Architecting a Data Lake on AWS

November 3, 2016

amazon web services | Partner Network

Today’s speakers

Rahul Bhartia
Ecosystem Solution Architect
Amazon Web Services

Mark Schreiber
General Manager, Cloudwick

Raza Shaikh
CTO, NorthBay

Mick Bass
CEO, 47Lining

Agenda

- What is a Data Lake?
- Benefits of a Data Lake
- Why deploy your Data Lake on AWS?
- Deploying your Data Lake with featured APN Big Data Competency Partners
  - Cloudwick
  - NorthBay
  - 47Lining
- Getting Started

What is a Data Lake?

Data Lake is a new and increasingly popular way to store and analyze massive volumes and heterogenous types of data in a centralized repository.

Solving big data challenges with a Data Lake

The volume, variety, and velocity at which data is being generated are leaving organizations with new questions to answer, such as:

“How can I collect data quickly from various sources and store it efficiently?”

“Why is the data distributed in many locations? Where is the single source of truth ?”

“How can I scale up with the volume of data being generated?”

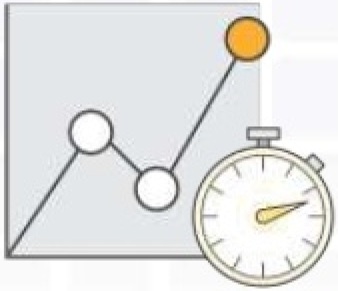
“Is there a way I can apply multiple analytics and processing frameworks to the same data?”

Benefits of a Data Lake

"How can I collect data quickly from various sources and store it efficiently?"

Benefits of a Data Lake

“How can I collect data quickly from various sources and store it efficiently?”



Quickly ingest data without needing to force it into a pre-defined schema.

Benefits of a Data Lake

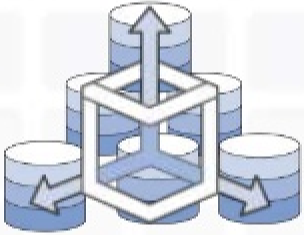
"Why is the data distributed in many locations? Where is the single source of truth ?"

Benefits of a Data Lake

"Why is the data distributed in many locations? Where is the single source of truth ?"

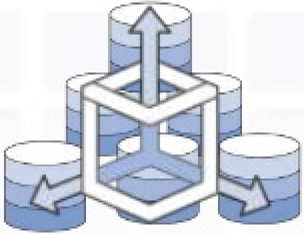
Store and analyze all of your data, from all of your sources, in one centralized location.

Benefits of a Data Lake

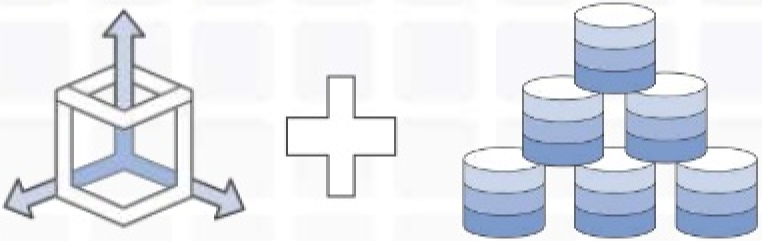


“How can I scale up with the volume of data being generated?”

Benefits of a Data Lake

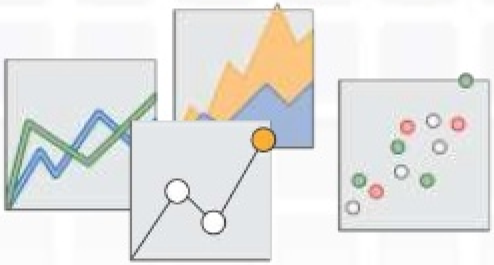


“How can I scale up with the volume of data being generated?”



