slalom | databricks

Slalom Data Lakehouse Accelerator

Accelerator Overview

Data Lakehouse Accelerator

slalom

The Lakehouse accelerator helps clients get going faster and provides a framework to grow

< 3 weeks

Time to stand up a production grade data lakehouse *

< 1 hour

Time to onboard a new data source to the platform**

**No Spark experience required

3

Curated use cases generated per month per data engineer

*Subject to architectural approval and access to the client environment

slalom

What is the Lakehouse Accelerator?

The accelerator is an all-in-one package for standing up a Data Lakehouse

Cloud Infrastructure

The accelerator comes with Terraform templates that deploy all the necessary infrastructure needed to build a Lakehouse including a Databricks workspace, networking, storage, key stores, and an orchestration tool.

DevOps

The accelerator has all the necessary DevOps components needed to run a development, testing, and production environment. It also includes pipelines to test, approve, and push deployments into each environment.

Data Pipelines & Orchestration

The accelerator includes all the needed parts to extract, transform, load, and curate datasets into a Lakehouse paradigm. It takes advantage of metadata driven pipelines to reduce the time and effort needed to ingest sources and extract value out of data.

ML & Curated Use Case Framework

The accelerator is equipped with a framework to rapidly develop machine learning and curated use cases in a sandbox environment and provides processes to deploy them into production when ready.

*Subject to architectural approval and access to the client environment

slalom

The flow of a dataset in the Lakehouse Accelerator

CSV JSON

DELTA LAKE

databricks

Power BI

tableau

Landing Zone

All data sources are ingested into the landing zone of the data lake. The files present in this zone would be an exact replica of its source type. For database sources, the data is ingested as parquet files.

Purpose: To store an exact copy of the ingested file for archival purposes in the data lake environment.

Raw Zone

All data in the landing zone is moved to the raw zone as a Delta Lake Table. This table will contain unaltered data from the sources and can be queried from Databricks Notebooks.

Purpose: To store unaltered data as a Delta Lake Table to allow for Databricks querying and processing. These tables are often used when troubleshooting the Processed Zone.

Processed Zone (For Power Users)

Data from the raw zone is processed (e.g. de-duplicated, trimmed, type constrained) and stored in the processed zone as a Delta Lake Table that can be queried from Databricks Notebooks.

Purpose: To store processed and structured data that can be easily used to create curated datasets. This processed data will be the main source for data scientists and machine learning engineers to create models.

Curated Zone (For Business Users)

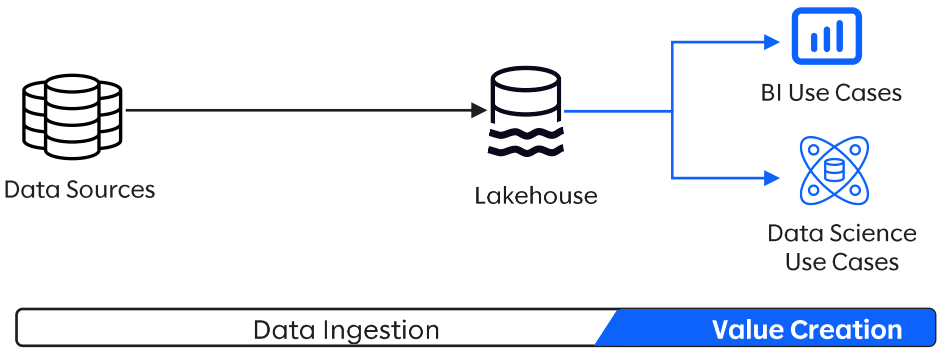
By using data from the Processed Zone, power users can create curated datasets by writing SQL/Python/Scala code and storing this data as Delta Lake Tables. Business users can view data in the curated zone using data visualization tool.

Purpose: To store combined like data from a variety of sources and/or aggregated data for the use of the business users. Data from this zone can also be used to feed a data warehouse.

