DATA+AI SUMMIT 2022

Simplifying Migrations to Lakehouse

The Databricks Way

Ram Venkat
Senior Field Engineering Manager, Databricks

Ron Guerrero
Lead Partner Solutions Architect, Databricks

ORGANIZED BY databricks

Product Safe Harbor Statement

This information is provided to outline Databricks' general product direction and is for informational purposes only. Customers who purchase Databricks services should make their purchase decisions relying solely upon services, features, and functions that are currently available. Unreleased features or functionality described in forward-looking statements are subject to change at Databricks discretion and may not be delivered as planned or at all.

Agenda

1. Challenges with legacy system
2. How we are reimagining migrations - The DBX way
3. The Migration journey at Databricks
4. Hadoop Migrations

Challenges of legacy platforms

Comparing Challenges

Just moving to cloud doesn't solve the problem!

Legacy Data & AI platform limitations:

- Proprietary
- Expensive
- Complex
- Inability to scale
- **Tightly coupled architecture**
- **Limited innovation**
- Siloed
- Product gaps to support future use cases

Cloud native Data & AI limitations

- Either proprietary or cloud specific
- Relatively expensive [# of services]
- Complex
- Scaling multiple services is challenging
- **Two-tiered architecture storing raw data in the Data Lake and then ingesting it into a Data Warehouse or a ML service**
- **Fragmented experience**
- Multiple services stitched together - no unified experience

Challenges by workloads & personas

What customers are tackling in a holistic migration

Data engineering workloads

- Platform is tightly coupled and difficult to scale
- **Multiple services and systems have to scale together to operationalize a system or a use case**
- Inability to innovate faster with unknown costs and performance of the data pipeline

Data warehouse workloads

- Huge data analytics backlog resulting from moving data across multiple platforms
- Lengthy data lifecycle
- Considerable amount of time spent in administering the platform
- **Error prone process when multiple steps involved in data movement**

Data science workloads

- **Performance comes at a huge cost**
- Operational overhead to support production system is overwhelming
- Unable to justify RoI for new use cases
- Business losing confidence in IT on keeping them above competition

Key factors driving migrations

Databricks comes out on top for Data & AI platform migration

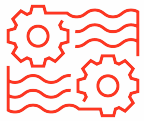
Key Factors	Maintain Status Quo [On prem or cloud]	Begin with cloud native and figure out later	Migrate to Databricks on cloud
Cost	High	Relatively High	Low
Simplicity	Low	Relative by service	High
Features	Low	High	High
Innovation	Low	High	High
Open source	No	No	Yes
Multi-cloud support	No	No	Yes
Out of box support for multiple workloads and Personas	No	Offered through multiple services	Yes

If **Data & AI** is the lifeline of your business, how many migrations are you willing to undertake to get it right and would you settle for a sub-optimal platform?

How we are reimagining migrations - The DBX way

The Databricks way – Do it once and do it right

The only Data and AI target platform migration you will ever need!

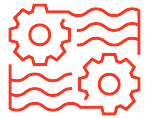


Migrate to Lakehouse architecture- the only true unified data & AI platform architecture
True multi-cloud, multiple persona support, open source

Support your Engine1 and Engine2
Migrate core business and drive new avenues of future growth

Predictable model – Early Value Realization and Guaranteed RoI
Minimize after-migration risk, Lower TCO, Eliminate Tech debt

*Engine 1 and Engine 2 framework concepts were introduced by Bain & Co and illustrated in HBR

