nClouds℠

Data Lake as Code on AWS Implementation Workshop

aws

www.nclouds.com

Special OFFERS

Free Data Lake Assessment
for all eligible attendees.

Get started in as little as
72 hours.

Data Lake as Code Implementation Workshop

PRESENTERS

Kireet Kokala
VP, Big Data & Analytics

nClouds

Paul Underwood
Principal Solutions Architect

aws

Carlos Rodriguez
Senior DevOps Engineer

nClouds

Amazon Web Services (AWS) is the world's most comprehensive and broadly adopted cloud platform, offering over 175 fully featured services from data centers globally. Millions of customers—including the fastest-growing startups, largest enterprises, and leading government agencies—are using AWS to lower costs, become more agile, and innovate faster.

nClouds is an AWS Premier Consulting Partner and award-winning provider of AWS and DevOps consulting and implementation services. We are an integrated team of skilled engineers, architects, developers, project managers, and sales & marketing pros who are passionate about client success, software excellence, and innovation. We enable our clients to deliver better products faster and create awesome customer experiences.

nClouds SUPERPOWERS

DevOps & Infrastructure Modernization

- CI/CD pipelines
- Containers & microservices
- DevOps-as-a-Service

Cloud Migration Services

- Migration Readiness Assessment
- From simple lift & shift (rehosting) to re-architecting and refactoring
- CloudChomp C3 Certified Partner
- VMware/Windows/Linux/Database

Data & Analytics

- DataOps: Athena, Aurora, Glue, QuickSight, Data Warehouse, Data Lake, Hadoop, Redshift, ETL / ELT
- ML & AI: SageMaker, AI, Deep Learning, Alexa

nOps (SaaS) Cloud Management

- AWS Well-Architected Reviews
- Cost optimization
- Security review

Trusted by INNOVATIVE BRANDS

nClouds

Data Lake as Code Implementation Workshop

AGENDA

DETAILS *(All times EDT)*

- **1:00 - 1:10 pm** - Intro & Workshop Objectives *by Kireet Kokala, nClouds*
- **1:10 - 1:20 pm** - Architecture: Data Lake via Cloud Formation *by Paul Underwood, AWS*
- **1:20 - 1:30 pm** - Data Lake Use Cases *by Carlos Rodriguez, nClouds*
- **1:30 - 1:50 pm** - Demo: Data Lake as Code on AWS *by Carlos Rodriguez, nClouds*
- **1:50 - 1:55 pm** - Getting Started: Data Lake Implementation Options *by Kireet Kokala, nClouds*
- **1:55 - 2:00 pm** - Q&A

Data Lake as Code Implementation Workshop

OBJECTIVES

Data Lakes on Amazon

Use Cases, Architecture, Demo

Implementation

Process, Cost Clarity, Timing

nClouds Data & Analytics

Services, Benefits, Identifying Next Steps

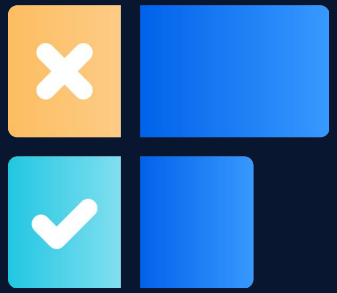
Data Lakes History

- A set of technologies enabling the storage and analysis of vast volumes of data in its raw, natural format.
  - 1970s: Data marts used for digital storage and enhanced sales efforts.
  - 2000s: Unofficial use of data lakes using HDFS services.
  - 2010: “Data lake” emerged via Pentaho onto the IT domain.
- Usually, the single store-platform for all enterprise managed data. Used for business intelligence (BI) tasks, like reporting and visualization, or advanced analytics ML.
- In 2020, nClouds used AWS Data Lake services to ingest U.S. Stock Market data, analyse, and visualize it.