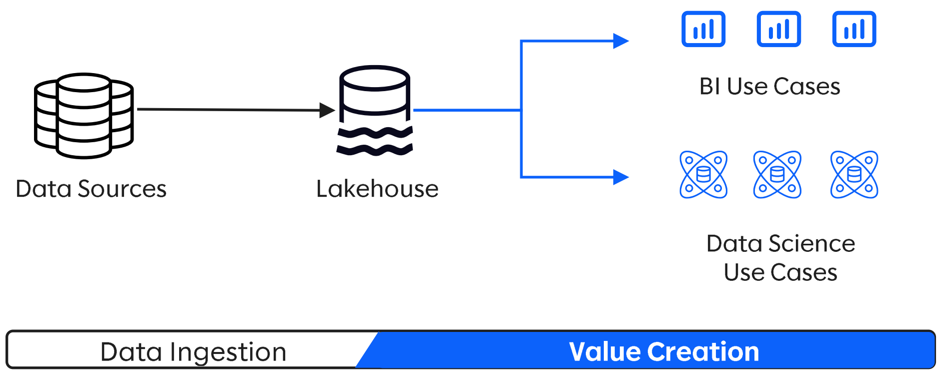
How the Accelerator Creates Value

The accelerator simplifies loading data into a Lakehouse so data teams can focus on what really matters - extracting value from data

Without the Lakehouse Accelerator - data teams spend more time loading data than creating value



With the Lakehouse Accelerator - this is reversed, allowing data teams to spend more of their time creating value



slalom

Metadata-driven Pipelines

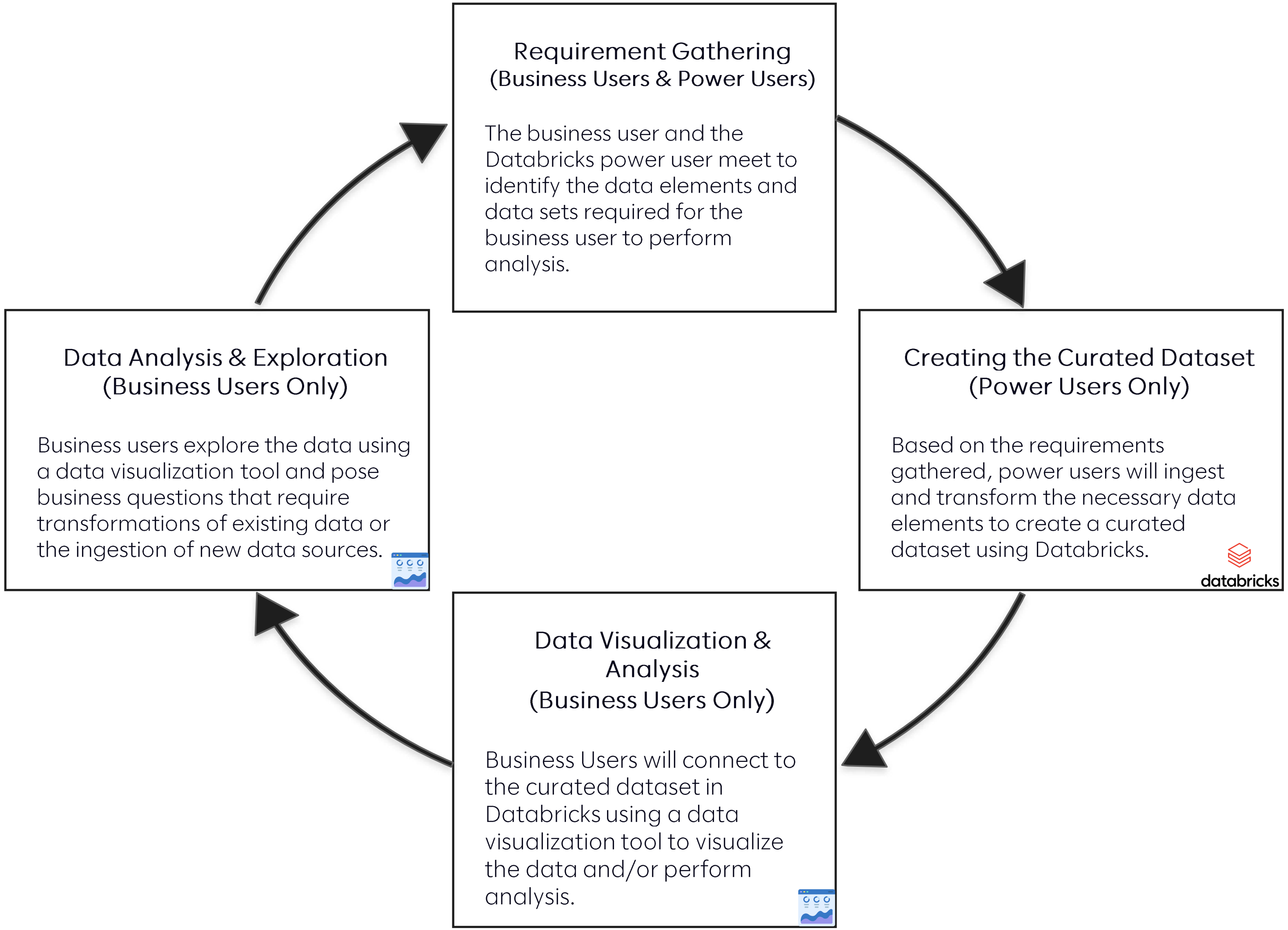
What are metadata-driven pipelines?

