aws | Pop-up Loft
JOHANNESBURG

aws | Pop-up Loft
JOHANNESBURG

Building a Serverless Data Lake

Tame Your Big Data with Kinesis Firehose, S3, Glue & Athena

Hasan-Basri AKIN, MSc., Exec-MBA
Amazon Web Services

2

Overview

Serverless Data Lake Workshop

Objective

- This workshop is meant to give customers a hands-on experience with mentioned AWS services.
- Serverless Data Lake workshop helps customers build a cloud-native and future-proof serverless data lake architecture.
- It allows hands-on time with AWS big data and analytics services including Amazon Kinesis Services for streaming data ingestion and analytics, AWS Glue for ETL and Data Catalogue Management, Amazon Athena to query data lake.

Target Audience

- Data analysts, IT Managers, Solutions Architects, Data Scientists.
- AWS hands-on knowlegde: AWS Console, IAM, S3, CloudWatch.
- Concept knowledge: Hadoop, NoSQL, ETL, Kafka Streams. HIVE, Presto engines. JSON, Parquet data formats.

Audience Check

Prerequisites

Lab Preparations

1. **Bring your own AWS Account**: You need to use your own AWS Account/user for this workshop. Give AdminRights to the user. (In the lab guide, there are detailed instructions to implement a fine grained policy). It is not recommended to do the workshop from your AWS production account/user.
2. **Install Kinesis Data Generator Tool**: install the KDG tool: Follow the KDG Guide on https://awslabs.github.io/amazon-kinesis-data-generator/web/help.html to install and configure the KDG on your AWS account. You simply need to use the CloudFormation template to create KDG to us-west-2 (Oregon) region
3. Download the Lab Guide & Code: https://github.com/CloudyVolla/aws-serverless-data-lake-workshop including the slides from the workshop

Expected Costs & Cleanup

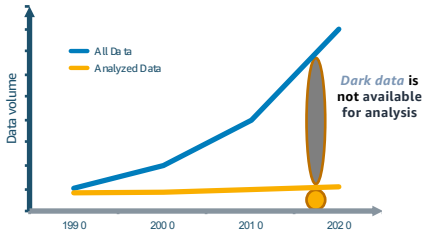
- You are responsible for the cost of the AWS services used while running this lab.
- As of the date of publication, the baseline cost for running this solution as-is should be around < USD 3.
- Make sure you delete the resources you created.
- The permissions shall either be removed for the user after the immersion day, or they shall be turned into more fine-grained permissions.

If you don't have AWS account, don't worry, pair with someone or simply watch as go through the hands-on-labs.

Hands-on Labs Overview

Ingestion to Data Lake Storage **S3** with **Kinesis Firehose**	**LAB** Module 1 1.1 - Create simulated data in real-time data with KDG 1.2 - Create Kinesis Firehose Delivery Stream to load Stream data on S3
Data Cataloging & ETL with **Glue**	**LAB** Module 2 2.1 - Cataloging a Data Source with Glue 2.2 - Transforming a Data Source with Glue
Ad-Hoc Analytics for a Public Dataset with **Athena**	2.4 – Working with a 170GB Public Dataset in Data Lake with Athena
Reporting & Visualization with **Quicksight** & 3PP BI Tools	**DEMO** Module 3 3.1 – Querying S3 Data Sources with Athena 3.2 – Data Visualization with QuickSight and with 3PP Tools using JDBC Endpoints

Big Data Challenges Lead to Dark Data



Sources:
Gartner: User Survey Analysis: Key Trends Shaping the Future of Data Center Infrastructure Through 2011
IDC: Worldwide Business Analytics Software 2012–2016 Forecast and 2011 Vendor Shares

Big Data still challenges most enterprises

Rigid data ingest
Inability to process new types of data

Data doesn't "connect"
Siloed data prevents single source of truth

Limited analytics platforms
Inability to generate new types of insights

Defining the AWS data lake

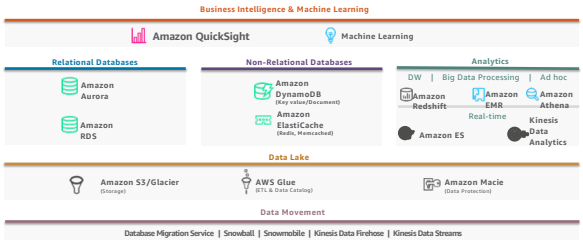
Data lake is an architecture with a virtually limitless centralized storage platform capable of categorization, processing, analysis, and consumption of heterogeneous data sets