Data Lake Value Proposition

- Complement your data mart or data warehouse. The functionalities of the data lake and the data warehouse become complementary.
- Perform mashups against unstructured, non-relational data in the data lake.
- Build more robust and fresh data lakes providing high-quality insights by enforcing schematization on data sets.
- Take control of data lakes via seamless ingestion and management of large analytical data sets over distributed file systems.

nClouds



Poll 1

aws

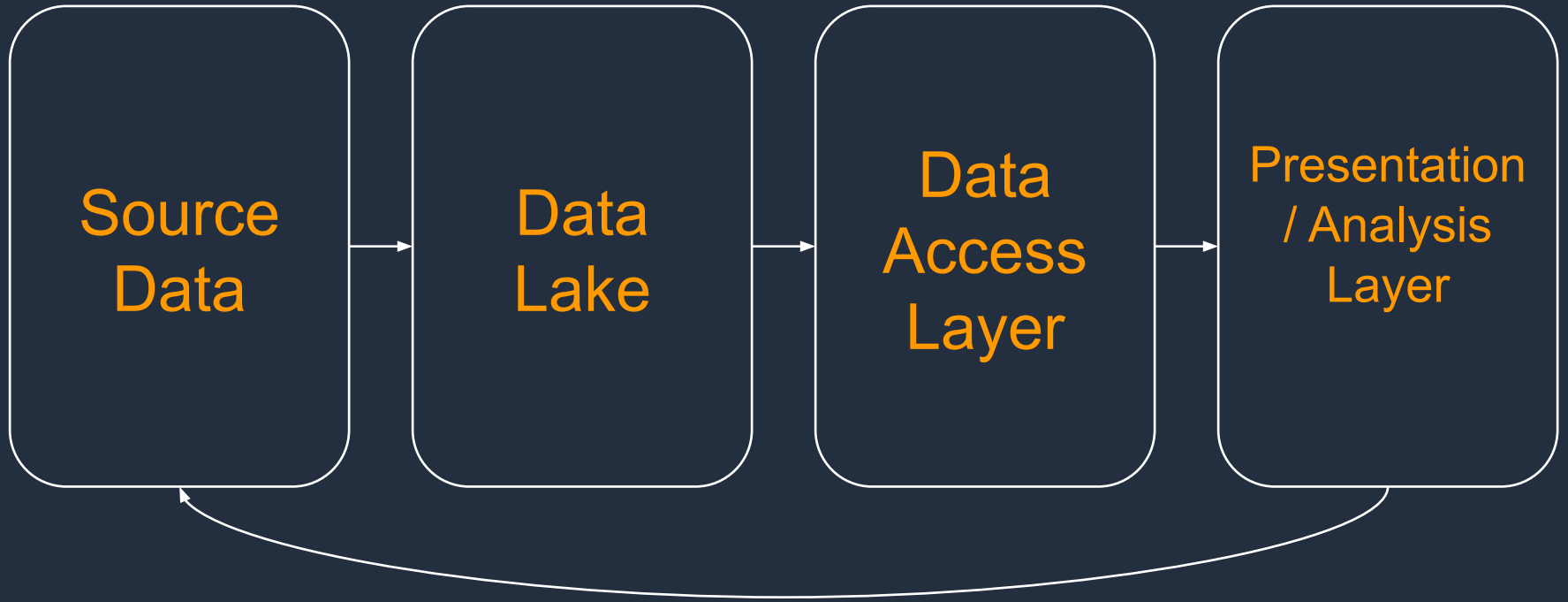
Data Lakes on on AWS

- **Data lakes** are easy to set up and integrate into Big Data, IoT, and related solutions using AWS services.
- **Comprehensive system** supports open file formats like Apache Parquet without the need to transform data; data can be stored in a standard format and analyzed using appropriate tools.
- **Security** and fine grain controls help you stay compliant with your policies and industry regulations. IAM and AWS Lake Formation help reduce the build times from months to days.
- **Cost effectiveness** and scalability are assured via auto-scaling, saving plans, and integration with Amazon EC2 Spot instances.

