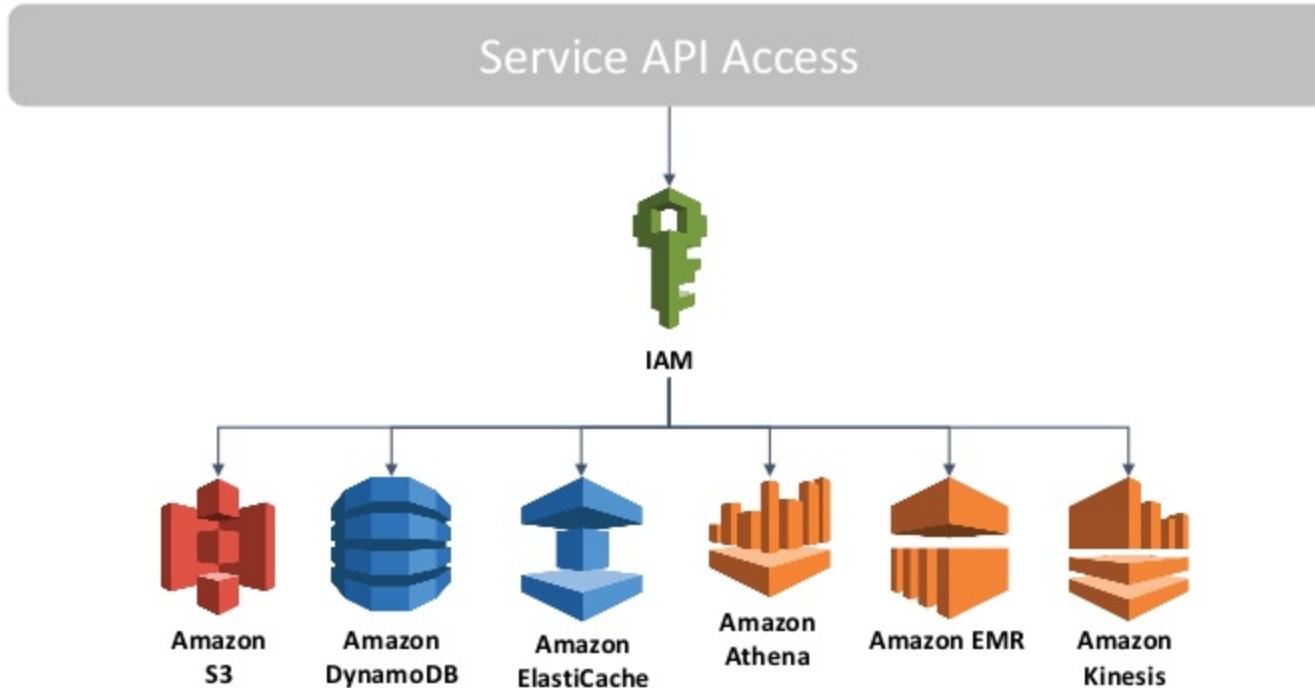


Identity and Access Management

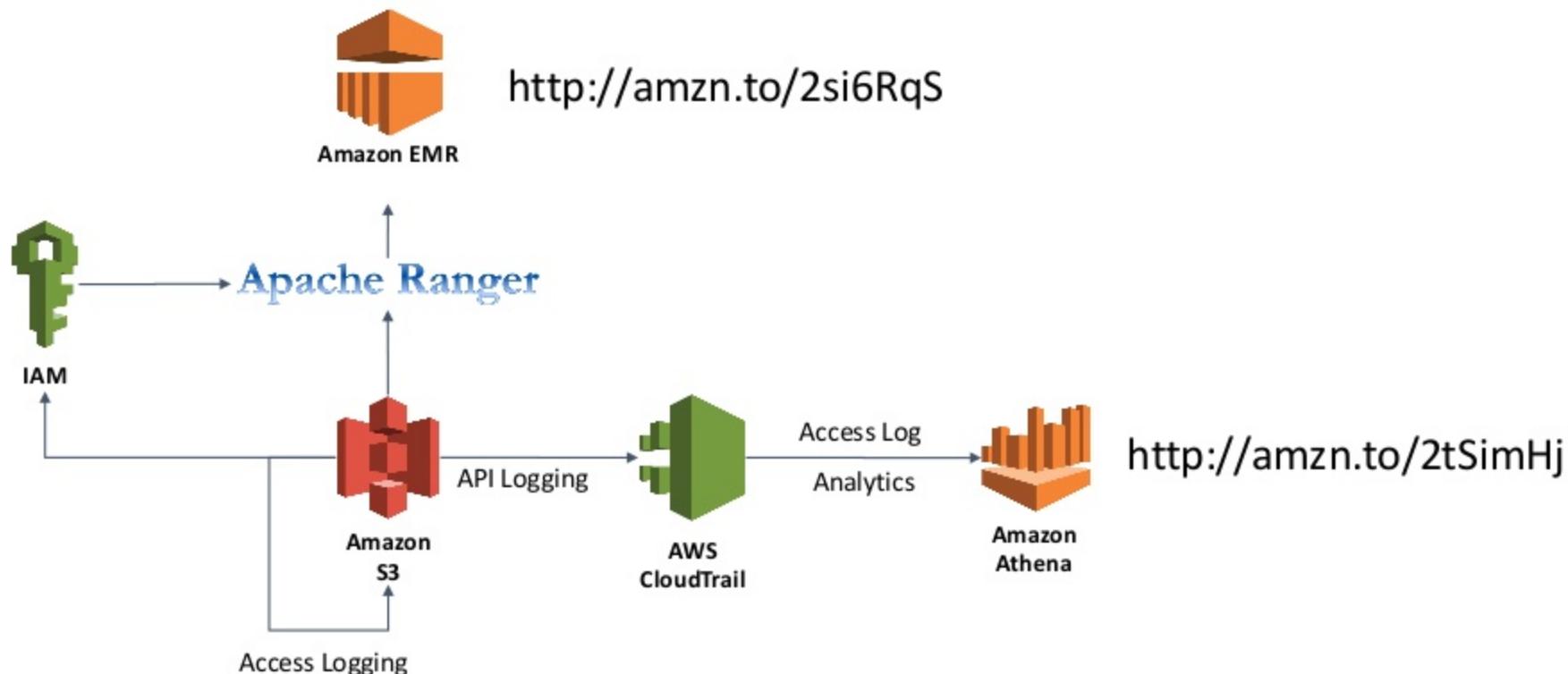


- **Manage** users, groups, and roles
- Identity **federation** with Open ID
- Temporary credentials with Amazon Security Token Service (Amazon **STS**)
- Stored **policy templates**
- Powerful **policy language**
- Amazon **S3 bucket policies**

Security at the Data Level



Third Party Ecosystem Security Tools



Storage Level Support for Access Logging and Audit



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

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

