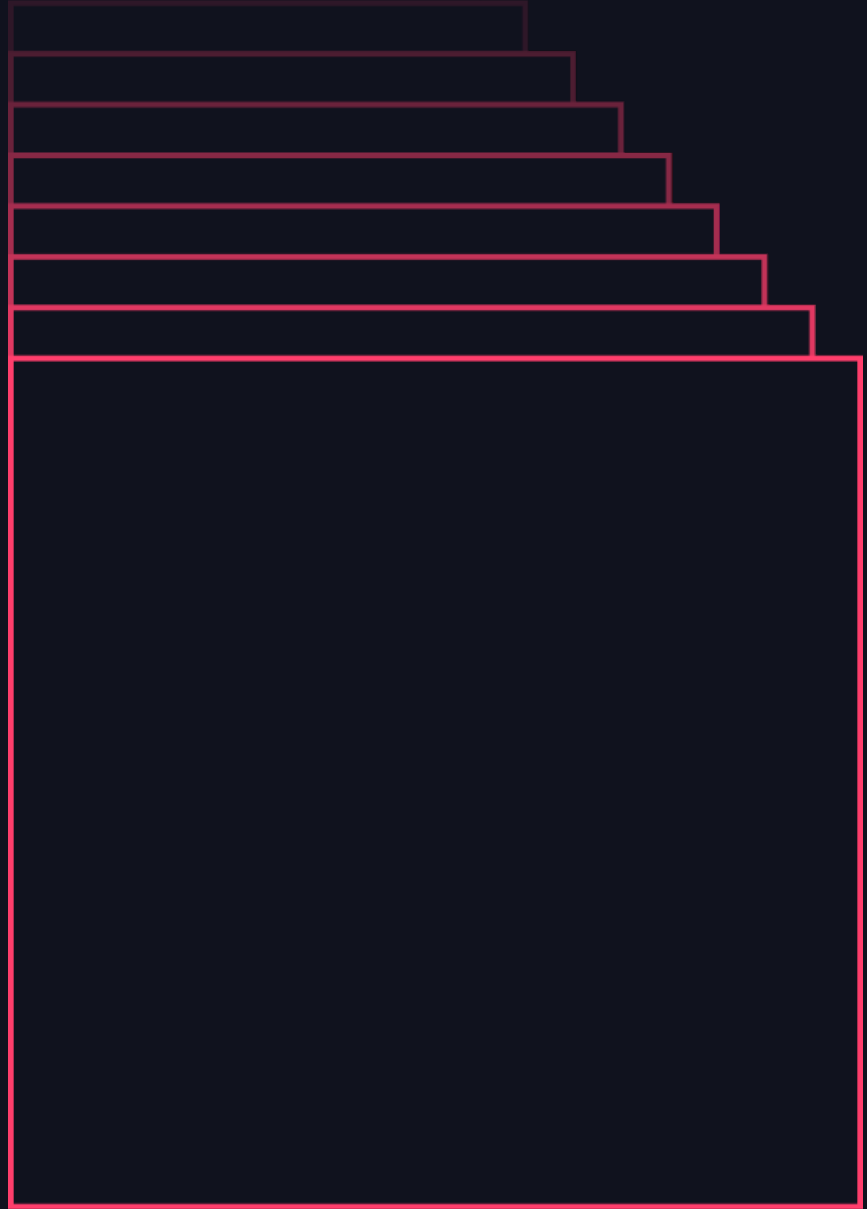


TREK BIKES ACCELERATES RETAIL ANALYTICS

Advait Raje & Garrett Baltzer
June 2024





TREK

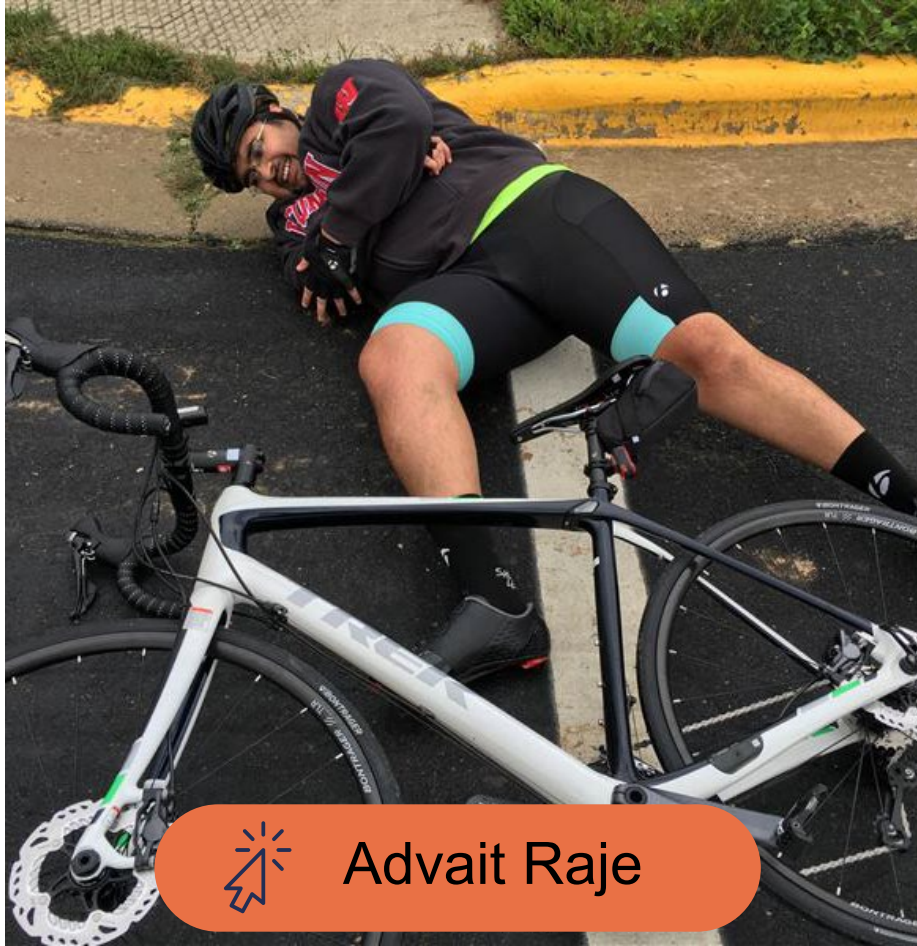
BIKES · SERVICE · GEAR


**Accelerating Retail
Analytics using Azure
Databricks and QLIK**

TREK


BIKES · SERVICE · GEAR

The speakers - Keep Pedaling!!