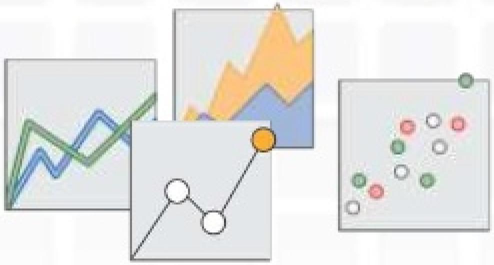
Separating your storage and compute allows you to scale each component as required

Benefits of a Data Lake



"Is there a way I can apply multiple analytics and processing frameworks to the same data?"

Benefits of a Data Lake



"Is there a way I can apply multiple analytics and processing frameworks to the same data?"

A Data Lake enables ad-hoc analysis by applying schemas on read, not write.

Important components of a Data Lake

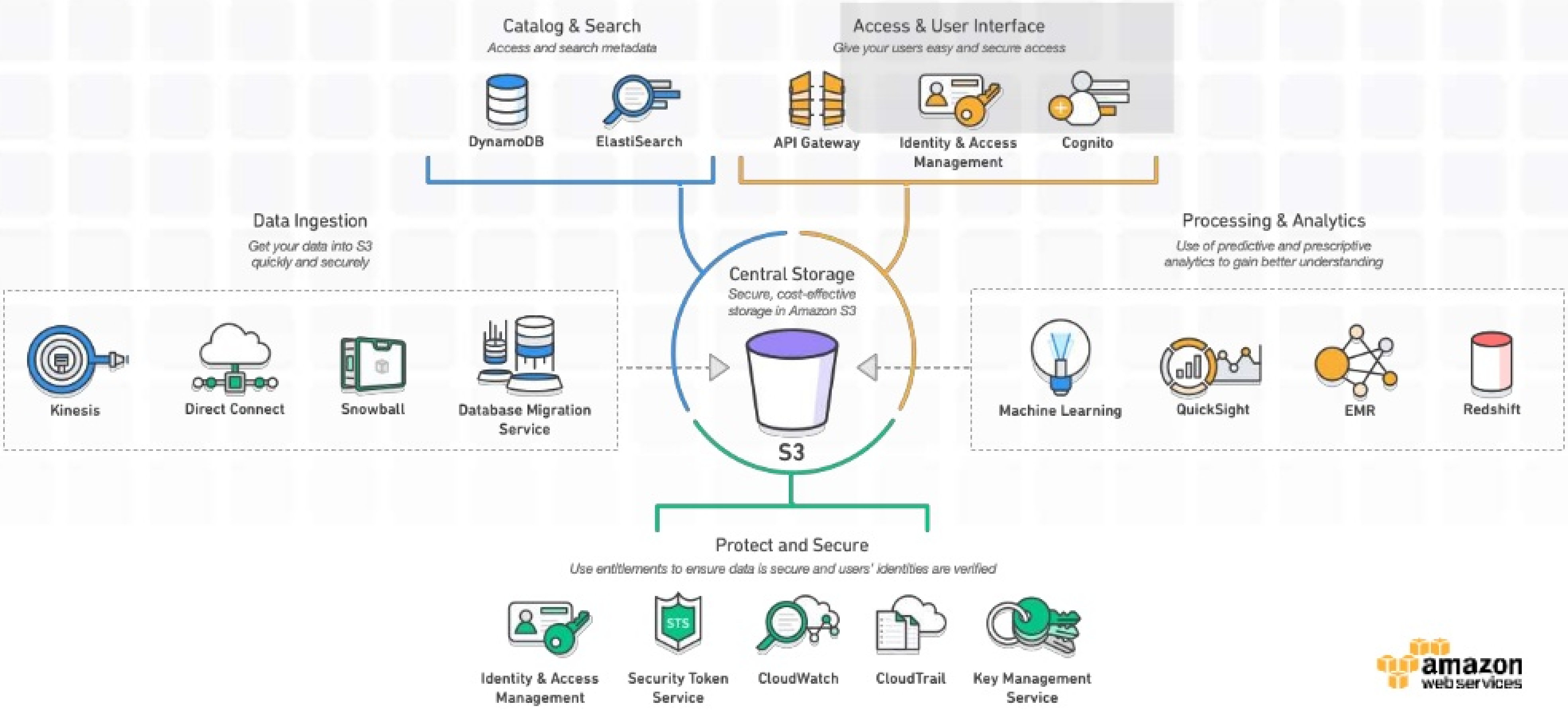
Data Ingestion

Catalogue & Search

Protect & Secure

Access & User Interface

Building a Data Lake on AWS



Catalog & Search

Access and search metadata

Access & User Interface

Give your users easy and secure access

Data Ingestion

Get your data into S3 quickly and securely

Central Storage

Secure, cost-effective storage in Amazon S3

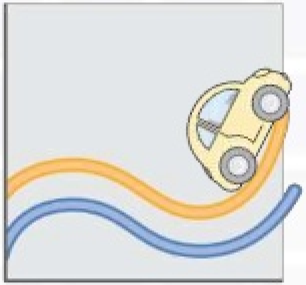
Processing & Analytics

Use of predictive and prescriptive analytics to gain better understanding

Protect and Secure

Use entitlements to ensure data is secure and users' identities are verified

Why a Data Lake on AWS?



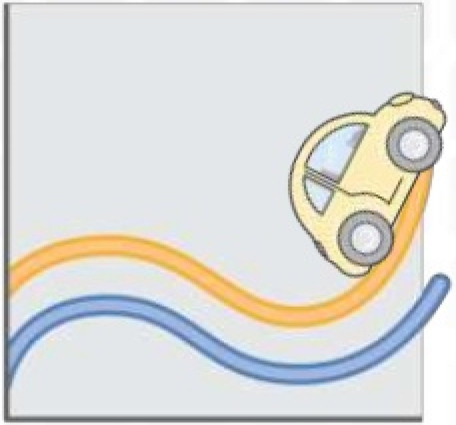
Flexibility and Agility

Security and Compliance

The Most Complete Platform for Big Data

Why a Data Lake on AWS?

Flexibility and agility



AWS provides a flexible platform that enables you to get the most out of your Data Lake:

- Store all of your data cost-effectively using Amazon Simple Storage Service (Amazon S3)
