brillio

# DATA ANALYTICS & ENGINEERING DATABRICKS CAPABILITY

**ACCELERATE WHAT MATTERS. NOW.**

# brillio | A STORY OF IMPACT, CREDIBILITY AND DIGITAL AT SCALE

**Headquartered in Silicon Valley**

**+18 Offices**

**Trusted By Fortune 1000 Customers Across Various Industries**

Current Upcoming




**North America**
Edison, NJ
San Ramon, CA
Seattle, WA
Chicago, IL
Atlanta, GA
Dallas, TX
Raleigh, NC
Vancouver, Canada

**Europe**
London, UK
Oradea, Romania

**India**
Bangalore
Pune
Hyderabad
Chennai





**Powered by 3,500+ Brillians**

- Enterprise Architects
- Data Scientists
- Biz-tech Consultants
- Platform Developers
- Design Thinkers

# Our capabilities and service offerings

**CONSULT**: Data & Technology Consulting Strategy

**ENGINEER & TRANSFORM**: Implementation Services

## DATA ENGINEERING

- Data Lake Setup & Data Migration
- Master Data Management
- Customer Data Platform
- Data Governance

## CX ENGINEERING & DIGITAL ANALYTICS

- Customer Insights
- Marketing Insights
- User Experience Insights
- MarTech Engineering

## DATA SCIENCE & AI

- Classic ML/ Behavioral Science
- Advanced AI (CV/NLP)
- AI Industrialization

## BI & ANALYTICS

- BI Modernization
- Self serve enablement
- Analytics @ Scale
- Operational Intelligence

**EVOLVE:** ML/AI , DevOps, DataOps, ModelOps

*Powered by* brillio one.ai

### Personalization

*A comprehensive analytics solution to help enterprises use their rich customer data to recommend next best action*

### Intelligent Migration Suite

*A suite of technical accelerators to smoothen journey from on-prem to cloud with accurate planning and estimation*

### Trust Suite

*A data governance solution ensuring Data catalogue, compliance, Data Lineage and Data audit.*

### Supply Chain 360

*A decision management solution to improve visibility and collaboration in the enterprise*

### Annotation (NLP/CV)

*An automated solution to annotate, augment and synthesize model training data*

### MLOps

*A ModelOps solutions framework to industrialize AI and help organizations incrementally improve their AI maturity*

Best Firms For Data Scientists To Work. 2021 | India

FORRESTER®

Recognized among top Computer Vision Service Providers & Consultancies

FORRESTER®

Featured for our work on Model Ops

zinnov

Leader in AI Engineering Services

FORRESTER®

Featured in NowTech for Customer Analytics 2021 in mid-size segment as leading service provider

Data Science CoE Leader featured as Top 10 Data Scientists in India

Data Science CoE Leader featured in 40 under 40 list 2021

Everest Group PEAK MATRIX®

Recognized as major contender for Data and Analytics service providers 2021

Microsoft

Awarded at MSP Prestige Award for Commitment to Community


DATABRICKS CAPABILITES

# brillio | Databricks Capabilities Overview

## Consulting Services

Consulting engagement to analyze and design:

- Understand and define data modernization use cases
- Design end to end data engineering pipeline
- Design end to end ML based pipelines

## Cluster Design and Workspace

Enable cost effective and highly performant infrastructure:

- Understand and define right cluster types and design
- Design and implement secure and performant Databricks workspace
- Design and implement right clusters and performance parameters for job, interactive types

## Engineering and Implementation

Build data engineering use cases and ETL/ELT jobs:

- Modern Data Platforms using Databricks
- Build Delta Lake enabled advanced data products for reporting

## Analytics and Data Science

Enable enterprise ML implementation:

- Build scalable enterprise grade feature stores and data science build products
- Build end to end Spark ML based ML pipelines
- Build advanced model validation and deployment frameworks

## Accelerators

### Problem Formulation & Solution Architecture

- Databricks cluster design frameworks
- Infra and Security Integration design frameworks for Azure, AWS
- Cluster cost calculation frameworks based on use cases

### Data Engineering

- Rule based data cleansing framework
- Real/near Realtime data validation framework
- Large scale data migration utilities
- Delta Lake framework for data model builds
- Azure, AWS orchestration frameworks for integration with cloud native services
- Delta Lake Update frameworks for large scale

### Analytics/Data Science

- Delta Lake based feature store frameworks
- End to end Spark ML model build/deploy frameworks for Azure, AWS implementation
- Real-time Model validation framework

**Clients**

Lowell, WELLA PROFESSIONALS, COTY, align

**Cloud Platform Expertise**

Azure, Google Cloud Platform, aws

# brillio | Our Experience with Databricks

## Lowell

**Objective:**

- Leveraging **Azure Databricks** implemented one of the largest **Delta Lake platform** in EU to handle ACID like **transformations** and built multiple data science products like data sampler and **feature stores**.

**Solution Highlights/Impact:**

- The solution processed **data loads** reaching **20 TB every week** using 896 core Databricks clusters enabling an **60% faster time to market** of data science model
- Aggregated data sets within **Synapse** used by business users for self-service reporting offering **70% reduction** in **Data preparation for report generation**

## COTY

**Objective:**

- Built an **advanced analytics platform** migrating data from SAP to enable business teams with **self-service insights** and uncover new data science models to drive operational efficiency.

**Solution Highlights/Impact:**

- Leveraged **Azure Databricks** for enabling Data Ingestion, Transformation and machine Learning while **Azure synapse** was used as data warehouse to enable **Self-service reporting using Power BI**
- The solution provided **81% reduction in time to insight generation** and **17% increased operational efficiency**

## align

**Objective:**

- Transfer third party **Loyalty data** from SharePoint through Azure Event Hub and push the same to Azure Blob in expected input format after performing required **validations and transformations using databricks**

**Solution Highlights/Impact:**

- Design/Architect/Develop flow to **handle** the **Loyalty** third party **data**.
- Leveraged **Azure Databricks** for **processing and transformation** of **massive** quantities of **data** and publish successful records
- As part of **audit, success and error** notifications/details were shared with relevant stakeholders

