How to Build a Big Data Analytics Data Lake

Neeraj Verma – AWS Solutions Architect
Saurav Mahanti – Senior Manager – Information Systems
Dario Rivera – AWS Solutions Architect

November 28, 2016

amazon web services

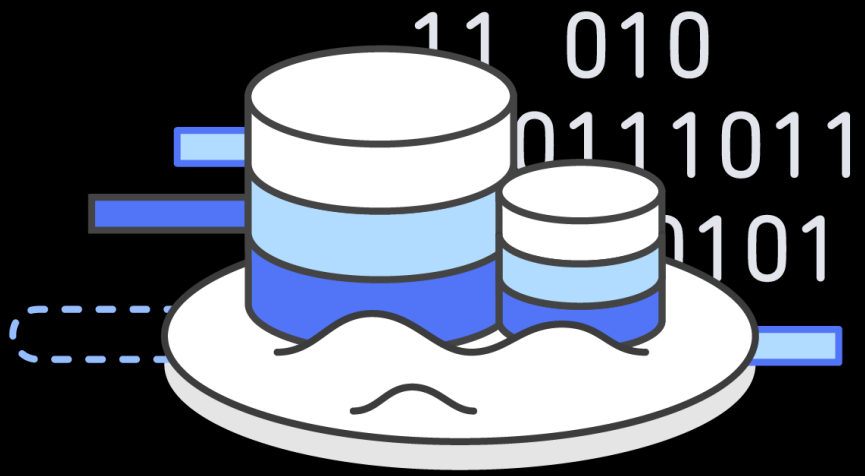
What to expect from this short talk

- Data Lake concept
- Data Lake - Important Capabilities
- AMGEN’s Data Lake initiative
- How to Build a Data Lake in your AWS account

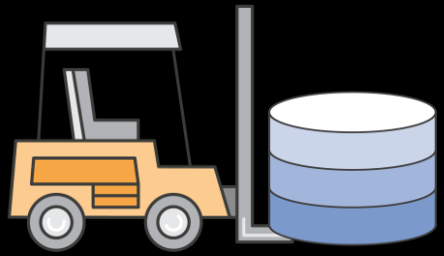
Data Lake Concept

What is a Data Lake?

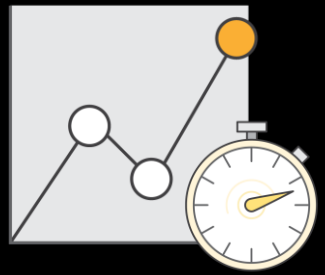
Data Lake is a new and increasingly popular way to store and analyze massive volumes and heterogenous types of data in a centralized repository.



Benefits of a Data Lake – Quick Ingest



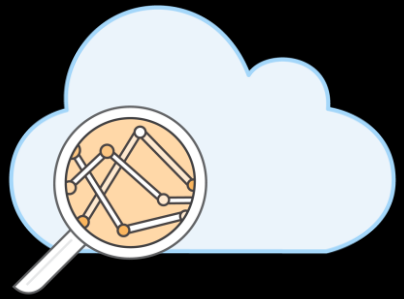
“How can I collect data quickly from various sources and store it efficiently?”



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

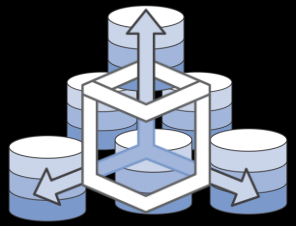
Benefits of a Data Lake – All Data in One Place

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

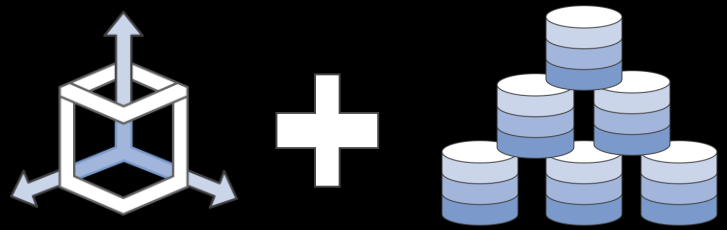


Store and analyze all of your data, from all of your sources, in one centralized location.

Benefits of a Data Lake – Storage vs Compute

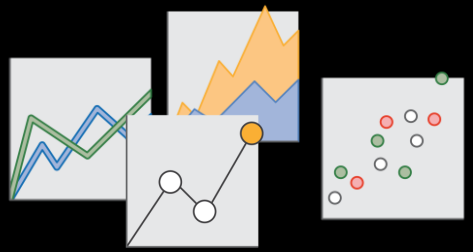


“How can I scale up with the volume of data being generated?”



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

Benefits of a Data Lake – Schema on Read



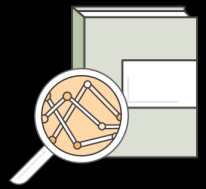
“Is there a way I can apply multiple analytics and processing frameworks to the same data?”

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

Important Capabilities of a "Data Lake"