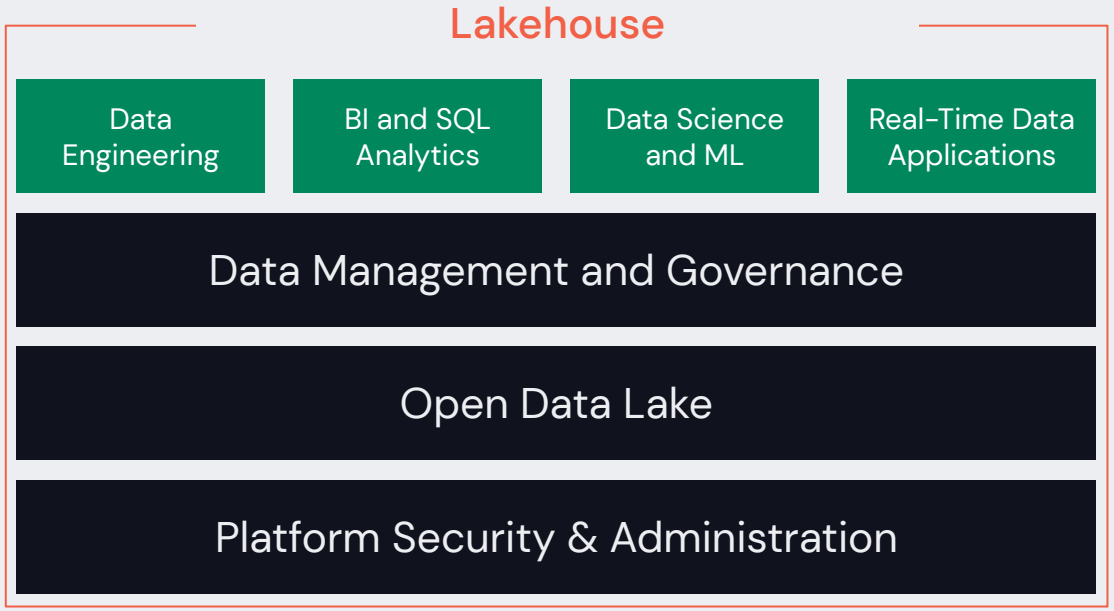


Migrate to Lakehouse

The Lakehouse

Simple. Open. Collaborative.

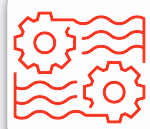
- **Lake-first approach** that builds upon where the freshest, most complete data resides
- **AI/ML** from the ground up
- **Multi-cloud & Inter-cloud** capabilities
- **Support for all migration use cases** on a single platform:
  - Data engineering
  - Data warehousing
  - Real time streaming
  - Data science and ML
- Built on **open source** and open standards



Unstructured, semi-structured, structured, and streaming data

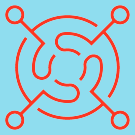
Microsoft Azure aws Google Cloud

Migrations expertise

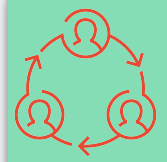


Migrate to Lakehouse

Optimal pathway with right technology mapping
Migration architecture is consistent & predictable across clouds supporting all workload migrations



Code compatibility and interoperability
Retain code or automate your code migration to your choice of programming language, bring your IDE's and Notebooks