 Advait Raje

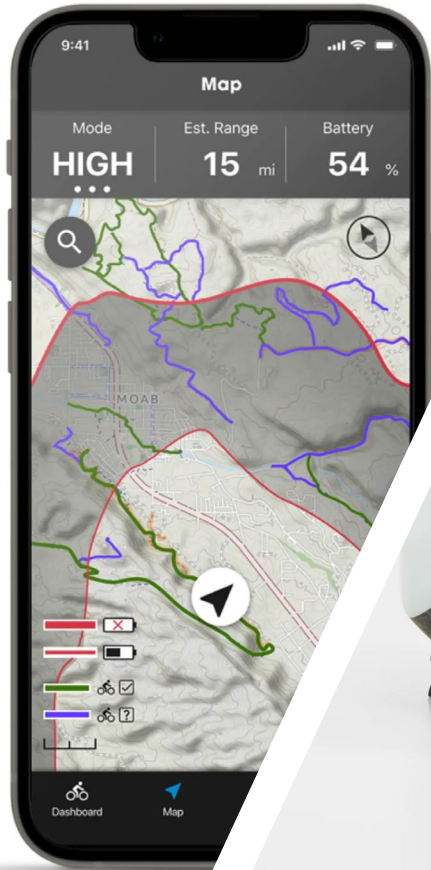


 Garrett Baltzer

Trek Bicycle Overview

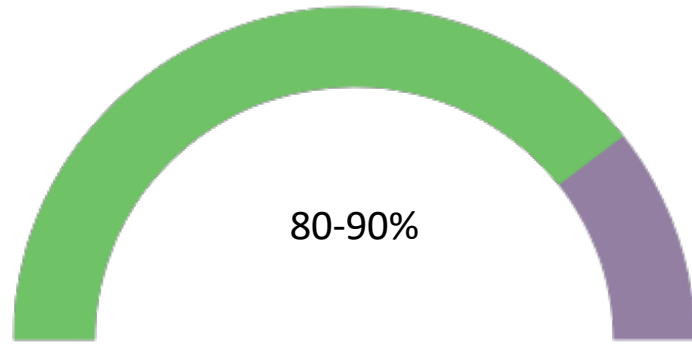
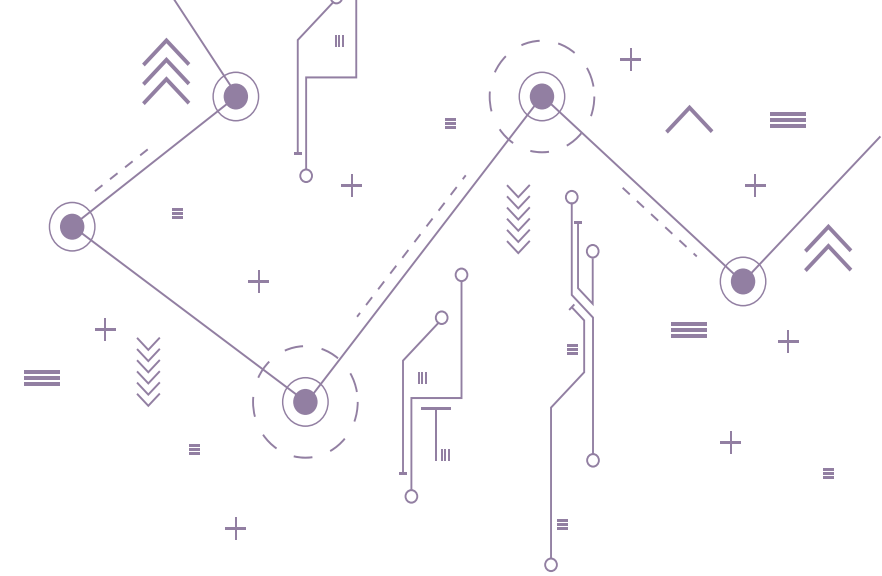
- Born in a Barn - 1976
- Global Leader in Bicycle Manufacturing, Distribution and Retail
- Significant number of company owned stores globally
- 2000+ independent retailer store network globally
- Trek's Mission:
 - Only build products we love
 - Offer incredible hospitality
 - Change the world for better by getting more people on Bikes





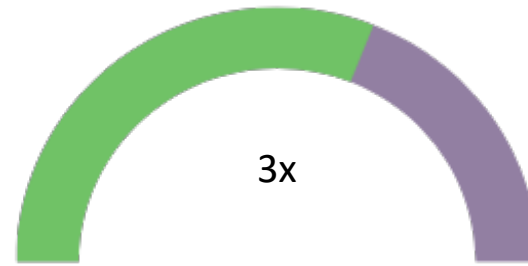
Awesome Trek
Products to Get
More People on
Bikes

INDUSTRY: Retail and Consumer Goods
SOLUTION: Real-Time Point-of-Sale Analytics
PLATFORM USE CASE: Qlik Replicate, Delta Lake, Databricks SQL, Delta Live Tables, Data Streaming
PARTNER: Qlik & Databricks
CLOUD: Azure



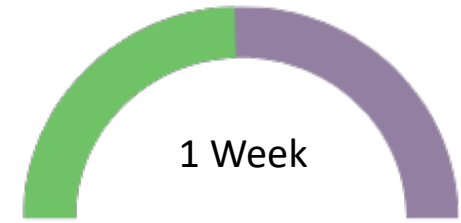
Acceleration

In runtime of retail analytics solutions globally



Increase

3X daily data refreshes using Azure Databricks



Reduction

Near real-time ERP Data Replication enabled by Qlik Replicate

Databricks Data Intelligence Platform, along with data replication to Databricks using Qlik, aligns perfectly with our broader cloud-first strategy, which demonstrates confidence in the adoption of this platform at Trek.”

— Steve Novoselac, Vice President, IT and Digital, Trek Bicycle

TREK

BIKES · SERVICE · GEAR

Agenda and Objectives

- Retail Data Growth
- Cost - Performance Matrix
- Old Architecture – ERP & POS
- Solution Overview
- Why Databricks?
- New Architecture – ERP & POS
- Delta Live Tables at Trek
- Databricks SQL at Trek
- Databricks Unity Catalog at Trek
- Lessons Learnt and Conclusion
- Q & A

Challenges Faced by Trek Bicycle

- Slow replication of ERP & POS data to data lake
- Legacy Technology Challenges
- Stale ERP & POS data in lake
- Rising data warehouse costs & performance issues
- Difficulty scaling analytics on Retail POS data
- Scaling analytics on Retail POS & ERP operational data
- Unified global data view necessity





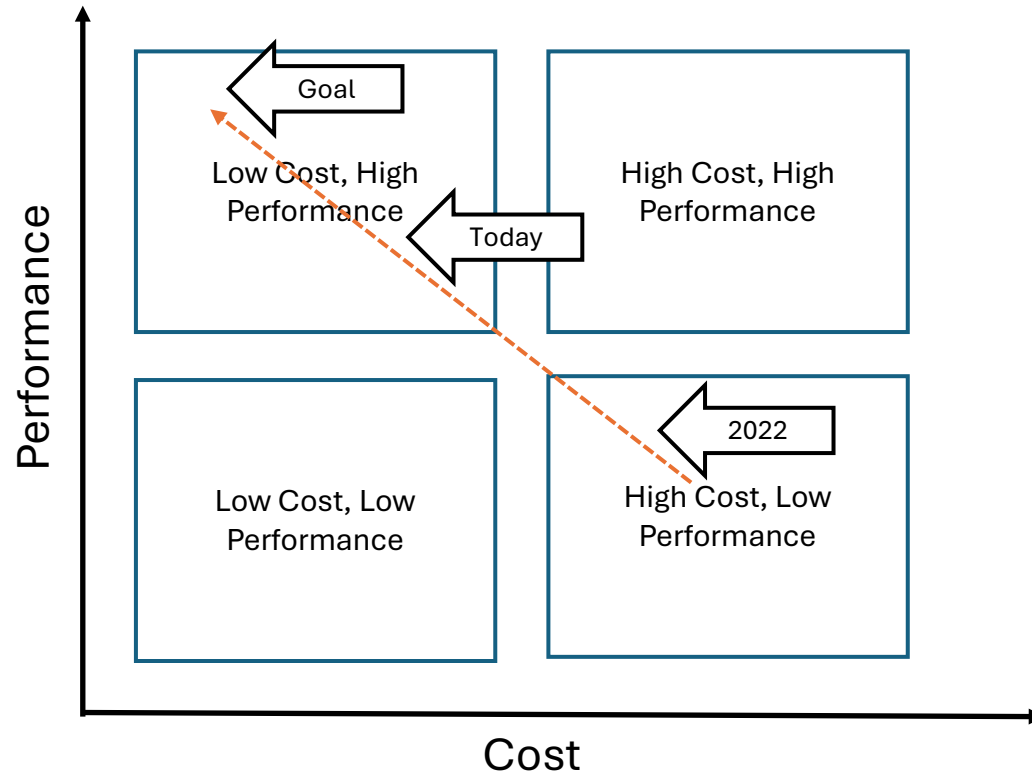
Rapid Data Growth

- B2C/B2B – TrekBikes.com
- Retail Focus
 - Significant Number of Company Owned Stores Globally
 - 2000 + Independent Bike Stores Globally
- 10 TB + of approx. size of “Silver” layer
- 500 GB + of approx. size of “Gold” layer, compressed and aggregated
- On-Premise/IAAS servers & PAAS on Azure were simply struggling to scale cost effectively