Metadata-driven pipelines abstract the process of creating data ingestion workflows by using an intuitive and reusable template. This template is populated with information about a data source, the pipeline then interprets this template and uses it to load data into the lakehouse in a standardized fashion. This process democratizes data ingestion and allows data teams to onboard sources at a faster pace.

Why use metadata-driven pipelines?

- **Reusability**
- **Faster Set-up**
- **Standardized Process**
- **Built-in Documentation**

The Curated Dataset Lifecycle



Requirement Gathering (Business Users & Power Users)

The business user and the Databricks power user meet to identify the data elements and data sets required for the business user to perform analysis.

Creating the Curated Dataset (Power Users Only)

Based on the requirements gathered, power users will ingest and transform the necessary data elements to create a curated dataset using Databricks.

Data Visualization & Analysis (Business Users Only)

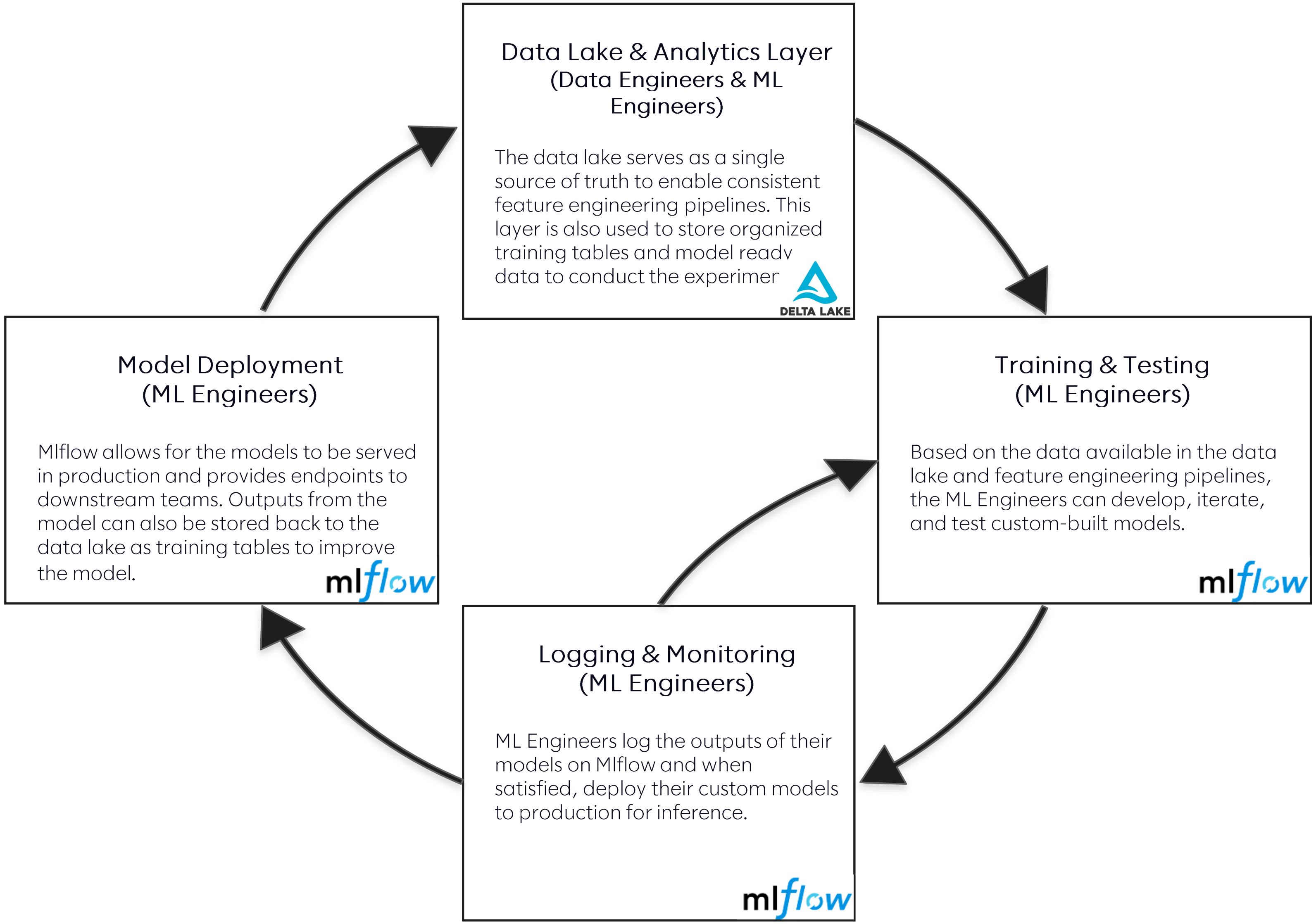
Business Users will connect to the curated dataset in Databricks using a data visualization tool to visualize the data and/or perform analysis.