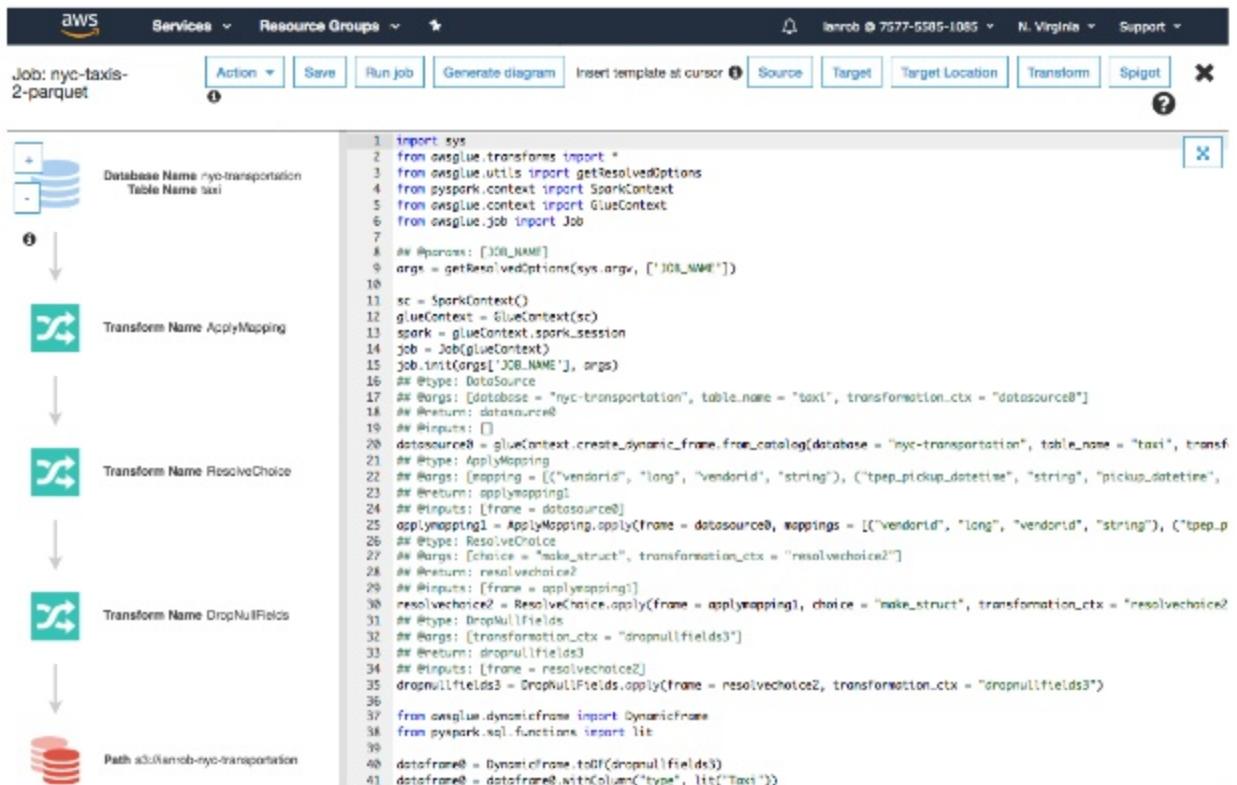


- Host multiple **versions** and stages of APIs
- Create and **distribute** API keys to developers
- Leverage AWS Sigv4 to **authorize access** to APIs
- **Throttle** and **monitor** requests to protect the backend
- Leverages **AWS Lambda**

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

The screenshot shows the AWS Glue Job Execution interface. At the top, there are buttons for "Add job" and "Action" (with a dropdown arrow), a search bar labeled "Filter by attributes", and navigation links for "Showing: 1 - 6" and "More".