TREK
BIKES · SERVICE · GEAR

Cost - Performance Matrix

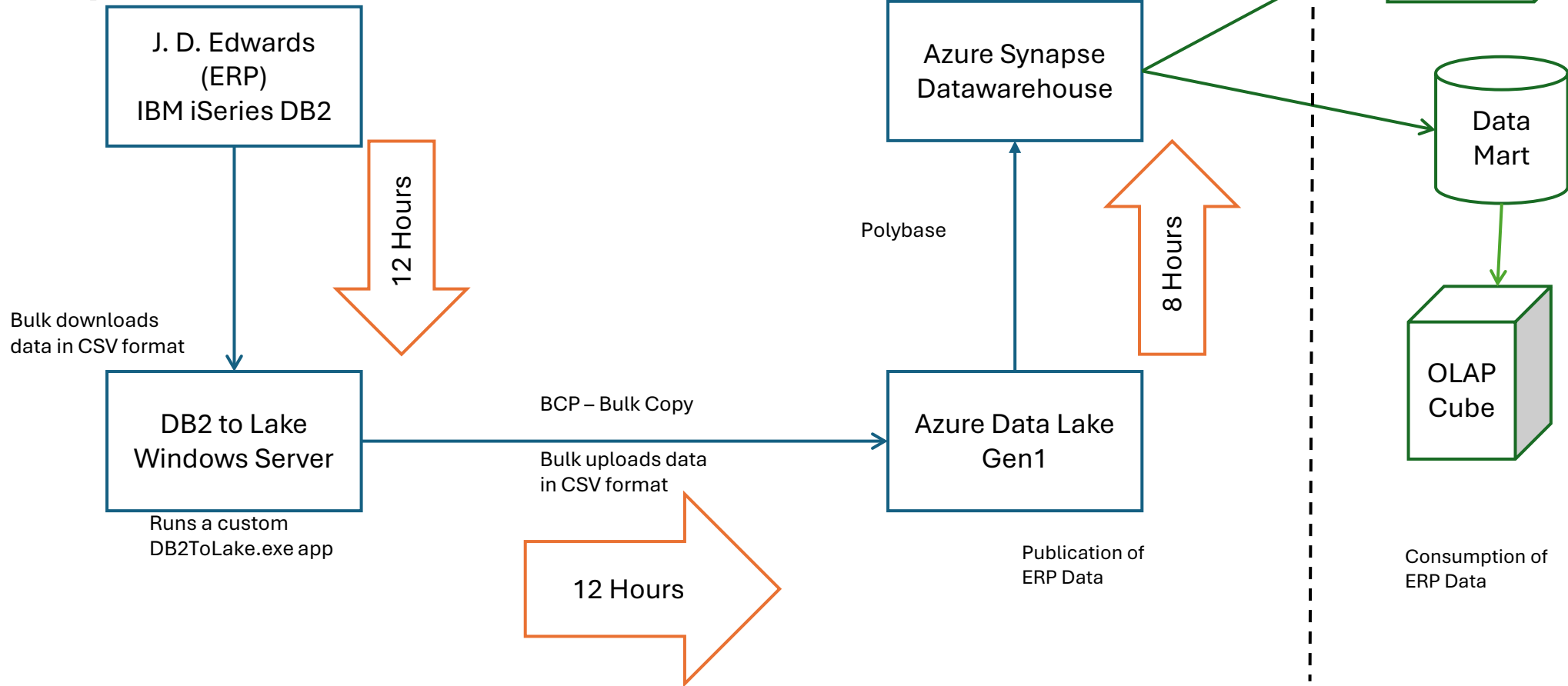
- Goal: Deliver high Datawarehouse performance at same or lower cost.



TREK

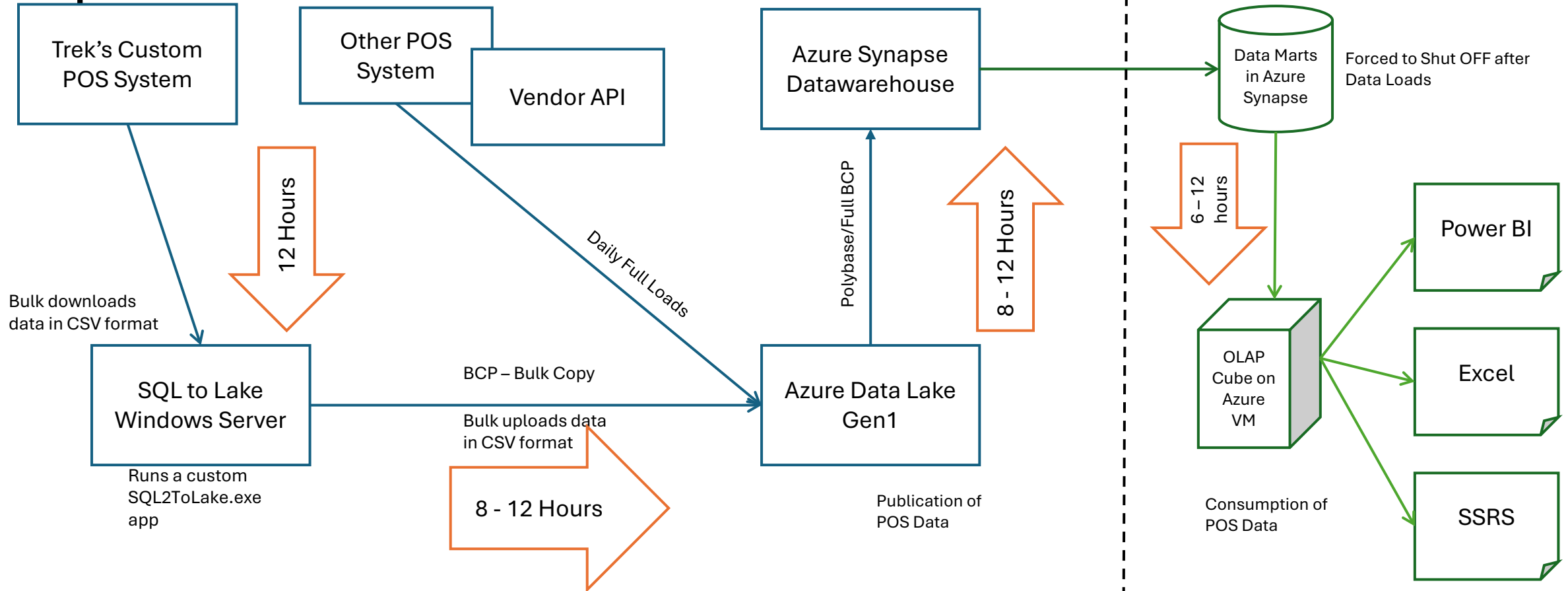
BIKES · SERVICE · GEAR

Old Architecture - ERP Replication



- Bulk Copy (BCP) architecture took 32-36 hours on good days and 48 hours during high volume days
- Reliance on custom stand-alone app needing a Windows Server
- Full loads every time – Wasted compute resources
- Reliance on expensive Azure Synapse Data Warehouse (now re-branded as dedicated SQL Pools in Synapse lingo)

