How CDC Integrated Complex Data to Drive Vaccination Forecasting with Databricks

Sheila Stewart, Senior Solutions Architect, Databricks
sheila.stewart@databricks.com

John Repko, Technical Program Manager, Peraton
John.Repko@Peraton.com

Jim Fetters, Cloud Solution Specialist, Microsoft Azure
Jamesfe@Microsoft.com

databricks

Agenda

- Introductions
- COVID Big Data Use Case
- Benefits of Cloud Services
- Benefits of a Data Lakehouse

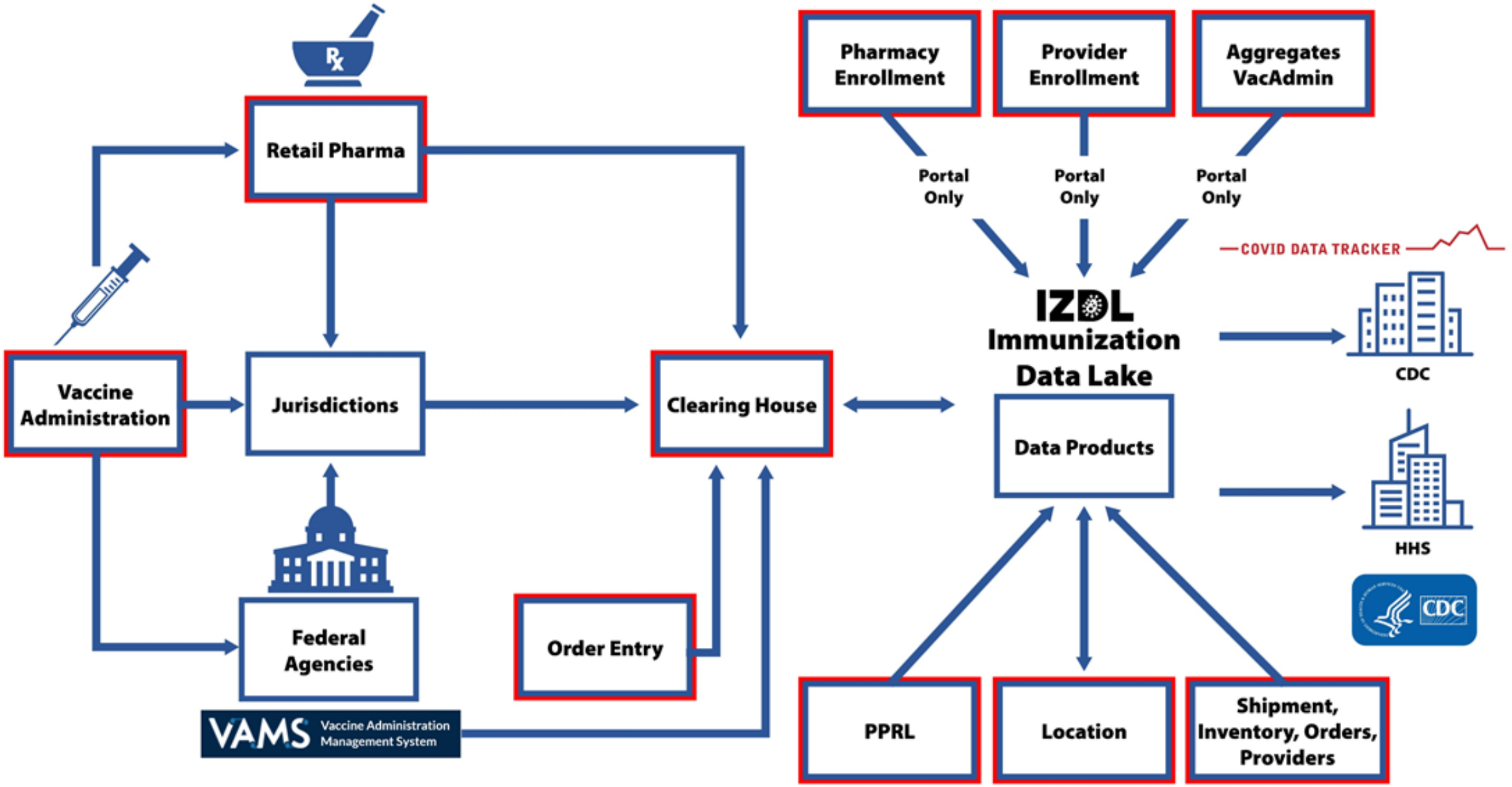
databricks

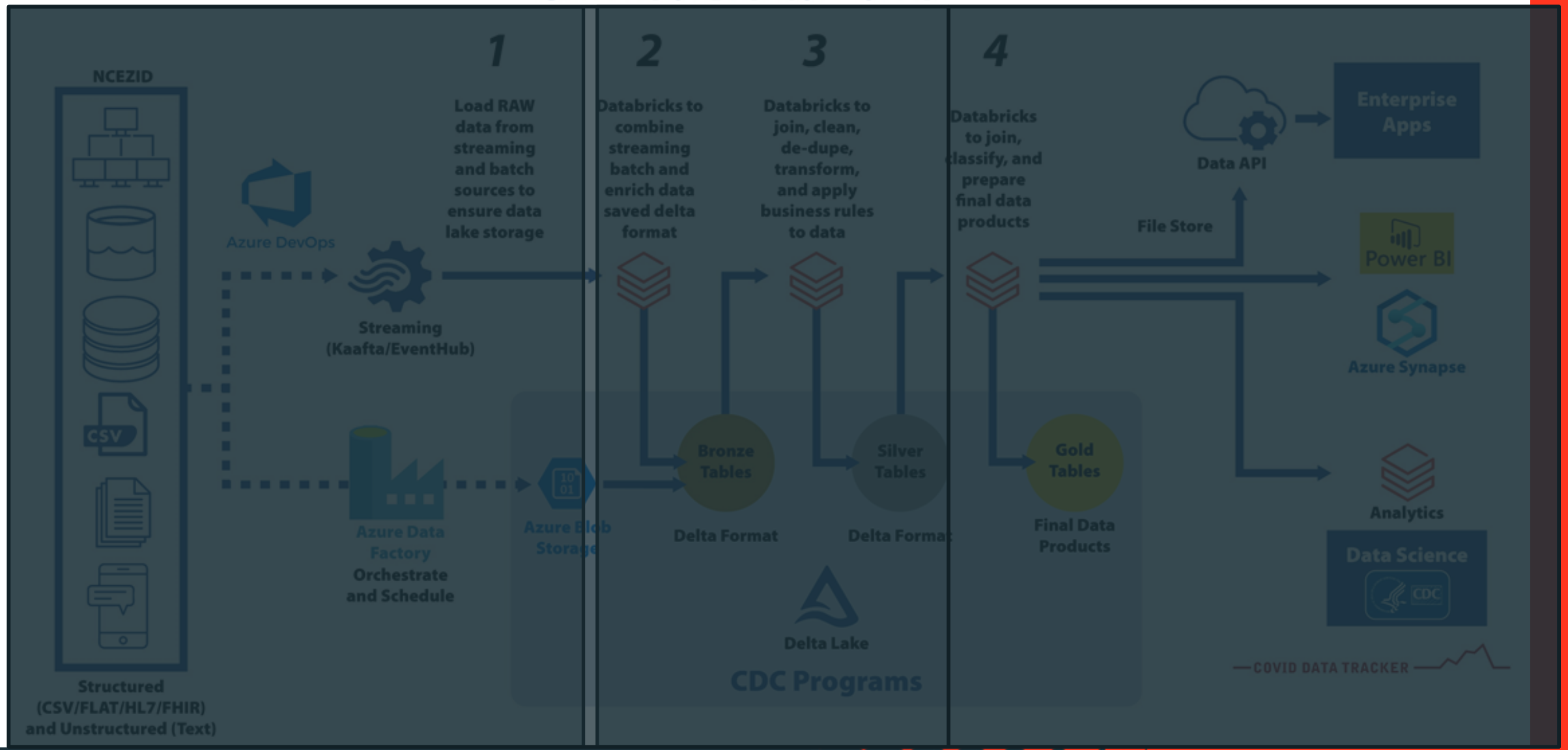
Peraton

Microsoft

COVID Big Data Use Case

Use Case


Pharmacy Enrollment
Provider Enrollment
Aggregates VacAdmin
Portal Only
Portal Only
Portal Only
COVID DATA TRACKER
IZDL Immunization Data Lake
Retail Pharma
Vaccine Administration
Jurisdictions
Clearing House
Data Products
CDC
HHS
Federal Agencies
Order Entry
VAMS Vaccine Administration Management System
PPRL
Location
Shipment, Inventory, Orders, Providers


Azure
Unified Delta Lake
databricks
1
2
3
4
NCEZID
Load RAW data from streaming and batch sources to ensure data lake storage
Databricks to combine streaming batch and enrich data saved delta format