Key data lake attributes

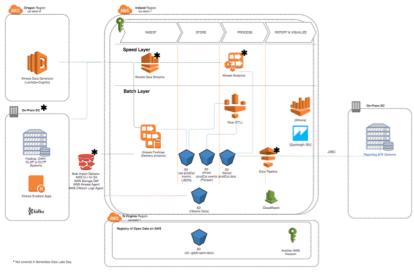
- Decoupled storage and compute
- Rapid ingest and transformation
- Secure multi-tenancy
- Query in place
- Schema on read

One platform for data and analytics

Broadest and deepest portfolio of purpose-built services



Workshop Solution Architecture



Architecture Highlights

- Serverless
- Hybrid

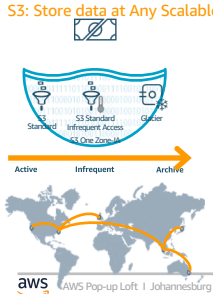
A note on Serverless

Hands-on Labs Overview

Ingestion to Data Lake Storage **S3** with **Kinesis Firehose**	**LAB** Module 1 1.1 - Create simulated data in real-time data with KDG 1.2 - Create Kinesis Firehose Delivery Stream to load Stream data on S3
Data Cataloging & ETL with **Glue**	**LAB** Module 2 2.1 - Cataloging a Data Source with Glue 2.2 - Transforming a Data Source with Glue
Ad-Hoc Analytics for a Public Dataset with **Athena**	2.4 – Working with a 170GB Public Dataset in Data Lake with Athena
Reporting & Visualization with **Quicksight** & 3PP BI Tools	**DEMO** Module 3 3.1 – Querying S3 Data Sources with Athena 3.2 – Data Visualization with QuickSight and with 3PP Tools using JDBC Endpoints

Service Used in this Lab

S3: Store data at Any Scalable with Unmatched Durability & Availability



- Built to store any amount of data
- Runs on the world's largest global cloud infrastructure
- Designed to deliver 99.999999999% durability
- Geographic redundancy & automatic replication
- Seamlessly replicates data between any region
- Tiered storage to optimize price/performance: Store data at $0.023/GB/month at S3 ($0.004/GB/month at Glacier)

Amazon Kinesis—Real Time

Easily collect, process, and analyze video and data streams in real time

Kinesis Video Streams
Capture, process, and store video streams for analytics

Kinesis Data Streams
Build custom applications that analyze data streams

Kinesis Data Firehose
Load data streams into AWS data stores

Kinesis Data Analytics
Analyze data streams with SQL

Service Used in this Lab

Amazon Kinesis Firehose

Capture and submit streaming data to Firehose

Firehose loads streaming data continuously into S3, Amazon Redshift, and Amazon ES

Analyze streaming data using your favorite BI tools

Zero administration: Capture and deliver streaming data to Amazon S3, Amazon Redshift, or Amazon Elasticsearch Service **without writing an app or managing infrastructure.**

Direct-to-data-store integration: Batch, compress, and **encrypt** streaming data for delivery **in as little as 60 seconds.**

Seamless elasticity: Seamlessly scales to match data throughput without intervention.

LAB 1

Summary of Lab 1

Congratulations! You have successfully created an ingestion pipeline without servers, which is ready to process huge amounts of data.

In the next lab, you will create the processing pipeline to convert, transform the data from the ingestion layer.

Hands-on Labs Overview

Ingestion to Data Lake Storage **S3** with **Kinesis Firehose**	**LAB** Module 1 1.1 - Create simulated data in real-time data with KDG 1.2 - Create Kinesis Firehose Delivery Stream to load Stream data on S3
Data Cataloging & ETL with **Glue**	**LAB** Module 2 2.1 - Cataloging a Data Source with Glue 2.2 - Transforming a Data Source with Glue
Ad-Hoc Analytics for a Public Dataset with **Athena**	2.4 – Working with a 170GB Public Dataset in Data Lake with Athena
Reporting & Visualization with **Quicksight** & 3PP BI Tools	**DEMO** Module 3 3.1 – Querying S3 Data Sources with Athena 3.2 – Data Visualization with QuickSight and with 3PP Tools using JDBC Endpoints

Service used in this Lab
AWS Glue: Data Catalog & ETL Service

- Automatically discovers data and stores schema
- Data is immediately searchable, and available to extract, transform, and load (ETL)
- Automatically generates customizable ETL code
- Schedules and runs your ETL jobs
- Serverless

AWS Glue: components

Data Catalog

- Hive metastore compatible metadata repository of data sources.
- Crawls data source to infer table, data type, partition format.

