FRAGMA DATA

Data Lake on Azure Databricks

Data Ingestion Pipeline

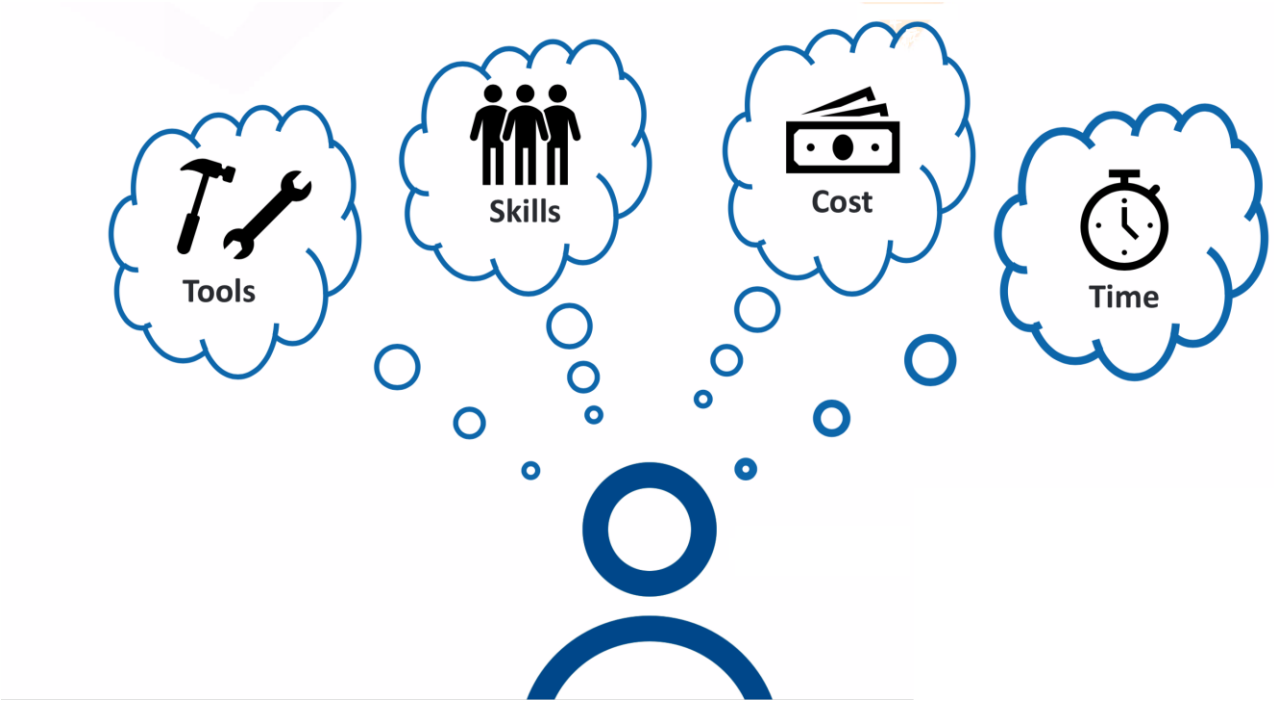
Traditional data warehousing and business intelligence approaches have been challenged as being too slow to respond. Reducing the time to value is a primary objective of a modern data architecture. In a modern data architecture, acquiring new data should be relatively easy so that new analysis can be conducted swiftly.
With a modern data lake, organizations can continue to leverage their existing investments, begin collecting data they have been ignoring or discarding, and ultimately enable analysts to obtain insights faster.

Data Migration to Cloud using Generic Data Ingestion Pipeline

- Secure flexible and easy to use
- Real Time Data Analytics
- Lower time to onboard to data lake using Generic Data Ingestion Pipeline Framework
- Performance Optimization

Advanced Analytics – Key Challenges

- **Multitude of Tools –** No unified interface for data pipelines, exploration, analytics and modelling
- **Skills –** To match each of the tools used in the stack
- **Cost –** Predictability and Manageability of the cost
- **Time –** For Envisioning to Rolling Out for consumption
- **Scattered Data –** Data split across multiple systems and data sources and challenges in federated queries



Data Ingestion Pipeline – Key Values

Data Ingestion Pipeline can help to adopt Databricks as the centralized analytics platform to address speed and ease-of-use concerns, improve product design, troubleshoot quickly, and fine-tune the performance of production systems.