Data Analysis & Exploration (Business Users Only)

Business users explore the data using a data visualization tool and pose business questions that require transformations of existing data or the ingestion of new data sources.

The ML Experiment Lifecycle

slalom



Data Lake & Analytics Layer (Data Engineers & ML Engineers)

The data lake serves as a single source of truth to enable consistent feature engineering pipelines. This layer is also used to store organized training tables and model readv data to conduct the experimer

Training & Testing (ML Engineers)

Based on the data available in the data lake and feature engineering pipelines, the ML Engineers can develop, iterate, and test custom-built models.

Logging & Monitoring (ML Engineers)

ML Engineers log the outputs of their models on Mlflow and when satisfied, deploy their custom models to production for inference.

Model Deployment (ML Engineers)

Mlflow allows for the models to be served in production and provides endpoints to downstream teams. Outputs from the model can also be stored back to the data lake as training tables to improve the model.

slalom | databricks

Representative Past Performance

customer story: Healthcare Information Exchange

Building an analytics platform for a non-profit that enables rapid response contact tracing

why

A healthcare information exchange (HIE) wanted to better use the mountains of data that it owned. The HIE initially partnered with Slalom to build a modern data platform; however when a global pandemic began, we collectively decided to pivot. The HIE and Slalom needed to automate and re-engineer the current processes that reported vital COVID19 patient data to the state in order to perform contact tracing and so that its citizens could remain informed.

what

Slalom was able to build the cloud infrastructure in about three weeks using our Lakehouse accelerator. Concurrently, our data engineers worked with the healthcare experts in the HIE to transfer their business logic for combining and standardizing many different data sources across different companies into curated outputs for the state government and contact tracers to use.

wow

Through our partnership with the HIE, we were able to produce and deliver hourly contact tracing datasets to the state's contact tracers, allowing them to perform their jobs with timely information. Additionally, Slalom was able to perform the original "pre-COVID" scope, allowing the HIE to have an extensible analytics environment that allows them to add new data sources to the platform in less than an hour.

solutions

Lakehouse Accelerator

Modern Data Architecture

Data Engineering

DataOps

customer story: National Non-profit

Modernizing reporting through a data lake

why

A national non-profit recognized their increasing need for a modernized data platform that would enable them to take advantage of valuable donor information.

The vision was to design a cost efficient and robust enterprise data management solution using cloud technologies to provide scalability with a lean implementation process.

Slalom was engaged to evaluate their current state, development a resilient architecture, and implement a data solution that would meet current needs while providing the ability to grow out as needed.

In addition, Slalom was engaged to develop an education and socialization strategy for the final product.

what

The Slalom project team developed a data lake architecture that would ingest the client's mission critical data sources, consolidate them, and expose them to BI tools for powerful insight creation.

Using our Lakehouse accelerator, the team was able to quickly stand up a platform that was highly performant and easily expandable.

Due to the well thought out architecture and the focus on long-term sustainment, the solution was handed off to the client’s development team with ease, allowing them to add additional data sources and continue to build out the platform.

wow

This solution instantly allowed the client to start drawing insights by blending and manipulating data from a single source.

Business units that were previously unable to access data due to organizational silos were provided with the ability to explore a wealth of donor knowledge.

Data refresh time for the client’s preferred BI tool was lowered from 70 minutes to 15 minutes, and the time required to add additional data sources was cut down from two weeks to just a few hours.

solutions

Lakehouse Accelerator

Modern Data Architecture

Data Engineering

DataOps

slalom

Thank you!