- Quickly and securely move data to AWS
- Provision and scale storage, compute, and networking capacity on-demand
- Eliminate challenges associated with formatting data and converting it to a pre-defined schema

Cloudwick

Mark Schreiber, General Manager

www.cloudwick.com

amazon web services | Partner Network

Architecting Data Lakes on AWS with Cloudwick

Powering the Digital Enterprise

Cloudwick is an AWS Advanced Consulting Partner with Big Data Competency designation that is one of the largest enterprise Big Data-as-a-Service providers, managing dozens of Hadoop, Cassandra, and Spark services for Global 1000 companies on AWS.

Cloudwick leverages a repeatable 3-phase approach to architecting a Data Lake and performing Big Data migrations:

1

Phase 1 – Analyzing your on-premises environment and developing Data Lake requirements

2

Phase 2 – Migrating your data workloads to AWS

3

Phase 3 – Managing your Data Lake and Big Data workloads on AWS

amazon web services Cloudwick

How Cloudwick helped a large, well-known healthcare company improve flexibility

About the Healthcare Company:

This large and well-known healthcare company leverages a unique business model employing doctors to develop software to make it easier to maintain Electronic Health Records (EHR).

The Challenge:

- The company's data ingestion process took 48 hours to complete
- Data ingestion needed to wait until weekends
- Business decisions were being made on week-old data