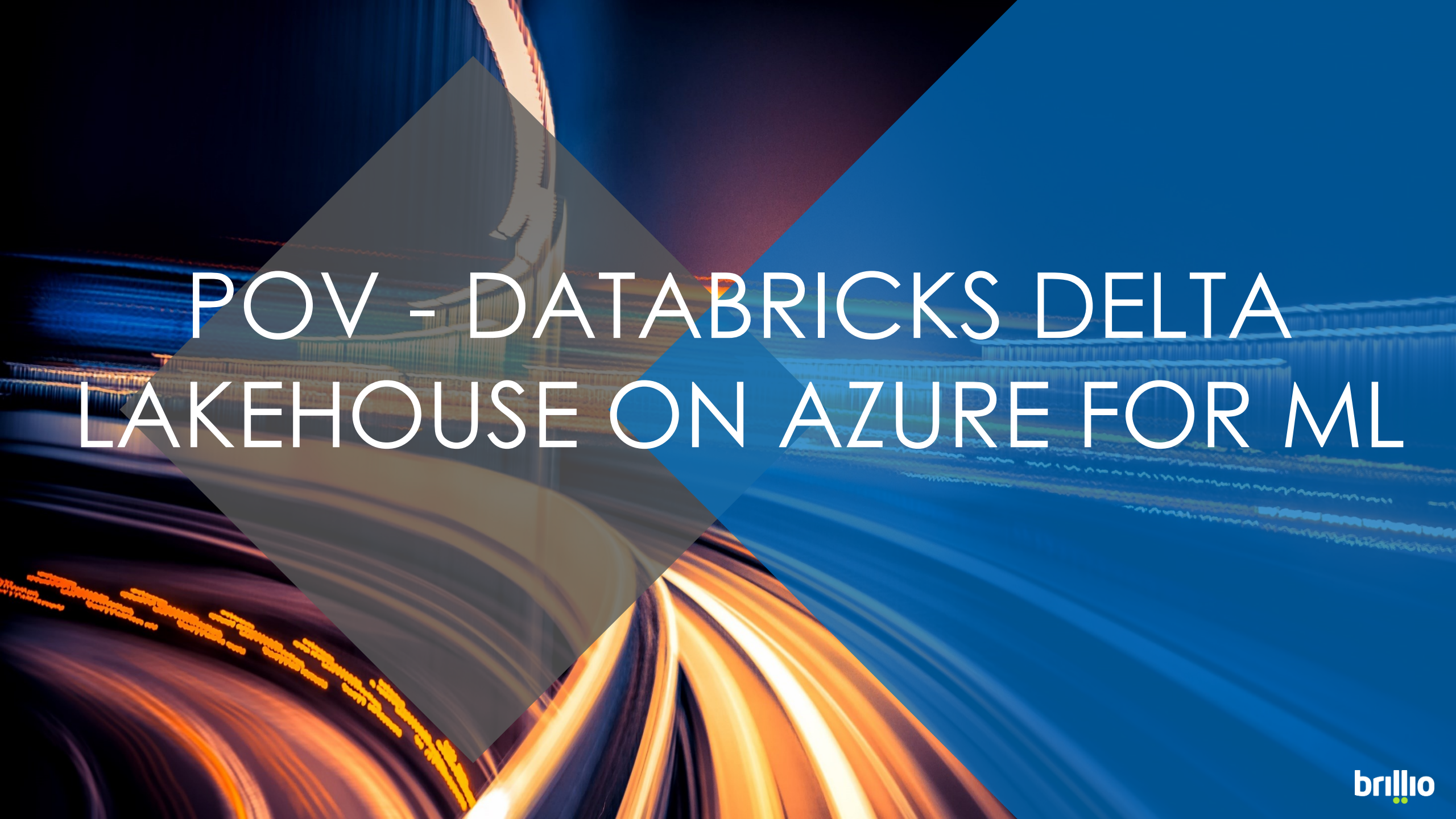
## RNL

**Objective:**

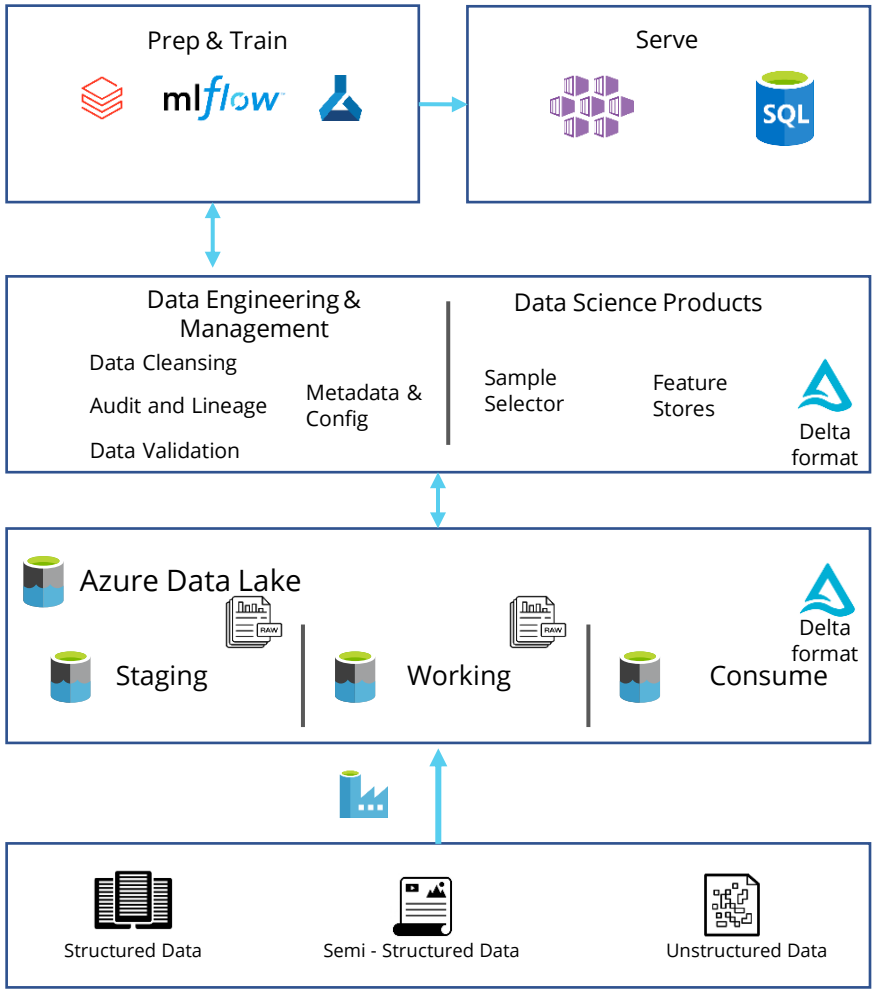
- Accelerate **transformation of Data Supply Chain** to drive efficiency and agility in end-to-end data pipeline through automation, innovation & technology refresh leveraging **Databricks** leading to **enhanced customer experience**