Job Authoring

- Generates Python code to move data from source to destination.
- Edit with your favorite IDE; share code snippets using Git.

Job Execution

- Runs jobs in Spark containers – automatic scaling based on SLA.
- Glue is serverless - only pay for the resources you consume.

Glue data catalog

Discover and organize your data sets

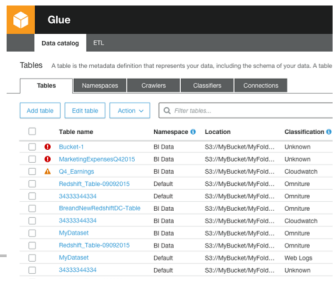
AWS Glue Data Catalog

Manage table metadata through a Hive metastore API or Hive SQL. Supported by tools such as Hive, Presto, Spark, etc.

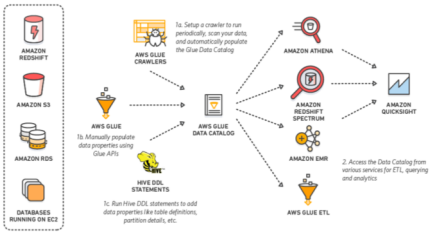
We added a few extensions:

- **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

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

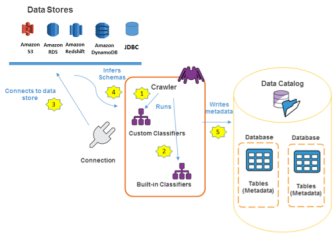


AWS Glue Data Catalog



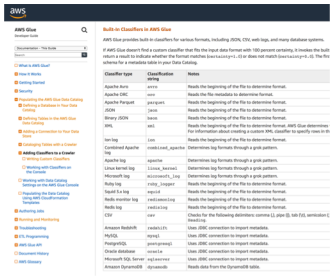
- A unified metadata repository across data sources & data formats
- Integrated with other AWS services

AWS Glue Workflow Overview



1. Use a crawler to infer the schema of your data.
2. Use Classifiers to Populate the AWS Glue Data Catalog with table definitions.
3. Crawler connects to the data store.
4. Schema inference occurs.
5. Define a job that describes the transformation of data from source to target.
6. Run your job to transform your data.
7. Monitor

AWS Glue Built-in Classifiers



Ref: AWS Glue Developer Guide

Crawlers: Auto-populate Data Datalog

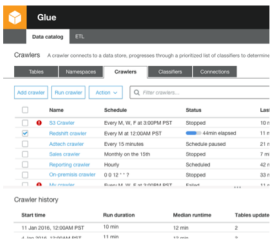
Automatic schema inference:

- Built-in **classifiers** detect file type and **extract schema:** record structure and data types.
- Add your own or share with others in the Glue community - It's all Grok and Python.

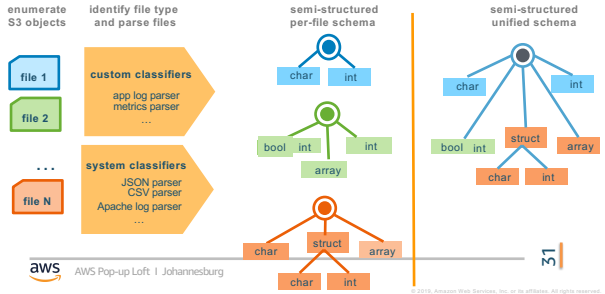
Auto-detects Hive-style partitions, grouping similar files into one table.

Run crawlers on schedule to **discover** new data and **schema changes.**

Serverless – only pay when crawls run.



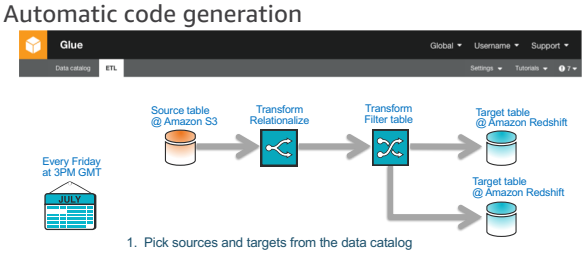
Crawlers: automatic schema Inference



Job authoring in Glue

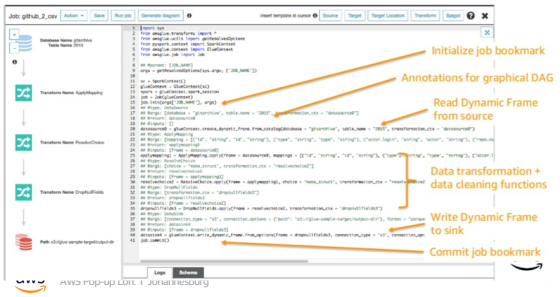
Make ETL job authoring like code development using your own tools