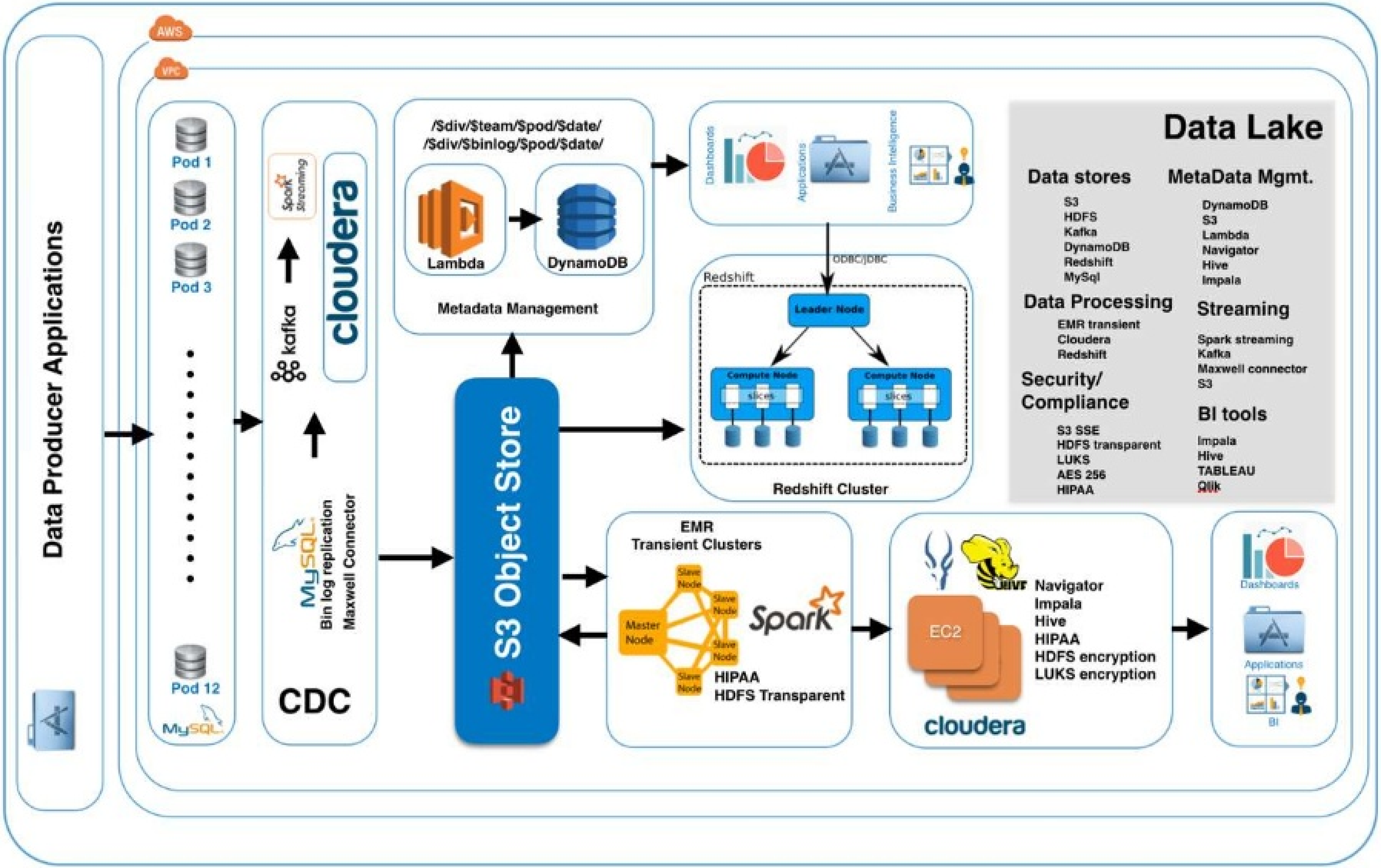
How Cloudwick helped a large, well-known healthcare company improve flexibility

The Cloudwick Solution:

- Architect a Data Lake capable of reducing ingestion time to 3 hours
- Develop a new Data Governance process
- Integrate the company's predictive analytics applications with their Data Lake on AWS

The Benefits:

- Improved metadata management
- Reduced data ingestion time
- HIPAA and PHI Compliance
- Near real-time analytics



Data Lake

Data stores

S3
HDFS
Kafka
DynamoDB
Redshift
MySql

MetaData Mgmt.