The first table lists jobs:

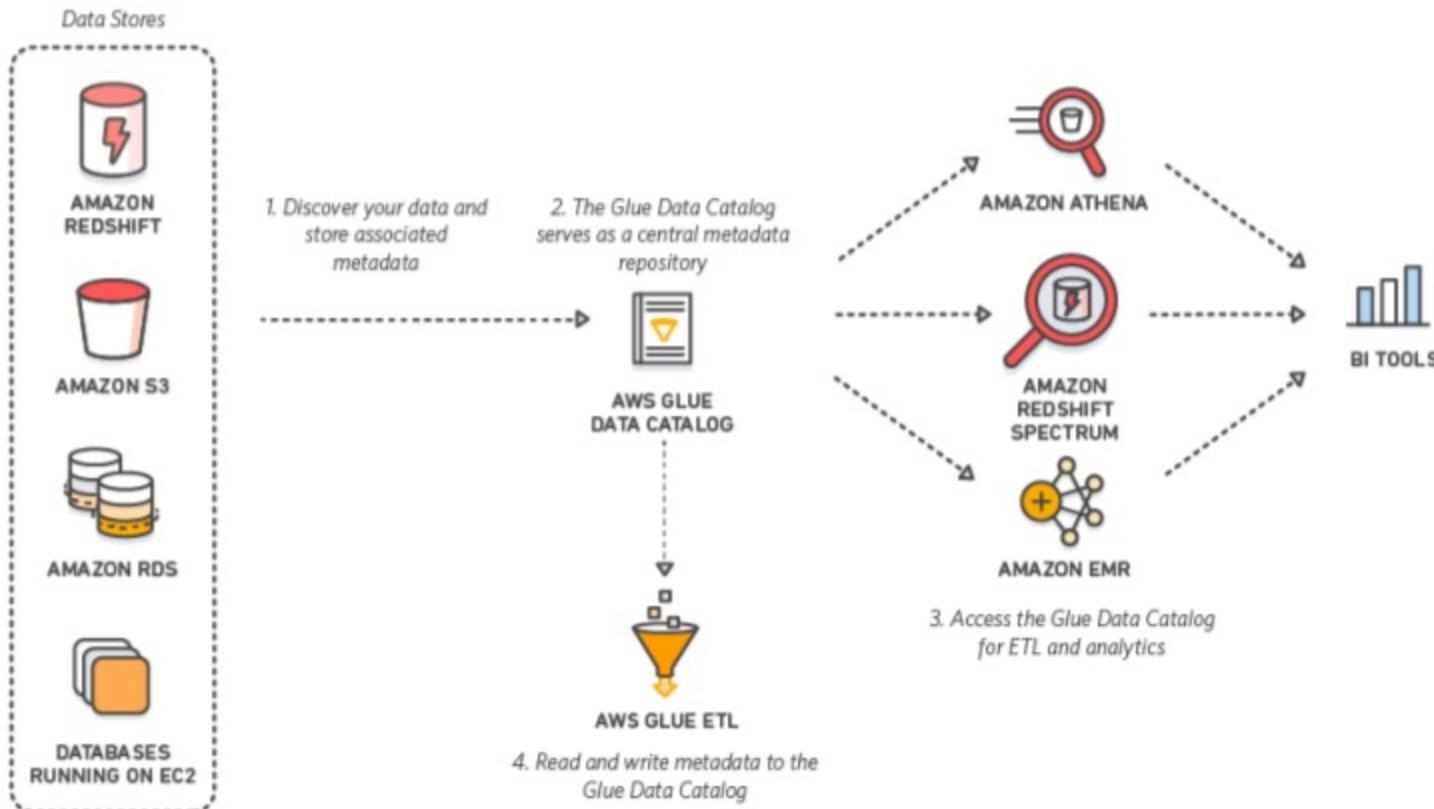
Name	Script location	Last modified	Job bookmark
<input checked="" type="checkbox"/> nyc-taxis-2-parquet	s3://aws-glue-scripts-75775585108...	13 September 2017 2:44 PM UTC+1	Disable
<input type="checkbox"/> Load_Dimension_SUPPLIER_DIM	s3://aws-glue-scripts-75775585108...	25 August 2017 1:35 PM UTC+1	Disable

The second table lists job runs:

Run ID	Retry attempt	Run status	Error	Logs	Error logs	Duration	Triggered by	Start time	End time
jr_3b03c82...	-	Succeeded		Logs		20 mins	LOAD_NYC...	19 Septemb...	19 Septemb...
jr_7308ce28...	-	Succeeded		Logs		16 mins	LOAD_NYC...	19 Septemb...	19 Septemb...
jr_b2c1cf01...	-	Succeeded		Logs		20 mins	LOAD_NYC...	15 Septemb...	15 Septemb...
jr_ddff0ffef0f	-	Succeeded		Logs		14 mins	LOAD_NYC...	13 Septemb...	13 Septemb...