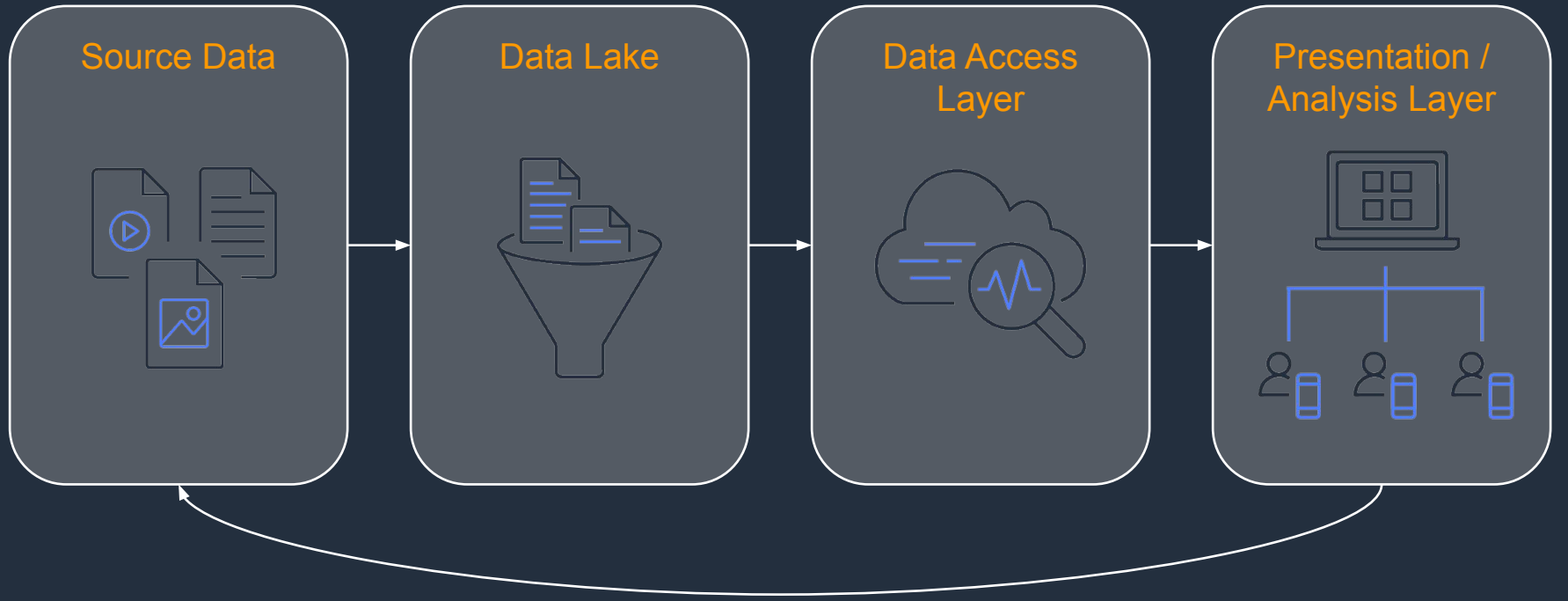
Architecture: Data Lakes on AWS

Paul Underwood
Principal Solutions Architect

What is a data lake? What is it not?



For example…


Source Data
Data Lake
Data Access Layer
Presentation / Analysis Layer

Metadata

Metadata

Databases
Tables
Columns
Locations

ETL

Security

Storage

ETL

Metadata

Databases
Tables
Columns
Locations

ETL

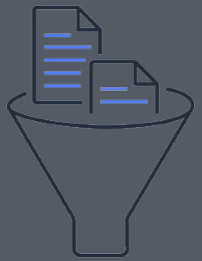
Extract Transform and Load

Security

Storage

Security

Data Lake



Metadata — Databases, Tables, Columns, Locations

ETL — Extract Transform and Load

Security — Permissions and Grants (DB/Table/Column)

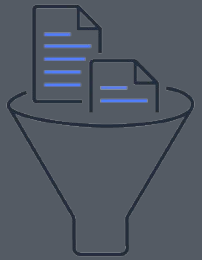
Storage

Storage

Metadata	Databases Tables Columns Locations
ETL	Extract Transform and Load
Security	Permissions and Grants (DB/Table/Column)
Storage	Cost-Effective, Long-Term Storage

Service Responsibilities

Data Lake



Metadata

AWS Glue Crawler

AWS Glue Data Catalog

ETL

AWS Glue Jobs

AWS Glue Workflows

Security

AWS Identity and Access Management

AWS Lake Formation

Storage

Amazon S3

Data Access Layer Responsibilities

Query engine – join data!