DynamoDB
S3
Lambda
Navigator
Hive
Impala

Data Processing

EMR transient
Cloudera
Redshift

Streaming

Spark streaming
Kafka
Maxwell connector
S3

Security/ Compliance

S3 SSE
HDFS transparent
LUKS
AES 256
HIPAA

BI tools

Impala
Hive
TABLEAU
Qlik

Why Data Lakes on AWS?

Security & Compliance

AWS enables organizations like yours to improve their security posture in the cloud:

- Run on top of the secure AWS data center infrastructure
- Encrypt data at rest and in-transit using 256-bit encryption
- Meet compliance standards including PCI DSS, HIPAA, FedRAMP, and more
- Manage user/group access with AWS Identity and Access Management (AWS IAM)

NorthBay

Raza Shaikh, CTO

www.northbaysolutions.com

Architecting Data Lakes on AWS with NorthBay

Teaching Old Data New Tricks™

NorthBay is an AWS Advanced Consulting Partner that has been implementing Data Lakes on AWS since 2013 and was among the first in the world to achieve the AWS Big Data and Mobile competencies.

When NorthBay works with an organization like yours to architect a Data Lake, they:

Develop a strategy that addresses your business and Big Data requirements

Design a Data Lake to that meets your compliance requirements

Quickly architect the Data Lake and migrate your data and applications to AWS

amazon web services NorthBay

How NorthBay Helped Eliza Corporation Deploy a Data Lake on AWS

About Eliza Corporation:

Eliza Corporation develops healthcare consumer engagement solutions to address some of the industry’s greatest challenges – from adherence, to prevention, to condition management, to brand loyalty and retention.

The challenges

Eliza Corporation analyzes more than 300 million interactions per year

Outreach questions and responses form a decision tree, and each question and response are captured as a pair