Important components of a Data Lake

Ingest and Store

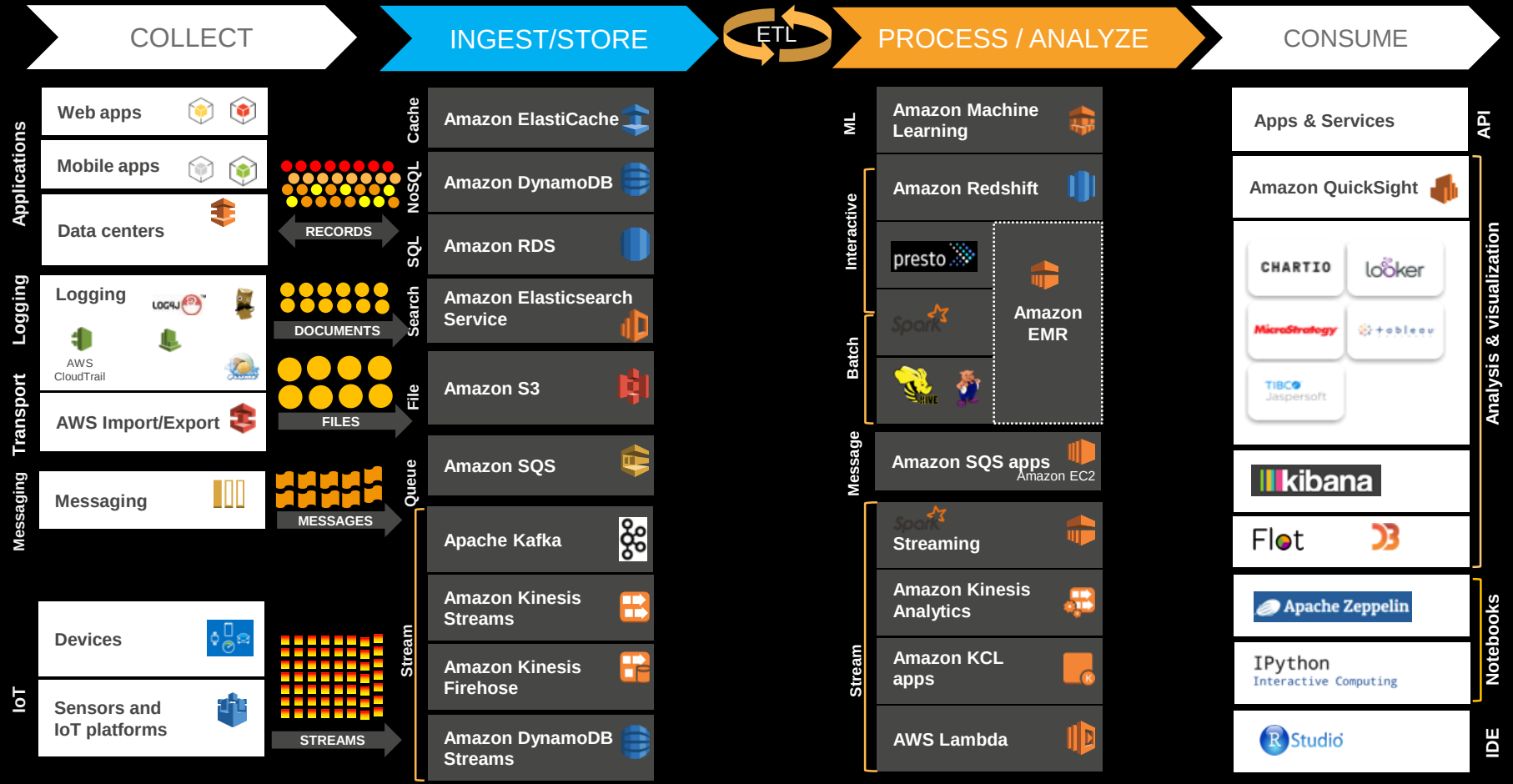


Catalogue & Search

Protect & Secure

Access & User Interface

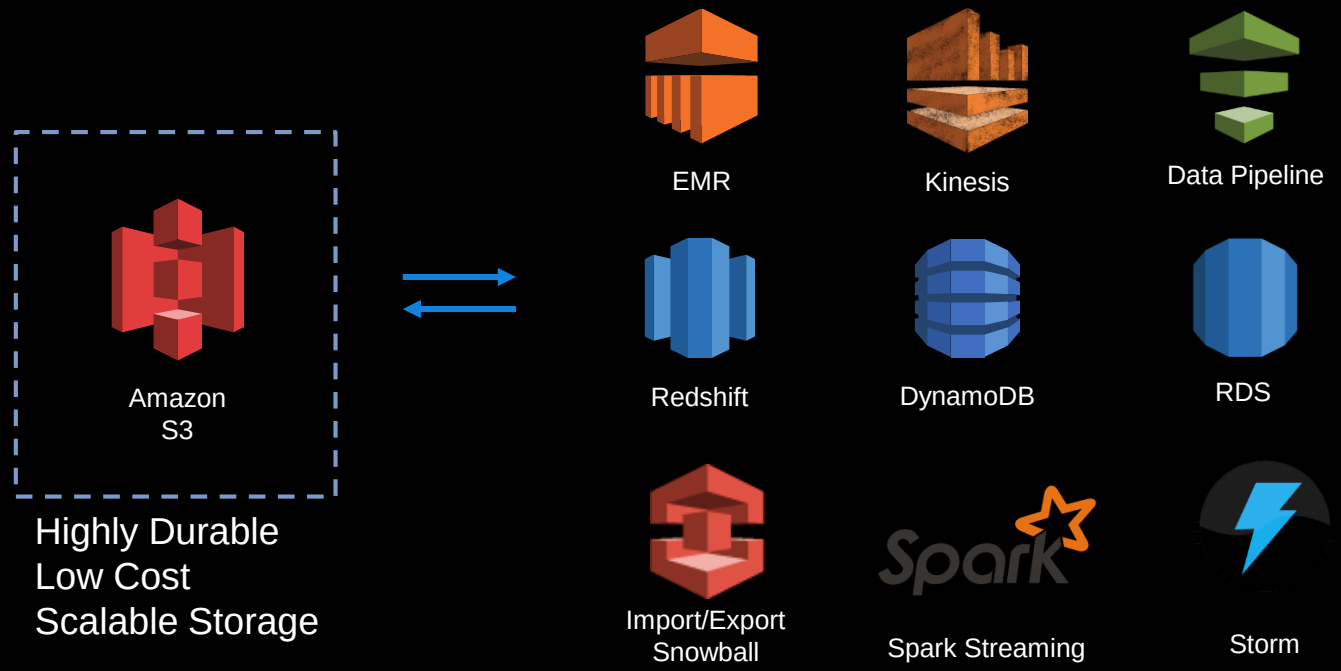
Many tools to Support the Data Analytics LifeCycle



Ingest and Store

Ingest real time and batch data
Support for any type of data at scale
Durable
Low cost