**Solution Highlights/Impact:**

- Setup technology platform on Azure with **automated workflow** for ingestion using **Databricks, auto scale, error handling**, **audit**, & end to end **data lineage**
- The solution enabled **lower operational costs** due to on-demand execution **flexibility of Databricks.**
- **40% reduction** in file **processing time** & **60% reduction** in pipeline **failure rate**


POV - DATABRICKS DELTA
LAKEHOUSE ON AZURE FOR ML
brillio

# DATABRICKS DELTA LAKEHOUSE ON AZURE FOR ML



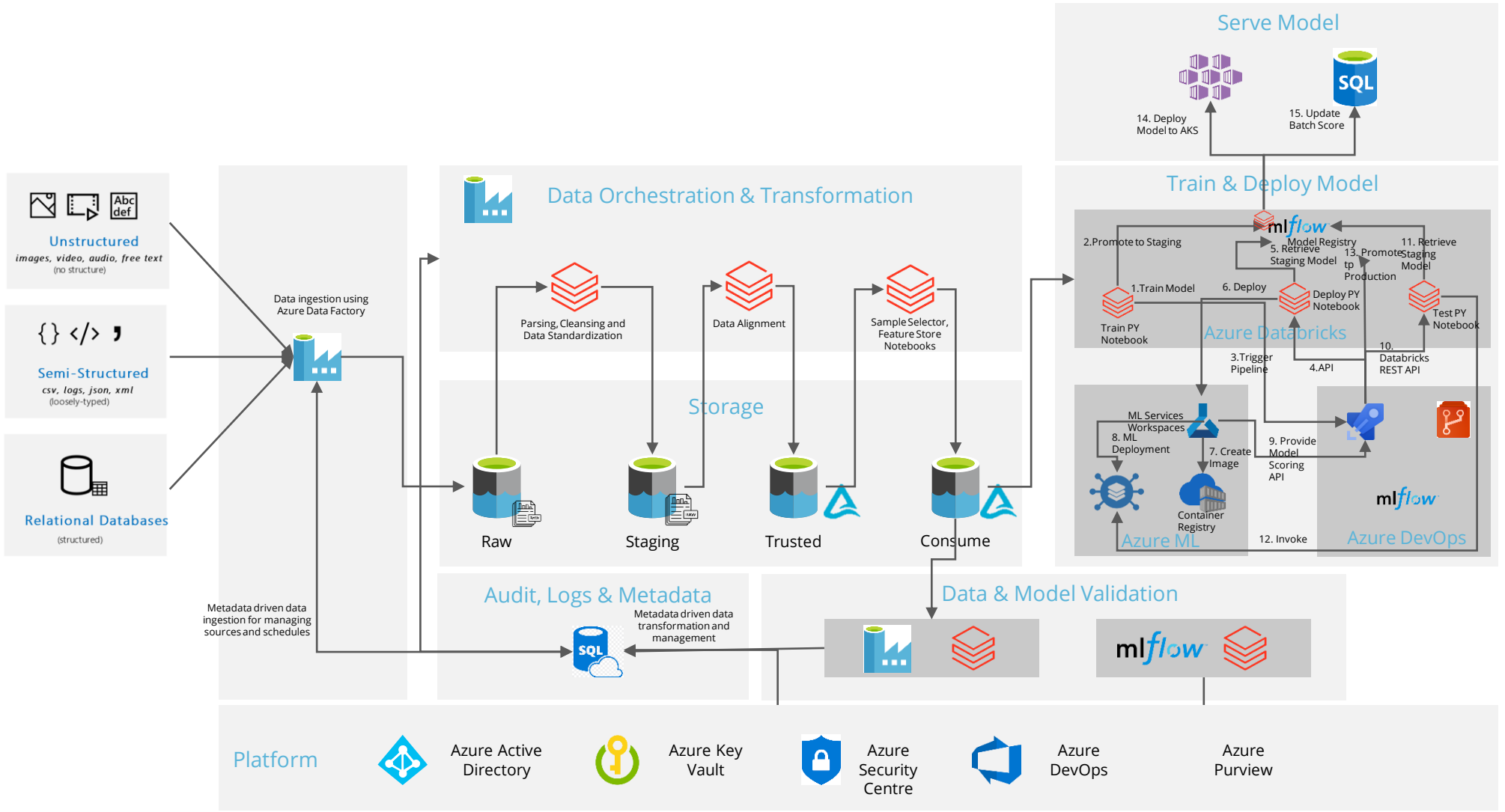

## Overview

- Enabling an end to end ML build, deploy and serve platform on Azure and Databricks
- Sample selector and feature stores will be supporting data products
- mlFlow on Databricks will be used to prep and train model, which are served over AKS for API based scoring and SQL Server for batch scoring

## Key Feature of Enterprise Feature Stores

- Organized based on specific business area and domain supported by current and future analytics needs
- Governed, catalogued and managed by data stewards
- Ability to add/edit features or behaviours withing existing stores
- Ability to enable time travel or retro capabilities based on changes in data sources
- Data monitoring frameworks monitors and validates data correctness of feature store on each refresh

# SOLUTION ARCHITECTURE



# DATA ENGINEERING PRODUCTS

"**Sample Selector**" provide data scientists with pre-built datasets for the most used sources for training testing and validation purposes

- **Aggregated tables on delta lake ensuring SCD behavior** using control columns
- Data scientists can join tables together directly so queries can be filtered at source, meaning faster run-times and only required data is extracted
- **Pre-configured and parameterized Databricks notebooks** for performs sample selector table creation on a batch intervals

**Feature stores** for Client on **Delta Lake** which calculates the state of different variables and maintains them over time.