JDBC and ODBC support

Glue Data Catalog integration

Ephemeral?

Data Access Layer
Where to start?

Amazon Athena

START HERE

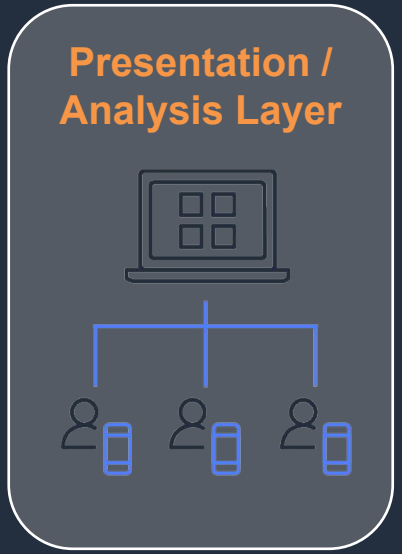
Amazon EMR

THEN HERE

Amazon Redshift

OR HERE

Presentation / Analysis Layer



Amazon QuickSight

Amazon EC2

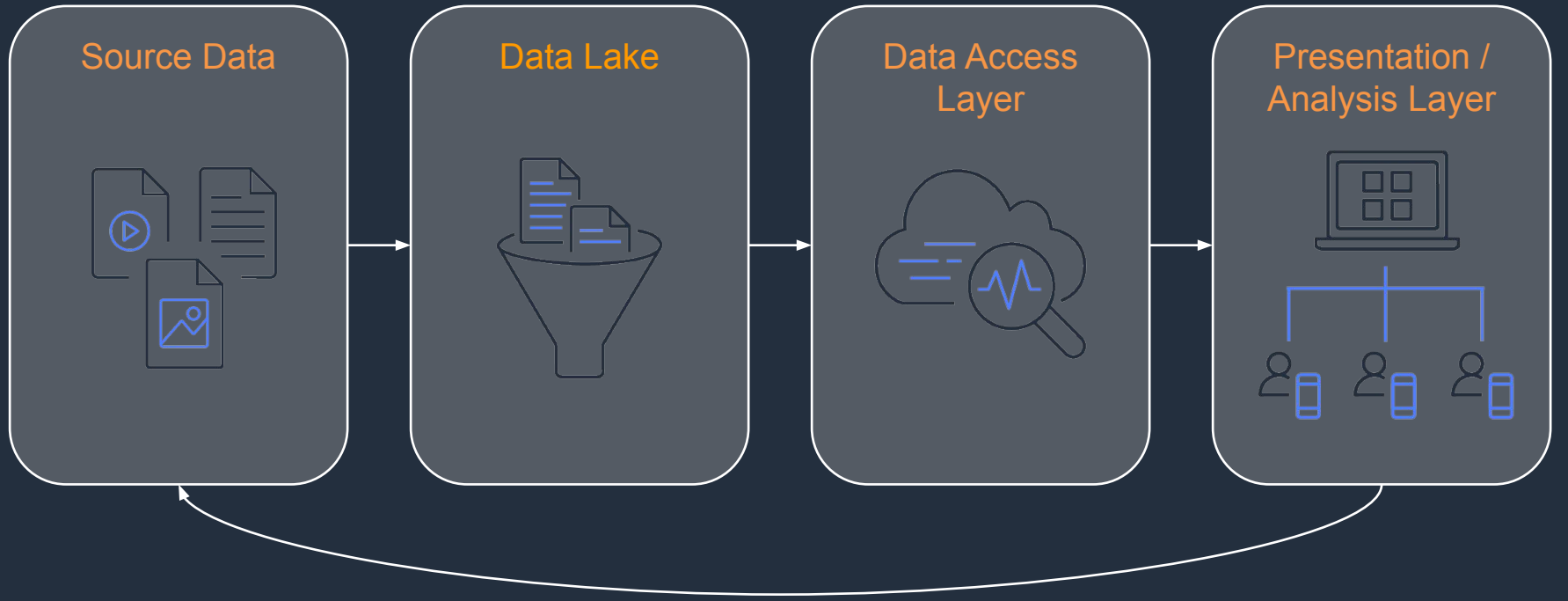
Jupyter
RStudio

AWS Management
Console

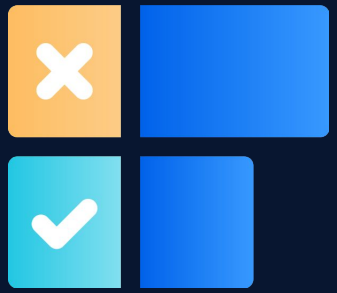
Your
Laptop

Tableau
Generic BI Tools
HPC Infrastructure
…

‘The loop’