Use S3 as Data Substrate – Apply to Compute as Needed

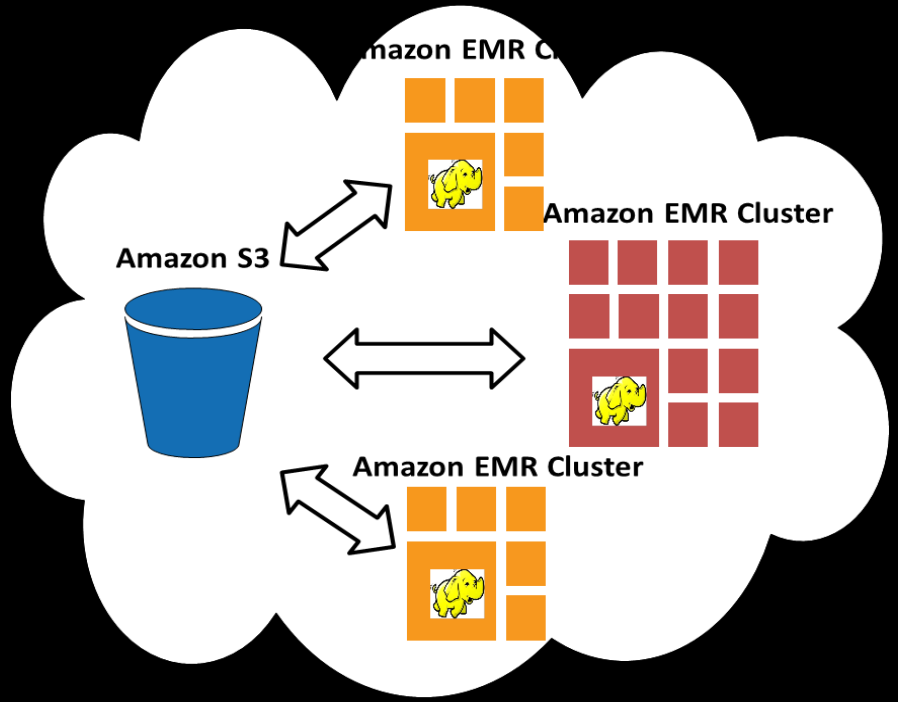


Amazon S3 as your cluster's persistent data store

Amazon S3
Separate compute and storage

Resize and *shut down* Analytics Compute Environments with *no data loss*

Point *multiple* compute clusters at same data in Amazon S3



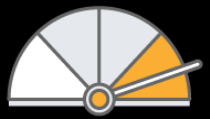
Data Ingestion into Amazon S3

AWS Direct Connect

AWS Snowball

ISV Connectors

Amazon Kinesis Firehose



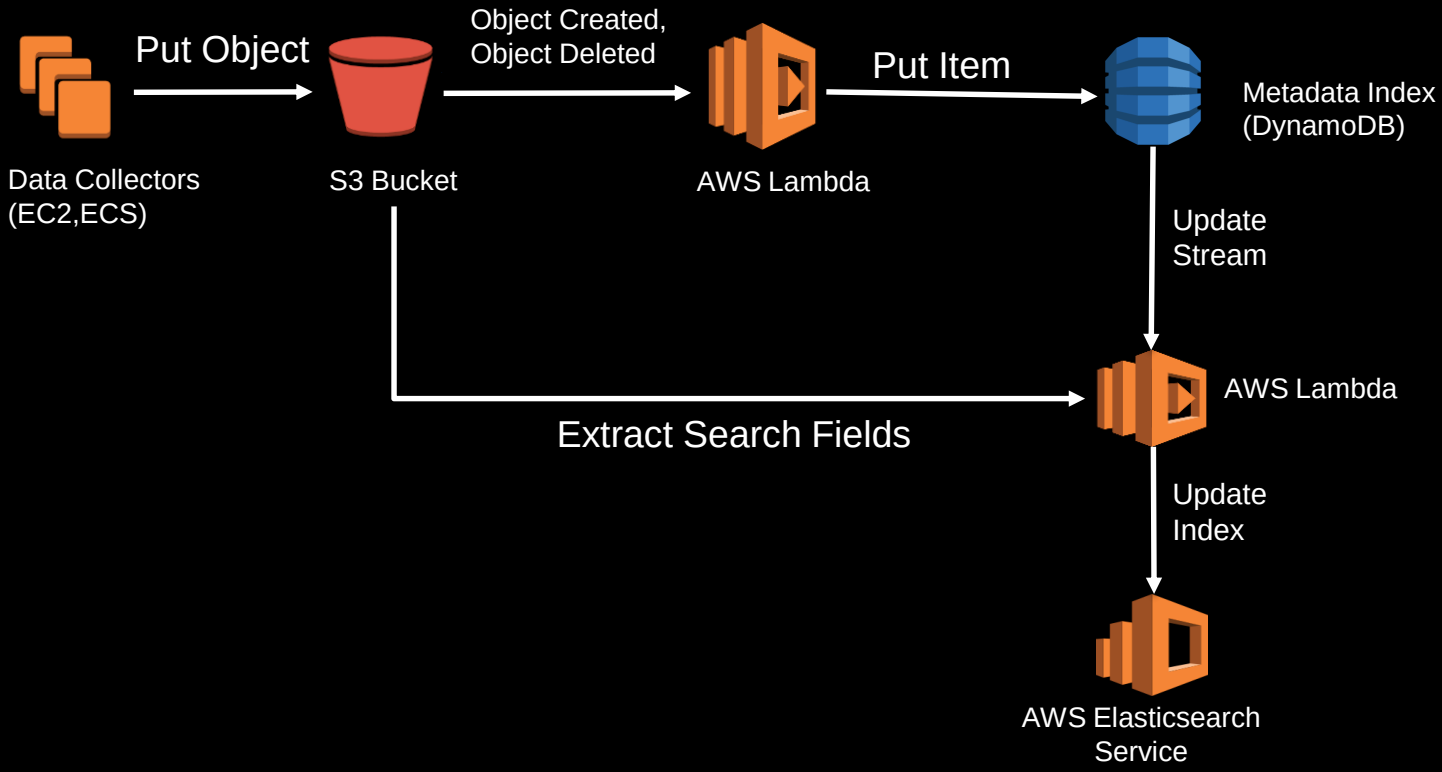
S3 Transfer Acceleration

AWS Storage Gateway

Catalogue & Search

Metadata lake
Used for summary statistics and data
Classification management
Simplified model for data discovery & governance

Catalogue & Search Architecture



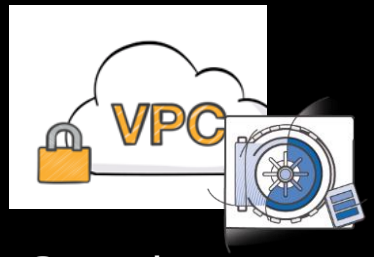
Protect and Secure

Access Control - Authentication & Authorization
Data protection – Encryption
Logging and Monitoring

Implement the right controls

Encryption

- SSL endpoints
- Server Side Encryption (SSE-S3)
- S3 Server Side Encryption with provided keys (SSE-C, SSE-KMS)
- Client-side Encryption



Security

- Identity and Access Management (IAM) policies
- Bucket policies
- Access Control Lists (ACLs)
- Query string authentication
- Private VPC endpoints to Amazon S3

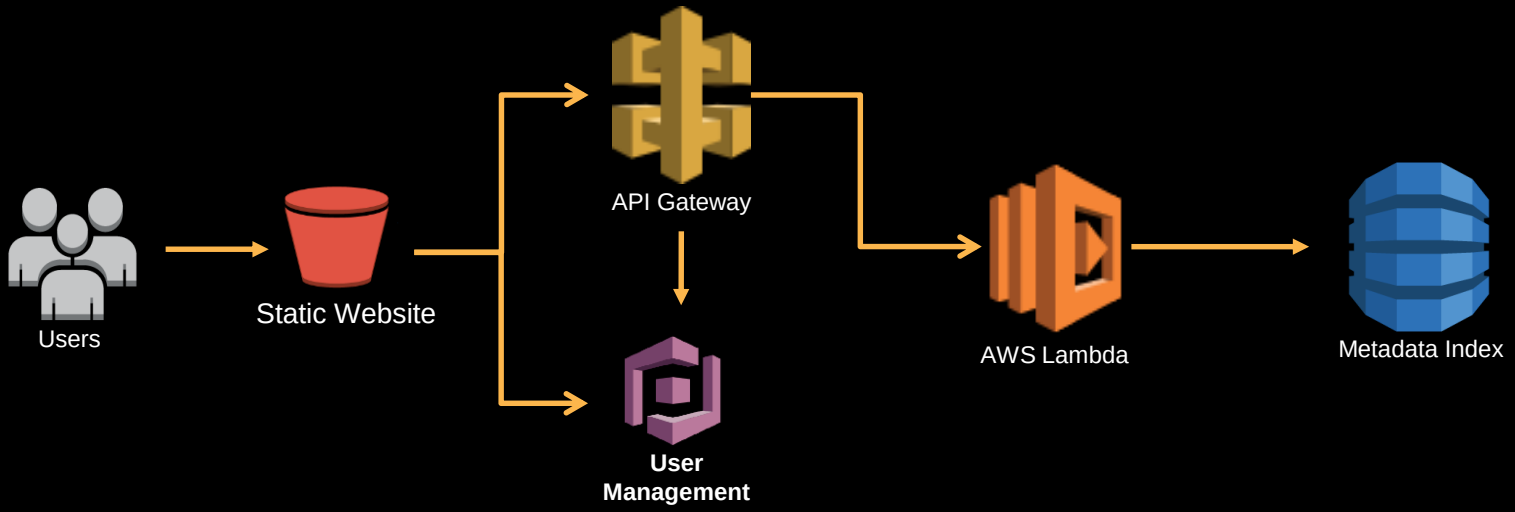
Compliance

- Buckets access logs
- Lifecycle Management Policies
- Access Control Lists (ACLs)
- Versioning & MFA deletes
- Certifications – HIPAA, PCI, SOC 1/2/3 etc.