- Organized based on specific business area and domain supported by current and future analytics needs
- Governed, catalogued and managed by data stewards
- Ability to add/edit features or behaviours withing existing stores
- Ability to enable time travel or retro capabilities based on changes in data sources
- Data monitoring frameworks monitors and validates data correctness of feature store on each refresh
- Delta lake provides ACID capabilities
- Time travel capabilities enabled using Delta lake provided Update capabilities
- Advance cluster management using a combination of job clusters and spot instances for optimized performance and cost management
- Data validation frameworks performs automated validation checks on feature store completion and report to stewards

# DATA SCIENCE PRODUCTS

"**Model Training**" enabled using Databricks and MLFlow

- Azure Databricks used to build and train machine learning models
- Makes use of pre-installed optimized libraries within Databricks to build and train models
- Azure Repos and DevOps pipelines used to move code across different environment for build, train purposes
- MLflow tracking used to capture the machine learning experiments, model runs, and results
- MLFlow also used as the model registry to stores, manage models and load them in production

**Model Deployment and Egress using Azure Kubernetes Service**

- Production ready model deployed within Azure Kubernetes for API based scoring
- Certain model which performs batch scoring are stored within Azure SQL Databases and consumed by other applications and users
- Azure DevOps pipelines used to manage the release of models into production including deployment to AKS
- Azure build-release pipelines with pre-configured environment variables will move the models and deploy them into production

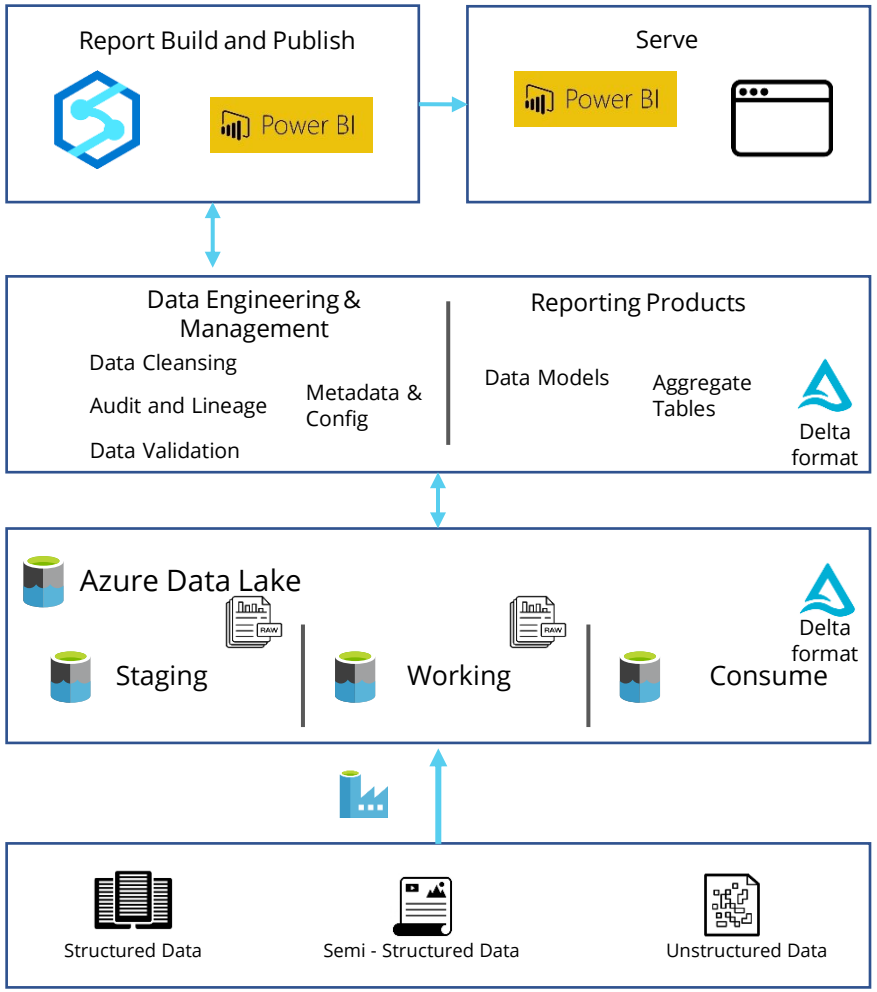
**Model Validation** enabled using Databricks and Azure Data Factory

- Model validation notebooks are run post model score generation or run at specific interval to check score correctness
- Validation notebooks are Databricks PY notebooks executed by ADF post completion of model run pipelines
- Certain validation notebooks looks for variances in the scores to set benchmarks while others are used to monitor certain score behaviours over a period of time.
- Based on the validation results, model training and improvements are performed