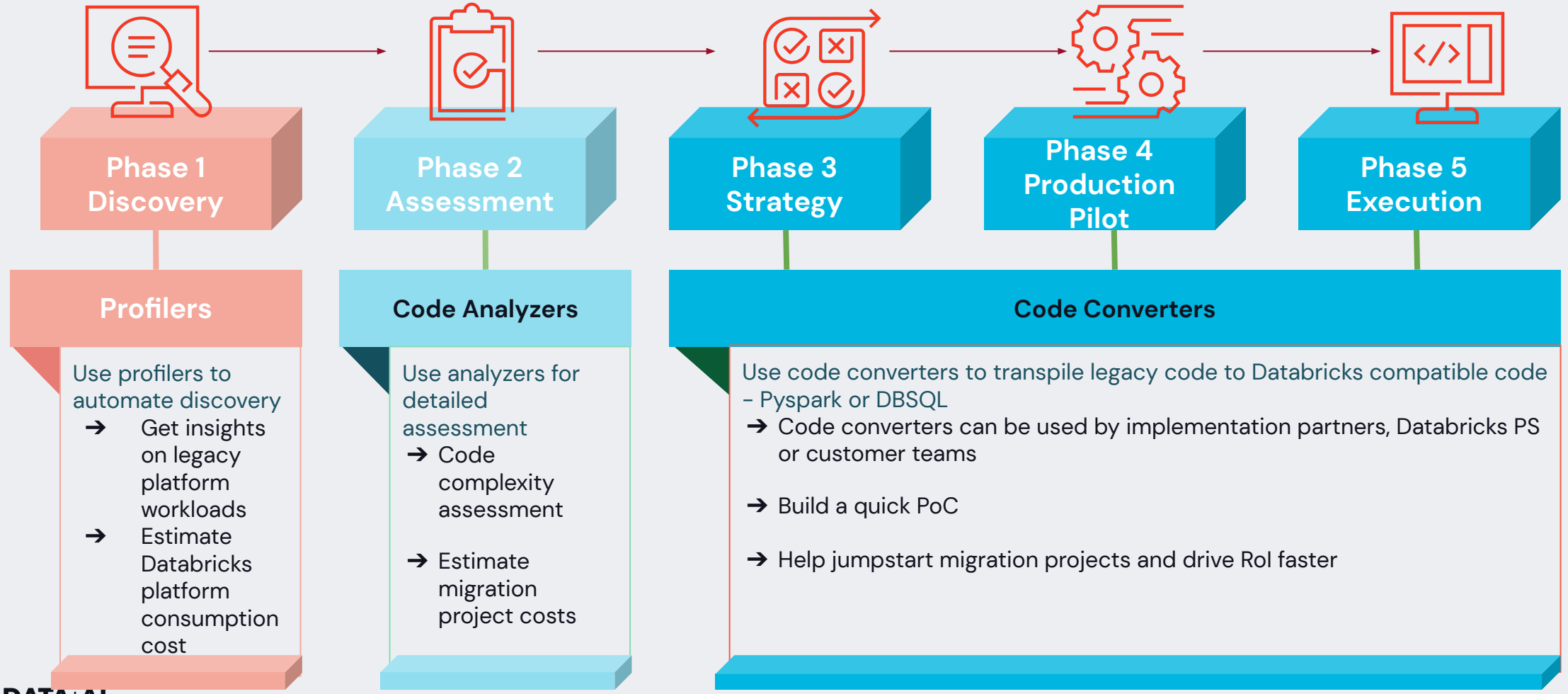


Tight integration with our Partner ecosystem
Augment or compliment with partner integration tools after migration for DE, DW or ML workloads

Databricks Migration Journey

Migration Methodology

Automation to accelerate different migrations phases



Hadoop Migrations

Databricks success in migrations

Our customer stories depicts the success

300+ Successful Hadoop Migrations to Databricks

600+ person years of Hadoop experience in-house

Methodical framework with automation and transformative migration capabilities

CBC

CBC/Radio-Canada uses Databricks Lakehouse to deliver audience-centric programs that delight and retain listeners.

