IntelliPaat

Designing and Implementing an Azure Data Solution

DP 200 and DP 201

IntelliPaat

Azure Data Bricks and Azure Stream Analytics

Agenda

01 What is Azure Databricks?

02 Azure Spark-based Analytics Platform

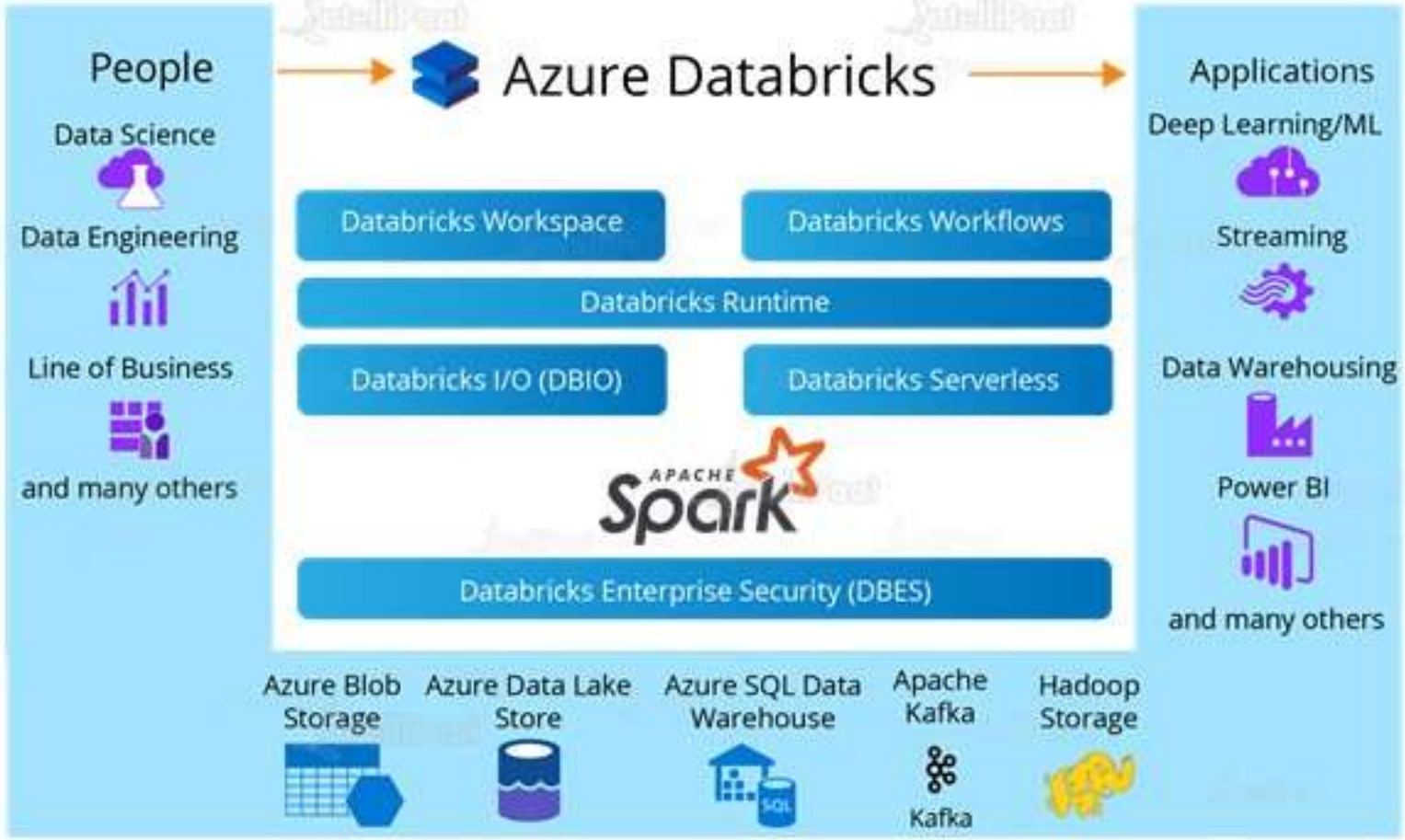
03 Apache Spark in Azure Databricks

04 Azure Stream Analytics

05 Stream Analytics Windowing Functions

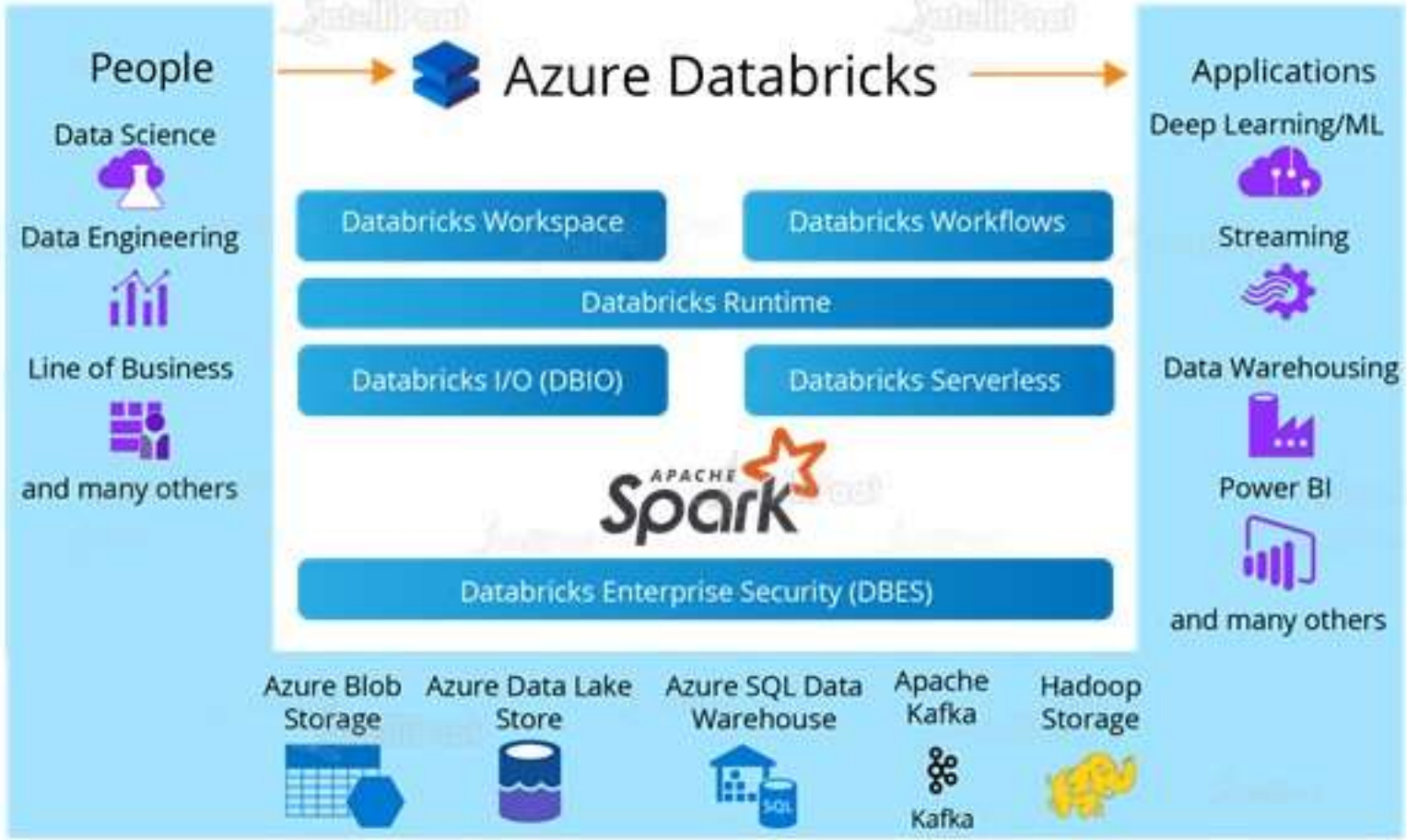
What is Azure Databricks?