# POV 2 – Delta Lakehouse for Enterprise Reporting

brillio

# Azure Databricks Powered Delta Lakehouse for Enterprise Reporting



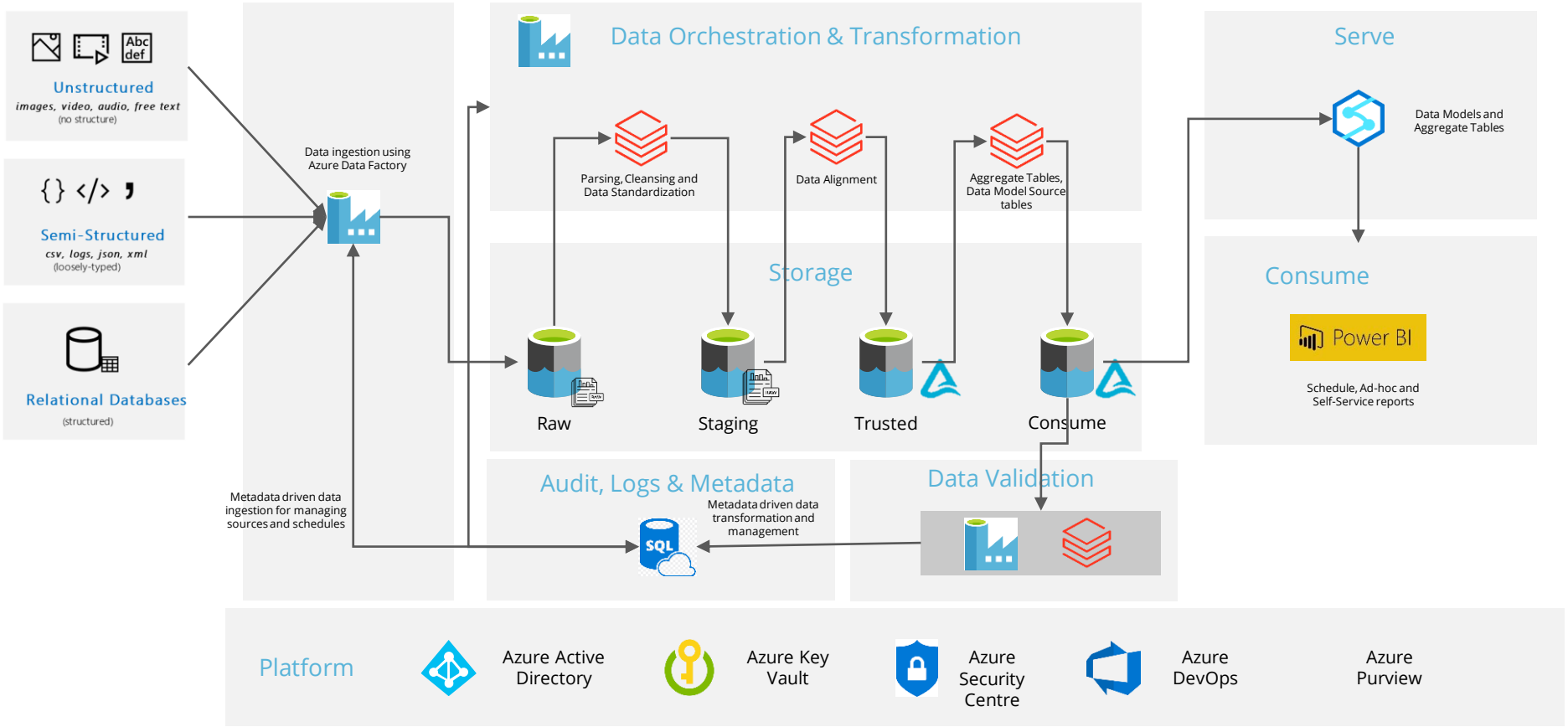

## Overview

- Data Models served for enterprise business reporting for separate business units
- Separate data marts holds a set of dimension models and aggregate tables used for building the PowerBI dashboards and reports
- Data Lake used as centralized data repository for all data reporting needs

## Enabling Databricks for Reporting

- Underlying data lake tables created as source tables and aggregate tables for building data model tables on Synpase
- Aggregate tables which are large tables with complex calculation are prepared using Databricks and then moved to Synapse for faster reporting
- PowerBI used for reporting, this include scheduled reports, ad-hoc and some self-service reports

# SOLUTION ARCHITECTURE



# DATA SCIENCE PRODUCTS

**Azure Data Factory and Databricks enabling single source of truth of data**

- ADF used for data loads from multiple sources onto data lake raw zone
- Dynamic data cleansing, parsing notebooks on Databricks performs data cleansing, parsing and data standardization of loaded data onto Staging
- Metadata within Azure SQL and Config files enables dynamic rule engine, data load configurations, so that new table loads, or existing table load rules can be modified
- Data Alignment notebooks align all tables into a single table Delta format rather than the data-based folder structure of data lake. This is also enabled by metadata and config so that future or existing table changes can be easily handled without code changes.

**Aggregate tables and Source tables for reporting**

- Aggregate tables , which creates a set of granular aggregations for reporting needs are built in Consume zone. This ensures that complex data transformation over large data sets are handled by Databricks and not a PowerBI layer.
- Source tables which are used to build the fact and dimension used for reporting is available in Consumer layer. Some of the tables will have filters in place so only data required for reporting needs are moved to consume layers.
- Data analysts, engineers who are looking to build future data models will use Trusted zone to identify and design future data models

**Synapse and PowerBI for reporting**

- The data models which are used for reports within PowerBI are transformed and persisted in Synapse
- PowerBI uses Synapse data sets as the source for all reporting needs

# Case Studies

A BRILLIO PRESENTATION

# Case Study 1: Built A High Performing Advanced Analytics Platform On Cloud, Powered By Azure Databricks For Enabling 4x Faster Go To Market

## Overview

The client team wanted to increase the speed of bringing highly accurate and reliable predictive AI/ML models to market. In a growing industry with increasing credit risk exposures and frauds, it was critical for Client to foresee and minimize these potential risks through effective data models in scale.

## Challenges

- **Lack of single source of truth for business decision making**
- **Extremely high time (8 + months) to build analytical models for due to manual work, lack of data prep etc.**
- **Lack of tooling for Self-Service MI/BI and Decision Science Predictive Analytics capability.**
- **Inconsistent data quality and governance impacting decision making**

**ASSESSMENT** → **MODERN ANALYTICS PLATFORM** → **SELF SERVICE BI ENABLEMENT** → **DATA SCIENCE**

Brillio provided a solution which addresses the current requirements as well as future proofs Client's aspirations of being a Self-Serve Analytics driven organization capable of leveraging industrial grade analytics models for business decision making

### ASSESSMENT

- Pre-migration planning across multiple geographies
- Landscape analysis and assessment of business requirements
- Target Architecture design
- Data mapping and cleaning

### MODERN ANALYTICS PLATFORM

- Single source of truth for data
- Standardization of processes and aggregation for data
- PII data treatment complying with GDPR and Client's code of conduct
- Enterprise-wide business glossary
- Strong data governance to ensure high platform adoption

### SELF SERVICE BI ENABLEMENT

- Standardization of KPIs through well cataloged data model
- Self-service reporting for internal and external users using Azure Synapse
- Centralized development & distribution of reports

### DATA SCIENCE

- Multiple data science products
- E2E Model Deployment & Consumption through MLOps process
- Reusability and Automation
- Model Testing & Monitoring process

## IMPACT DELIVERED

- **ROI of 3.7x times the investment (Approx. profit of $17M)**
- **Reduced model development time by 60% thereby, doubled the number of models delivered**
- **70% reduction in cost of data storage**
- **40% faster new data source on-boarding provisioned**