Use Case

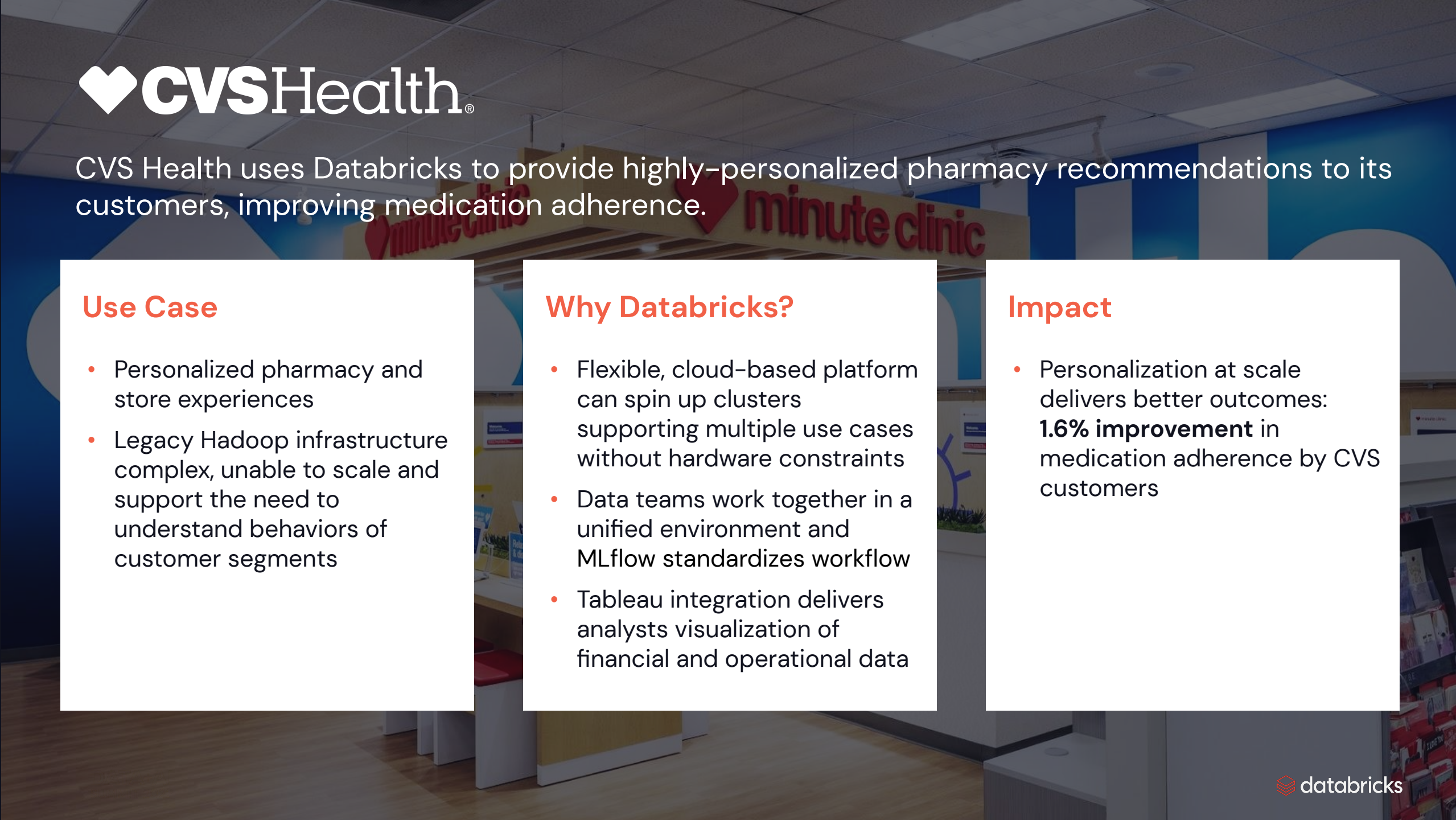
- Customer Retention
- Customer Segmentation
- Rigid Hadoop system led to uncontrollable costs

Why Databricks?

- Lakehouse allows “data warehouse-like” interaction with tables, enabling streamlined workflows
- Delta Lake provides a common data layer to bridge gaps between engineers & analysts
- Databricks SQL enables new insights into their digital audiences

Impact

- **50% reduction** in time to insight
- More visibility into digital audiences
- Ability to develop strategies and services that boost engagement and retention



CVSHealth®

CVS Health uses Databricks to provide highly-personalized pharmacy recommendations to its customers, improving medication adherence.

Use Case

- Personalized pharmacy and store experiences
- Legacy Hadoop infrastructure complex, unable to scale and support the need to understand behaviors of customer segments

Why Databricks?

- Flexible, cloud-based platform can spin up clusters supporting multiple use cases without hardware constraints
- Data teams work together in a unified environment and MLflow standardizes workflow
- Tableau integration delivers analysts visualization of financial and operational data

Impact

- Personalization at scale delivers better outcomes: **1.6% improvement** in medication adherence by CVS customers

databricks

Journey from Hadoop to Databricks

Automated Profiling and Analysis

Open Source Compatibility

Automated Code Migration

Supports JAR files and Notebook code
95%+ SQL compatibility with Hive/Impala*
SQL Interface for non-spark users
3rd party tooling for Data and Code migration

* observed with previous migrations

Key Tenets of Migration

Workload	Hadoop component	Databricks component
Data engineering and machine learning	Spark on YARN	Spark on Databricks
ETL via SQL	Hive/Impala	Spark (SQL notebook) on Databricks
BI/Analytics	Hive/Impala	Databricks SQL
Stream processing	Spark DStream/Storm	Spark Structured Streaming
Batch processing data	MapReduce	Spark on Databricks
Machine learning	Zepplin/Cloudera Data Science Workbench	Databricks Notebook + ML Runtime + MLflow + Horovod