Instantly Query Your Data Lake on S3



AWS Batch



Fully managed



Dynamic provisioning
and scaling

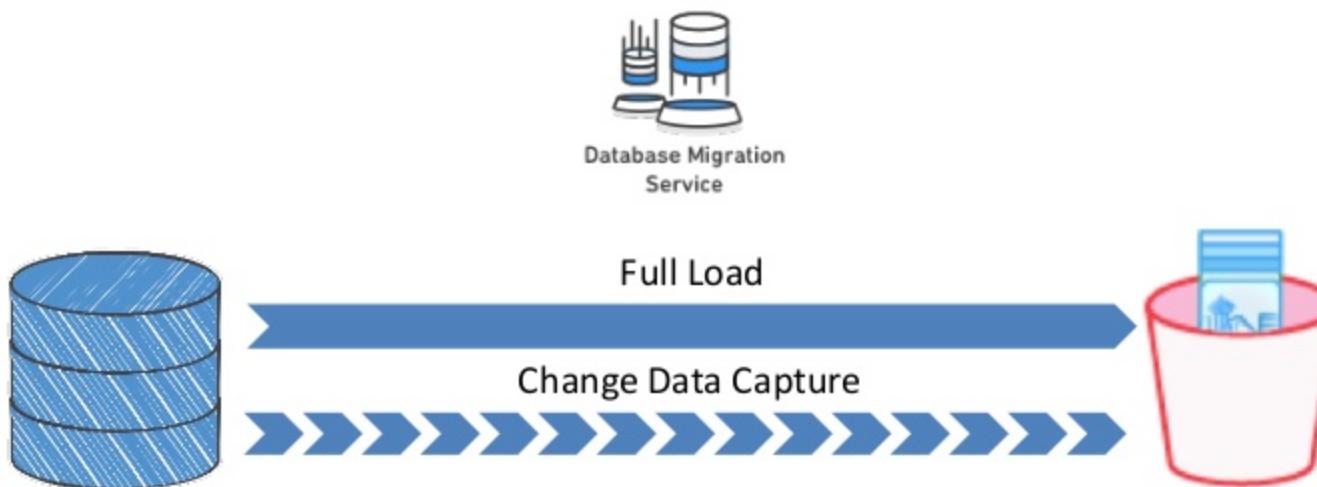


Cost optimization
through EC2 Spot fleet



Priority-based queues
and scheduling

Write Database Changes to S3 with DMS



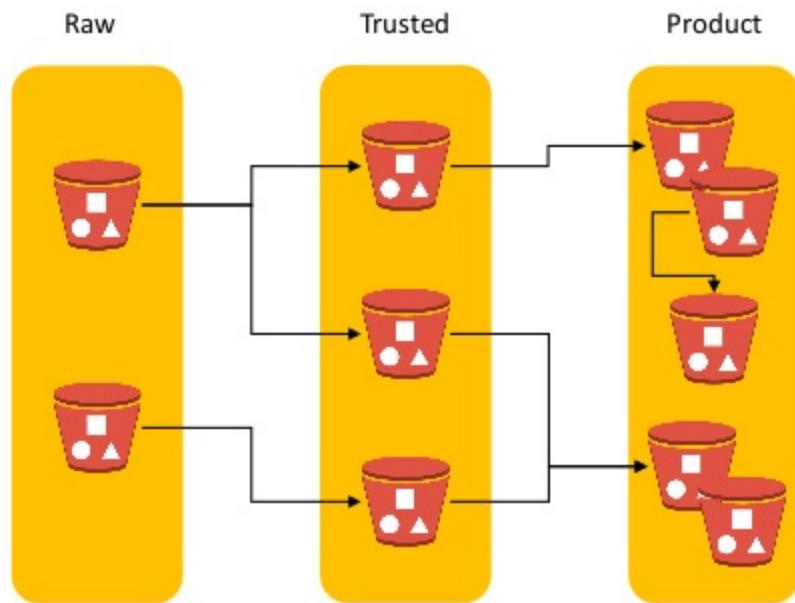
```
<schema_name>/<table_name>/LOAD001.csv  
<schema_name>/<table_name>/LOAD002.csv  
<schema_name>/<table_name>/<time-stamp>.csv
```