Source Data
Data Lake
Data Access Layer
Presentation / Analysis Layer

nClouds



Poll 2

aws

nClouds℠

Data Lakes Use Cases

Carlos Rodriguez
Senior DevOps Engineer
nClouds℠

Use Case

POC

- Our primary PoC use case is to understand how global events have impacted the U.S. stock market and enable financial decisions based on the most recent stock market data.

Data Lakes Use Cases

ETL Process

- Glue crawlers grab the data set from S3 and transforms it before ingesting it into Redshift.

Querying and Visualization

- Data in Redshift can be queried by Athena using a custom Lambda Function integration.
- At the end of the pipeline we create visuals from the Redshift database using AWS Quicksight.

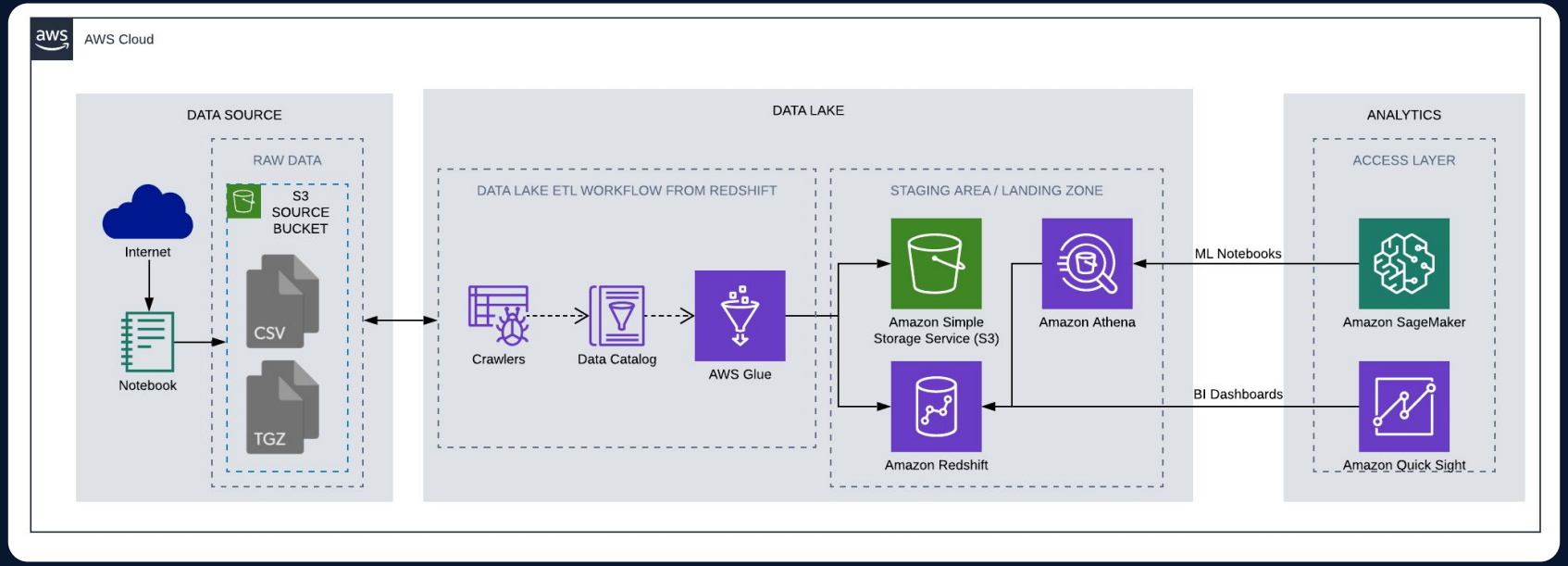
S3 Data Sources

- Custom data ingestion into S3 using Jupyter Notebook.

