Serverless Data Lake
Immersion Day

What is the Serverless Data Lake Immersion Day?

The Serverless Data Lake Immersion Day workshop is prepared to assist you ingest, store, transform, create insights on unstructured data using AWS serverless services.

We will build a cloud-native and future-proof serverless data lake architecture using Amazon Kinesis Firehose for streaming data ingestion, AWS Glue for ETL and Data Catalogue Management, S3 for data lake storage, Amazon Athena to query data lake and provide JDBC Connectivity to external BI tools, and finally Amazon Quicksight for data visualization.

Benefits of the Serverless Immersion Day

Serverless Data Lake Day helps customers create an end-to-end, cloud-native and future-proof data lake pipeline without servers. It allows hands-on time with AWS big data and analytics services including Amazon Kinesis, AWS Glue, Amazon Athena and Amazon Quicksight. The focus is on the batch processing layer of the Lambda Architecture data-processing design pattern.

To demonstrate the power of data lake architectures, we will demonstrate how to catalogue an open dataset at AWS Open Data Registry using AWS Glue and query using Amazon Athena. A centralized security governance approach is also demonstrated by creating IAM roles & policies needed for the labs using CloudFormation templates.

Modules | Data Movement

Data Ingestion & Central Storage

The data ingestion step comprises data ingestion by both the speed and batch layer, usually in parallel. For the batch layer, historical data can be ingested at any desired interval. For the speed layer, the fast-moving data must be captured as it is produced and streamed for analysis.

For fast data ingestion Kinesis Data Streams is the recommended service to ingest streaming data into AWS. Customers can use Amazon Kinesis Agent, a pre-built application, to collect and send data to an Amazon Kinesis stream or use the Amazon Kinesis Producer Library (KPL) as part of a custom application.

For batch ingestions, customers can use AWS Glue or AWS Database Migration Service to read from source systems, such as RDBMS, Data Warehouses, and No SQL databases. Amazon Simple Storage Service (Amazon S3) forms the backbone of such architectures providing the persistent object storage layer for the data lake.

Modules | Catalog Data

Data Cataloging and ETL

Data cataloging is the ability to understand what data is in the lake through crawling, cataloging, and indexing of data. ETL is performing data engineering on the data.

The cataloging & ETL of the data in S3 can be performed using using AWS Glue, a fully managed ETL service on the AWS platform.

This workshop includes hands-on labs on the main topics covered.

aws partner network | immersion days