API & User Interface

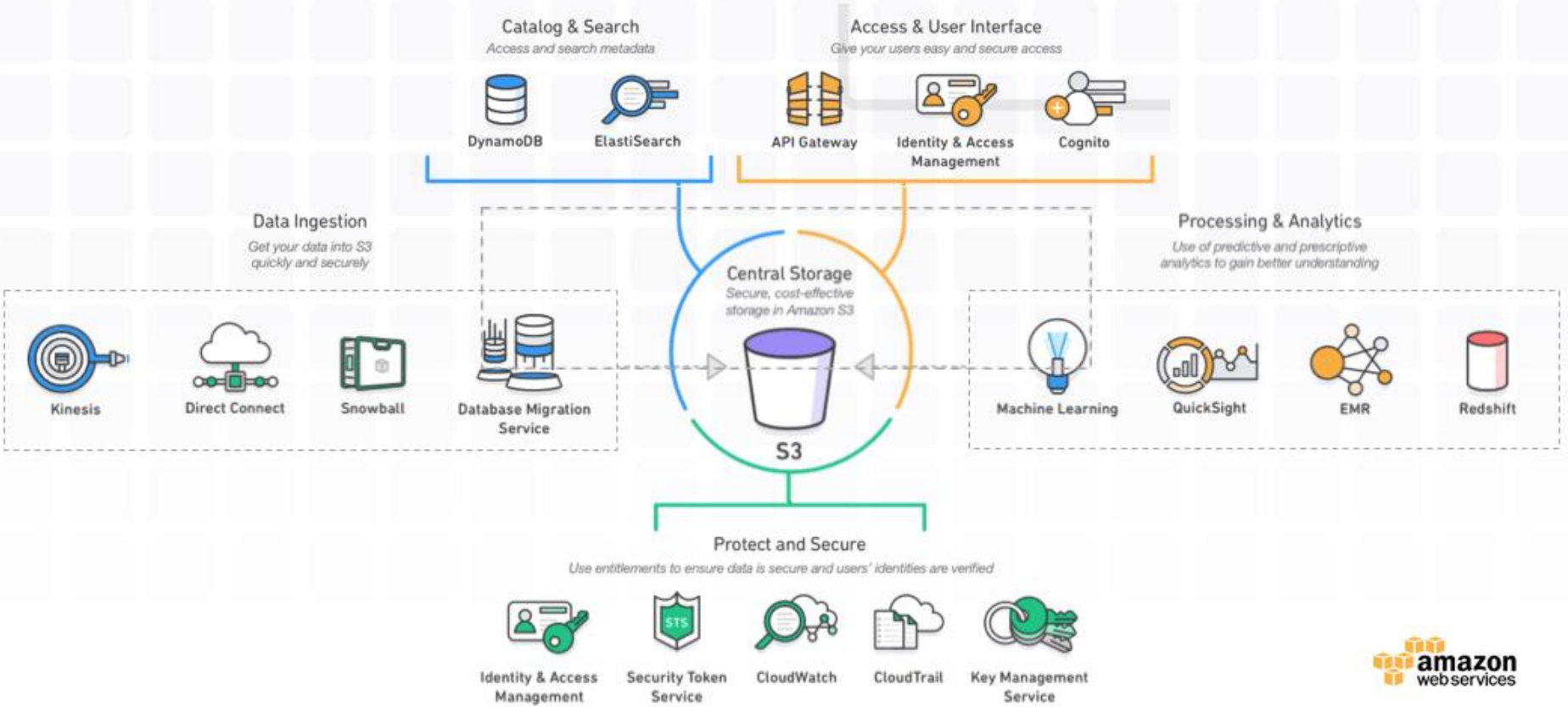
Exposes the data lake to customers
Programmatically query catalogue
Expose search API
Ensures that entitlements are respected

API & UI Architecture



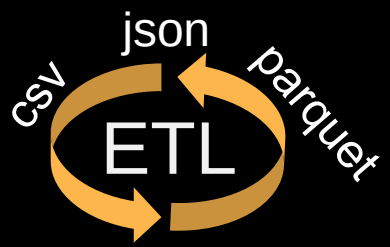
Putting It All Together

Building a Data Lake on AWS



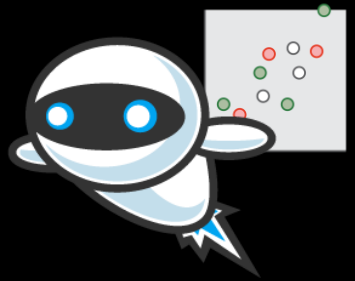
amazon web services

Common Add-On Capability to Data Lake



Backend ETL

Organizational Control Gates for Dataset Access



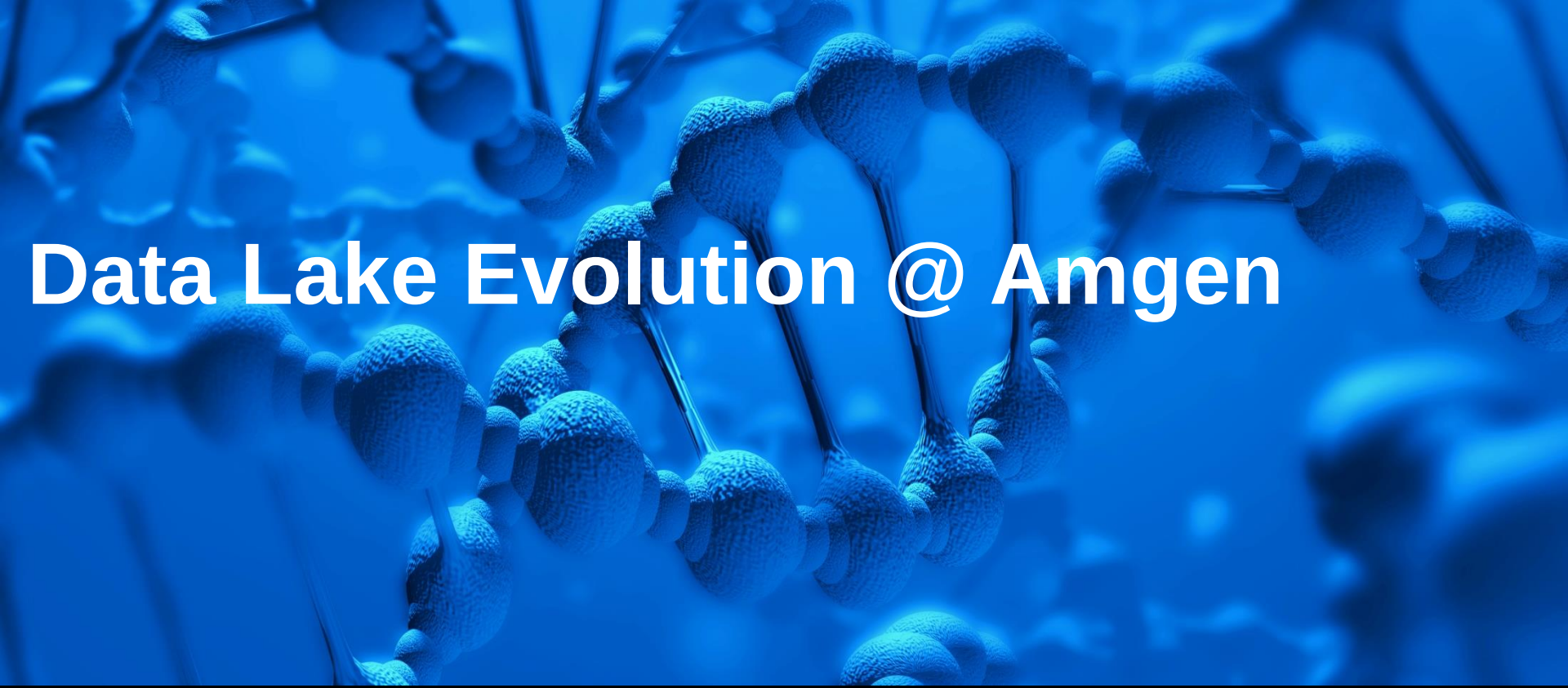
Data Correlation Identification via Machine Learning