Automatic code generation

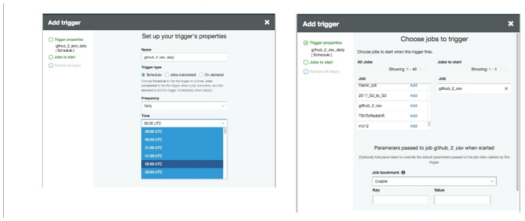


1. Pick sources and targets from the data catalog
2. Glue generates transformation graph and *Python code*
3. Specify trigger condition

Anatomy of a Generated ETL Job Script



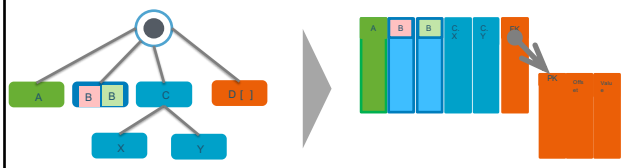
Scheduling Jobs



Several event types

Pass parameters

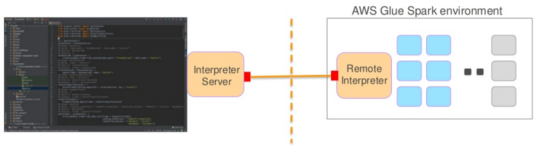
Glue transformations are flexible and adaptive



- Flatten semi-structured objects with arbitrary complexity into relational tables, on-the-fly.
- Pivot arrays and other collection types into a separate table, generating key-foreign key values.
- Modify mapping as the source schema changes, and modify the target schemas as needed.

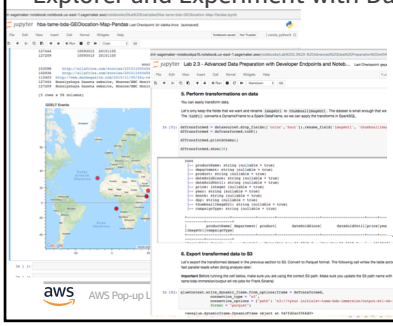
Kahoot!

Edit & Test with Dev-Endpoints



Connect your IDE to an AWS Glue development endpoint
Environment to interactively develop, debug & test ETL code.

Explorer and Experiment with Data



Connect your notebook (e.g. Jupyter, Zeppelin) to an AWS Glue development endpoint.

Interactively experiment and explore datasets and data sources.

Deploy to production
- Push scripts to S3
- Create or register with ETL job

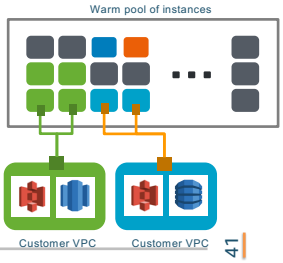
Orchestration & resource management

Fully managed, serverless job execution

Serverless Job Execution

There is no need to provision, configure, or manage servers

- Warm pools: pre-configured fleets of instances to reduce job startup time
- Auto-configure VPC and role-based access
- Automatically scale resources to meet SLA and cost objectives
- Only pay for the resources you consume per-second billing (10-minute min)



Apache Spark & AWS Glue ETL

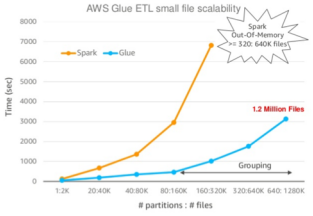
What is Apache Spark?