Azure Databricks



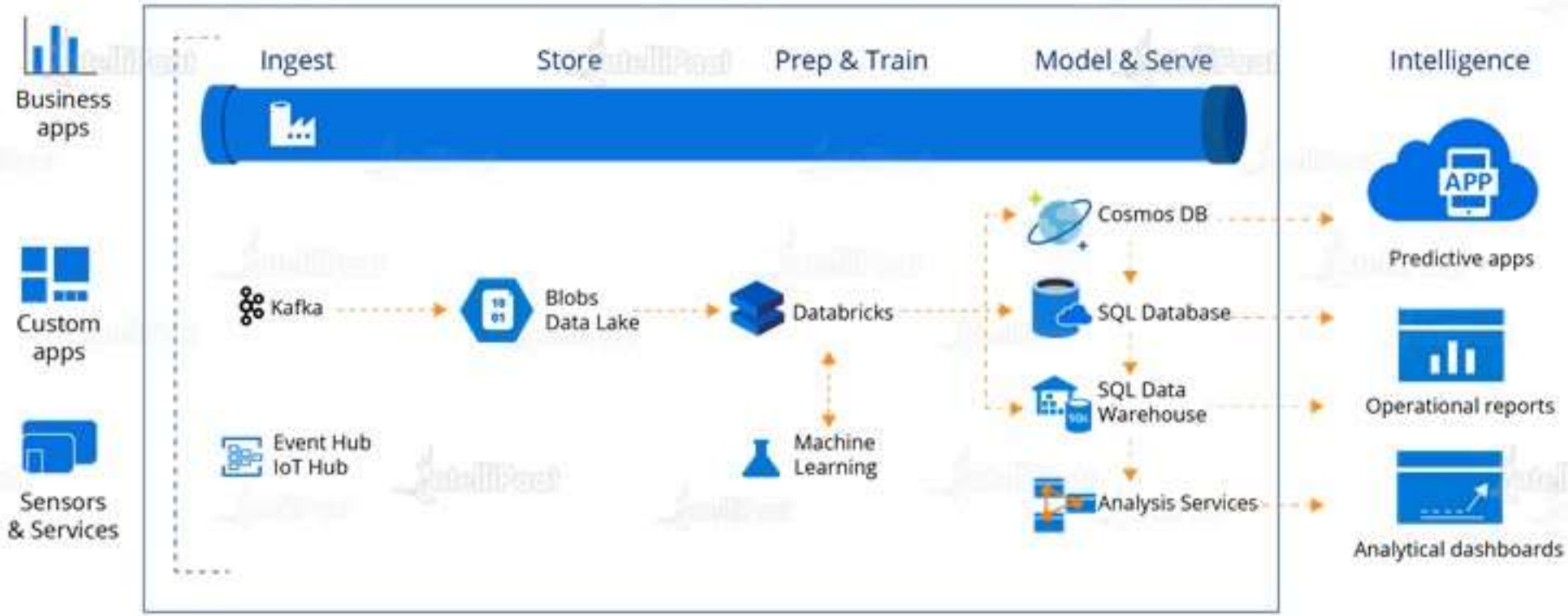
- An Apache Spark-based analytics platform optimized for the Microsoft Azure cloud services platform
- Designed by the founders of Apache Spark

Azure Databricks

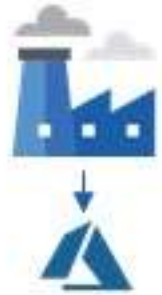


Databricks is integrated with Azure to provide a one-click setup, streamlined workflows, and an interactive workspace that enable collaboration between Data Scientists, Data Engineers, and Business Analysts

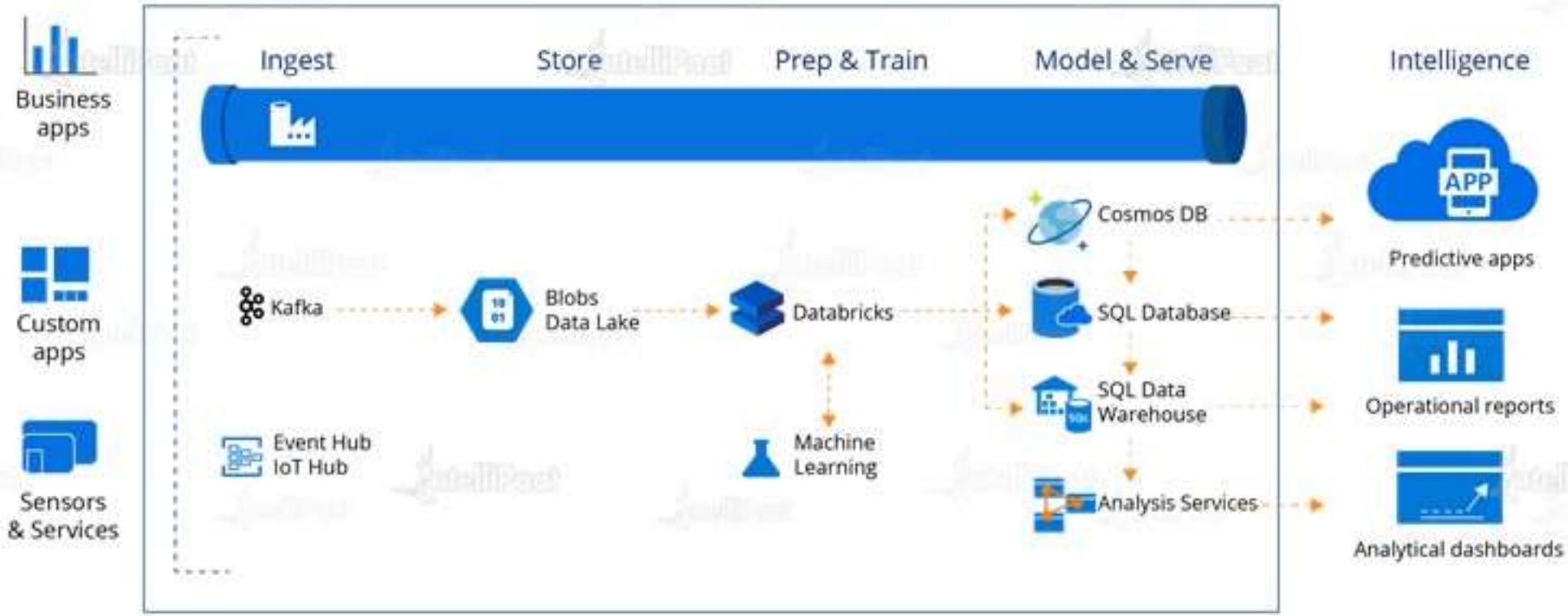
Azure Databricks



For a Big Data pipeline, the data (raw or structured) is ingested into Azure through Azure Data Factory in batches or streamed in near real-time using Kafka, Event Hub, or IoT Hub

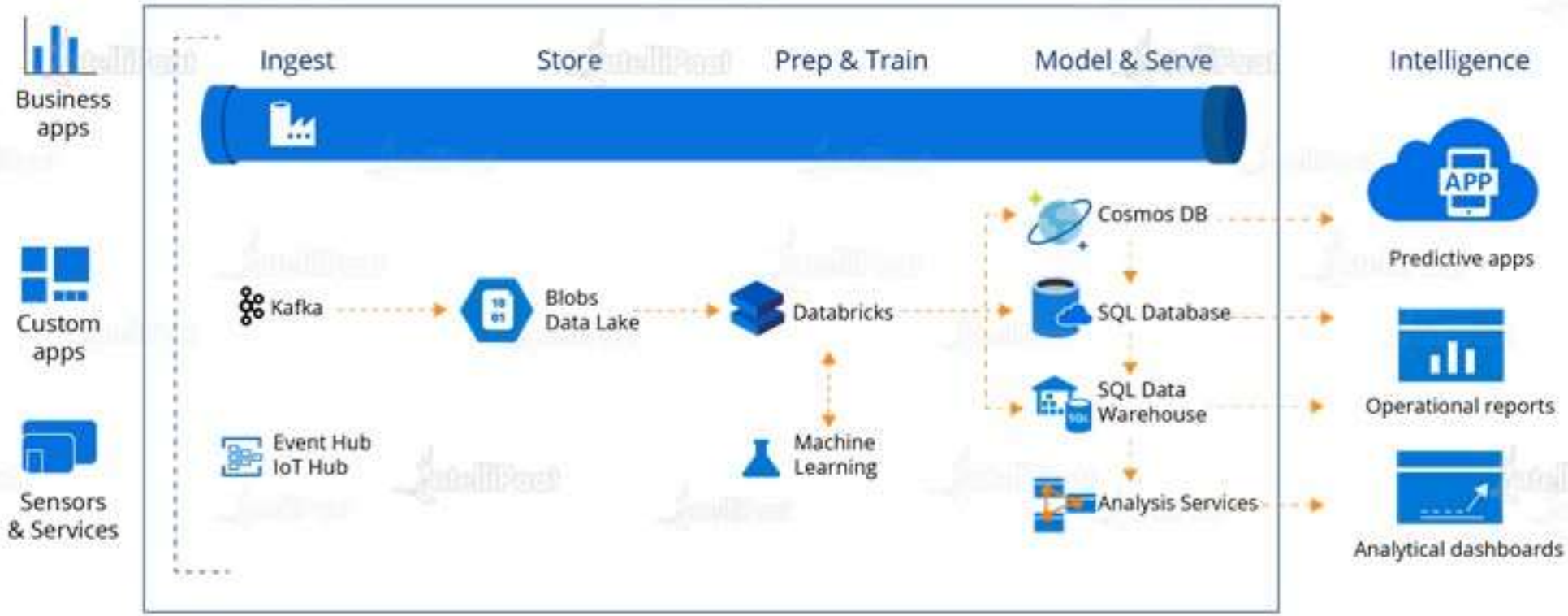


Azure Databricks



This data lands in a data lake for a long-term persisted storage, in Azure Blob Storage or Azure Data Lake Storage

