Dynapt.ai

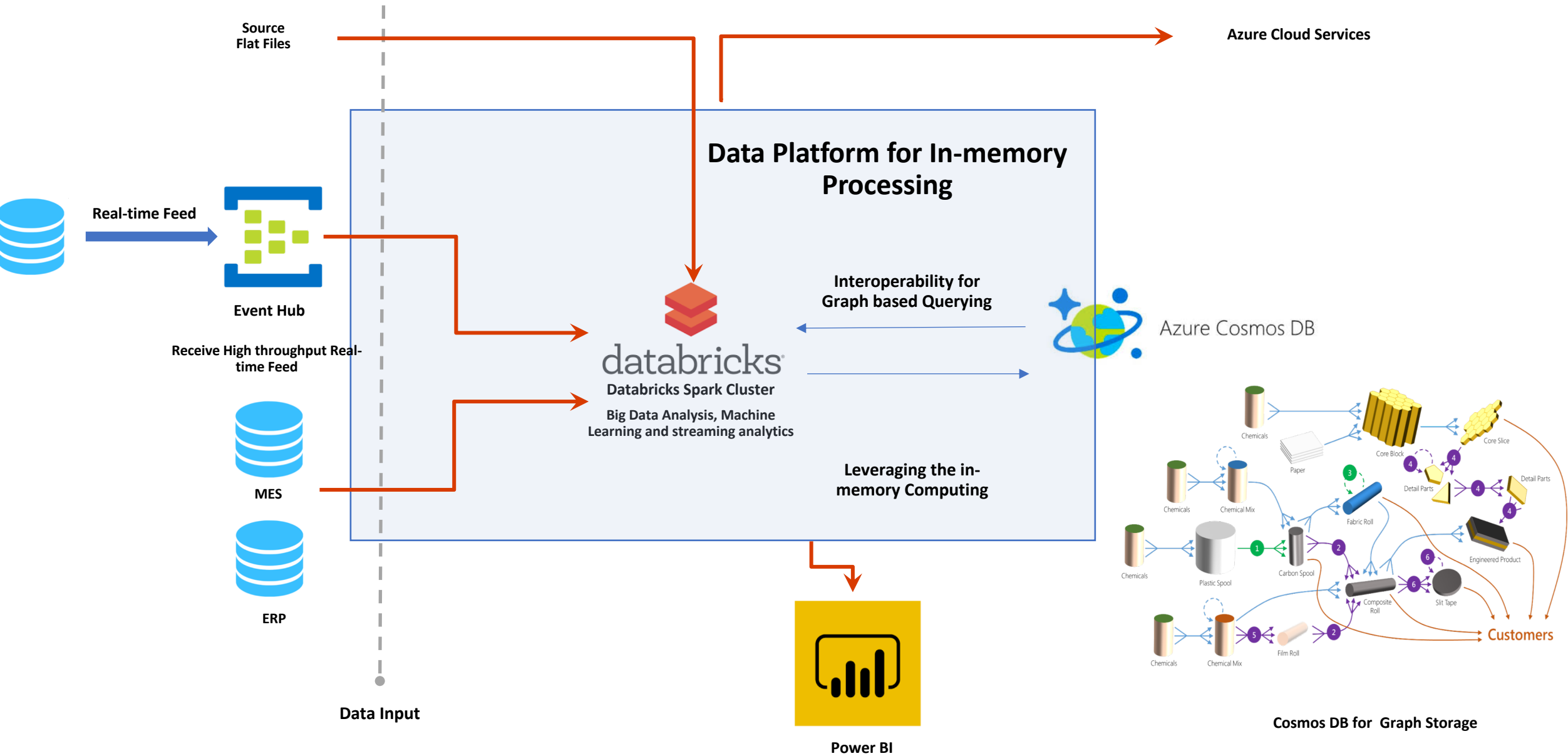## Batch Genealogy Solution on Azure for Discrete Manufacturing

Background

- Customer produces material in several manufacturing stages

- Data is attributed as material manufacturing traverses through several layers of a varying BOM hierarchy

- Analysing such dataset in SQL like structure is not feasible, especially with large volume (A conventional query for such analysis can generate large temp. data)

Current Landscape

- Data sets from ERP and MES systems feed the relational data base

- Information is stored in traditional RDBMS with regular SQL queries

- Databricks (Apache Spark) is being explored for Analysis and Data Science

Proposed Architecture Option

Source Flat Files

Azure Cloud Services

Data Platform for In-memory Processing

Real-time Feed

Event Hub

Interoperability for Graph based Querying

Azure Cosmos DB

Receive High throughput Real-time Feed

Databricks Spark Cluster

Big Data Analysis, Machine Learning and streaming analytics

MES

Leveraging the in-memory Computing

ERP

Data Input

Power BI

Cosmos DB for Graph Storage

Architecture Components

Azure Databricks

- Azure Cloud-based Big Data platform on top of Spark to make big data simple.
- Designed in collaboration with the founders of Apache Spark.
- Native integration with Azure services (CosmosDB, PowerBI, SQL DW, ADL, Blob Storage)
- Enterprise-grade Azure security (Active Directory integration, compliance, enterprise-grade SLAs)
- Interactive workspace (Notebooks) that enables collaboration between data scientists, data engineers and business analysts.
- One-click setup; streamlined workflows

Cosmos DB

- Globally distributed, multi-model database service
- Supports Open Graph APIs (based on the on the Apache TinkerPop specification, Apache Gremlin).
- Elastically and independently scale throughput and storage on demand and worldwide
- Build highly responsive and mission-critical applications

Event Hub

- Hyper-scale telemetry ingestion service that collects, transforms, and stores millions of events
- Distributed streaming platform that gives low latency and configurable time retention
- Enables you to ingress massive amounts of telemetry into the cloud and read the data from multiple applications using publish-subscribe semantics.

Proposed Architecture

- Consume data from Source Apps (ERP, MES) into Azure Databricks (via Batch and Streaming)

- Databricks In-Memory Engine would process data such that it could be consumed by a Graph DB (Cosmos) [Data Conversion across Schemas with Scala]

- CosmosDB would allow Graph Queries on attribute data between any two stages of a manufacturing process on the fly

- Interoperability between Azure Databricks and CosmosDB

- Azure Databricks would also support Data Science workloads

Scope

- Setup Landscape

- 4 Data feeds [batch] into Azure Databricks

- 30 Million Records (Edges)

- Data Transformation (Relational to Graph)

- Data Movement to Cosmos DB

- 2 Graph queries on Cosmos DB

Outcome

- Scheduled Data Loads

- Data Transformation

- Query Performance (Target < 5 Sec)

- Update times for DB (1 Record/Edge)

- Ability to query across different stages of manufacturing process by leveraging Graph queries on Cosmos DB

Thank You