Databricks to join, clean, de-dupe, transform, and apply business rules to data
Databricks to join, classify, and prepare final data products
Azure DevOps
Streaming (Kaafta/EventHub)
Azure Data Factory
Orchestrate and Schedule
Azure Blob Storage
Bronze Tables
Delta Format
Silver Tables
Delta Format
Gold Tables
Final Data Products
Delta Lake
CDC Programs
Data API
Enterprise Apps
File Store
Power BI
Azure Synapse
Analytics
Data Science
CDC
COVID DATA TRACKER
Structured (CSV/FLAT/HL7/FHIR) and Unstructured (Text)

Success

- **IZDL- stood up in 6 weeks from build out to Operational at CDC**

  8 Inbound Data Streams, 10 Outbound Data Streams and 45 Data Products

- **Limitless Spark Processing Power**

  At any scale for Streaming and Batch
  Volume – 5M+ new records per day, Billions Analyzed per day

- **Realtime Data Streaming Pipelines**

  Velocity – near Realtime – Hourly- Daily
  For very Large-Scale high throughput Parallel Processing (100K per sec)
  Full provenance (raw to aggregate)
  “What If” scenarios (Time Travel)
  Update/Deletes as Public Health Events

Success (cont)

- **Horizontal scaling for storage**

  Blob, Synapse -> Petabytes, Parquet (blazingly fast consumption)

- **Delivering Data to Public Health and multiple CDC Centers at Scale**

  COVID Data Tracker & Data.CDC.Gov
  Bulk Exports -> Terabyte Slices of Data
  Data APIs (Synch and Asynch)
  IZDL, CELR, DAART, EZDX, PHLIP, – In production or nearly so

Benefits of Cloud Services

Azure capabilities for public health

Data Interoperability

Standardizing, structuring and persisting data according to leading health standards (HL7 FHIR, DICOM)

Data governance

Discovering, protecting, and controlling data

Data modernization

Establishing future-proof platform that connects all data

Cloud native applications

Supporting high performance apps of any scale

Analytics and insights

Understanding customer behaviours to generate insights

Data science

Powering AI experiences

Microsoft

Azure Cloud Services

Azure Data Factory
Synapse Pipelines

Azure Purview
Data Governance
Lineage, MDM

Sources

Geospatial DB

DBs, files, etc

Ingest

Azure Data Factory

Synapse Pipelines

(raw) store

Azure SQL DB (Spatial)

Azure DB for PostgreSQL (PostGIS)

Azure Data Lake Storage Gen 2 (CDM)