Old Architecture - POS Replication



- Bulk Copy (BCP) architecture took 32-36 hours on good days and 48 hours during high volume days
- Reliance on custom stand-alone app needing a Windows Server on On-Prem and Azure
- Full loads every time – Wasted compute resources – Full replace of CSV files on Azure Data Lake Gen1
- Reliance on expensive Azure Synapse Data Warehouse (now re-branded as dedicated SQL Pools in Synapse lingo)
- Reliance on OLAP Multi-dimensional cubes on expensive Azure VM

Solution Overview

- ~~On-Prem SQL Server, SSAS, SSIS~~
- ~~Custom .Net apps~~
- ~~Azure Data Lake Gen1~~
- ~~Azure Synapse~~
- Azure Data Lake Gen2
- Azure Databricks
 - Apache Spark
 - Delta Tables
 - Photon
 - Lakehouse Architecture
 - Databricks SQL
 - Spark Streaming, Auto-Loader and Delta Live Tables
 - Unity Catalog
- CDC: QLIK Replicate, Five Tran, Azure Data Factory
- Semantic Layer: Power BI, At Scale

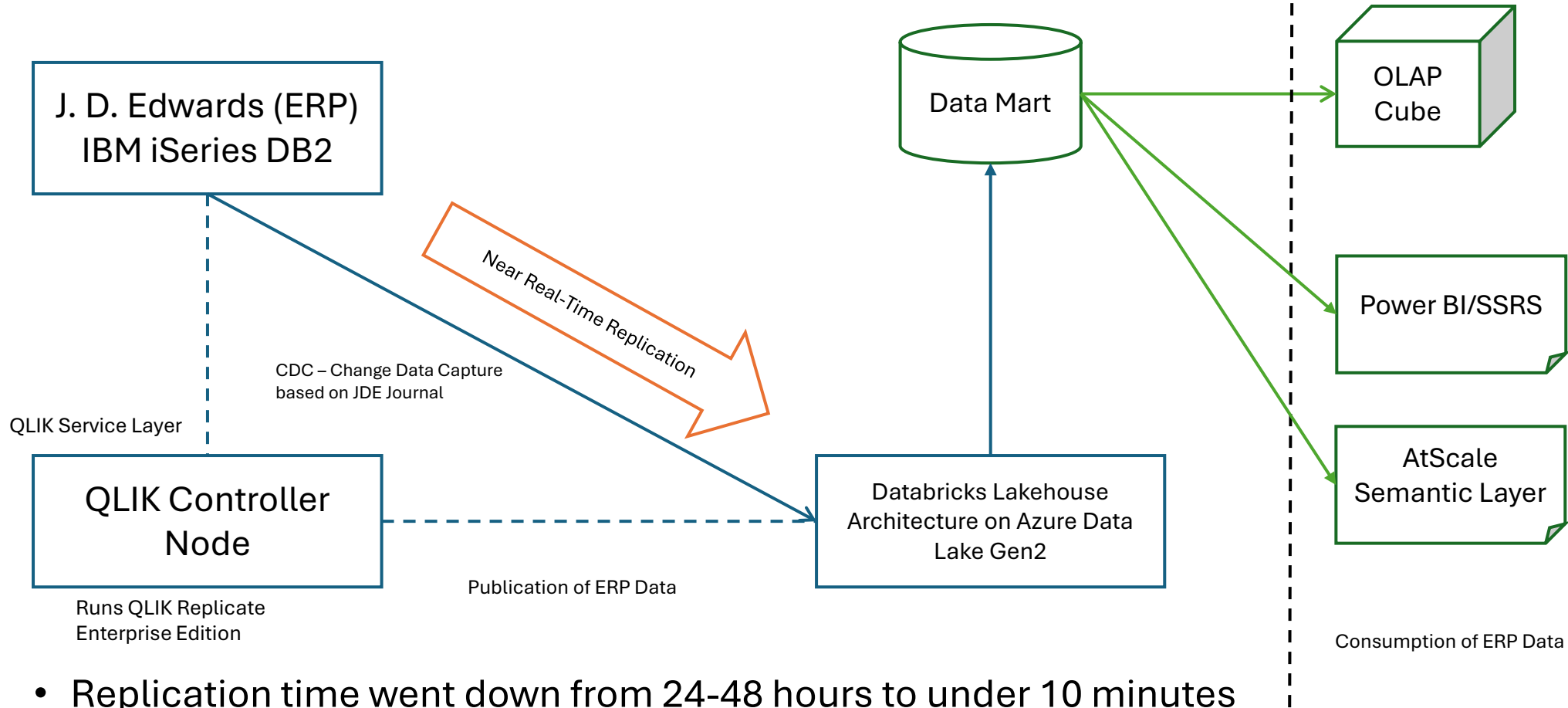
TREK
BIKES · SERVICE · GEAR

Why Databricks?

- Embrace the Open-Source paradigm:
 - Apache Spark
 - Delta universal format
 - Python/Scala
- Separation of Compute and Storage
- Cloud Agnostic
- A move away from vendor locked-in data format
- Lake House Architecture – Best of data lake and data warehouse
- A need for unified analytics platform