# Built Advanced Analytics Platform Leveraging Brillio Accelerators

## Key Platform Highlights

### Data Encryption

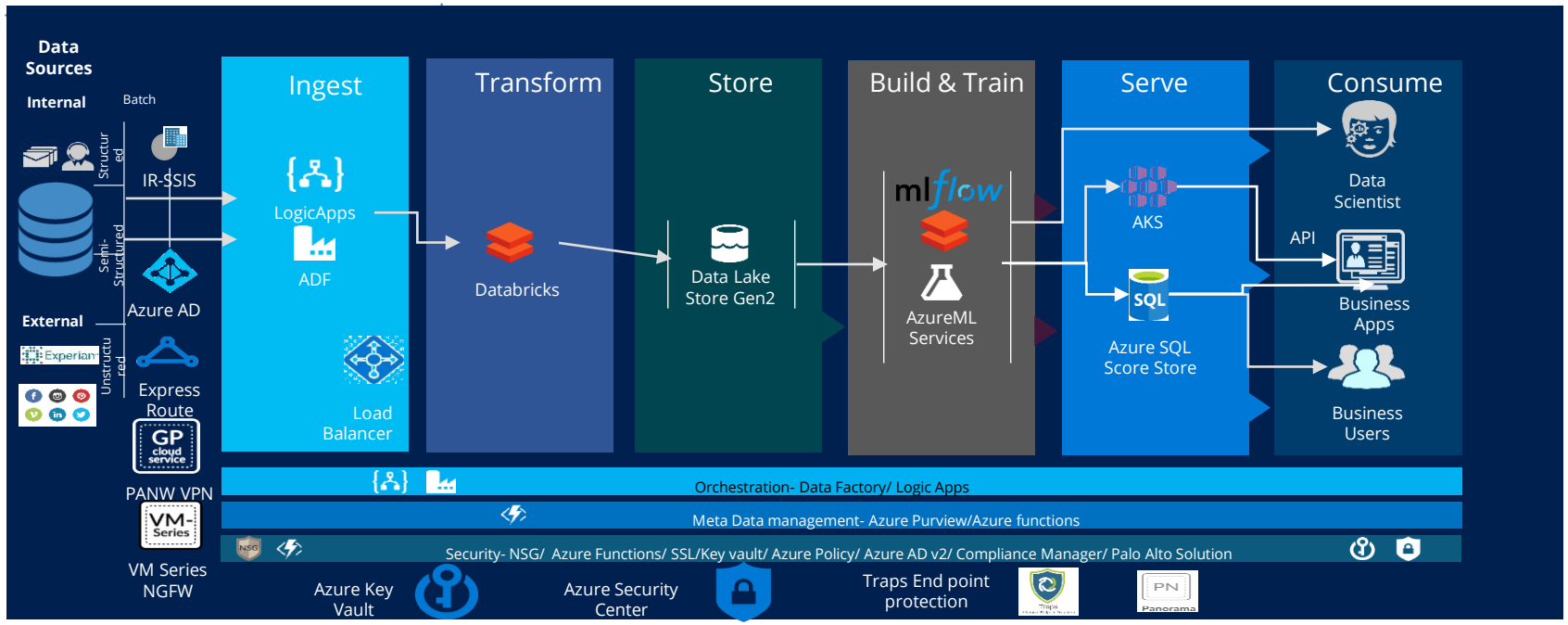
- ✓ Customized data tokenization framework for GDPR and PII Data compliance
- ✓ Detokenization framework for consume applications

### Data Governance

- ✓ End to end data lineage for higher trust on data
- ✓ Data catalogue with Azure Purview catalog and Brillio Metadata Solution
- ✓ Row level security

### End to End ML Enablement

- ✓ Delta Lake powered ML products – Sampler & Feature store
- ✓ MLOps enabled using Azure DevOps and Databricks




## Implementation Highlights

- Optimized the architecture to process high volumes of **near real time data totaling 400+ TB and increasing at 20TB per week**
- **One of the largest Delta lake implementation leveraging Azure Databricks and in EU** to handle ACID like transformations in data lake for feature stores and sampler with data processing loads reaching 20TB every week using 896 core Databricks clusters
- ML-driven **Data Cleansing framework** to self-learn & resolve DQ issues aided by **APTA - Brillio accelerator library**

# brillio | Enabled Data Scientist Toolkit For Accelerated Outcomes

Brillio built "**Sampler**" for Client to provision pre-built datasets for the most used sources for training testing and validation purposes

- 12 **aggregated tables on delta lake ensuring SCD behavior** using control columns
- Can join tables together directly so queries can be filtered at source, meaning faster run-times and only required data is extracted
- Delta Lake based tables enables faster query time on cost effective interactive cluster enabling data scientists to build sample data sets faster

Brillio built **Feature stores** for Client on **Delta Lake** which calculates the state of different variables and maintains them over time.

- Databricks notebook containing script to build and append all variables
- Powered by Delta Lake weekly runs perform retro operation over **1000TB+ source data** performing Delta updates for feature stores
- **Readily available data**; Multiple views of bureau data; full view of historic PINs and data as well as snapshot of current PINs and data backdated. Keeping historical retros means restated records at the bureau are captured and included
- **Stable**, no lost time re-running failed queries due to run errors

**MLOPS enablement** for reduced time, effort, costs in productionizing ml projects

- **Self service ML platform** for Fully staged deployment models: from dev to test to preproduction to production
- Access to **model libraries** – best of open source and algorithms trained for domain specificity
- Data discovery – **ML led effective data search** , use of data marketplaces

**90% reduction in query execution time on a like-for-like basis**

**18000+ variables – feature store for future data scientist use**

**Reduction in data sampling time by around 80%**

**Reduction in model development time from 7 months to 2 weeks**

# Client Testimonial

*" We needed to reduce time to market by increasing automation and model performance for superior customer experience and insights along with enabling self-serve BI and predictive analytics. Brillio was brought on board to build the data engineering strategy and institutionalizing a modern data lake platform. This transformational initiative using data and analytics cements and advances our position as a pan-European leader in the credit management industry."*

***- Gary Edwards, Global CIO, Lowell***

# Case Study 2: Predictive Analytics & Accelerated Insight Enablement for a Global CPG Client

## Overview

The client, one of the world's largest multinational beauty company with **77+ brands**, wanted to **enable its business teams** – Sales & marketing, supply chain and finance with self-serve insights for **business decision making** and **leverage data science** for new use cases in SCM and cashflow management for optimized resource utilization to drive up the **operational efficiency**.

## Challenges

- Lack of single source of truth across **4 business systems (SAP BW/4HANA etc.) and 35+ 3rd party data sources**
- Manual data management processes
- Lack of standardized reporting and KPIs
- Lack of data science capabilities impacting business operations

**ASSESSMENT** → **MODERN ANALYTICS PLATFORM** → **SELF SERVE ANALYTICS** → **DATA SCIENCE USECASES**

Brillio's solution addressed the current requirements along with future proofing Coty's aspirations of being a Self-Serve Analytics driven organization capable of leveraging industrial grade analytics models for business decision making.

- Assess 4 BU's spread across 6 countries with 2K + users in **4 weeks**
- **Current state analysis** (architecture, data practices, integrations, governance, reporting etc.)
- Architectural Recommendations with ML use cases prioritization

- Created **Azure Data lake** as Single source of truth
- **Azure Databricks** for enabling Data Ingestion & Transformation
- **ADF and Custom connectors** for integration with varied SAP & non-SAP data sources.