Challenging to process and analyze data

Diverse downstream consumption requirements

How NorthBay helped Eliza Corporation Deploy a Data Lake on AWS

Create next generation data architecture

Decouple Storage and Compute

Ability to process old & new data streams

Achieve HIPAA compliance

Ingest & store original datasets

Allow both real-time & batch processing

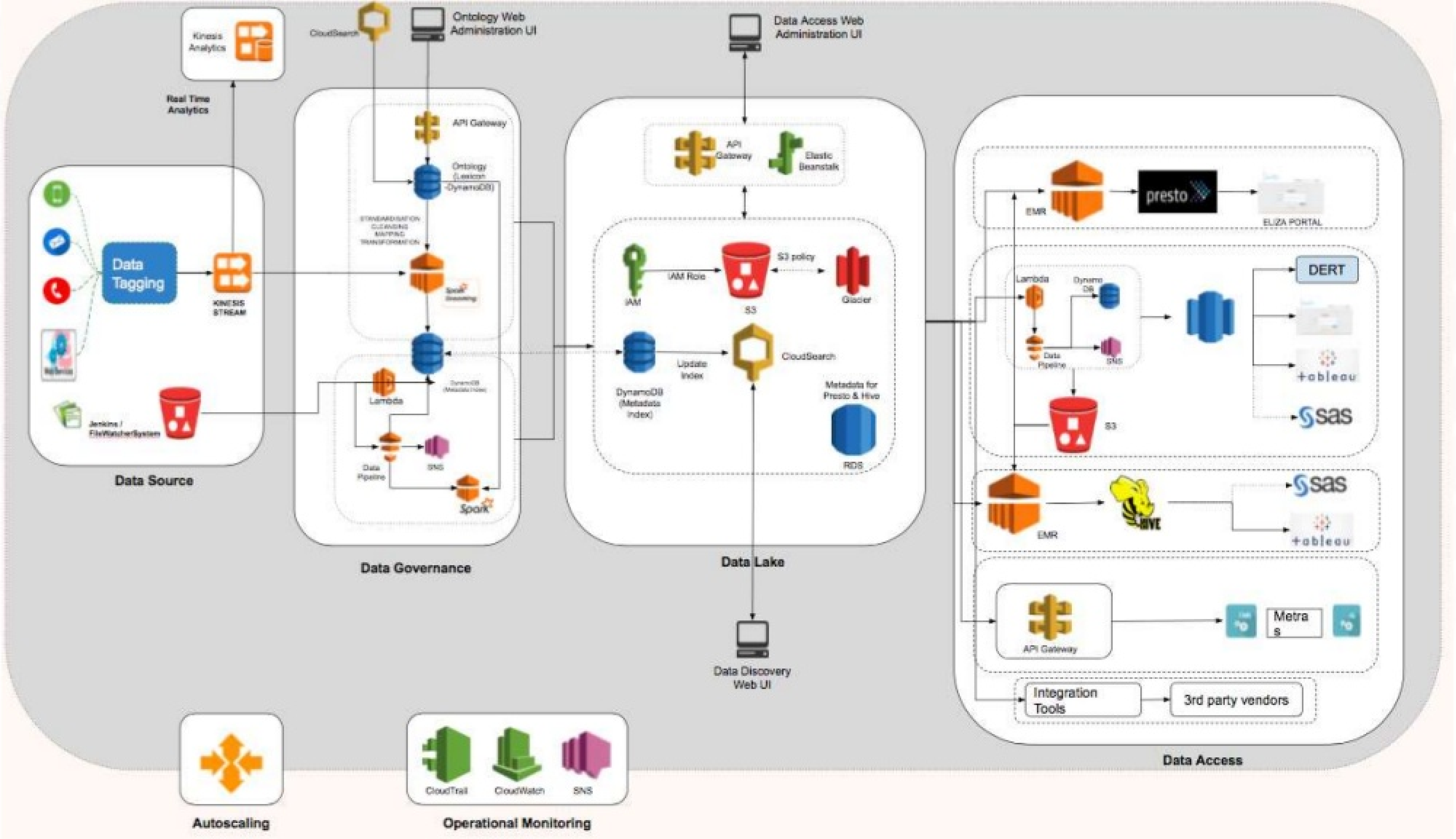
Enable access through entitlements and governance

Increase self-service for end-users

amazon web services NorthBay

Benefits of the Data Lake on AWS

- Streamlined data load process by enabling schema on read
- Improved business agility
- Improved cost management by more clearly separating costs
- Provided ability to provision and scale resources on-demand
- Reduced end-to-end client analytics time


Kinesis Analytics
CloudSearch
Ontology Web Administration UI
Data Access Web Administration UI
Real Time Analytics
API Gateway
Ontology (Lexicon -DynamoDB)
STANDARDISATION CLEANSING MAPPING TRANSFORMATION
Spark Streaming
DynamoDB (Metadata Index)
Lambda
Data Pipeline
SNS
Spark
Data Tagging
KINESIS STREAM
Jenkins / FileWatcherSystem
Data Source
Data Governance
API Gateway
Elastic Beanstalk
IAM
IAM Role
S3
S3 policy
Glacier
DynamoDB (Metadata Index)
Update Index
CloudSearch
Metadata for Presto & Hive
RDS
Data Lake
Data Discovery Web UI
EMR
presto
ELIZA PORTAL
Lambda
Dynamo DB
Data Pipeline
SNS
S3
DERT
tableau
sas
EMR
HIVE
sas
tableau
API Gateway
Metra s
Integration Tools
3rd party vendors
Data Access
Autoscaling
CloudTrail
CloudWatch
SNS
Operational Monitoring

Why Data Lakes on AWS?

The Most Complete Platform for Big Data

AWS offers the most complete platform for Big Data

- Support any workload regardless of volume, velocity, or variety of data
- Use a variety of descriptive, predictive, and prescriptive analytics for business insights
- Immediately available and easy to scale

47Lining

Mick Bass, CEO

www.47lining.com

amazon web services | Partner Network

Architecting Data Lakes on AWS with 47Lining

Breathtaking Big Data

47Lining is an AWS Advanced Consulting Partner with Big Data Competency designation that develops Big Data solutions and delivers Big Data managed services on AWS.