Technology Mapping

Code Compatibility and Interoperability

Open Source

Partners

DATA+AI SUMMIT 2022

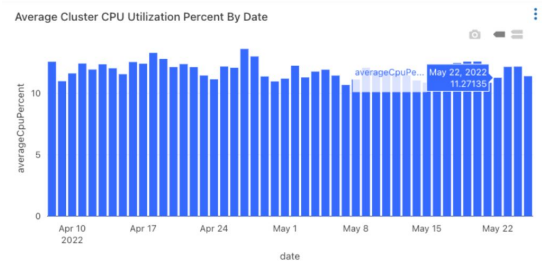
Hadoop Profiler

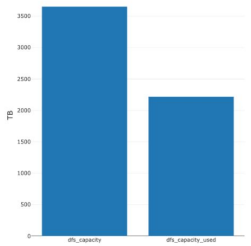
https://github.com/databricks-migrations/hadoop-profiler

What we collected

Cluster:

- Metric data from **2022-05-09 to 2022-06-2**
- CDP Version: **7.1.7**
- Total Cluster Nodes: 143
- Worker Nodes: 101
- Total Worker Node Vcores: 8048
- Key workload types: MR, Spark, Hive(Tez)



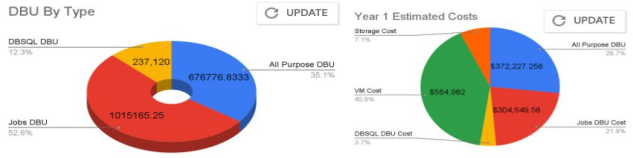


Workload Breakdown

Workload Type	Unique Job Names	Associated Users	Databricks Equivalent	Notes
Spark	5565	8	Databricks Notebooks/JAR	Minor change, Spark Version 2-3
Hive	86149	1	Databricks Notebooks DBSQL	Minimal syntax changes Recommend Delta format
MapReduce	7 Includes Scoop Job	3	Databricks Notebooks / JAR	Requires refactor
Oozie	Shell: 2546 Hive: 1 Hive2: 14681 Spark: 5	47	ADF Airflow	Database Multi-task Jobs recommended for Databricks workloads only (notebooks, jars, etc)

Full Migration - Databricks Projected Usage - Year 1

Migration Timeline - 6 months
60/40 - Automated vs Interactive workloads
List Price for VMs, DBUs, Storage
2x performance gain