- **Standardization and aggregation** of KPIs
- Enabled **Azure Synapse** as Datawarehouse with Self-service reporting using Power BI
- Secure data access to consumption layer leveraging **dynamic tables**

- Enabled multiple data science use cases with model re-training capabilities to address business needs for finance and sales functions etc.
- Enabled **ML@Scale** framework
- **MLOPS** for data science enablement

## IMPACT DELIVERED

- **Reduced time to insight generation by 81%**
- **17% increase in operational efficiency**
- **60% faster new data source on-boarding provisioned**
- **2X Better Operating Cashflow Ratio**

# Solution Approach - Advanced Analytics Platform Leveraging Brillio Accelerators

## Solution Highlights

### CLOUD PLATFORM SET-UP

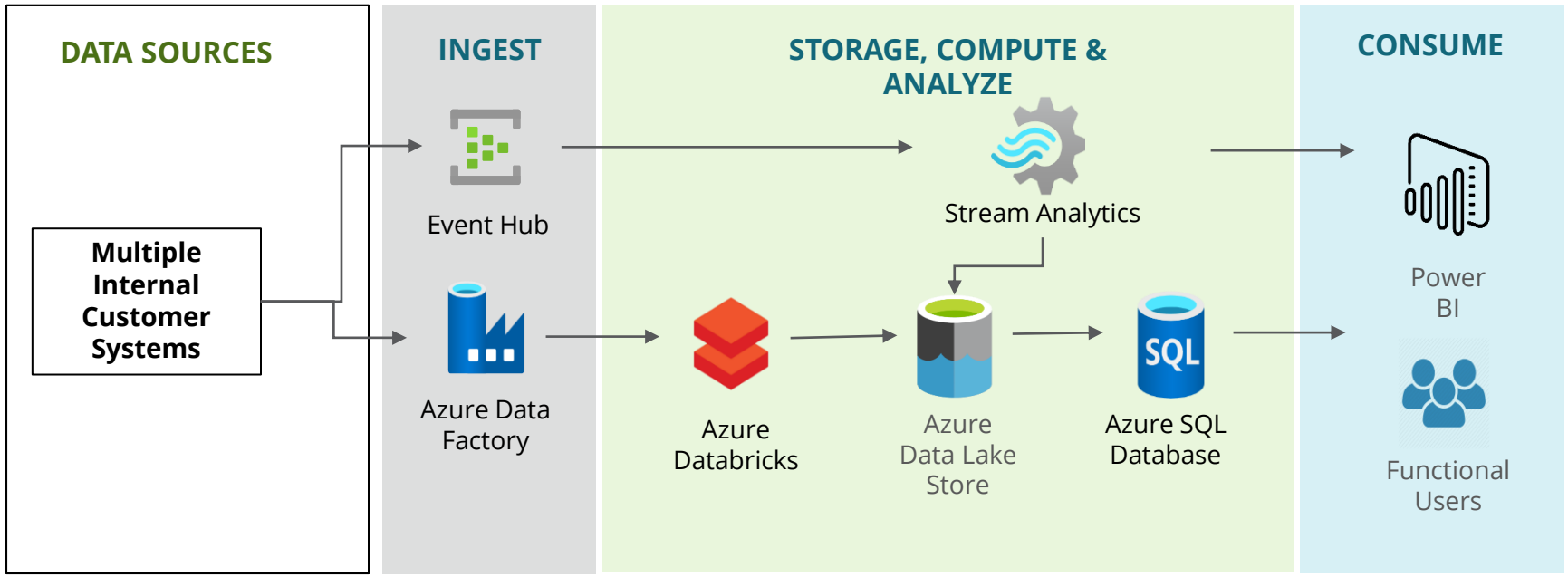
- ✓ Ingesting data (~**25+TB**) from **SAP BW4HANA** into Azure using MDX, SAP webi queries
- ✓ Azure DLS, Databricks, SQL DB, VM, Logic Apps, Devops, ADF following security and folder structures

### DATA TRANSFORMATION

- ✓ **Azure Databricks** for enabling **Data Transformation and Machine Learning**
- ✓ Data catalogue with Azure data catalog and Brillio Metadata Solution
- ✓ Automated ETL monitoring and validation

### ML DRIVEN INSIGHTS

- ✓ Self serve insights through Power BI with dynamic row level security using Azure Synapse
- ✓ Used Pyspark, **Azure Databricks**, Azure ML, ML Lib package, and pipeline feature for repeatable codes




## Implementation Highlights

- **Scalable architecture with HA & DR** with integration of data from both SAP and Non-SAP sources
- **Automation of data staging** into Azure via Open Hub and MDX connectors
- **Multi-tiered security** across the data and consumption layer
- **Data Catalog Implementation**
- **Automated Model retraining**

## Enabled by Brillio Accelerators

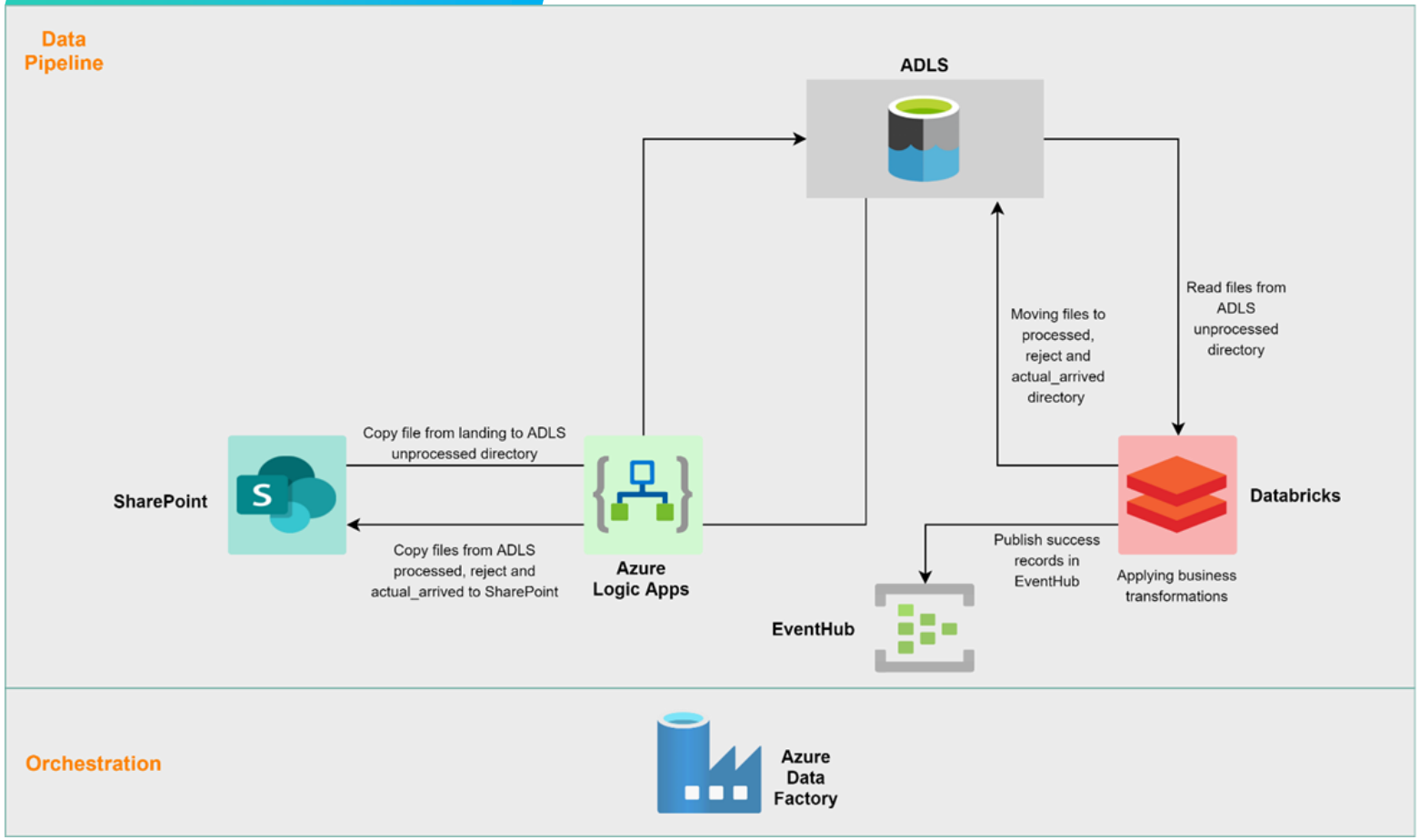
**Brillio DASH™** **Brillio CRED™** **Brillio CLIP™** **Brillio Script Centre** **Brillio ML@Scale** **Brillio APTA™**