Azure Databricks

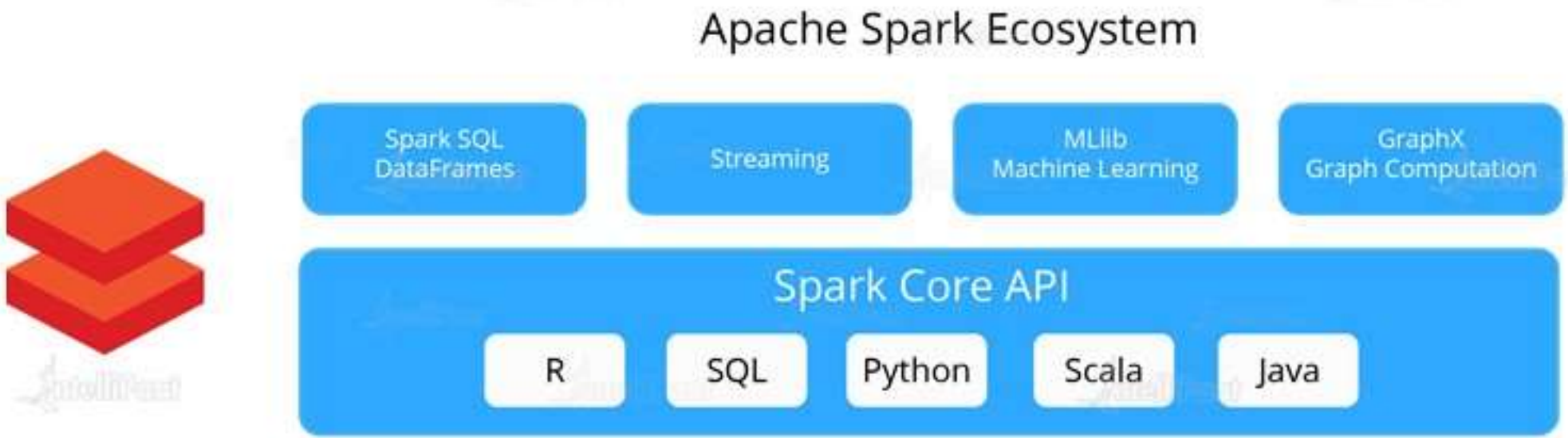


We use Azure Databricks as part of our analytics workflow to read data from multiple data sources, such as Azure Blob Storage, Azure Data Lake Storage, Azure Cosmos DB, or Azure SQL Data Warehouse, and turn it into breakthrough insights using Spark

Apache Spark-based Analytics Platform

Apache Spark-based Analytics Platform

Azure Databricks comprises the complete open-source Apache Spark cluster technologies and capabilities. Spark in Azure Databricks includes the following components:



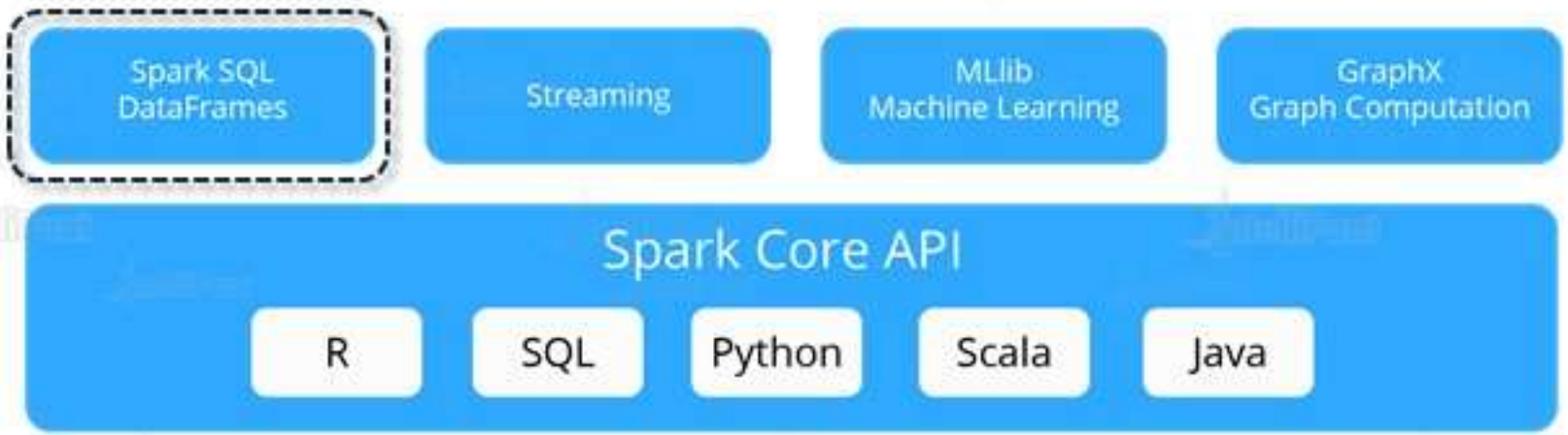
Apache Spark-based Analytics Platform

Spark SQL and DataFrames

Spark SQL

Spark SQL is the Spark module for working with the structured data. A DataFrame is a distributed collection of data organized into named columns. It is conceptually equivalent to a table in a relational database or a data frame in R/Python

Apache Spark Ecosystem



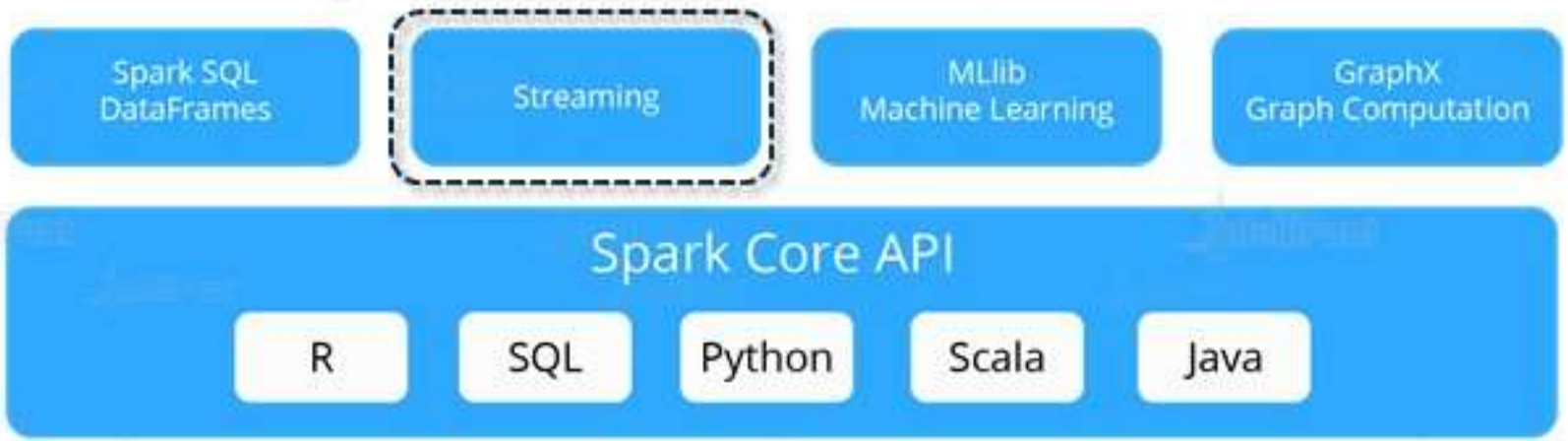
Apache Spark-based Analytics Platform

Streaming

kafka

Spark Streaming is the real-time data processing and analysis for analytical and interactive applications. It integrates with HDFS, Flume, and Kafka