Value Delivering

- Real Time Data Analytics
- Enterprise level security
- Deploy big data technologies easily
- Cost Effective
- Store data of any size, shape and speed with Azure Data Lake.
- High speed connector to Azure Data Storage.

Azure Data Lake can handle any data in their native format, as is, without requiring prior transformations. Data Lake does not require a schema to be defined before the data is uploaded, leaving it up to the individual analytic framework to interpret the data and define a schema at the time of the analysis. Being able to store files of arbitrary size and formats makes it possible for Data Lake to handle structured, semi-structured, and even unstructured data.

CIO IDG Research Survey Results - Advanced Analytics

- 90% of the corporates are investing heavily in Advanced Analytics
  - However, only 1/3rd of the projects make their way to production
  - Long wait time for rollout to production averaging more than 6 months
- Challenges faced with:
  - Data Preparation – 56 %
  - Data Exploration – 56 %
  - Deploying Models – 53 %
  - Siloed Data – 80 %
  - Multiple Tools – Avg of 7 tools for a single project
  - 96% of the adopters face these challenges

Unified Data Platform – Azure DataBricks

- Azure Databricks is a Unified Data Platform providing fast, collaborative analytics using Apache Spark.
- Azure Databricks handles:

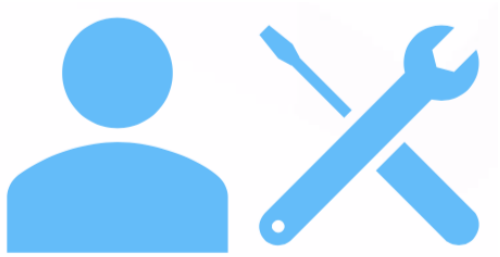
Batch Processing

Stream/Event Processing

Interactive Analytics

Machine Learning

Single Tool for Multiple Users – Azure DataBricks



Data Engineer

Create jobs for batch or streaming/realtime data load, transformation. Scheduling the Jobs

Data Scientist

Explore Data, Create and Deploy Machine Learning Models and Perform Other Advanced Analytics Tasks

Business Analyst

Write SQL Queries, Analyze and Visualize Data in Notebooks or external tools like PowerBI