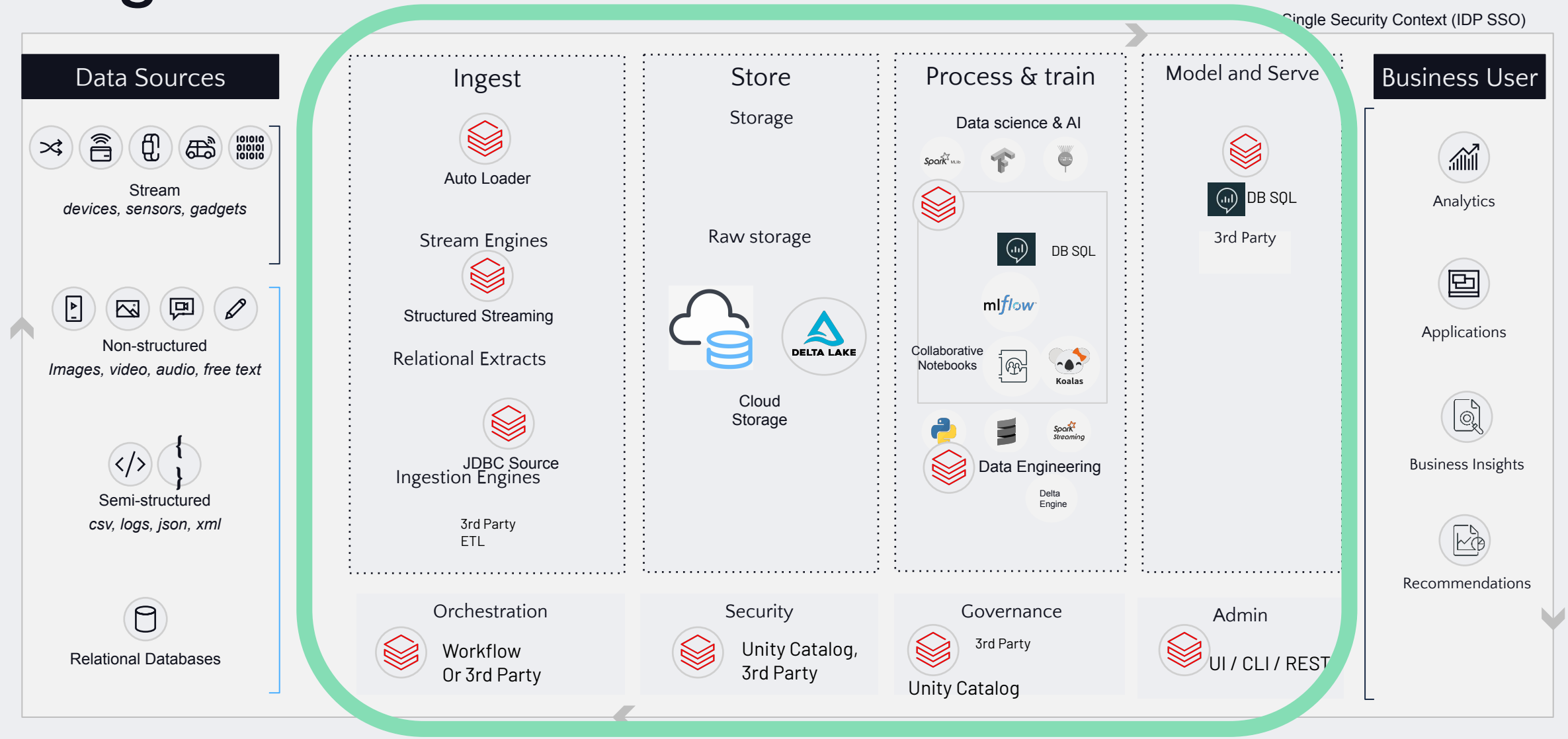


~$12M savings, for a 300 node cluster

Target State Architecture

databricks

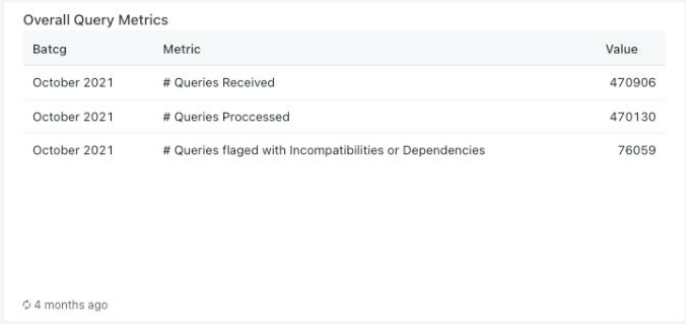
Target State Architecture

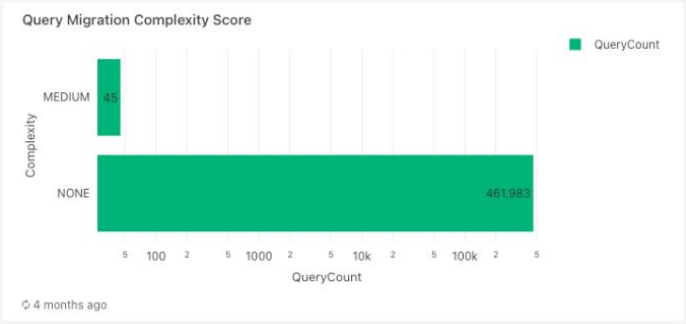


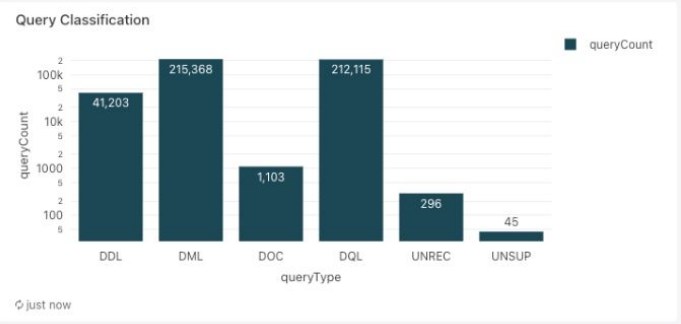
databricks

Unified Platform