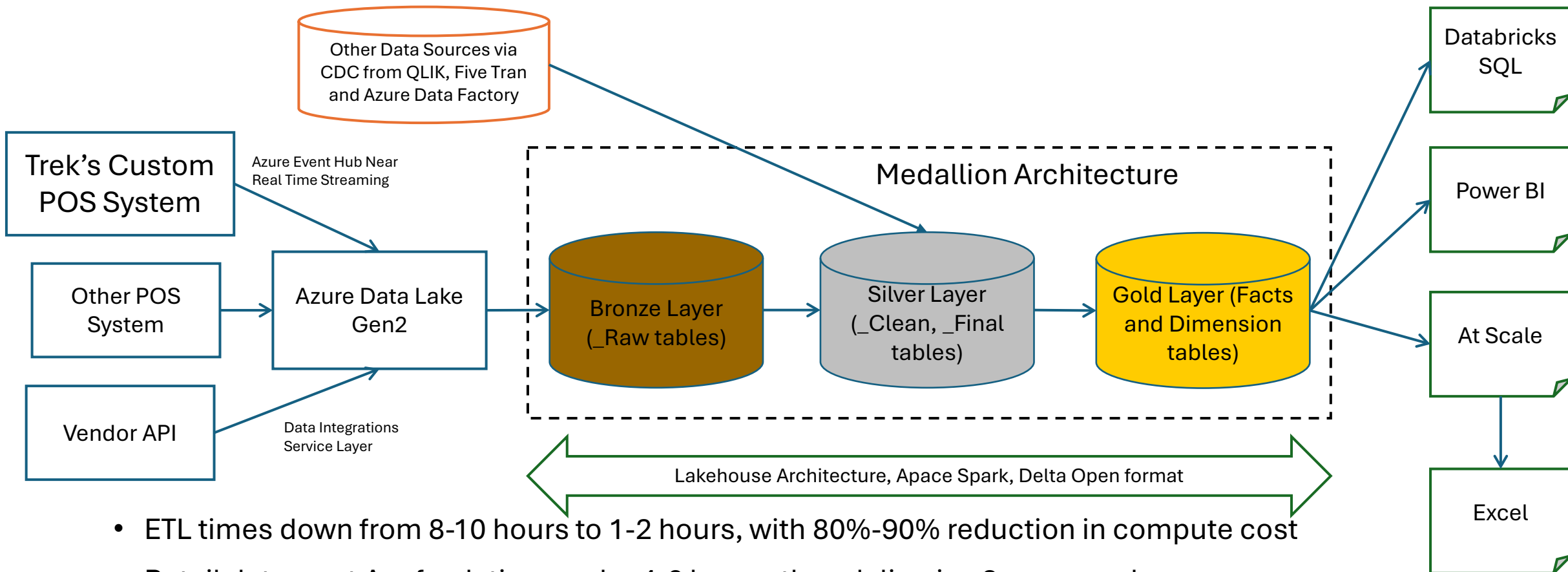
TREK
BIKES · SERVICE · GEAR

New Architecture - ERP Replication



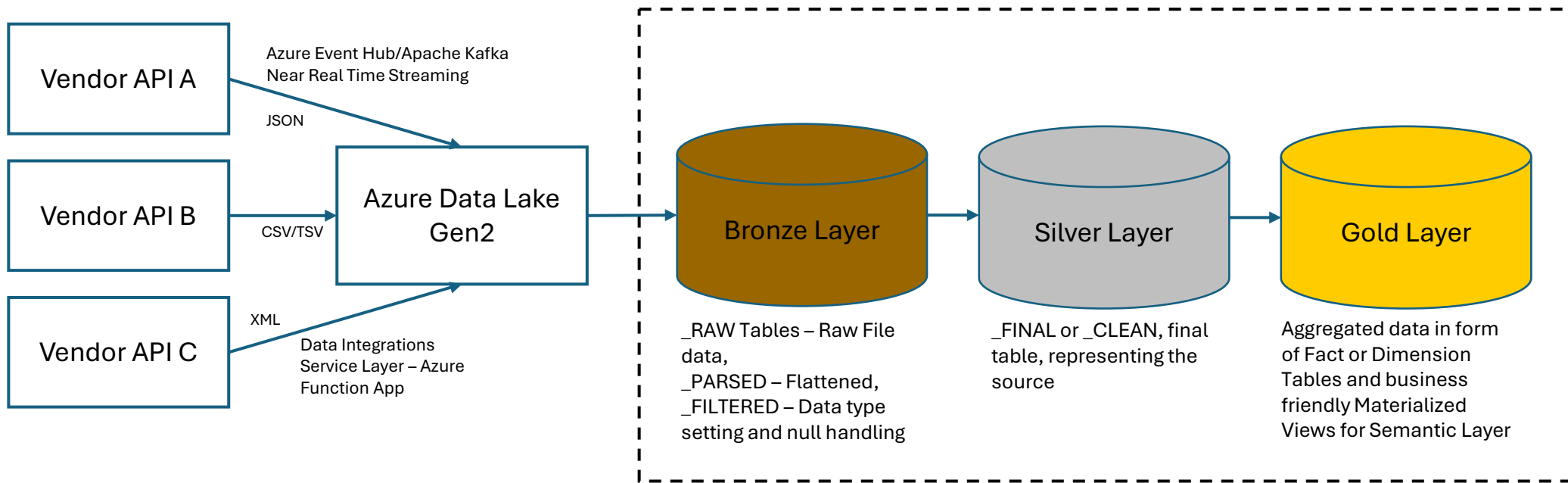
- Replication time went down from 24-48 hours to under 10 minutes
- Full supported QLIK Replicate Enterprise Application
- Great UI for management of CDC Pipelines
- True CDC system, Full Loads only on-demand
- Eliminated reliance on expensive Azure Synapse Data Warehouse
- Eliminated reliance on expensive Azure VM for BCP Application

New Architecture - POS Replication



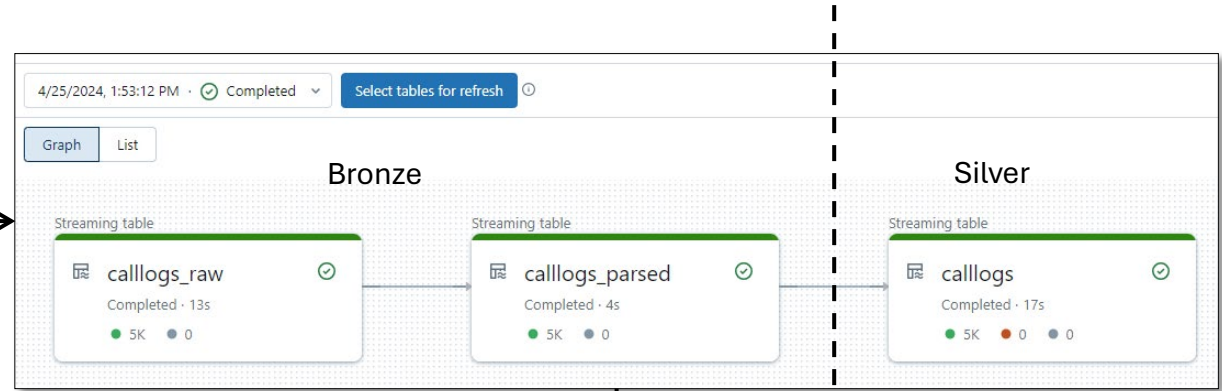
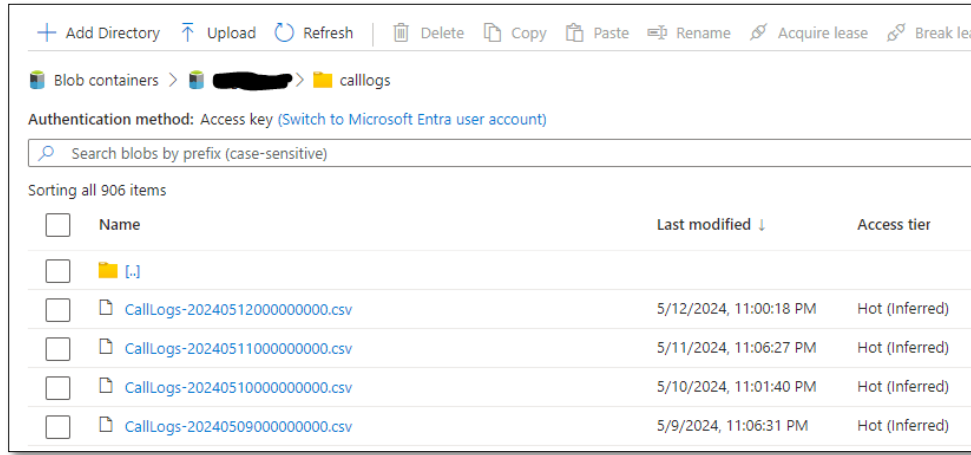
- ETL times down from 8-10 hours to 1-2 hours, with 80%-90% reduction in compute cost
- Retail data mart A refresh time under 4-6 hours, thus delivering 3 runs per day.
- Retail data mart B refresh time down from 48 hours to 6-8 hours, thus enabling daily runs.
- Eliminated reliance on expensive Azure Synapse Data Warehouse
- Eliminated reliance on expensive Azure VM for BCP SSAS OLAP Cubes
- Eliminated reliance on expensive On-Prem VMs for SSAS and SSIS