Azure Databricks – Formal Introduction

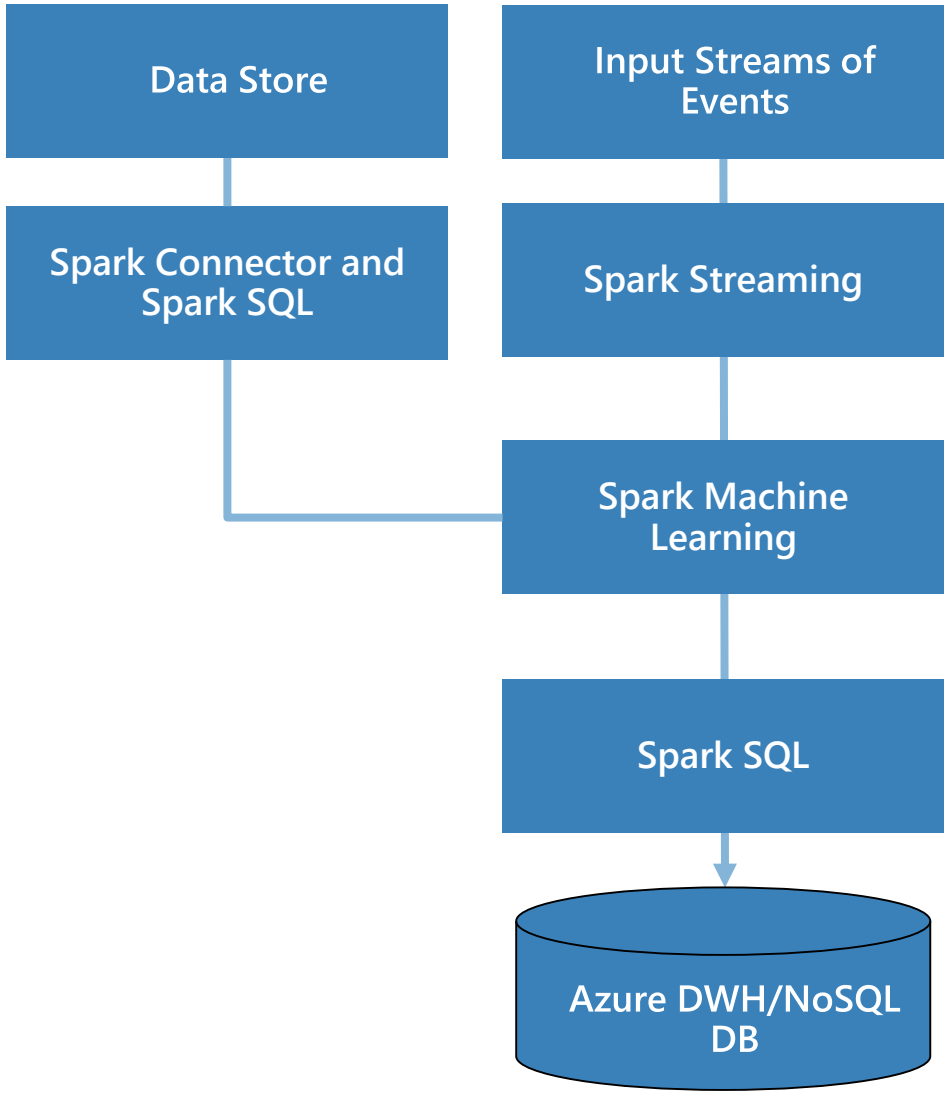
- Azure Databricks is a **first party** service on Azure.
  - Unlike with other clouds, it is not an Azure Marketplace or a 3rd party hosted service.
- Azure Databricks is integrated seamlessly with Azure services:
  - Azure Portal: Service an be launched directly from Azure Portal
  - Azure Storage Services: Directly access data in Azure Blob Storage and Azure Data Lake Store
  - Azure Active Directory: For user authentication, eliminating the need to maintain two separate sets of uses in Databricks and Azure.
  - Azure SQL DW and Azure Cosmos DB: Enables you to combine structured and unstructured data for analytics
  - Apache Kafka for HDInsight: Enables you to use Kafka as a streaming data source or sink
  - Azure Billing: You get a single bill from Azure
  - Azure Power BI: For rich data visualization
- Eliminates need to create a separate account with Databricks.