Analysis of Data

- Data analytics done with Amazon SageMaker and visualizations with Amazon Quicksight.

Data Lake
POC Architecture Review



High-level view of the end-to-end architecture

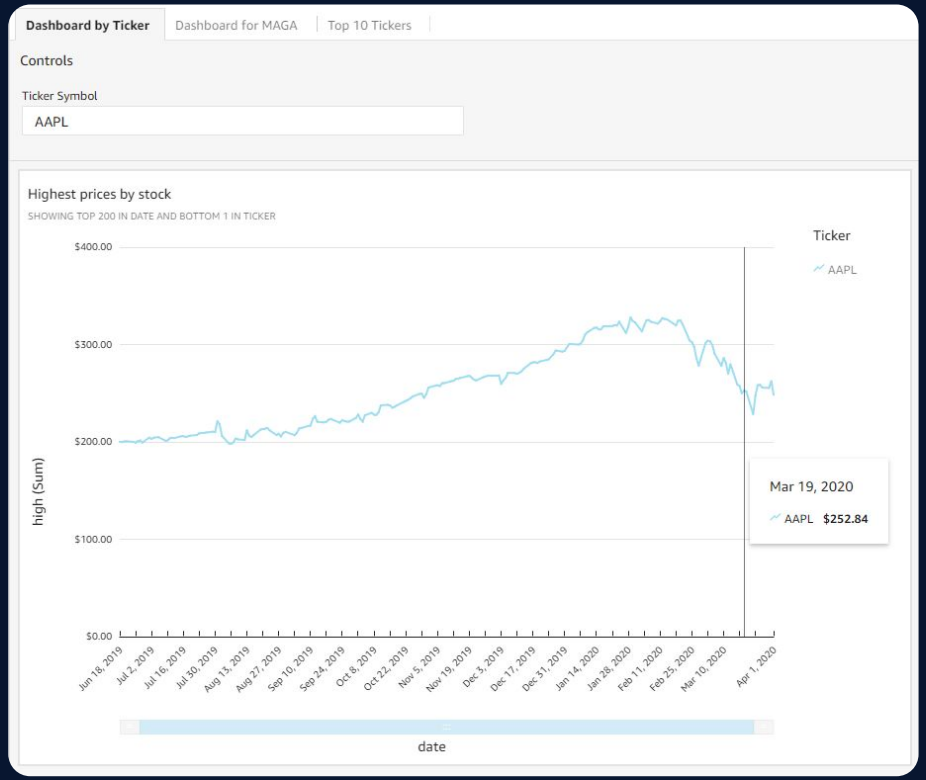
Data Lake
POC Dataset

We ingested a publicly available dataset from Kaggle and Nasdaq. Then we automated an Amazon SageMaker Studio notebook to download the data and store it in the appropriate Amazon S3 location.

As a result, we had a more useful folder structure, which enabled us to use Athena to preview the data and then AWS Glue services to make it available in the analytics section of our PoC.