Code Assessment - SQL Analyzer

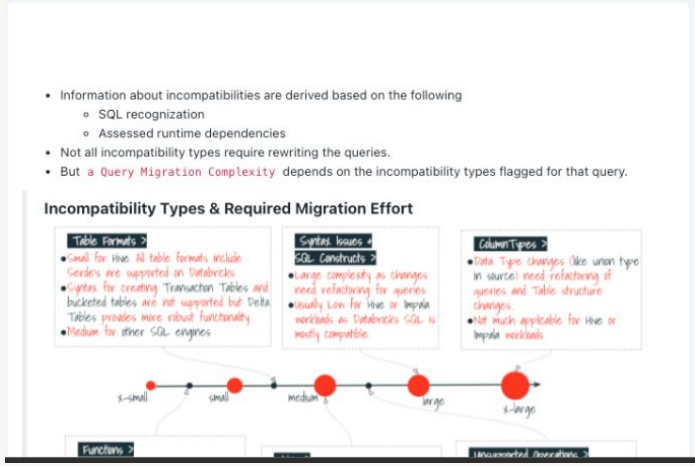




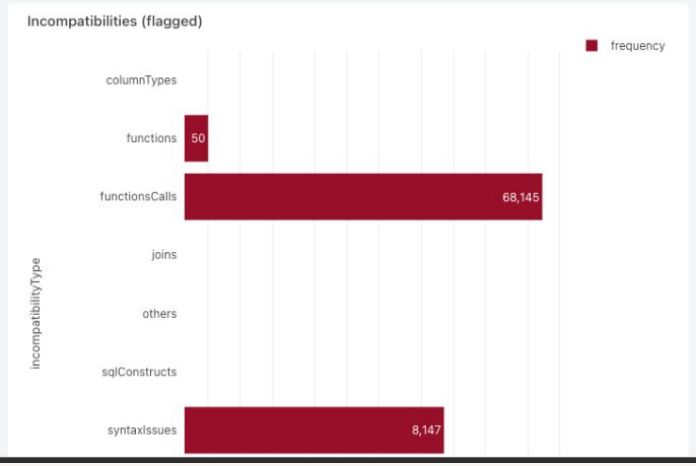


Incompatibilities

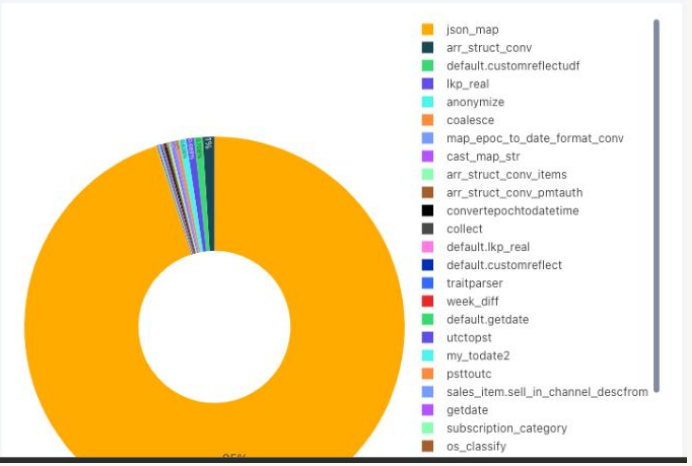
Migration Complexity Model Overview



Flagged Incompatibilities/Dependencies



Missing Functions (Not Built-in)