Example use cases include:

- **Propensity to Buy**
- **Making More Timely Business Decisions**
- **Customer Churn**
- **Risk Management and Mitigation**

How 47Lining helped The Howard Hughes Corporation leverage the most complete platform for big data

About The Howard Hughes Corporation:

- The Howard Hughes Corporation (HHC) owns, manages and develops commercial, residential and mixed-use real estate throughout the country.

The Challenge:

- Struggled to fuse on-premises and third party data
- Had difficulty leveraging data to answer their most interesting business questions

Benefits from The Howard Hughes Corporation Data Lake

The 47Lining Solution:

- Architect a Data Lake that ingests third party and on-premises data on AWS
- Implement a lead-scoring model using Amazon Machine Learning
- Turn the customer's Big Data practice into an AWS-based managed service

The Benefits:

400%

increase in the number of qualified leads in their pipeline

10x

reduction of acquisition cost per lead

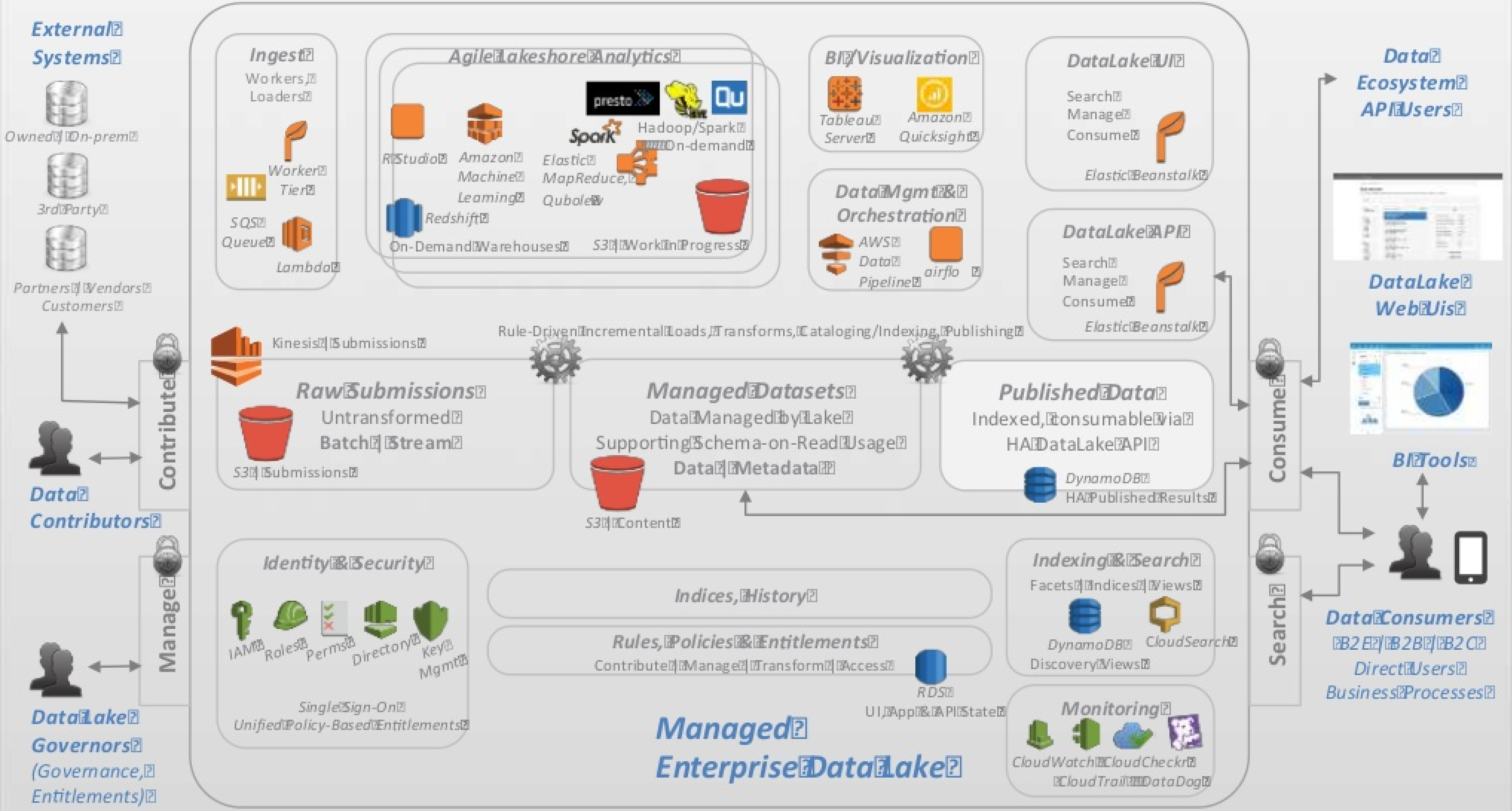
amazon web services 47Lining

How partnering with 47Lining helped The Howard Hughes Corporation reduce time to value

“Our Data Lake is enabling us to answer previously unanswerable questions through on-demand data fusion and analytics. We are leveraging data for strategic advantage in real estate in ways that are groundbreaking for our industry.”

– Daryan Dehghanpisheh,
Senior Vice President, The Howard Hughes Corporation

amazon web services 47Lining



External Systems

Owned | On-prem

3rd Party

Partners | Vendors
Customers

Data Contributors

Data Lake Governors
(Governance,
Entitlements)

Contribute

Manage

Ingest
Workers,
Loaders
Worker Tier
SQS Queue
Lambda

Agile Lakeshore Analytics
R Studio
Amazon Machine Learning
Elastic MapReduce, Qubole
Hadoop/Spark On-demand
Redshift
On-Demand Warehouses