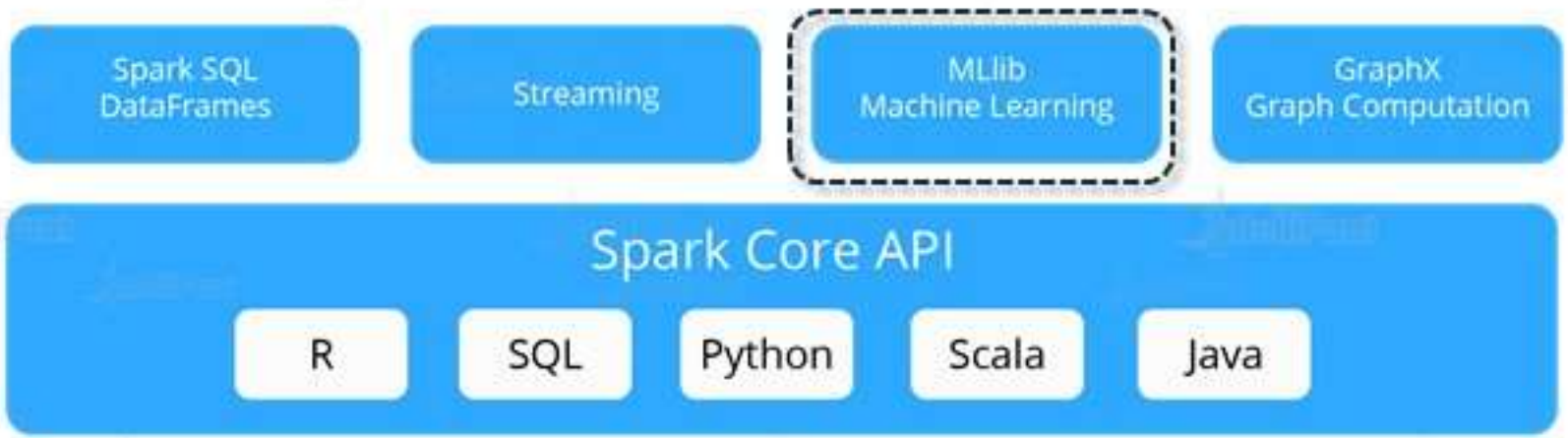
Apache Spark Ecosystem



Apache Spark-based Analytics Platform

MLlib is the Machine Learning library consisting of common learning algorithms and utilities, including classification, regression, clustering, collaborative filtering, dimensionality reduction, and underlying optimization primitives

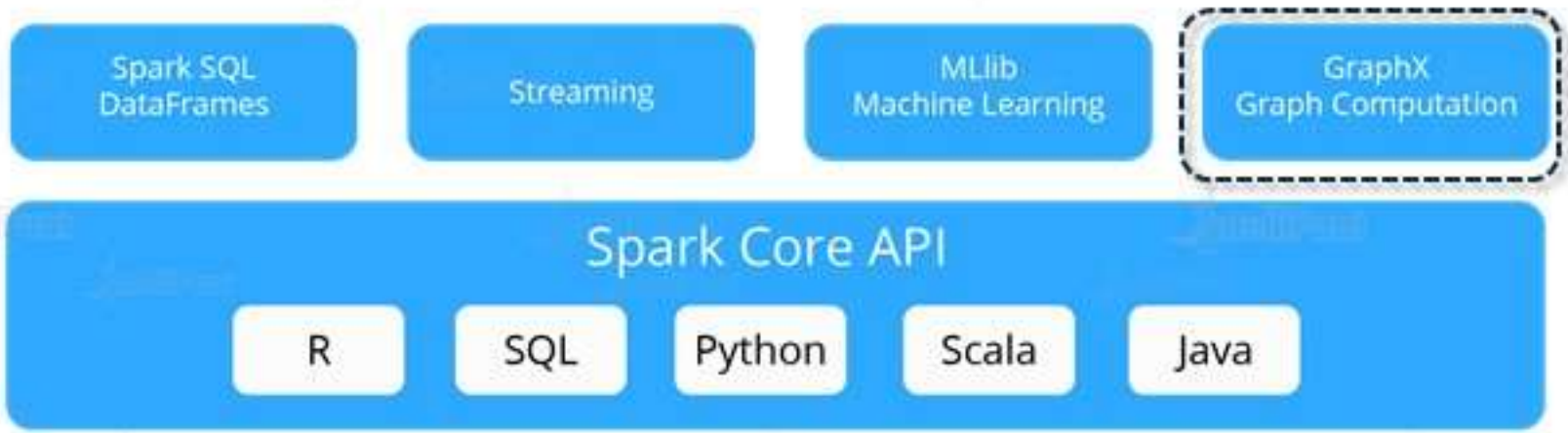
Apache Spark Ecosystem



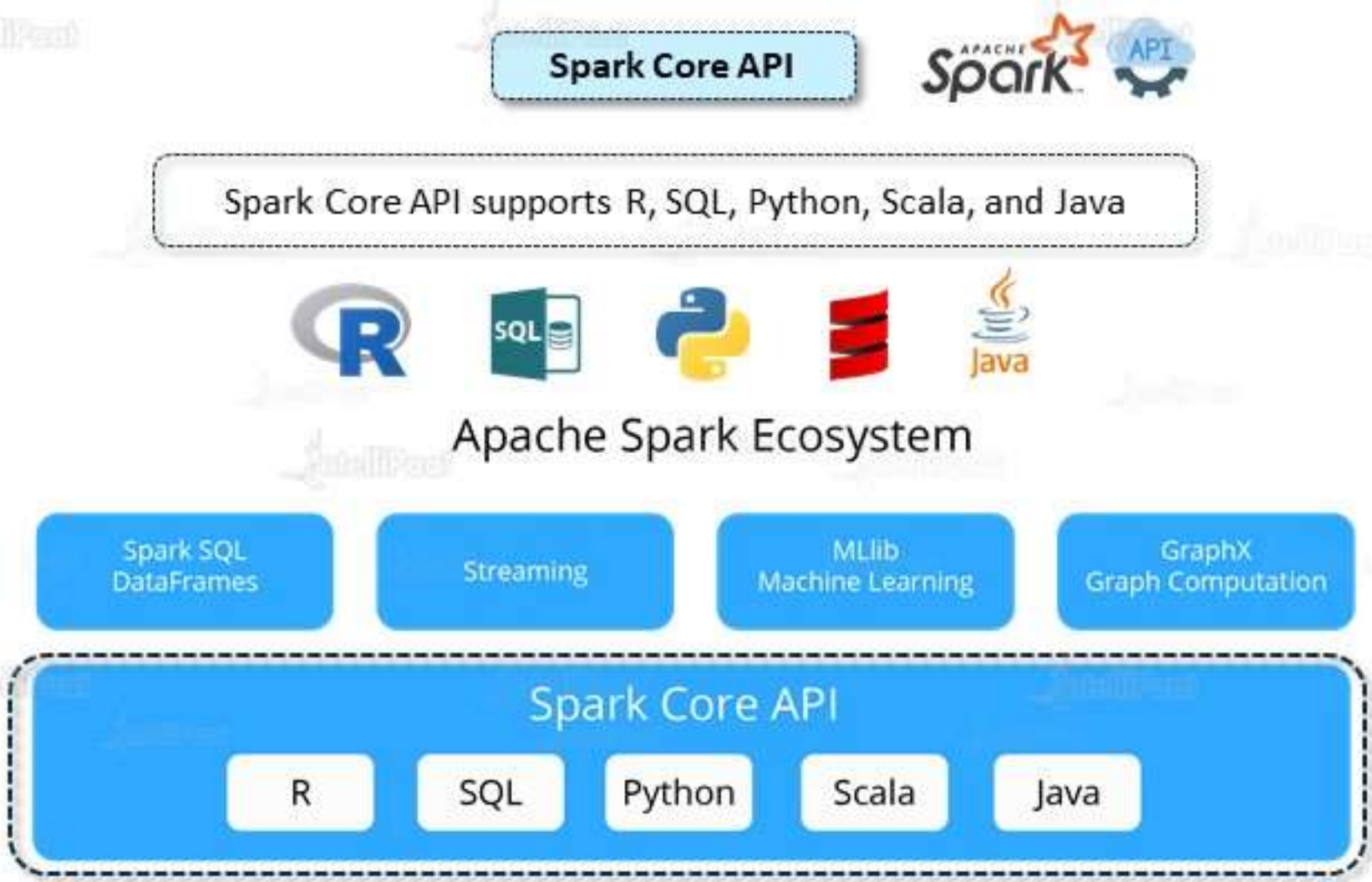
Apache Spark-based Analytics Platform

GraphX provides graphs and graph computation for a broad scope of use cases from cognitive analytics to data exploration