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

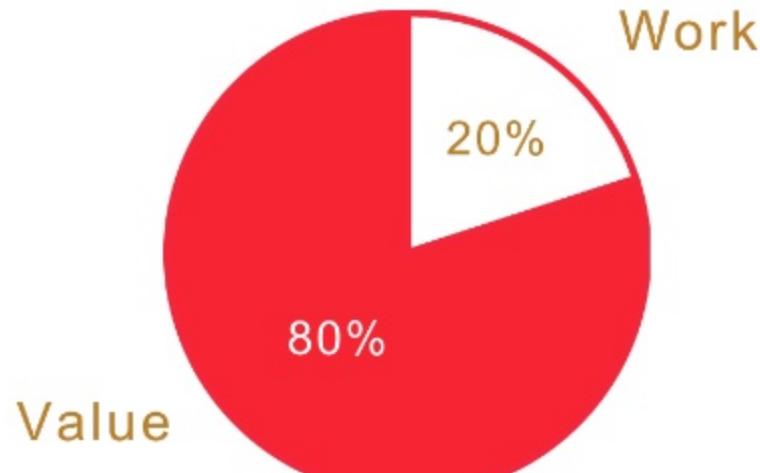
AWS CloudTrail and S3 Access Logging



Shifting Analytics
So It's

80%
Analytics

And Only
20%
Prep



Building a Data Strategy on AWS



Analytics Capabilities



Data
processing



Data
warehousing



Reporting

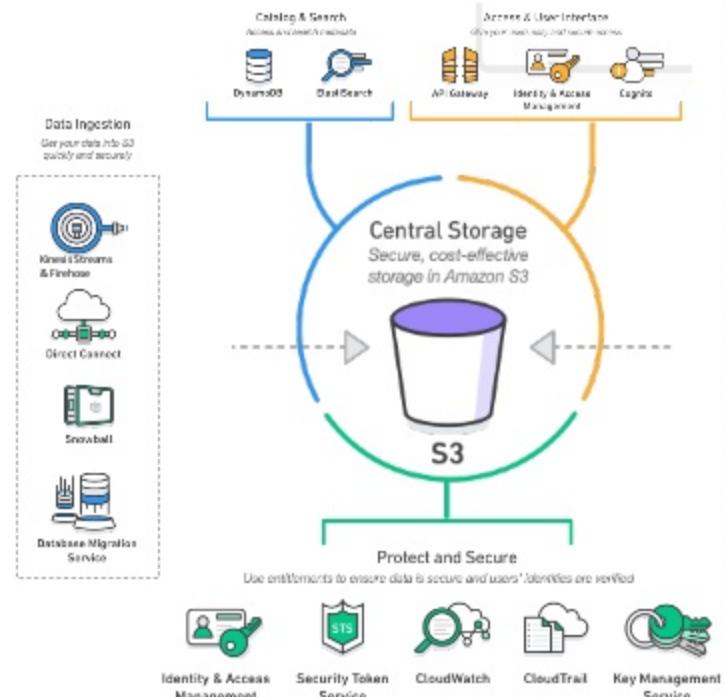


Real-time
processing



Predictive
analytics

Processing & Analytics



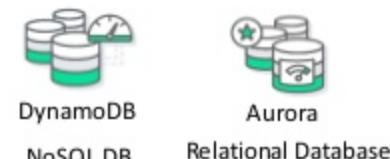
Real-time



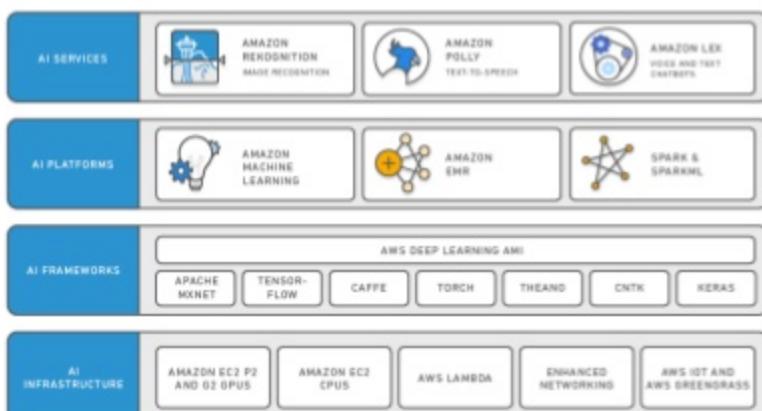
Batch



Transactional & RDBMS



Predictive



BI & Data Visualization



AWS Partner Ecosystem

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

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

The screenshot shows the top portion of the AWS Big Data Partner Solutions page. It features a blue header with the AWS logo and the text "AWS Big Data Partner Solutions". Below the header is a navigation bar with links: Big Data Management, Data Integration, Data Governance, Advanced Analytics, Data Analysis & Machine Learning, and Consulting Services.

The screenshot shows the "Data Integration" section of the AWS Big Data Partner Solutions page. It features a blue header with the text "Data Integration" and a sub-header "Reduce the effort to extract, transform, load data and incorporate streaming workflows". Below this is a grid of partner logos and descriptions:

- alooma**: Allooma Data Pipeline. They help you integrate and manage schema changes. [Learn more](#).
- alteryx**: Alteryx. Simplify analysis by building in Alteryx. Built for leading big data and machine learning. [Learn more](#).
- ATTUNITY**: ATTUNITY. Help you manage and analyze data, and build data-driven insights for strategy and decision-making. [Learn more](#).
- b.you**: b.you. Data integration made easy. [Learn more](#).
- informatica**: Informatica. Accelerate data integration and management. [Learn more](#).
- ironSource**: ironSource. IronSource's Data Integration Platform. [Learn more](#).
- marketo**: Marketo. Marketo's Data Integration Platform. [Learn more](#).
- snapLogic**: snapLogic. Simplify data integration with snapLogic. [Learn more](#).
- talend**: talend. Talend's Information Cloud & Talend Data Fabric. [Learn more](#).