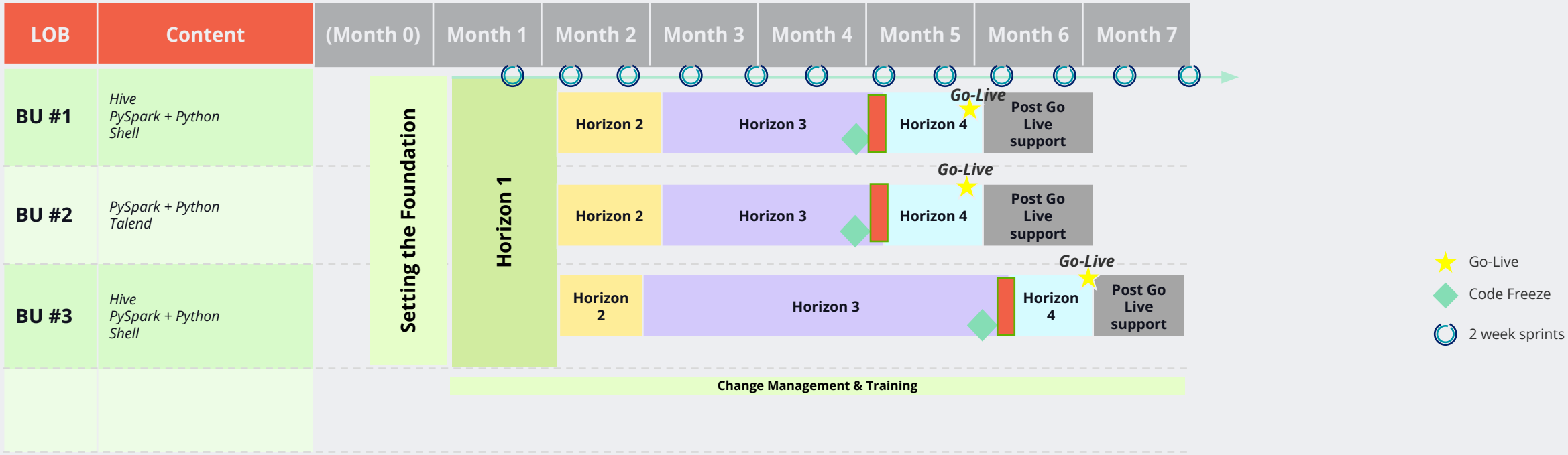
Code Migration

- Databrick Professional Services
- Partner SI Tooling

70% Automated Conversion Rate

Sample migration plan

The migration typically takes 4 to 7 months. Horizon 1 is common across all the LOBs, during which Hadoop workloads associated to LOBs would be prioritized to stagger the migration. The timelines include 1 month of parallel run and reconciliation prior to Go Live. 1 month of post Go Live support is provided to resolve migration/conversion related issues that may be detected after cutover.



Key Activities across Horizons

Horizon 1

- Documentation and Setup
- Review and Validation
- Sequence and Prioritize
- Architecture & Planning

Horizon 2

- Create migration design standards and templates for all workloads on Hadoop.
- Construct data ingestion, processing and consumption patterns for workloads based on migration roadmap

Horizon 3

- Build data ingestion, processing and consumption patterns for workloads based on migration roadmap
- Historical data migration
- End to end validation

Horizon 4

- Develop a release playbook
- Parallel run and go-live support
- Cut-over activities

Post Go-Live Support

- Provide support for all components released in production including defect triage and fixing.

Hadoop T-Shirt sizing guidance

Sizing guidance				
Workloads	**Small**	**Medium**	**Large**	**Custom**
Jobs	**<3000 Jobs**	**3000-15000 Jobs**	**15000-50000 Jobs**	**>50000 Jobs**
Data volume	**<100 TB**	**100-500 TB**	**500TB - 1PB**	**>1PB**
Node count [Baseline of 24 Vcores per node]	**<50 Nodes**	**50-150 Nodes**	**150-500 Nodes**	**>500 Nodes**
Avg % workload distribution	**80% Spark and Hive/Impala 15-18% Mapreduce 2-5% Sqoop**	**80% Spark and Hive/Impala 15-18% Mapreduce 2-5% Sqoop**	**80-90% Spark and Hive/Impala 8-15% Mapreduce 2-5% Sqoop**	**80-90% Spark and Hive/Impala 8-15% Mapreduce 2-5% Sqoop**
Timing Estimates	**12-18 weeks**	**16-24 weeks**	**24-48 weeks**	**24+ weeks**

Delivery Framework

Partners

Professional Services for
Migration Assurance

Migration Guide
Blog Posts
Notebooks

Call to Action

Attend

Explore

Engage

DATA+AI
SUMMIT 2022

Thank you

Ram Venkat - ram.venkat@databricks.com
Ron Guerrero - ron.guerrero@databricks.com