Apache Spark Ecosystem



Apache Spark-based Analytics Platform



Apache Spark in Azure Databricks

Apache Spark in Azure Databricks

Azure Databricks builds on the capabilities of Apache Spark by providing a zero-management cloud platform that includes:

Fully managed Spark clusters

An interactive workspace for exploration and visualization

A platform for powering our favorite Spark-based applications

Apache Spark in Azure Databricks

Fully Managed Apache Spark Clusters in the Cloud

Azure Databricks has a secure and reliable production environment in the cloud, managed and supported by Spark experts. We can:

01 Create clusters in seconds

02 Dynamically, auto-scale clusters up and down, including serverless clusters, and share them across teams

03 Use clusters programmatically by using the REST APIs

04 Use secure data integration capabilities built on Spark that help us unify data without centralization

05 Get instant access to the latest Apache Spark features with each release

Apache Spark in Azure Databricks

Databricks Runtime

The Databricks runtime is built on top of Apache Spark and is natively built for the Azure cloud

With the Serverless option, Azure Databricks completely abstracts infrastructure complexity and the need for specialized expertise to set up and configure a data infrastructure

The Serverless option helps Data Scientists iterate quickly as a team

Apache Spark in Azure Databricks

Databricks Runtime

For Data Engineers, who care about the performance of production jobs, Azure Databricks provides a Spark engine that is faster and performant due to its various optimizations at the I/O layer and the processing layer (Databricks I/O)

Hands-on: Running a Spark Job on Azure Databricks Using Azure Portal

Hands-on: The ETL Operation Using Azure Databricks

Hands-on: Streaming Data into Azure Databricks Using Event Hubs

Azure Stream Analytics

Azure Stream Analytics

It is a real-time analytics and complex event-processing engine that is designed to analyze and process high volumes of fast streaming data from multiple sources simultaneously

Patterns and relationships can be identified in the information extracted from a number of input sources including devices, sensors, clickstreams, social media feeds, and applications

These patterns can be used to trigger actions and initiate workflows such as creating alerts, feeding information to a reporting tool, or storing transformed data for later use

Azure Stream Analytics

Scenarios Using Azure Stream Analytics

1. Real-time analytics on Point-of-Sale data for inventory control and anomaly detection

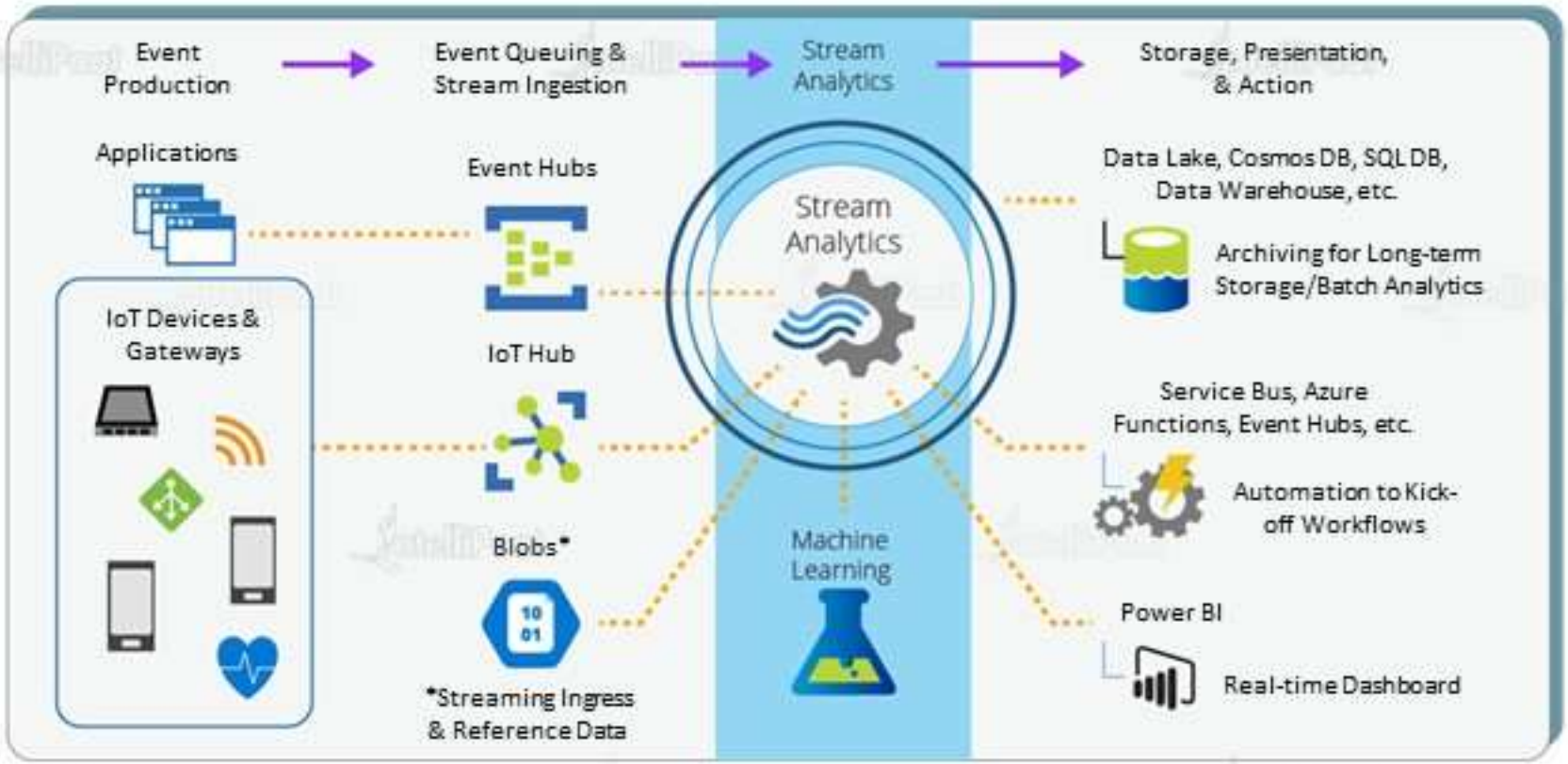
2. Analyzing real-time telemetry streams from IoT devices