AWS DATA LAKES

Kiran Tamana, EMEA Head of Solutions Architecture, Datapipe



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

DATA LAKES – A MARITIME USE CASE

120,000

Tracking vessels

163,500

Shipping companies

2,800

Ports

19,000

Terminals

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

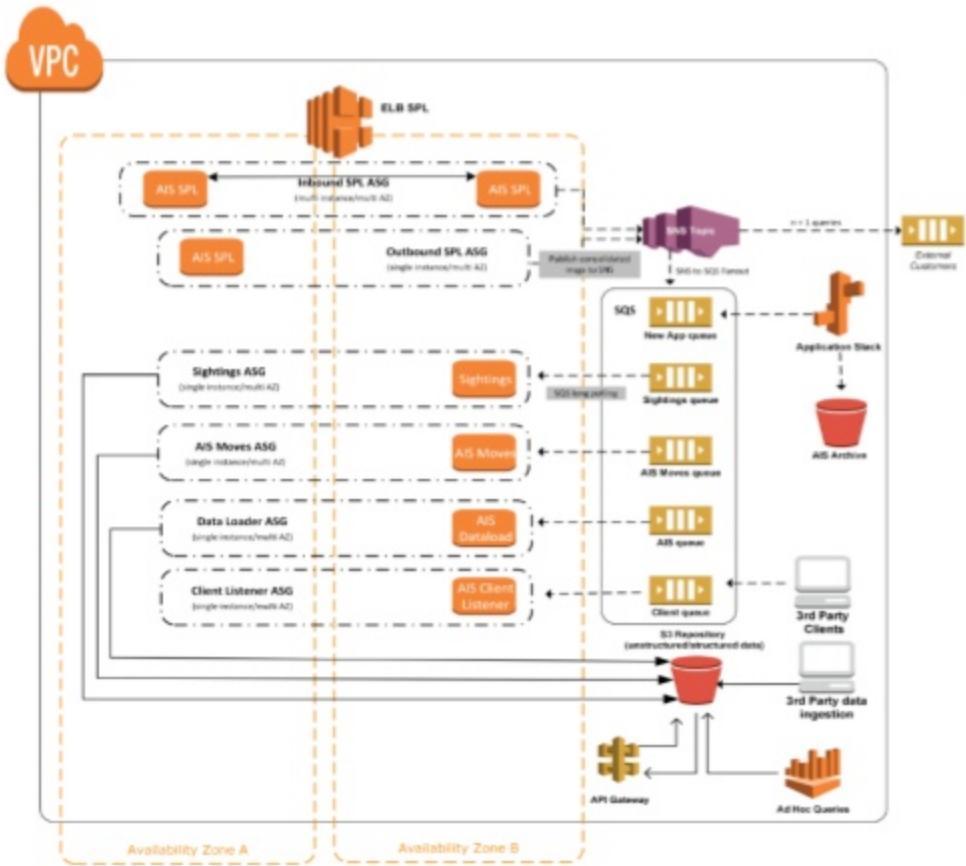
- Vessel tracking via satellite and terrestrial ID data
- Vessel ownership & characteristic data
- Vessel registration, safety and insurance data
- Casualty, arrest and vessel inspection data
- Fleet development data
- Global Ports and Terminal data



DEVELOP A MIGRATION STRATEGY



DATA LAKE ARCHITECTURE



CLIENT METHODOLOGY:

INGEST



STORE



DELIVER



Best Practices for Building Your Data Lake on AWS



*Presented by
Derwin McGeary
Solutions Architect, Cloudwick EMEA*



Contact us: services@cloudwick.com

Cloudwick

Webinar Agenda

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



Contact us: services@cloudwick.com

Cloudwick

About Cloudwick

7

Founded
in 2010

3

Global Offices
US | EMEA | APAC

150+

Professionals
US | EMEA | APAC

100+

Data Lake
Engagements



400+

Certifications

Big Data, Analytics, Cybersecurity
Visualization & Cloud Partnerships



Global
1000

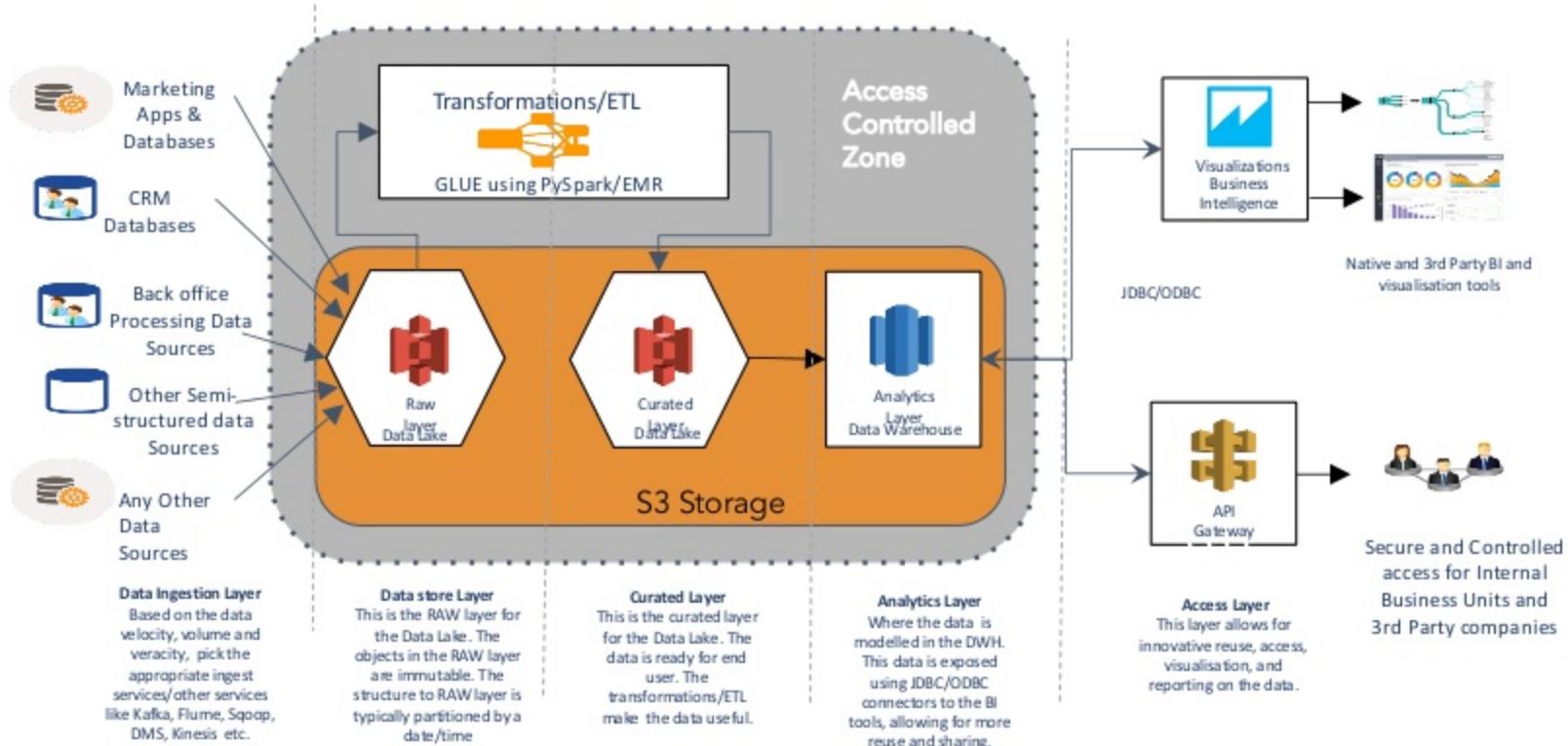
Proven Success



Contact us: services@cloudwick.com

Cloudwick

Reference Data Lake Architecture on AWS



Contact us: services@cloudwick.com

Cloudwick

AWS Data Lake Implementations



UK Data Service



Contact us: services@cloudwick.com

Cloudwick

SGN Use Case : Background & Challenges

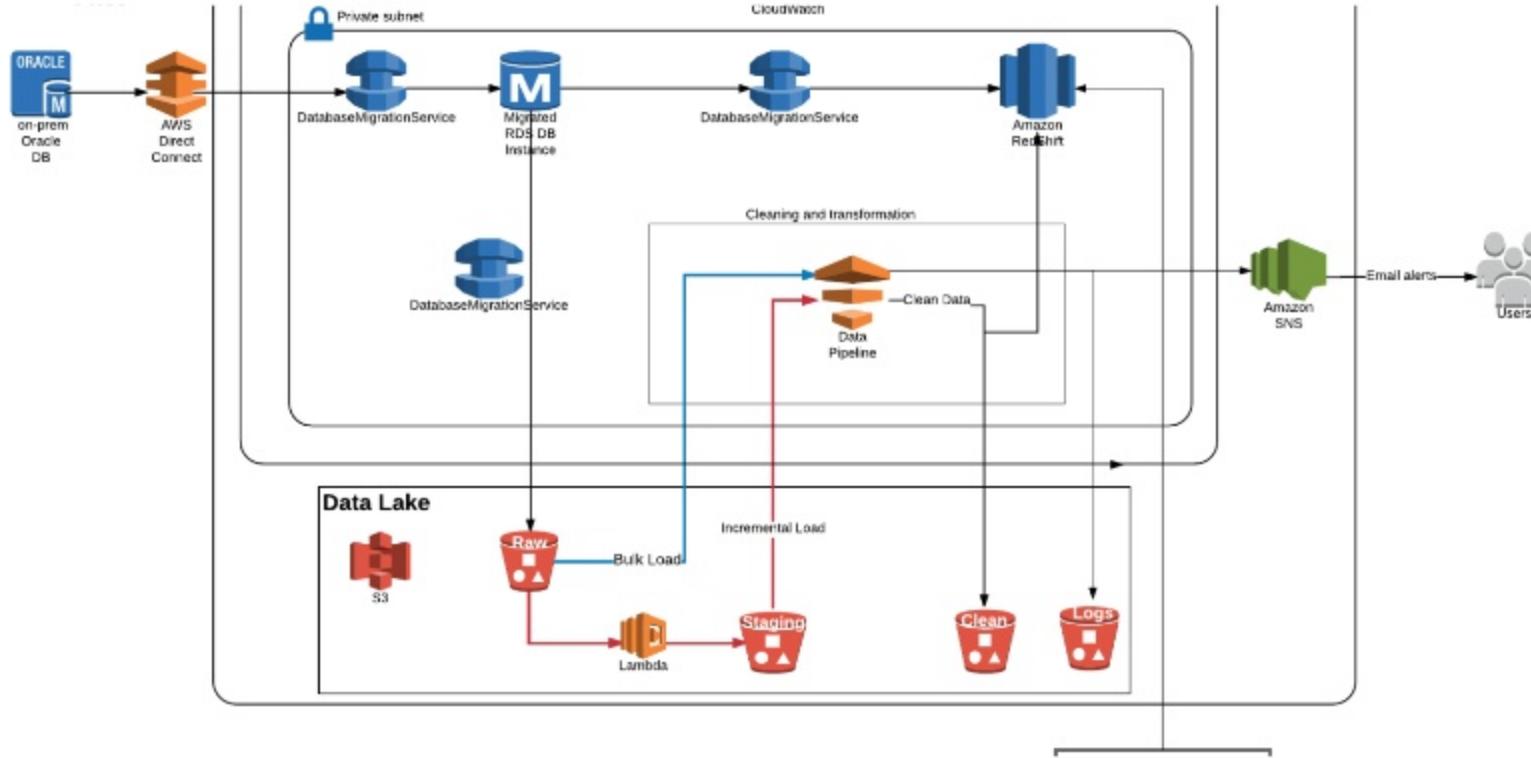
- One of the Largest Gas Distribution Company
- Reporting Responsibilities
- Data Silos
- Challenges
 - Growing Data Sources
 - Existing ETL process
 - Too many silos
 - Data Provision for different use cases
 - Operational and Analytical Reporting