Dataset Updates to Dependent Compute Environments

Data Lake is a Journey

There are Multiple implementation Methods for Building a Data Lake

- Using a Combination of Specialized/Open Source Tools from Various Providers to build a Customized Solution
- Use various managed services from AWS such as S3, Amazon ElasticSearch Service, DynamoDB, Cognito, Lambda, etc as a Core Data Lake Solution
- Using a DataLake-as-a-Service Solution



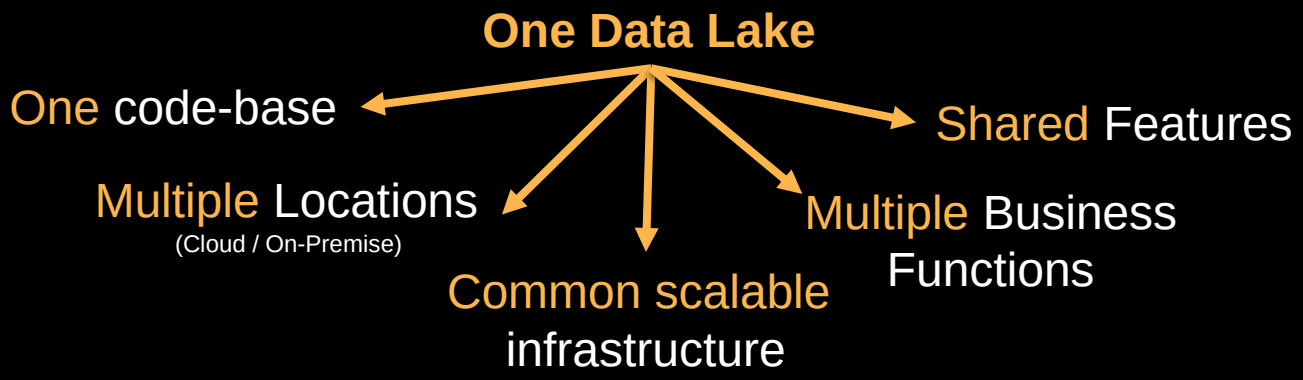
Data Lake Evolution @ Amgen

SAURAV MAHANTI

Senior Manager - Information Systems

AMGEN®

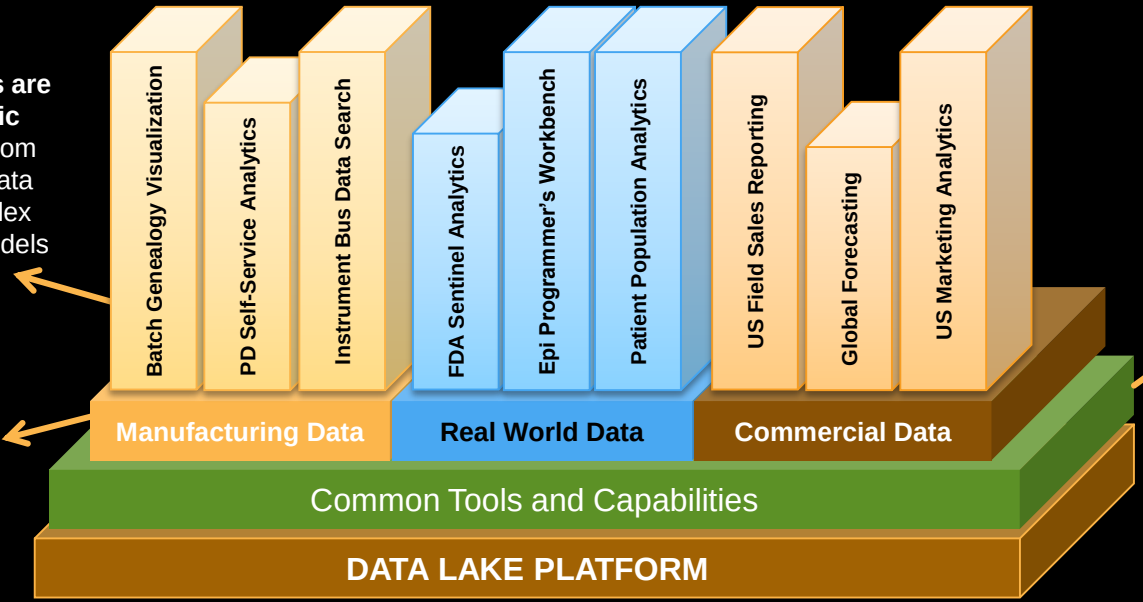
Pioneering science delivers vital medicines™


One Data Lake
One code-base
Shared Features
Multiple Locations
(Cloud / On-Premise)
Multiple Business Functions
Common scalable infrastructure

A CONCEPTUAL VIEW OF THE DATA LAKE

Winner
Best Practices Awards
★
Bio IT World 2016

Business applications are built to meet specific information needs, from simple data access, data visualization, to complex statistical/predictive models



Functions can manage their own data, while **contributing to the common data layer**

Innovative new tools and capabilities are reused across all functions (e.g. search, data processing & storage, visualization tools)