3. Remote monitoring and predictive maintenance of high-value assets

4. Web log/clickstream analytics

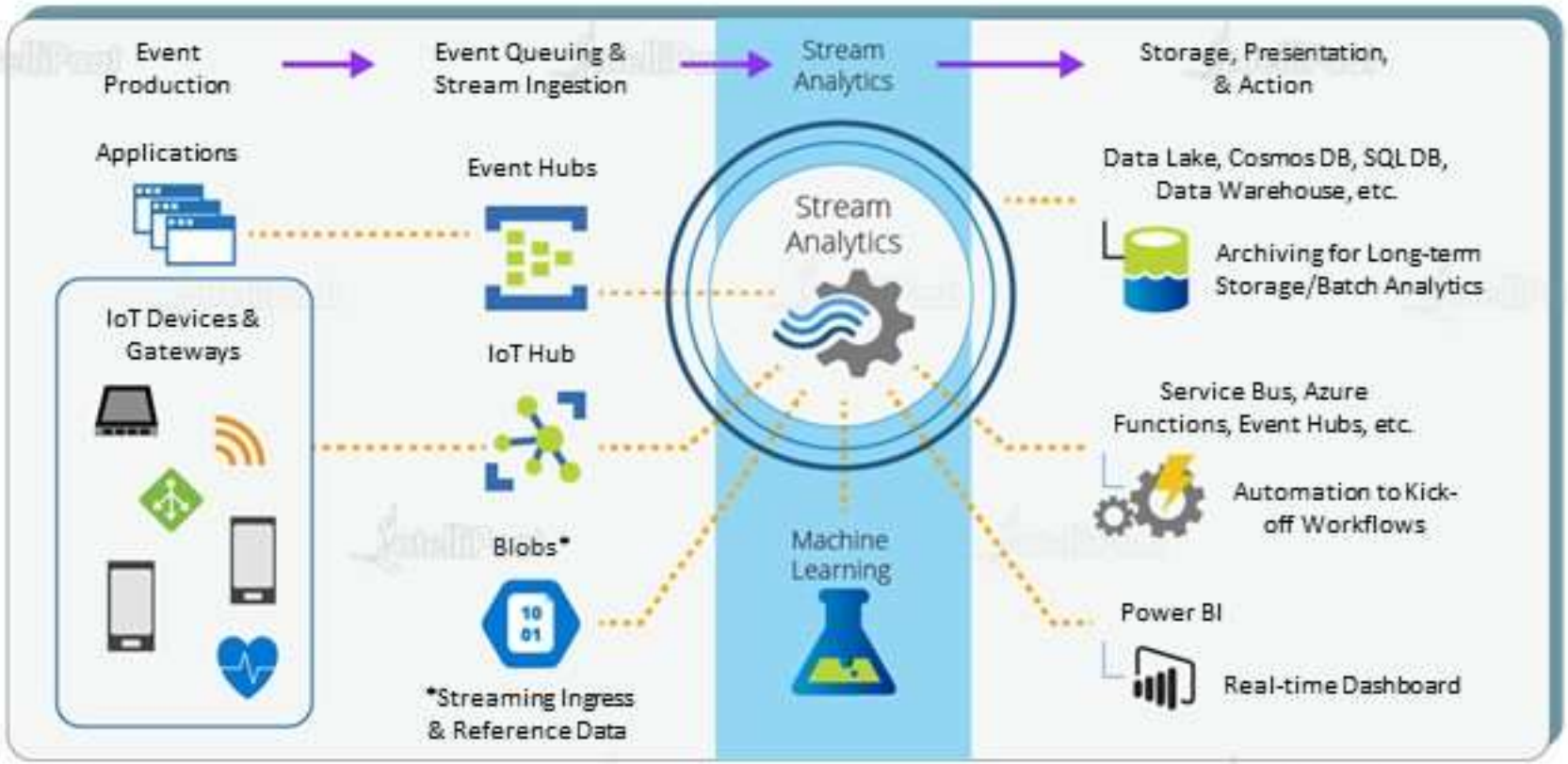
5. Geospatial analytics for fleet management and driverless vehicles

Working of Stream Analytics



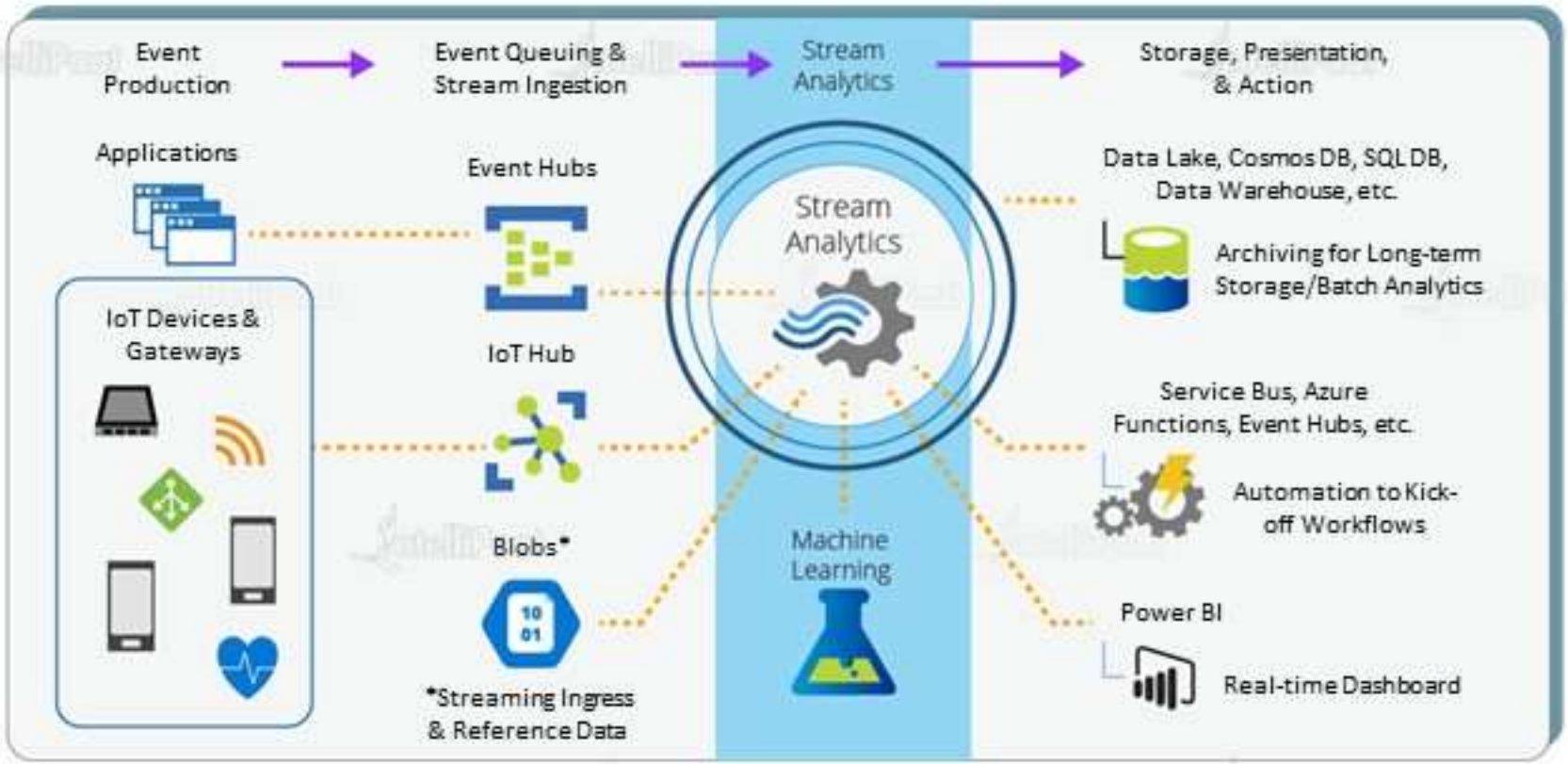
Stream Analytics ingests data from Azure Event Hubs, Azure IoT Hub, or Azure Blob Storage