Transform/ Analyze

Spark
Azure Databricks
Synapse Spark Pool

ADF Mapping Data Flow

Azure Batch

FME on VM

Load / Serve

Azure SQL DB
Azure DB for PostgreSQL

Synapse DWH

Cosmos DB

Azure Data Explorer

Publish

Azure Maps Data API

Consume

Client

Azure Maps: Web SDK

Azure App Service: Web App

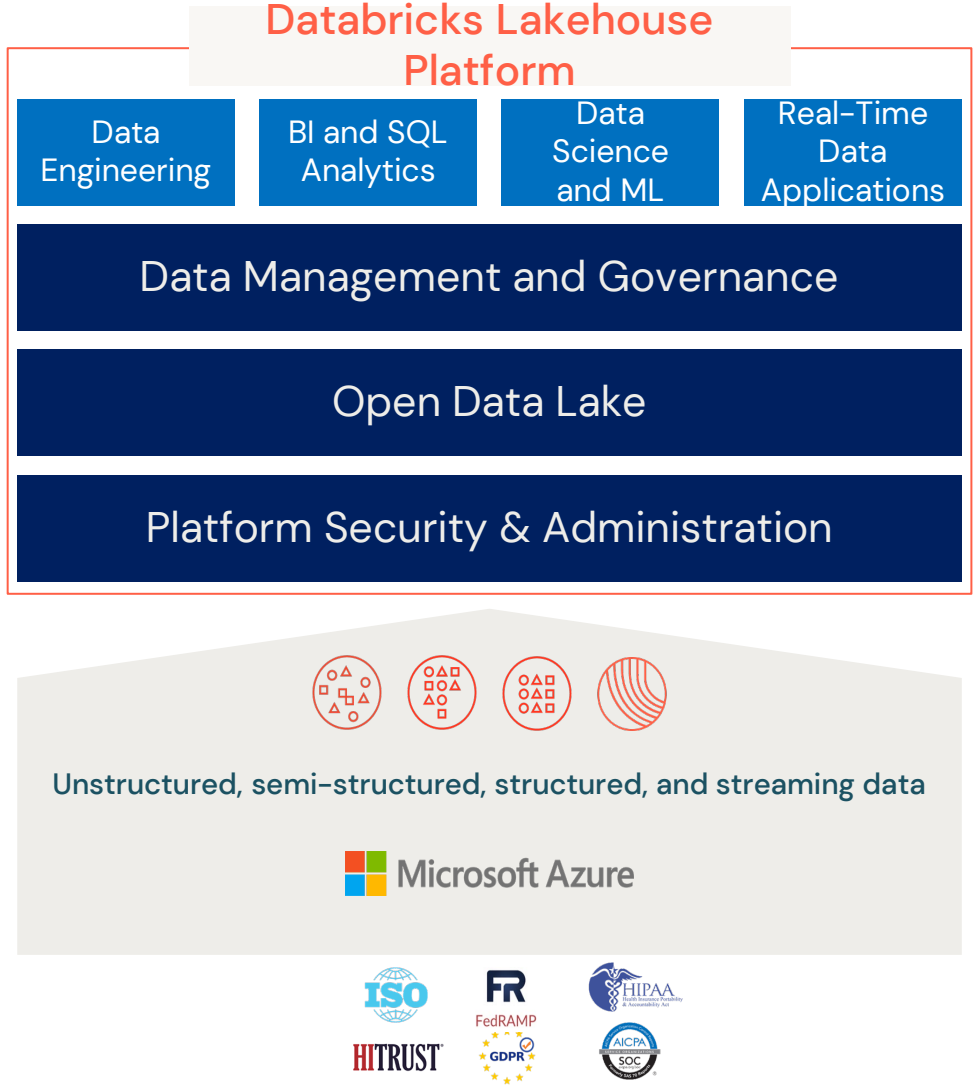
Power BI

Microsoft

Benefits of a Data Lakehouse

The Databricks Lakehouse Platform

- ✓ Simple
- ✓ Open
- ✓ Collaborative



What is a Lakehouse?

- One simple platform to unify all of your data, analytics, and AI workloads by combining the best features of data warehouses and data lakes.

databricks

Takeaways

Close teamwork between Databricks, Peraton, and Azure produced a very successful, quick-turnaround, reproducible cloud data management strategy

- Successful Forecasting is based on Rich, Complete, and Curated Data
- Whether historical data or real time collection is used for forecasting, data management needs planning and implementation
- It always takes a team of different technologies, processes, and dedicated people to be successful

- **THANK YOU to for the support from the CDC, Peraton, Microsoft Azure, and Databricks**

Questions?

Please come visit us at our booths!