Streaming Use Case: Delta Live Tables

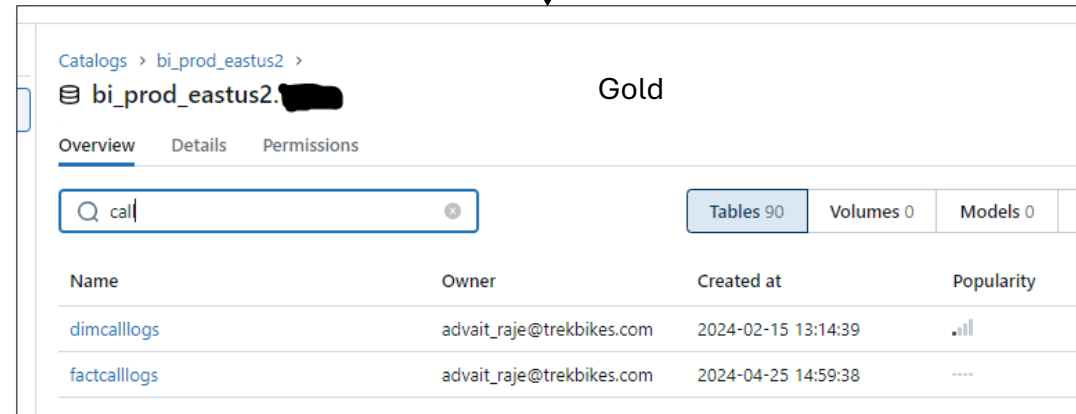


- When files (JSON, XML, CSV, TSV) arrive in Batch and/or Real or Near Real Time
- Checkpointing allows incremental Loading
- Smart auto scaling
- Ability to re-trigger a full load
- Ability to handle column drift and merge new schema
- SQL and Python Notebook friendly
- Unity Catalog enabled

Streaming Use Case: Delta Live Tables

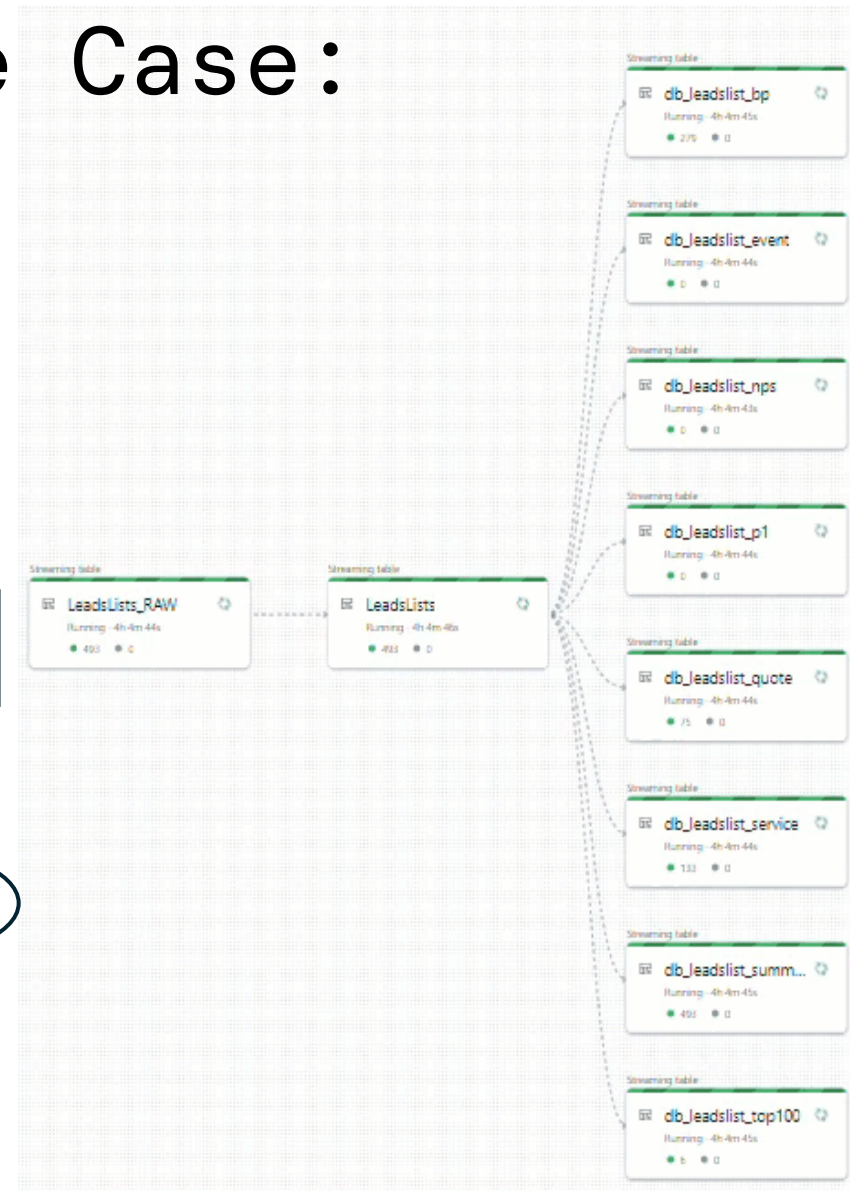
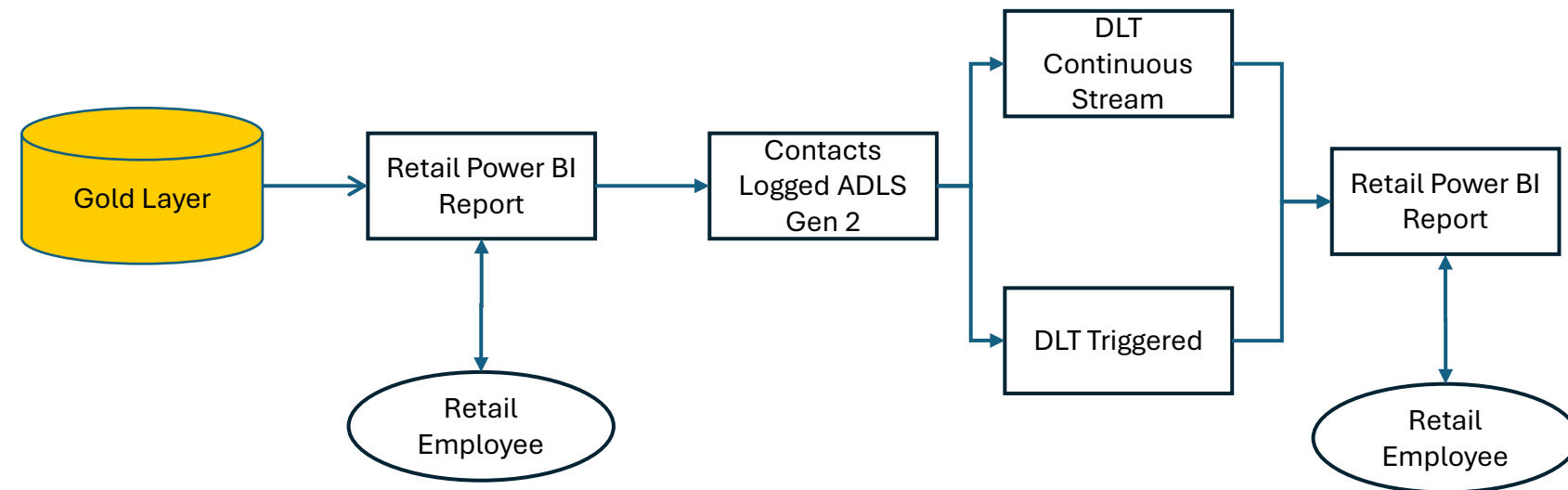


- Files arrive once a day (Batch)
- Arrival triggers a DLT Pipeline Job
- Bronze and Silver tables are created
- Job to build Gold Tables (Kimball style Fact and Dimension Tables) is triggered
- Power BI Semantic refresh