Working of Stream Analytics



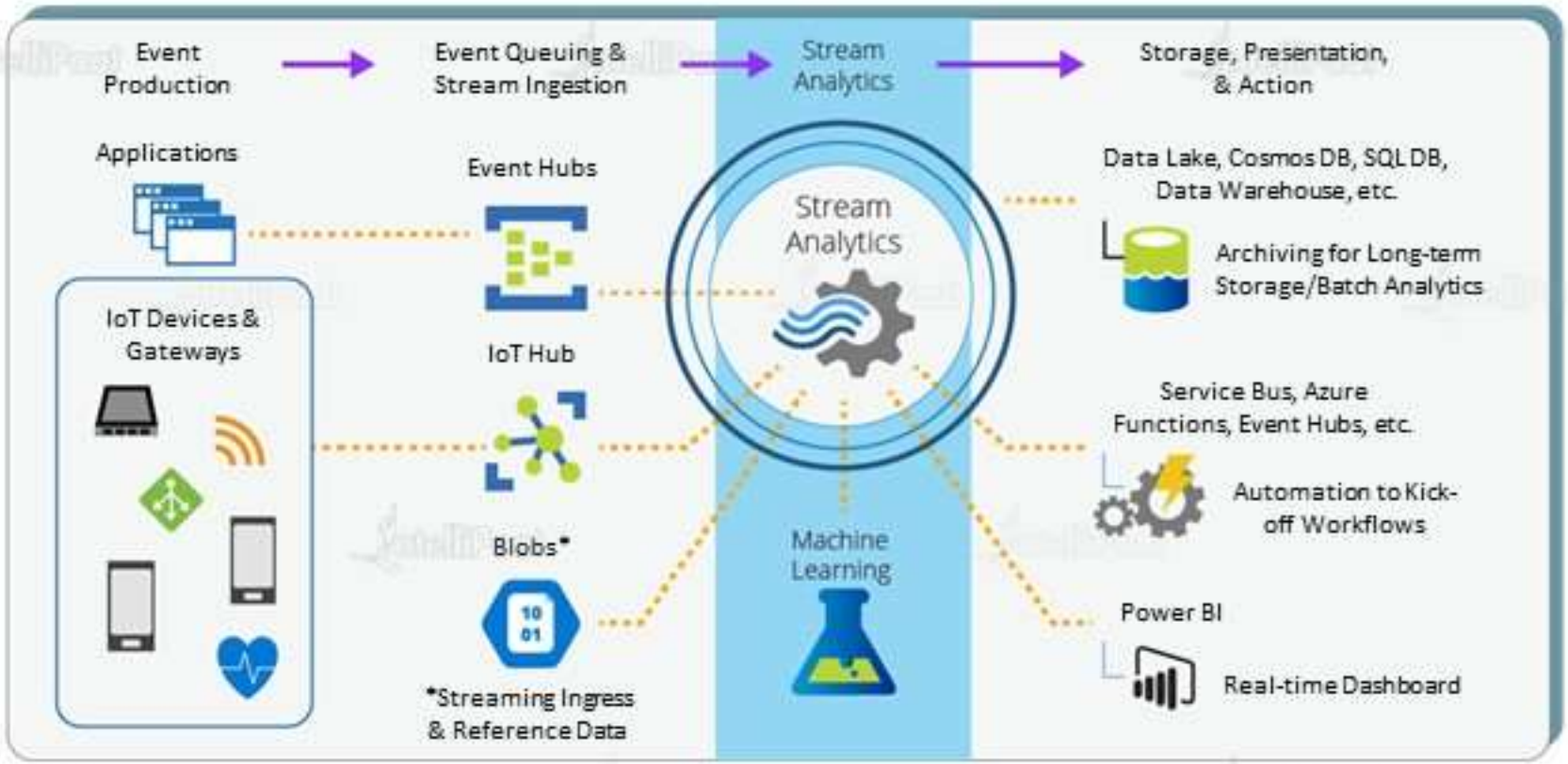
Each job has an output for the transformed data, and we can control what happens in response to the information we have analyzed

Working of Stream Analytics



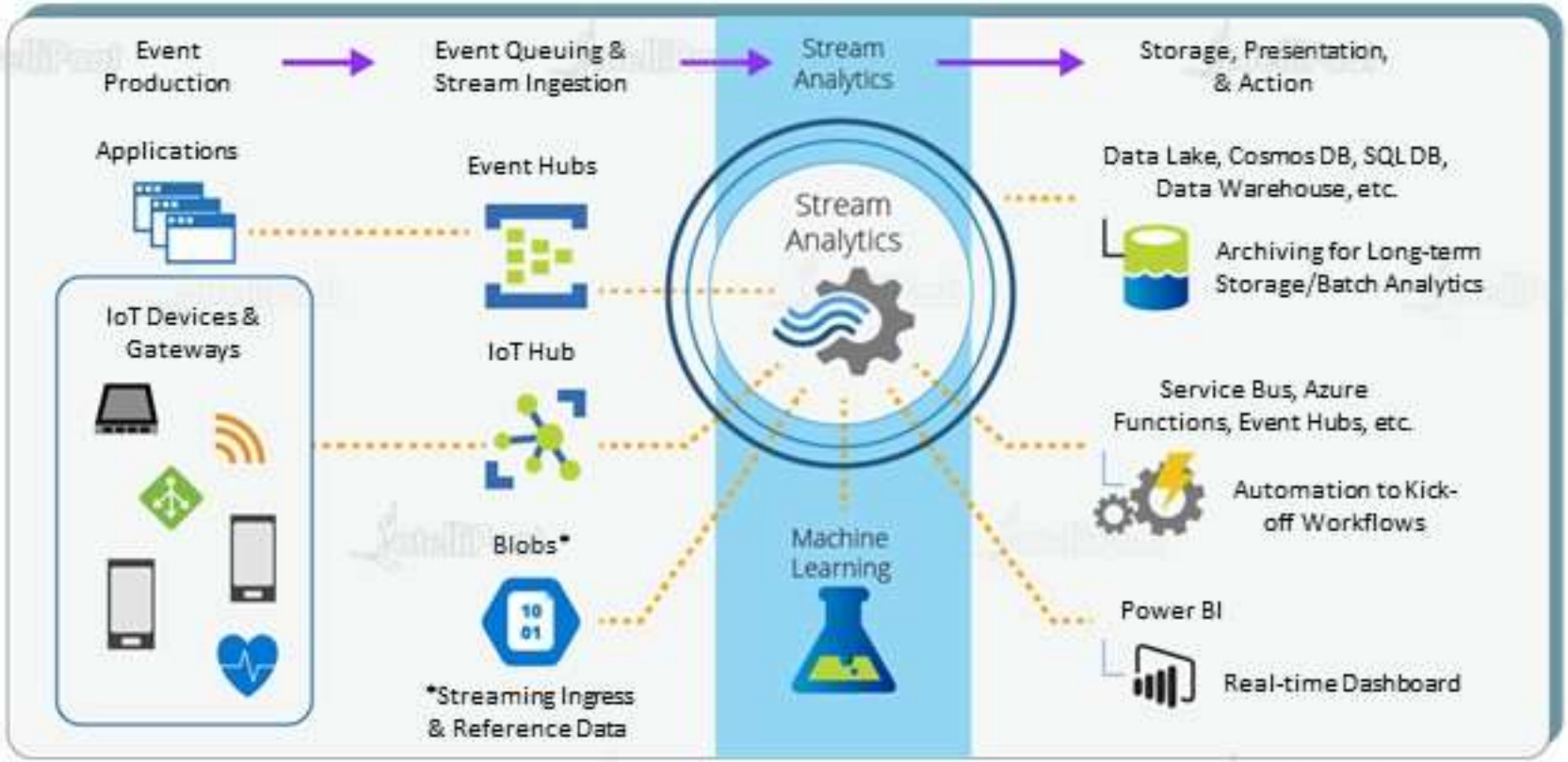
We can send data to services such as Azure Functions, Service Bus Topics, or Queues to trigger communications or custom workflows downstream

Working of Stream Analytics



We can send data to a Power BI dashboard for real-time dashboarding

Working of Stream Analytics



We can store data in other Azure Storage services to train a Machine Learning model based on historical data or perform batch analytics

Key Capabilities and Benefits

Azure Stream Analytics is designed to be easy-to-use, flexible, reliable, and scalable to any job size. It is available across multiple Azure regions

Hands-on: Analyzing Phone Call Data with Stream Analytics & Visualizing Results in Power BI Dashboard