One Data Lake

One code-base

Shared Features

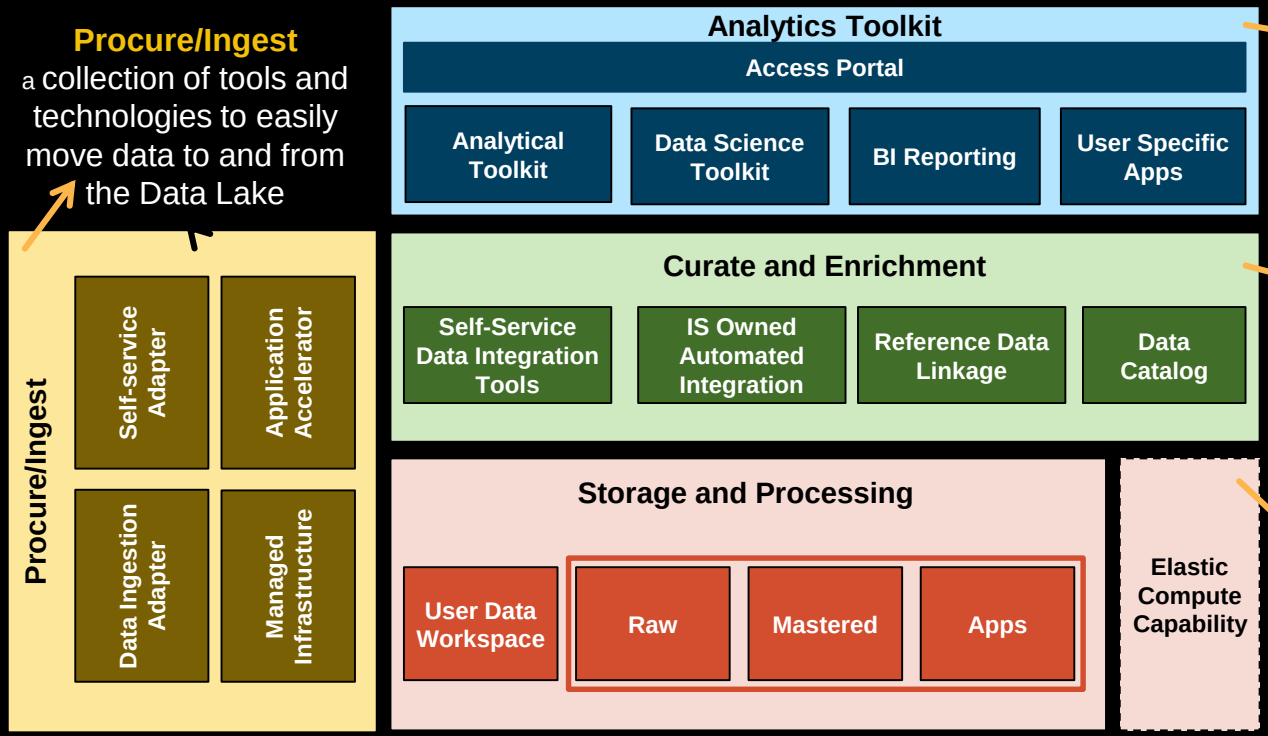
Multiple Locations
(Cloud / On-Premise)

Multiple Business Functions

Common scalable infrastructure

HIGH LEVEL COMPONENT ARCHITECTURE of the data lake

Procure/Ingest
a collection of tools and technologies to easily move data to and from the Data Lake



Centralized Analytics Toolkit
provides different reporting and analytics solutions for end-users to access the data on the Data Lake

Curate and Enrichment Layer
Enhances the value and usability of the data through automation, cataloging and linking

Data processing & storage layer
the core of the Enterprise Data Lake that enables its scale and flexibility

DATA PROCESSING AND STORAGE LAYER

The combination of Hadoop HDFS and Amazon S3 provides the right cost/performance balance with unlimited scalability while maintaining security and encryption at the data file level

HIVE and Impala provide SQL over structured data;
HBASE is used for NoSQL/Transactional jobs
Solr is used for search capabilities over documents
MarkLogic and Neo4J provide semantic and graph capabilities

Mark Logic

Powerful execution engines like YARN, Map Reduce and SPARK bring the "compute to the data"

Storage and Processing

User Data Workspace

Raw

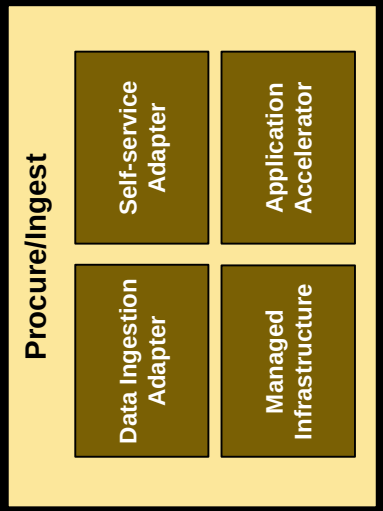
Mastered

Apps

Elastic Compute Capability

Amazon RedShift and Amazon EMR low-cost elastic computing with the ability to spin up clusters for data processing and metrics calculations

PROCURE AND INGEST - Pre-built and configurable common components to load any type of data into the Lake



Structured Data Ingestion - A common component for scheduled production data loads of incremental or full data. It uses Python and native Hadoop tools like Scoop and HIVE for efficiency and speed.

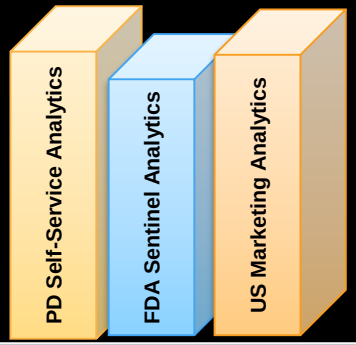
Real-Time Data Ingestion for real time or streaming data. It uses the Kafka messaging queue, Spark streaming, Hbase and Java.

Unstructured Data Pipeline Morphlines Document Pipeline for document ingestion, text processing and indexing. It uses Morphlines, Lily indexer and HBase.

Cloud Data Integration tools like SnapLogic enable data analysts end users to build data pipelines into the Data Lake using pre-built connectors to various cloud hosted services like Box and make it easy to move data set between HDFS, S3, sFTP and fileshares

CURATE AND ENRICHMENT

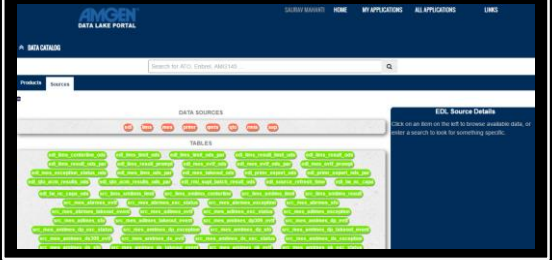
Analytical Subject Areas – Build targeted applications using Data Integration tools like SnapLogic or deploy packaged applications or data marts that transform the data in the Lake for consumption by end user tools



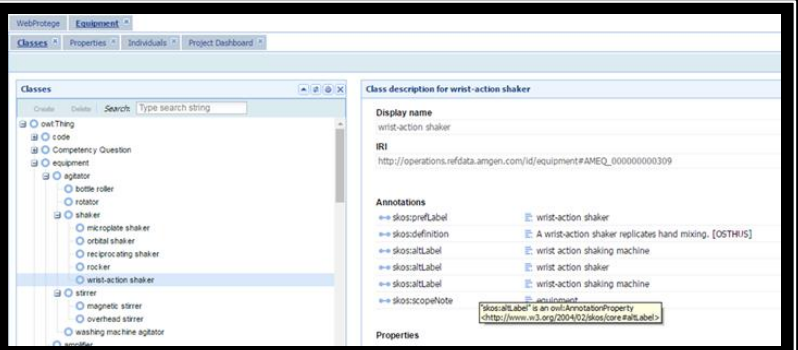
Curate and Enrichment

- **Self-Service Data Integration Tools**
- **Analytical Subject Areas**
- **Reference Data Linkage**
- **Data Catalog**

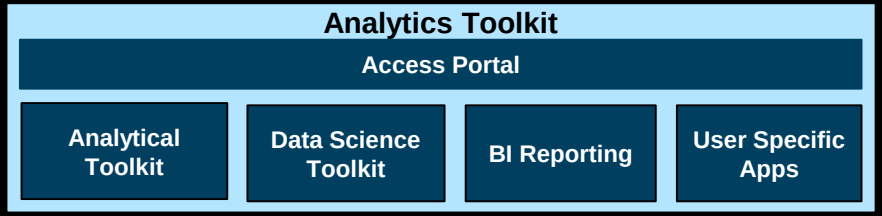
Data Catalog is an enterprise-wide metadata catalog that stores, describes, indexes, and shows how to access any registered data asset