Azure Databricks – Multi Language and Framework Support

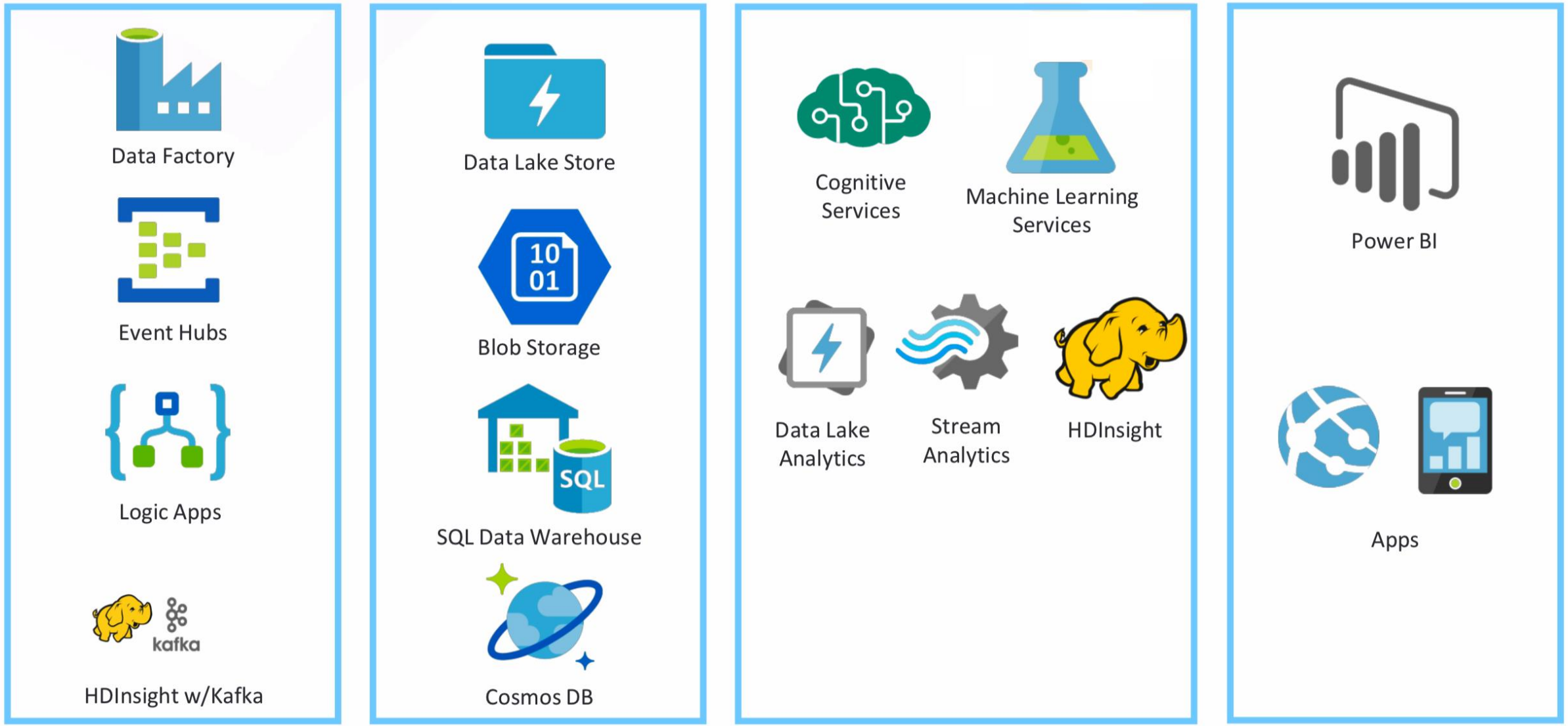
- Write ETL, Train Machine Learning Models, Perform Data Exploration using any of the following languages
  - Scala
  - Python
  - R
  - SQL
- Works with Popular Data Science and Deep Learning Frameworks:
  - Spark MLlib
  - TensorFlow
  - PyTorch
  - Scikit Learn
  - Horovod

Advantages of Unified Platform

- Improves developer productivity—a single consistent set of APIs
- All different systems in Spark share the same abstraction – RDDs (Resilient Distributed Datasets) and DataFrames
- Developers can mix and match different kind of processing in the same application. This is a common requirement for many big data pipelines.
- Performance improves because unnecessary movement of data across engines is eliminated. In many pipelines, data exchange between engines is the dominant cost