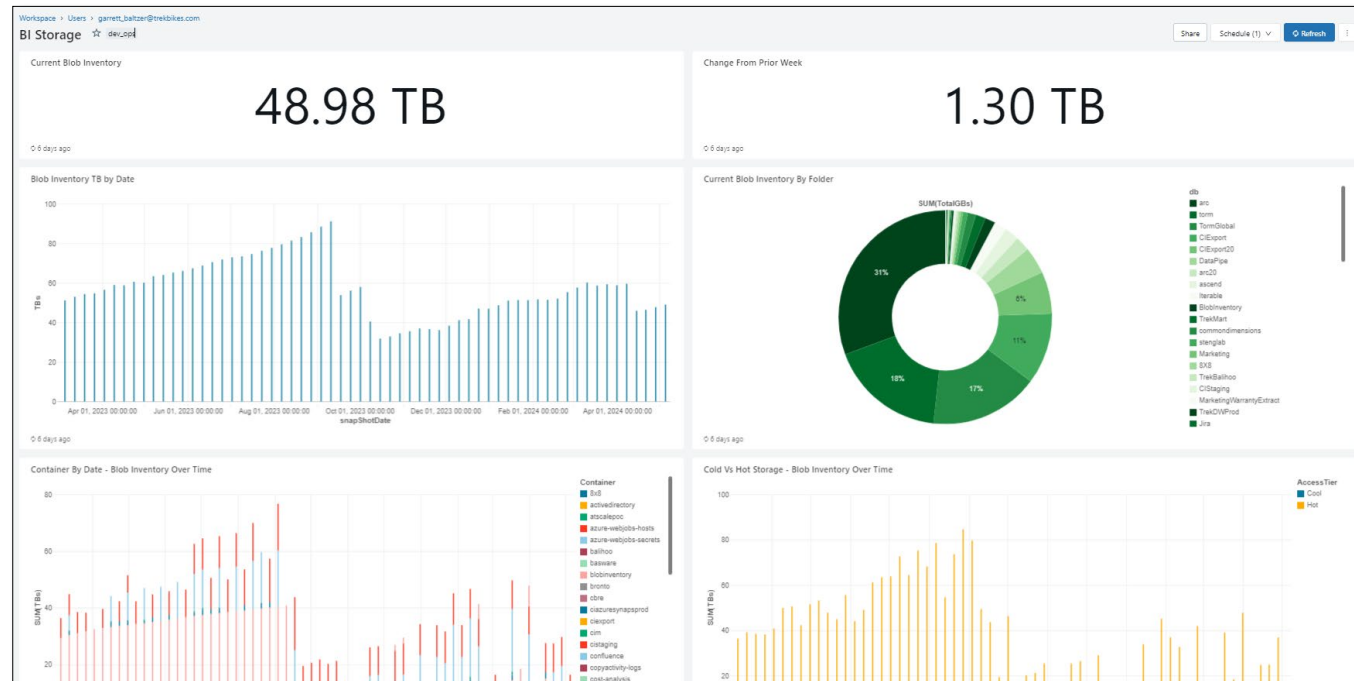
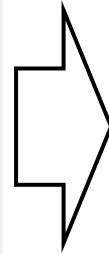
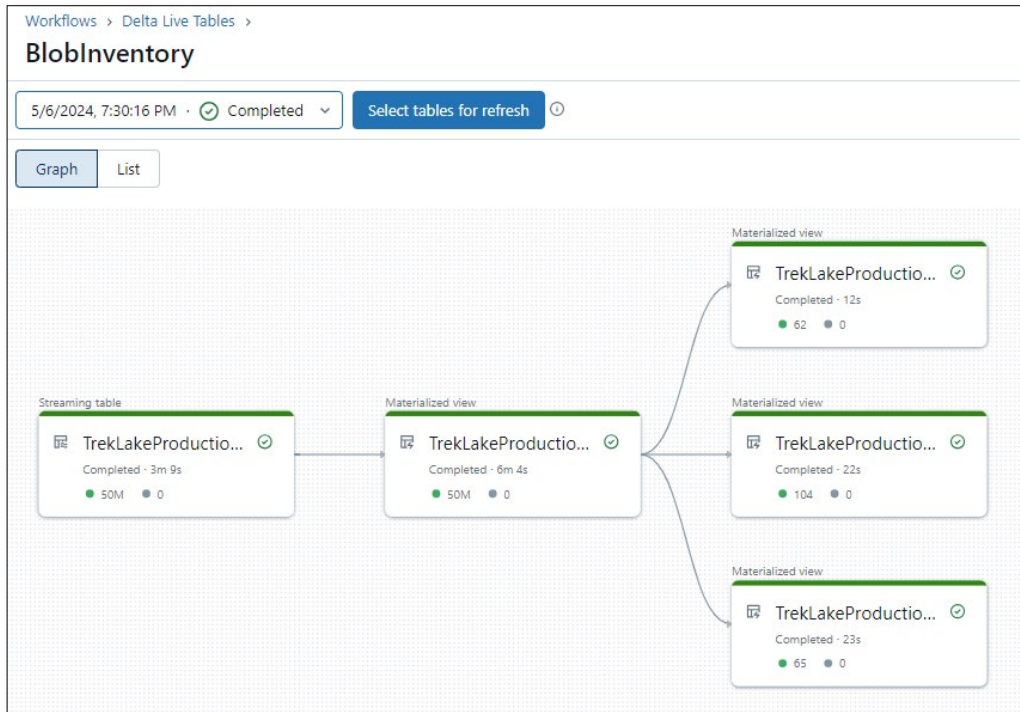
Hybrid Streaming & Batch Use Case: Delta Live Tables

pipelines.tableManagedByMultiplePipelinesCheck.enabled - false



- Batch Information is provided to our stores via Power BI
- Retail Employees across all company retail locations then action on the information
- Dedicated Classic SQL Warehouse with 5000+ Interactions per month
- Report updates in real-time from 8:00 AM to 5:00 PM CST, then every 3 hours thereafter.