- Parallel, scale-out data processing engine
- Fault-tolerance built-in
- Flexible interface: Python scripting, SQL
- Rich eco-system: ML, Graph, analytics, …

AWS Glue ETL libraries

- Integration: Data Catalog, job orchestration, code-generation, job bookmarks, S3, RDS
- ETL transforms, more connectors & formats
- New data structure: Dynamic frames

Performance: Lots of small files



Lots of small files, e.g. Kinesis Firehose

Vanilla Apache Spark (2.1.1) overheads

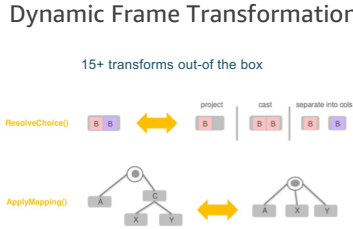
- Must reconstruct partitions (2-pass)
- Too many tasks: task per file
- Scheduling & memory overheads

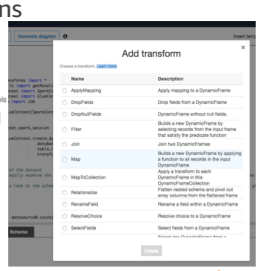
AWS Glue Dynamic Frames

- Integration with Data Catalog
- Automatically group files per task
- Rely on crawler statistics

Dynamic Frame Transformations

15+ transforms out-of the box

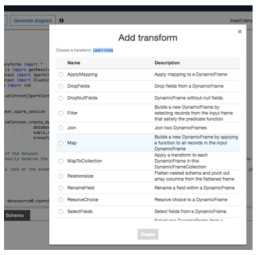




Useful AWS Glue Transforms

- toDF(): Convert to a Dataframe
- Spigot(): Sample data of any Dynamic Frame to S3
- Filter(), map(): Apply python UDFs to Dynamic Frames
- Join(): Join two Dynamic Frames

and more…



Hands-on Labs Overview

Ingestion to Data Lake Storage **S3** with **Kinesis Firehose**	**LAB** Module 1 1.1 - Create simulated data in real-time data with KDG 1.2 - Create Kinesis Firehose Delivery Stream to load Stream data on S3
Data Cataloging & ETL with **Glue**	**LAB** Module 2 2.1 - Cataloging a Data Source with Glue 2.2 - Transforming a Data Source with Glue
Ad-Hoc Analytics for a Public Dataset with **Athena**	2.4 – Working with a 170GB Public Dataset in Data Lake with Athena
Reporting & Visualization with **Quicksight** & 3PP BI Tools	**DEMO** Module 3 3.2 – Data Visualization with QuickSight and with 3PP Tools using JDBC Endpoints

LAB 2.4

Service used in this Lab

Amazon Athena—Interactive Analysis

Interactive query service to analyze data in Amazon S3 using standard SQL

No infrastructure to set up or manage and no data to load

Ability to run SQL queries on data archived in Amazon Glacier (coming soon)

Query Instantly	Pay per query	Open	Easy
Zero setup cost; just point to S3 and start querying	Pay only for queries run; save 30–90% on per-query costs through compression	ANSI SQL interface, JDBC/ODBC drivers, multiple formats, compression types, and complex joins and data types	Serverless: zero infrastructure, zero administration Integrated with QuickSight

Public Dataset

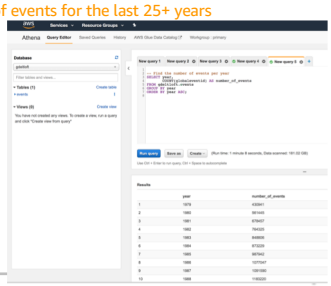
The Global Database of Events, Language and Tone (GDELT) Project monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organisations, counts, themes, sources, emotions, counts, quotes, images and events driving our global society every second of every day. The data set v1.0 is publicly available in S3 in the Registry of Open Data on AWS.

Ad-Hoc Query from the Data Lake:

Finding the aggregate number of events for the last 25+ years

Notice the data amount scanned? The results are returned by scanning 170+ GB of data from 4000+ uncompressed CSV files on S3.

That's the power of HIVE, Presto and other Hadoop Technologies simplified by Athena Service.



HIVE DDL Commands

- We are creating a schema definition in our Glue service, in our Data Catalogue.
- The actual data is in another AWS account, and in another AWS region (us-east-1: Northern Virginia)
- You can Access this data, because it is a public dataset located in 's3://gdelt-open-data/events/folder, and is open to everyone.
- Although we are creating TABLEs, there is no database. Events table is a representation of thousands of TSV (Tab Seperated Files) files stored in S3.
- Technologies like Apache HIVE and Presto enables accessing them using SQL like expressions.

Summary of Lab 2

Congratulations. You have successfully catalogued your data, created serverless ETL jobs for transforming your data.

You have also performed an ad-hoc query of thousands of uncompressed CSV files (containing hundreds of millions of lines, more than 170 GB) from data lake storage using Athena.