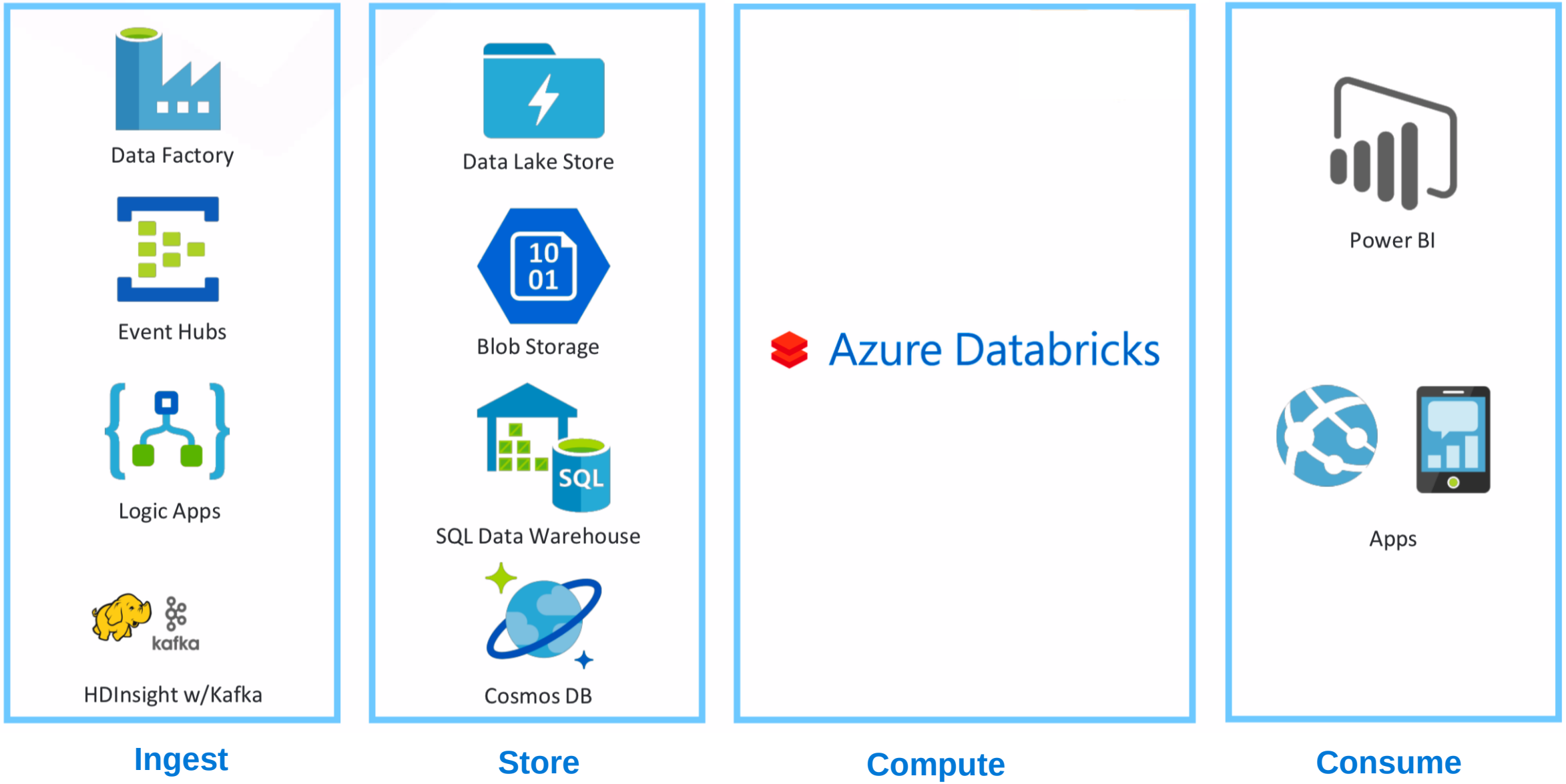


Azure Databricks – Simplifying the Azure Ecosystem



Ingest **Store** **Compute** **Consume**

Azure Databricks – Simplifying the Azure Ecosystem



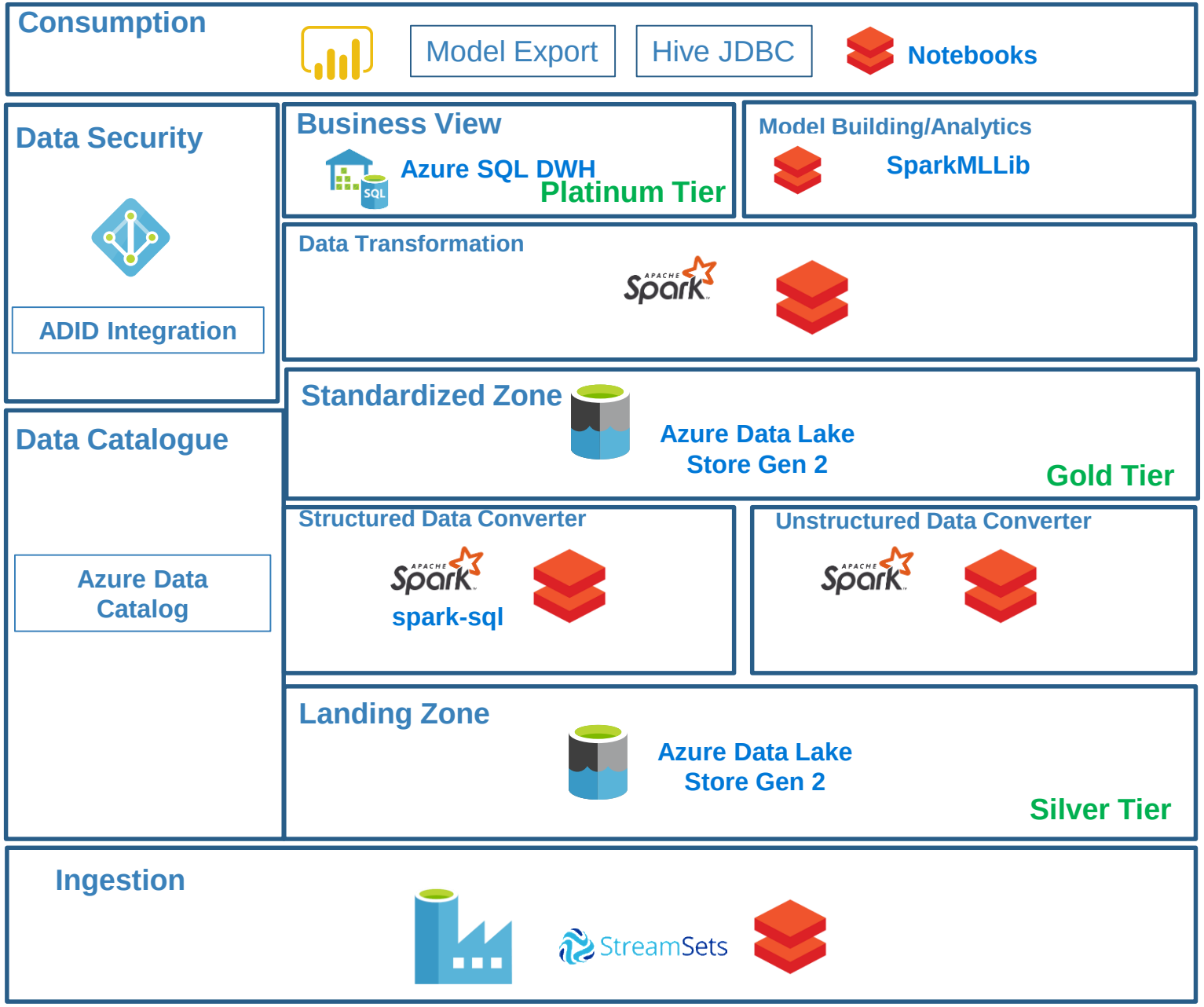
Data Lake Key Requirements

- Enterprise Scale Data Storage
- Support Ingestion and Processing of Structured, Semi-Structured and Unstructured Data
- Massive Data Processing Architectures
- Support Batch and Stream Processing
- Support for Exploratory Analysis, Model Building and Deployment
- Authentication and Authorization, Granular Access Control
- Scalability
- Metadata Management using Intelligent Cataloging
- Support for Downstream Reporting

Data Lake Solution - Key Components

- Data Extraction
- Data Ingestion
- Data Catalogue
- Data Transformation
- Data Exploration and Analysis
- Data Consumption
- Data Security
- Support for Downstream Reporting

Datalake Component Mapping



Use Case : Bureau Time Series Segmentation for a leading NBFC

- **Problem Statement**

  Segment the customer base depending upon internal and external bureau data for default propensity

- **Challenges**

  Sheer amount of size of data ~ 1 TB

  Application of segmentation model in SAS took 21 days

- **Solution Approach**

  All the data sources were made available in Azure Data Lake Storage by setting up data pipelines.

  Implemented the segmentation model in Spark.

- **Benefits**

  The segmentation run came down to 4 hours, thus giving business instant feedback.

Use Case : Real Time Processing of Events from Construction Machinery

- **Problem Statement**

  Customer has ~ 16,000 on field construction machinery generating about 1,000 events/second.

- **Challenges**

  Currently they were processing the IoT events as batches, recomputing metrics for whole day. The computations were getting delayed and was not scalable

- **Solution Approach**

  Consuming the IoT events directly from IoTHub and processing using structured streaming to update the metrics

- **Benefits**

  Almost realtime availability of the metrics from the latest events being obtained. Scalable solution with increase in devices

Thank You

connect@fragmadata.com

+91 97423 47119

www.fragmadata.com

FRAGMA DATA

From Data to
Insights