Contact us: services@cloudwick.com

Cloudwick

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

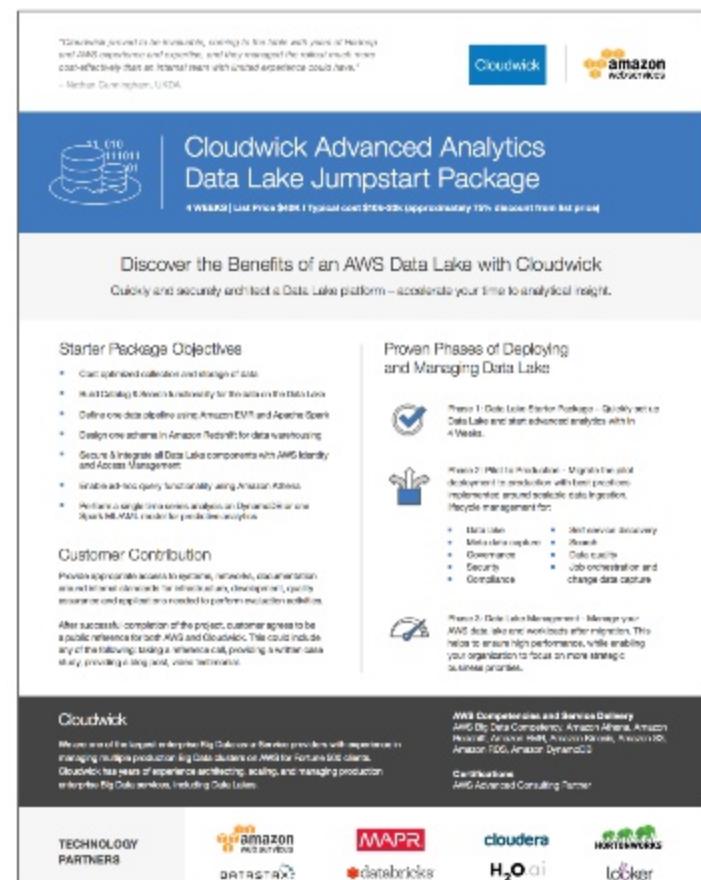
Cloudwick can help you get started:

Cloudwick Quickstart
<http://tinyurl.com/CloudwickQuickStart>

Cloudwick Jumpstart
<http://tinyurl.com/CloudwickJumpStart>



Contact us: services@cloudwick.com



The screenshot shows the Cloudwick Advanced Analytics Data Lake Jumpstart Package page. It features a header with the Cloudwick logo and the Amazon Web Services logo. Below the header, there's a section titled "Discover the Benefits of an AWS Data Lake with Cloudwick". This section includes a sub-section "Starter Package Objectives" with a bulleted list of goals, and a "Proven Phases of Deploying and Managing Data Lake" section with three phases: Phase 1 (Data Lake Starter Package), Phase 2 (ML/IA Production), and Phase 3 (Data Lake Maturity). At the bottom, there's a "Cloudwick" section with a brief description and logos for various technology partners.

TECHNOLOGY
PARTNERS



DATAGRAX



databricks



H2O.ai



locker

Cloudwick



Thank you!