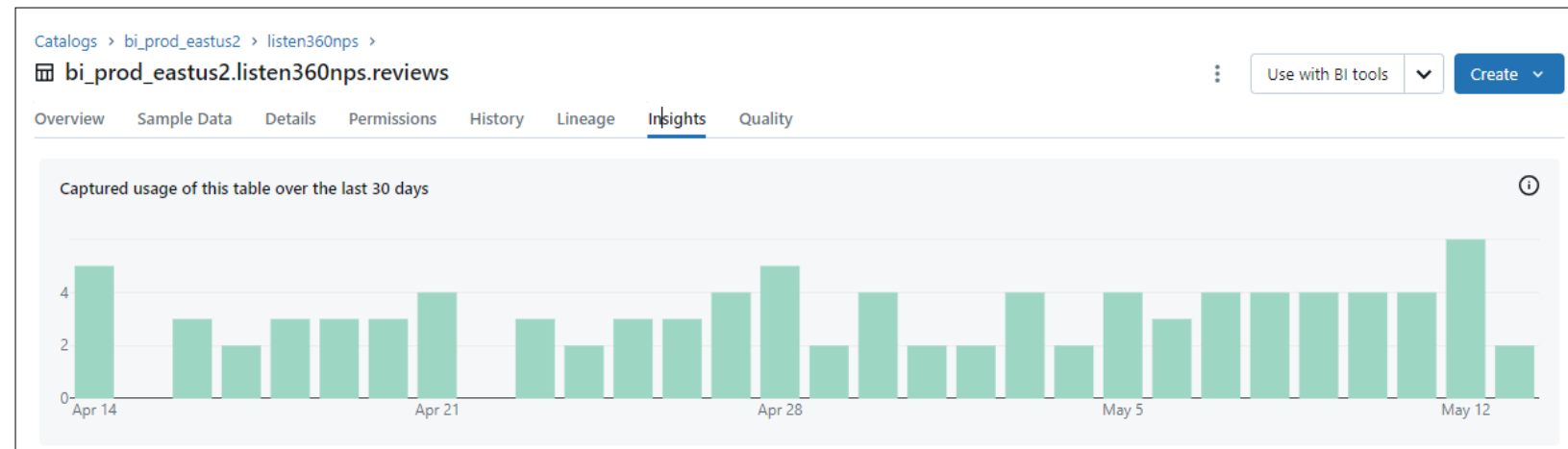
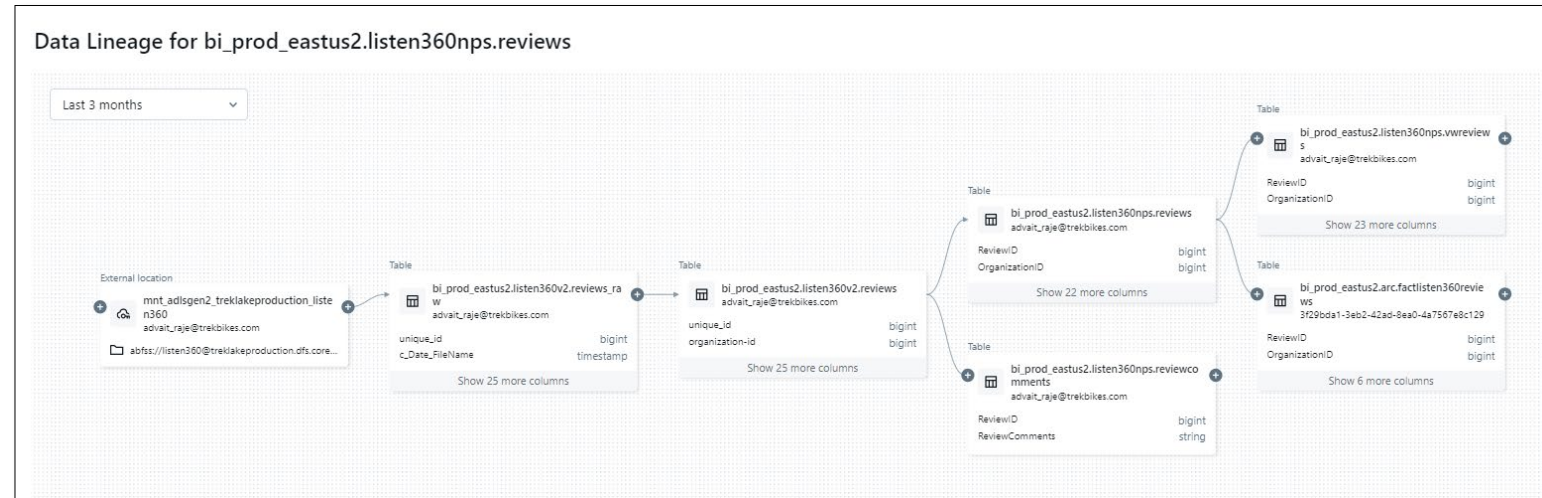
Databricks SQL



- DLT pipeline captures telemetry around Azure Storage Containers
- Databricks SQL for aggregations
- Databricks SQL Dashboards for visualizations

Unity Catalog

- Multiple Catalogs for Dev and Prod
- Separate workspaces for Analysts and Data Scientists
- External Sources for Storage Locations on Azure
- External Connections to Azure SQL Databases and other Databricks Workspaces



Key Principles for Effective Technology Integration:

1. Alignment with Company Values and Mission
2. Executive Support and Sponsorship
3. Strong Team Skills: SQL, Python, Scala, Pipelining
4. Relationship Building with Account Managers and Solutions Architects
5. Understanding Distributed Computing Theory
6. Optimized Delta Tables: Clustering, Deletion Vector, Auto-Optimization
7. Smart Use of Job Clusters and Cluster Tagging
8. Efficient Use of Photon & AQE
9. Utilization of Reserved Instances
10. Cost Awareness: Storage Management and Vacuuming
11. Emphasis on CI/CD Implementation
12. Simplified Data Warehouse Architecture: Kimball Date Modeling
13. Prioritize Data Quality and Ownership over AI/ML
14. Actionable Insights for Front-line Employees
15. Empower Front-line Staff with Data-driven Decision Making"

TREK

BIKES · SERVICE · GEAR

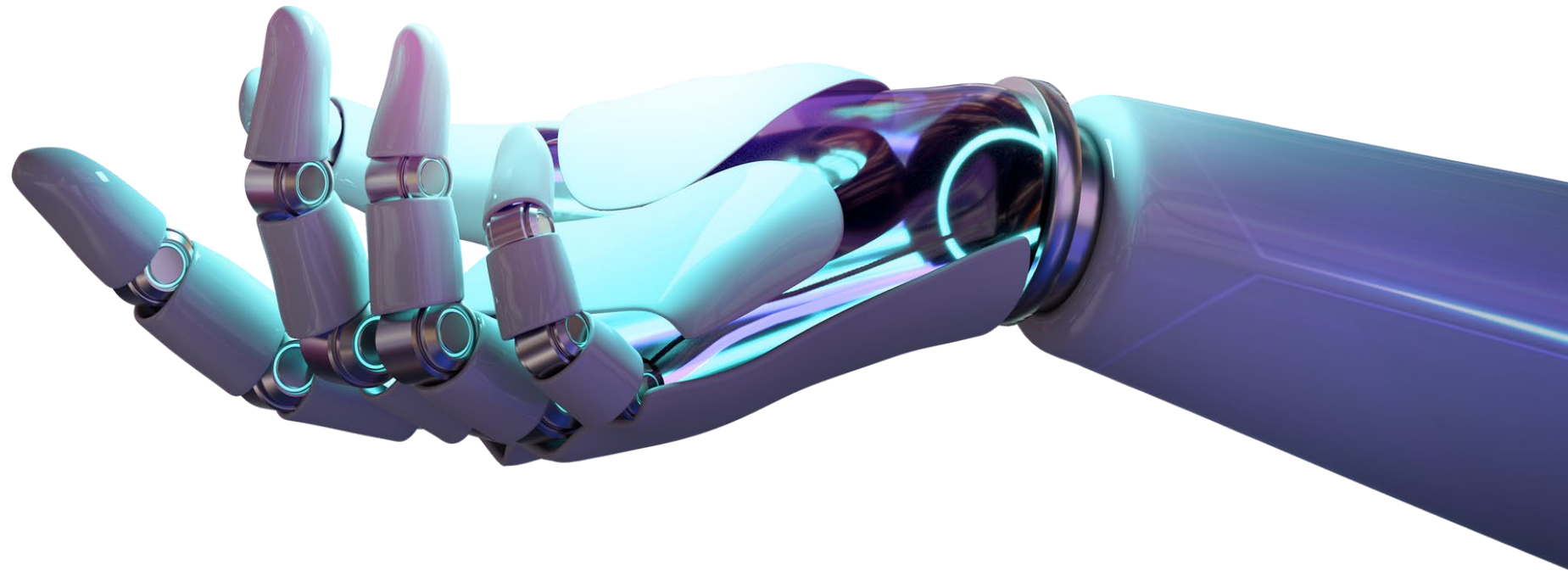
Conclusion - Databricks, QLIK Replicate

- 100% Adoption Achieved
- Near Real-time ERP & POS Data Replication
- Elimination of Daily or Weekly Lag(0) in Data Availability
- Substantial Reduction (75-85%) in Azure Spending
- Drastic Improvement (80-90%) in Runtime and SLA Compliance
- Increased Workload Frequency: From 1 to 3 times daily
- Enhanced Operational Efficiency: Workloads now run daily instead of weekly
- Improved Support with Enhanced Quality of Life



Q & A

ASK AWAY!



DATA+AI SUMMIT