# Case Study 3: Transfer Loyalty Data From SharePoint To Azure Event Hub for a Global Orthodontics Devices Company

**Objective**

The client, one of the world's largest medical devices company, partnered with Brillio to help them define and design the stages for handling **Loyalty data** received from **marketing at SharePoint** and push the same to **Azure Event Hub & Blob** in expected input format after performing required validations & **transformations** leveraging **Azure Databricks**

## Solution Architecture




## Solution Approach

- Design/Architect/Develop flow to handle the **Loyalty** third party **data**.
- Validate records against business provided **validation** rules and apply **transformations** using **Azure Databricks** wherever necessary.
- Created csv files for **successful validations** of records against business and push it to **Blob for storage**
- A separate csv file was also created to **record unsuccessful validations** which failed to comply with business rules were pushed to the same Blob
- Send emails to **share error notifications** wherever necessary

# Case Study 4: Transformation Of Data Supply Chain for One of the Largest Education Enrolment Service Providers

## Objective

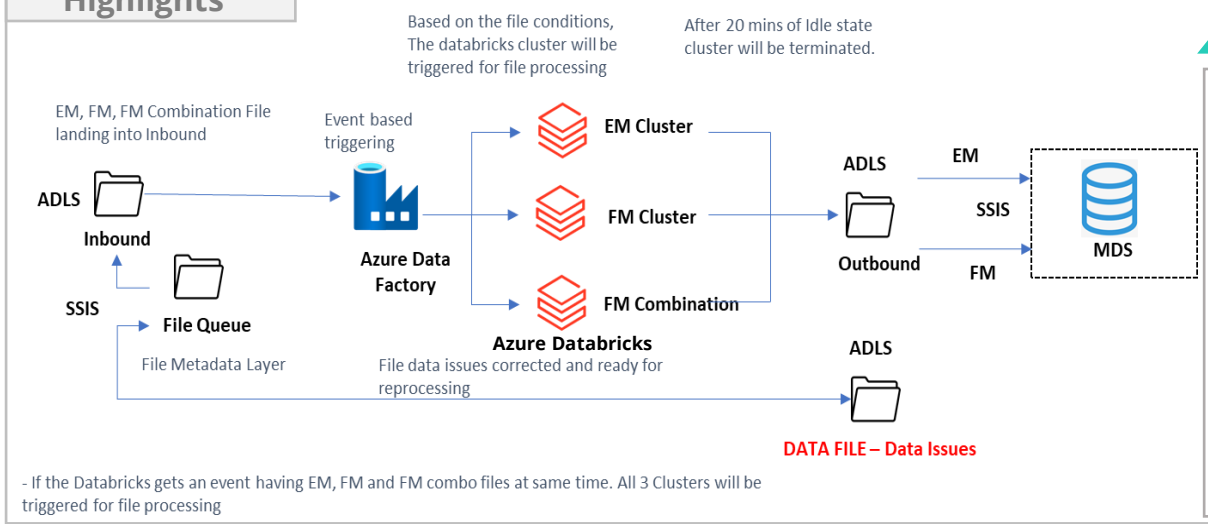
Brillio partnered with client to accelerate transformation of Data Supply Chain to drive efficiency and agility in end-to-end data pipeline through automation, innovation & technology refresh leveraging Azure **Databricks** leading to significantly enhanced customer experience across multiple Business Units.

## Challenges

- Longer data pipeline processing times and manual intervention
- High number of failures in file/data processing
- High operational costs due to HDInsight running in always on mode
- Upcoming end of support for HDInsight 3.6 that was in use

## Solution Highlights

- Event based triggering to process the files as it arrives – Low cost / High transfer rate
- Enabling Auto scale using databricks to add on demand instance
- Platform as a service – Serverless




### CURRENT STATE ANALYSIS

- Understanding of platform, script, configuration, tools used and interfaces, cost structure
- Identifying backlog for relevant **automation** and **tech debt opportunities**

### PLATFORM ANALYSIS & SELECTION

- Comparison of HDInsight 4.x new version and Databricks
- Identifying the **key differentiating factors** of decision making on the platform
- Execution of proof of concept to showcase **Databricks** capabilities

### AUTOMATION & TRANSFORMATION

- Setup new technology platform on Azure with automated workflow for ingestion using **Databricks**, error handling, audit, and end to end data lineage
- Enabling downstream consumption – dashboards and ML algorithms

## IMPACT DELIVERED

- 40% reduction in file processing time
- 60% Reduction in pipeline failure rate
- Lower operational costs due to on-demand execution flexibility of Databricks


LET'S BUILD SOMETHING
AMAZING TOGETHER...
brillio