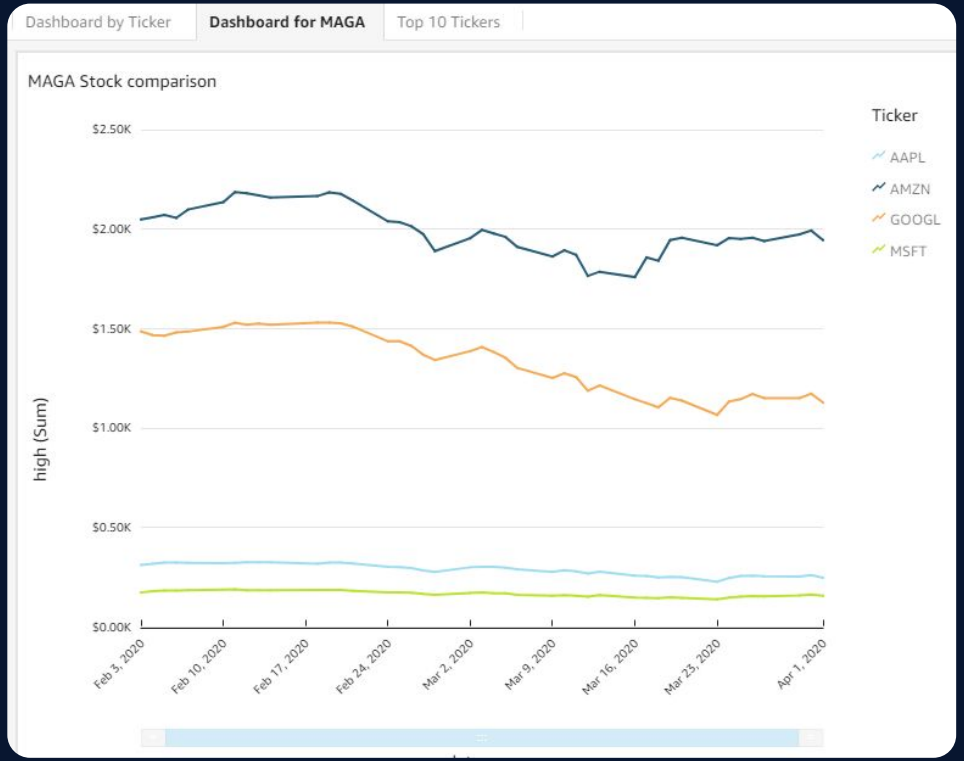
Related Blog: How to Create Custom Partitions in Amazon Athena

Data Lake POC Visualization



- The data shows the jump in the highest prices by stock, which peaked in February 2020 before they crashed in March and April of the same year.

Data Lake POC Visualization



- The dashboard, MAGA Stock comparison, proved very useful in our analysis. This dashboard filtered those stocks from the year 2020 and displayed their individual and group performances from February to April 2020.

nClouds

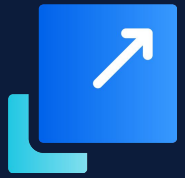
Data Lake as Code on AWS

Carlos Rodriguez
Senior DevOps Engineer
nClouds

Data Lake POC Demo

nClouds℠

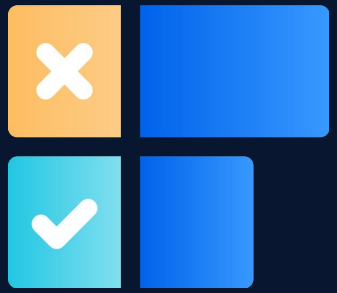
Setup Overview



Data Sets Used

Results

nClouds℠



Poll 3

aws

nClouds℠

Demo

Getting Started: Data Lake as Code
Implementation & Process

Kireet Kokala
VP, Big Data & Analytics
nClouds℠

Data Lake Implementation & Process

Readiness

1. **Assessing current footprint**
   - A combination of architectural, DevOps tools such as nOps, and technical interviews to determine technical landscape.
   - Often via discovery / assessment to determine t-shirt sizing.

2. **Data in play**
   - Analyze customer data sets around velocity, variety, volume.
   - Determine data security (of data at motion and rest).
   - Technology landscape review for
     - Data Movement (ELT) via AWS Glue and Partner Ecosystem (Upsolver, Fivetran, etc.)
     - Data Curation (transformations)
     - Data Analytics and Science (ML, Augmented AI)
     - Any BI and Dashboards for Data Visualization

Data Lake Implementation & Process

Process

3. AWS Data Lake Process, Cost Clarity, Timing

- Based on our customer's footprint, the nClouds assessment, and the nature of the data of your use case, we provide a SOW within 48 hours.
  - Customer footprint assessment.
  - Recommended high-level solutions.
  - Professional services to design and implement a data lake into your solution.
  - Professional services to improve the performance of your existing data lake while increasing cost savings.

nClouds Data & Analytics

Assessments

- Analytics Ecosystem (Tooling)
- Business Intelligence Strategy
- Solution Architecture Modernization
- Data Lakes and Data Warehouses

Data Lakes / Data Warehouse

- POC
- Enablement and Implementation
- BI Integration

Data Movement

- ETL / ELT
- Implementation
- Reporting Integration

Machine Learning

- POC
- Sagemaker Migration
- Services Integration
- Solution Acceleration

nClouds

Q&A

aws