Hands-on Labs Overview

Ingestion to Data Lake Storage **S3** with **Kinesis Firehose**	**LAB** Module 1 1.1 - Create simulated data in real-time data with KDG 1.2 - Create Kinesis Firehose Delivery Stream to load Stream data on S3
Data Cataloging & ETL with **Glue**	**LAB** Module 2 2.1 - Cataloging a Data Source with Glue 2.2 - Transforming a Data Source with Glue
Ad-Hoc Analytics for a Public Dataset with **Athena**	2.4 – Working with a 170GB Public Dataset in Data Lake with Athena
Reporting & Visualization with **Quicksight** & 3PP BI Tools	**DEMO** Module 3 3.2 – Data Visualization with QuickSight and with 3PP Tools using JDBC Endpoints

Demo

*Due to time constraint, we will make a demo of Lab 3.
You are encouraged to do it yourself after the session using the Lab Guide documents.

Amazon QuickSight

Amazon QuickSight is a fast, cloud-powered business intelligence (BI) service that makes it easy for you to deliver insights to everyone in your organization.

Data Scientist (Author)

Give power users and analysts the freedom to do their own self-serve data discovery and analysis on governed data you control

Dashboard Creator (Author)

Create and publish rich, interactive dashboards to all of your users

End User (Reader)

With the new Reader Role, you can provide everyone in your organization secure, easy access to interactive dashboards and reports, on any device

Why Amazon QuickSight?

No servers to manage

Amazon QuickSight has no servers or software to manage, maintain, deploy, upgrade or migrate. We do the heavy lifting so you don't have to.

Native AWS integration

Amazon QuickSight securely integrates with your data sources and AWS services like Amazon Simple Storage Service (Amazon S3), Redshift, Amazon Athena, Amazon Aurora, Amazon Relational Database Service (Amazon RDS), AWS Identity and Access Management (IAM), AWS CloudTrail, Amazon Cloud Directory and more - providing you with everything you need to build an end-to-end BI solution.

Easily scale from 10 users to 10,000

QuickSight automatically scales with your usage and activity, with no need for additional infrastructure. QuickSight will grow with your organization's needs from a few users to tens of thousands of users.

Pay only for what you use

Provide read-only access to interactive dashboards and pay only when your users access them with Pay-per-Session pricing. With Amazon QuickSight there are no upfront costs, no annual commitments and no charges for inactive users.

You are Invited to the next Loft Session about QuickSight on Monday

Serverless BI Using Amazon QuickSight

11th of March 2019
14:00-15:00 @Johannesburg

aws

Kahoot! time

Summary
AWS helps unlock Big Data in three ways

Future-proofed for new data types

Ingest **new data for differentiated experiences:** image/video, social media, IoT sensors, and more

Single source of truth (aka "data lake")

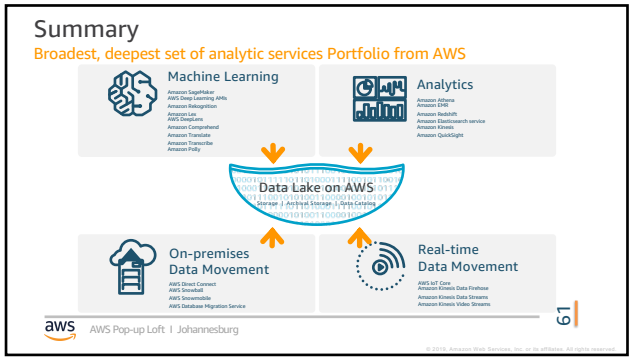
Secure and scalable backbone for all incoming data and all analytics – **no more data siloes**

The right analytics tool for every job

New analytics possibilities – machine learning, natural language processing, Hadoop-as-a-service, and more

Summary
Benefits of Data Lakes from AWS

- Open and comprehensive
- Secure
- Scalable and durable
- Lowest cost


Summary
Broadest, deepest set of analytic services Portfolio from AWS
Machine Learning
Analytics
Data Lake on AWS
On-premises Data Movement
Real-time Data Movement
AWS Pop-up Loft | Johannesburg
61


More Data Lakes & Analytics on AWS than Anywhere Else
AWS Pop-up Loft | Johannesburg


aws SUMMIT
Cape Town, 11 July 2019, CTICC


AWSOME DAY
ONLINE CONFERENCE
26.03.19
In partnership with intel
Kick start your journey to the AWS Cloud
Register for free:
http://bit.ly/awsomeday2019

Please help us improve
Complete the session survey!

Building a Serverless Data Lake
Tame Your Big Data with Kinesis Firehose, S3, Glue & Athena

7th of March 2019
13:00-15:00 @Johannesburg

aws