Reference Data linkage – Connect Datasets to Ontologies and Vocabularies to get more relevant and better results



CENTRALIZED ANALYTICS TOOLKIT - provides reporting and analytics solutions for end-users to access the data in the Lake



Analytics and Visualization tools are the most common methods used to query and analyze data in the Lake

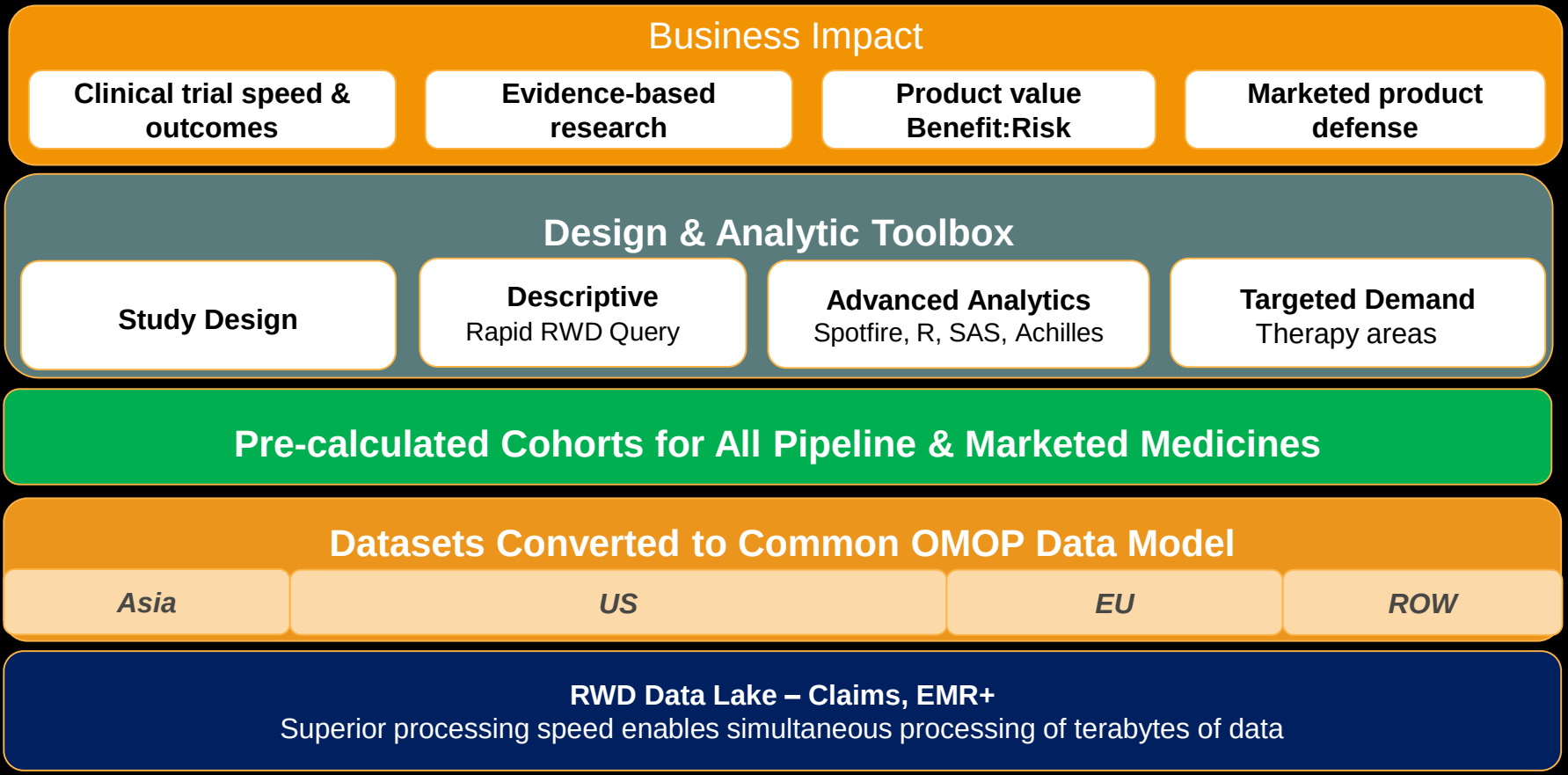
Reporting and Business Intelligence tools like MicroStrategy and Cognos can be deployed either directly on the Lake or on derived Analytical Subject Areas

Data Lake Portal provides a user-friendly, mobile-enabled and secure portal to all the end-user applications

Data Science toolkit enables analysts and data scientists use tools like SAS, R, Jupyter (Python notebooks) to connect to their analytics sandbox within the Lake and submit distributed computing jobs

Focused applications that target a specific use-case can be built using open source products and deployed to a specific user community through the portal

Real World Data Platform: Shared capability used by multiple business functions from R&D and Commercial



That’s a lot of Work! – Where do I even Start?!

Introducing:

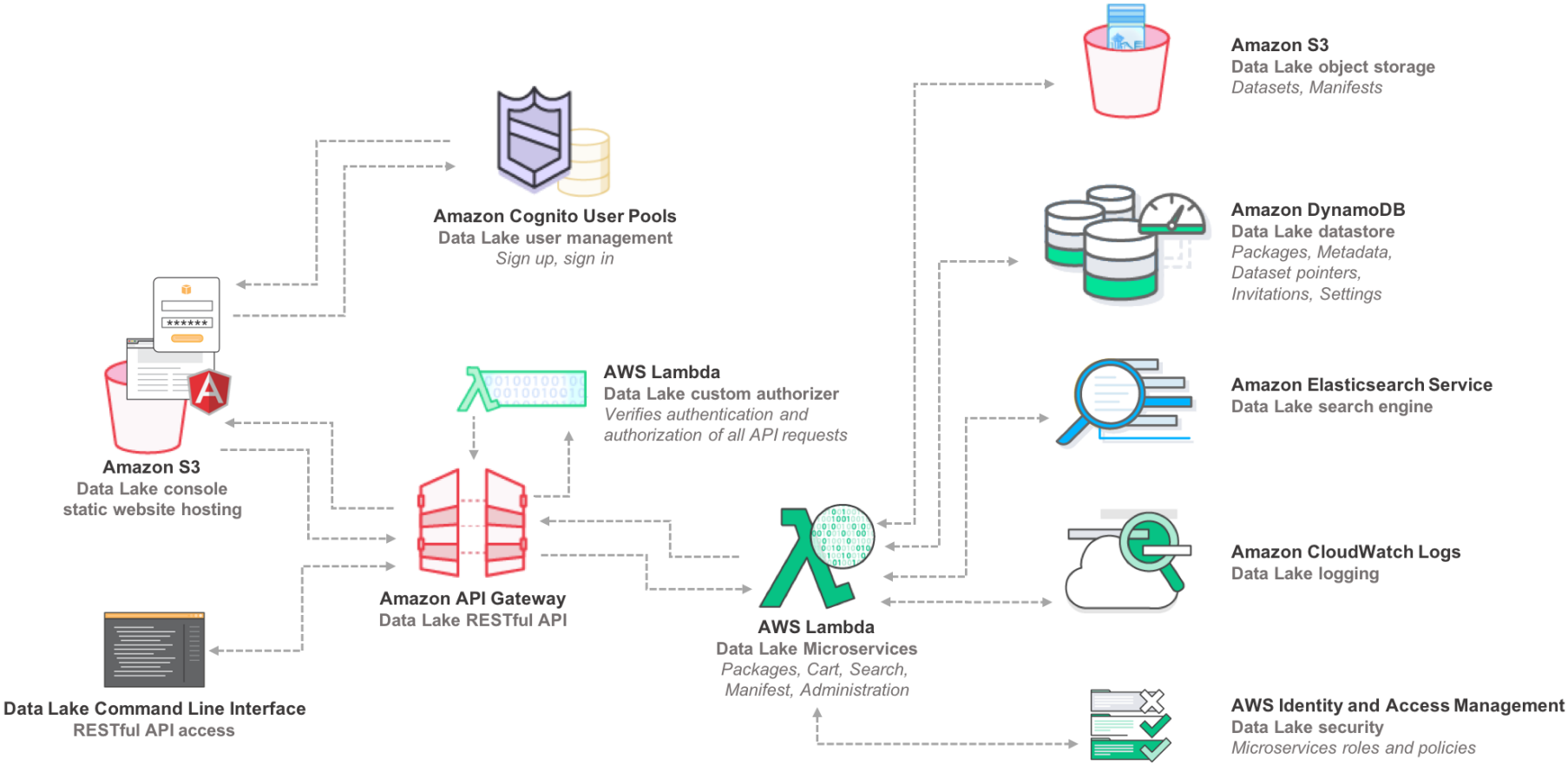
Data Lake Solution

Presented by: Dario Rivera – AWS Solutions Architect
Built by: Sean Senior – AWS Solutions Builder

Data Lake Solution

Package of Code – Deployed via CloudFormation into your AWS Account

Architecture Overview

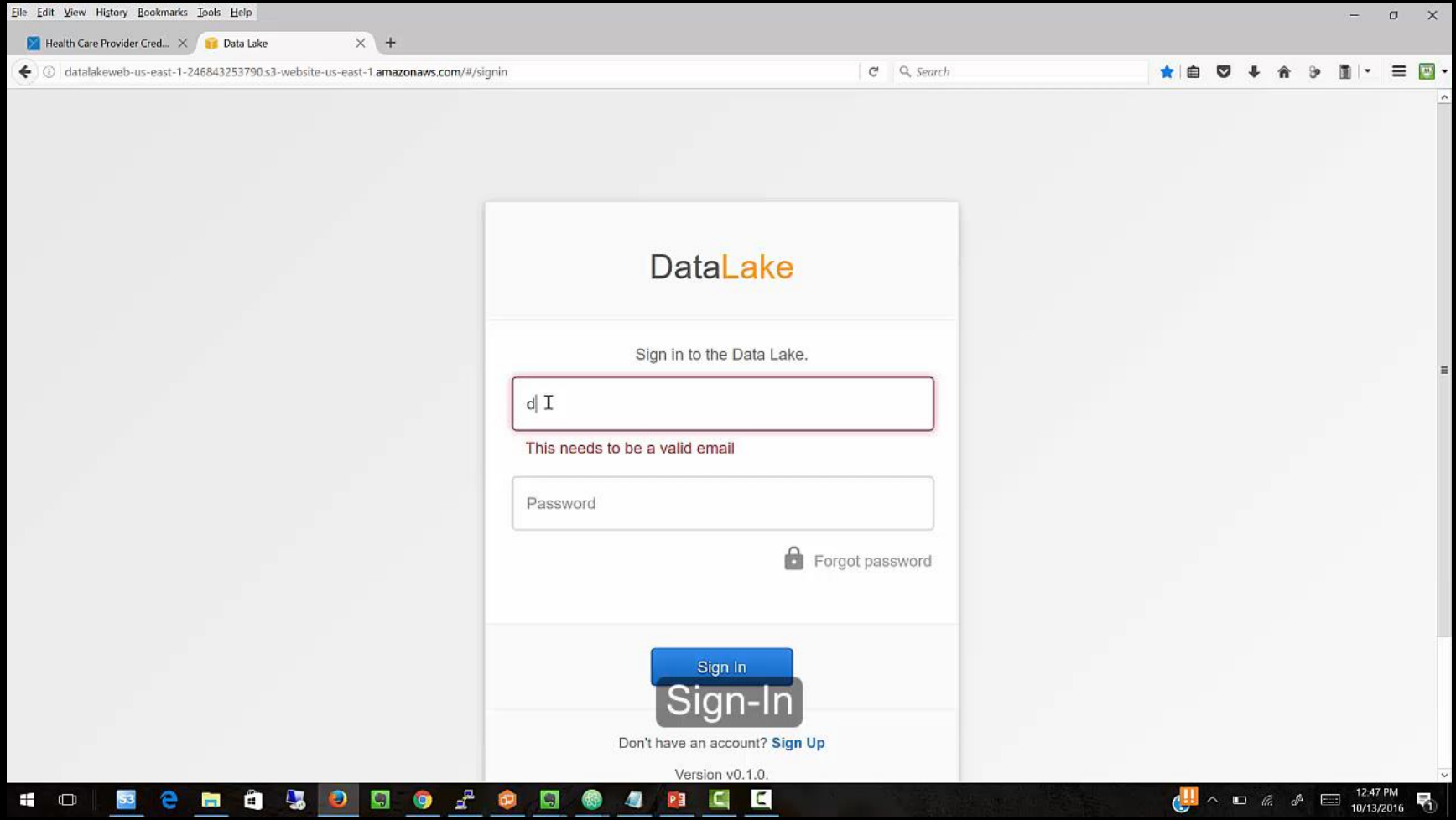


Data Lake Server-less Composition

- Amazon S3 (2 buckets)
  - Primary data lake content storage
  - Static Website Hosting
- Amazon API Gateway (1 RESTful API)
- Amazon DynamoDB (7 Tables)
- Amazon Cognito Your User Pools (1 User Pool)
- AWS Lambda
  - 5 microservice functions,
  - 1 Amazon API Gateway custom authorizer function,
  - 1 Amazon Cognito User Pool event trigger function [Pre sign-up, Post confirmation]
- Amazon Elasticsearch Service (1 cluster)
- AWS IAM (7 policies, 7 roles)

Demo of
AWS Data Lake Solution

http://tinyurl.com/DataLakeSolution


File Edit View History Bookmarks Tools Help
Health Care Provider Cred...
Data Lake
datalakeweb-us-east-1-246843253790.s3-website-us-east-1.amazonaws.com/#/signin
Search
DataLake
Sign in to the Data Lake.
d
This needs to be a valid email
Password
Forgot password
Sign In
Sign-In
Don't have an account? Sign Up
Version v0.1.0.
12:47 PM
10/13/2016

All of this and it costs less than a $1 / hour to run the Data Lake Solution

* Excluding Storage and Analytics Environment Costs

Data Lake Solution will be available End of Q4.

Available via AWS Answers

https://aws.amazon.com/answers/

Thank you!

Neeraj Verma – rajverma@amazon.com
Saurav Mahanti – smahanti@amgen.com
Dario Rivera – darior@amazon.com

amazon web services