S3 | Work In Progress

BI/Visualization
Tableau Server
Amazon Quicksight

Data Mgmt & Orchestration
AWS Data Pipeline
airflo

DataLake UI
Search
Manage
Consume
Elastic Beanstalk

DataLake API
Search
Manage
Consume
Elastic Beanstalk

Kinesis | Submissions

Rule-Driven Incremental Loads, Transforms, Cataloging/Indexing, Publishing

Raw Submissions
Untransformed
Batch | Stream
S3 | Submissions

Managed Datasets
Data Managed by Lake
Supporting Schema-on-Read Usage
Data | Metadata
S3 | Content

Published Data
Indexed, consumable via
HA DataLake API
DynamoDB
HA Published Results

Identity & Security
IAM
Roles
Perms
Directory
Key Mgmt
Single Sign-On
Unified Policy-Based Entitlements

Indices, History

Rules, Policies & Entitlements
Contribute | Manage | Transform | Access

RDS
UI, App & API State

Indexing & Search
Facets | Indices | Views
DynamoDB
CloudSearch
Discovery Views

Monitoring
CloudWatch
CloudCheckr
CloudTrail
DataDog

Managed Enterprise Data Lake

Consume

Search

Data Ecosystem API Users

DataLake Web UIs

BI Tools

Data Consumers
B2E | B2B | B2C
Direct Users
Business Processes

Summary

What is a Data Lake?

- A new and increasingly popular way to store and analyze all of your data, regardless of source or format, in a centralized repository for use with analytics.

What are the benefits?

- Accelerated time to value and reduced TCO
- Increased storage flexibility and agility
- Ad-hoc analysis that delivers business insights more quickly

Summary

Why AWS for a Data Lake?

- Flexibility and agility
- Security and compliance
- Most complete platform for Big Data

Who can help me implement a Data Lake on AWS?

- Featured APN Partners
- Cloudwick - http://www.cloudwick.com/
- NorthBay - http://northbaysolutions.com/
- 47Lining - http://www.47lining.com/

To learn more and schedule a POC, visit our partners' landing pages