Stream Analytics Windowing Functions

Stream Analytics Windowing Functions

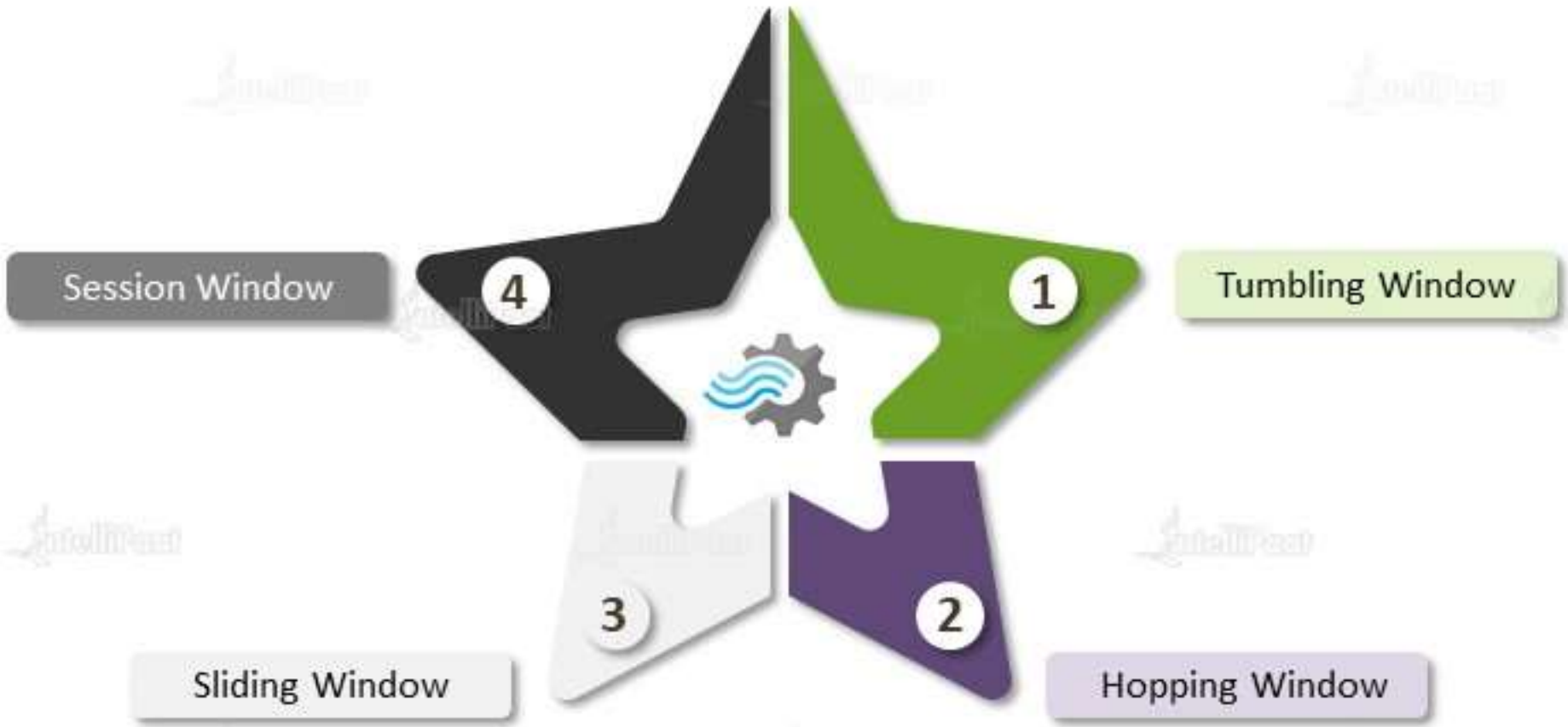
Support for Windowing Functions

Stream Analytics has native support for windowing functions, enabling developers to author complex stream processing jobs with minimal effort

Performing operations on the data contained in temporal windows is a common pattern in time-streaming scenarios

Stream Analytics Windowing Functions

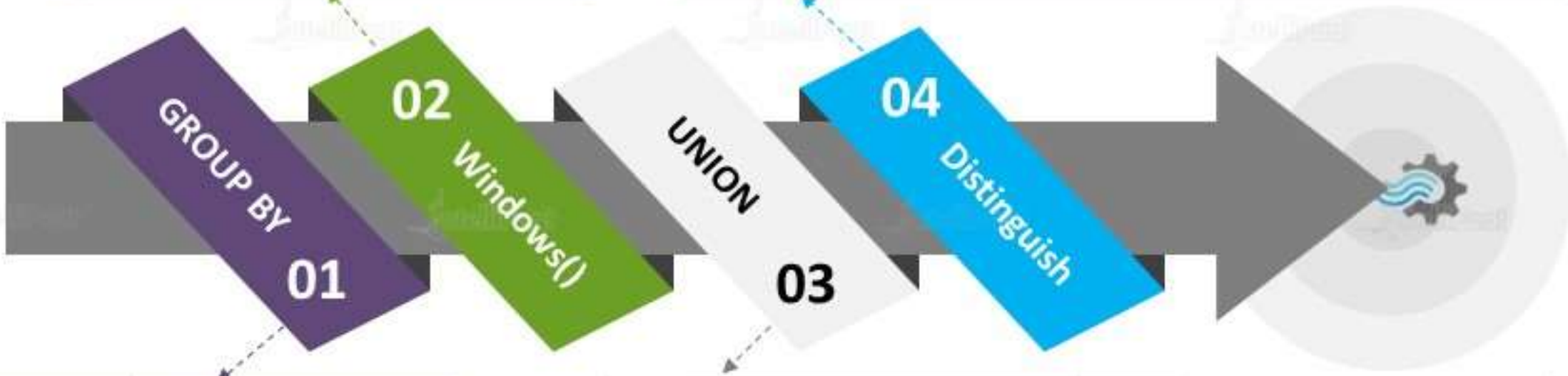
There are four kinds of temporal windows to choose from:



Stream Analytics Windowing Functions

We can aggregate events over **multiple windows** using the **windows()** function

To **distinguish** between different results, we can assign an identity to each window that can be accessed using the system function **System.Window().Id**. It returns a record with the ID as its field



We use the windows() function in the **GROUP BY** clause of the query syntax in our Stream Analytics jobs

Query logic is computed for each of these window definitions, and the result is a **UNION** of all window results

Stream Analytics Windowing Functions

There are two ways to define Windows:

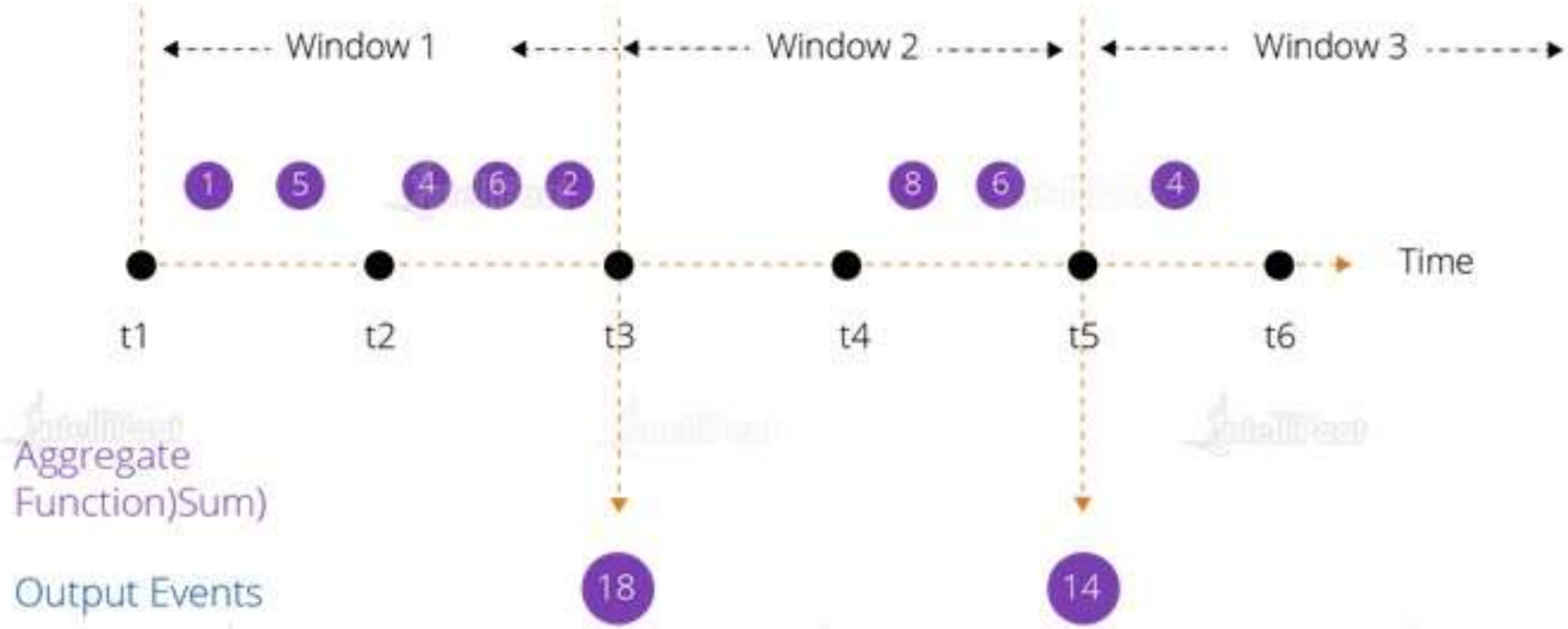
Assign unique identities using the **windows()** function, **Window (ID, window_definition)**, where ID is an identity of the **window_definition** and is a unique **varchar(max)** value within the Windows construct

1. Window (ID , window_definition)
2. System.Window().Id

Without identities, in which case **System.Window().Id** results in a null value

Stream Analytics Windowing Functions

- All windowing operations output results at the end of the window
- The output of the window will be a single event based on the aggregate function used
- The output event will have the time stamp of the end of the window, and all window functions are defined with a fixed length

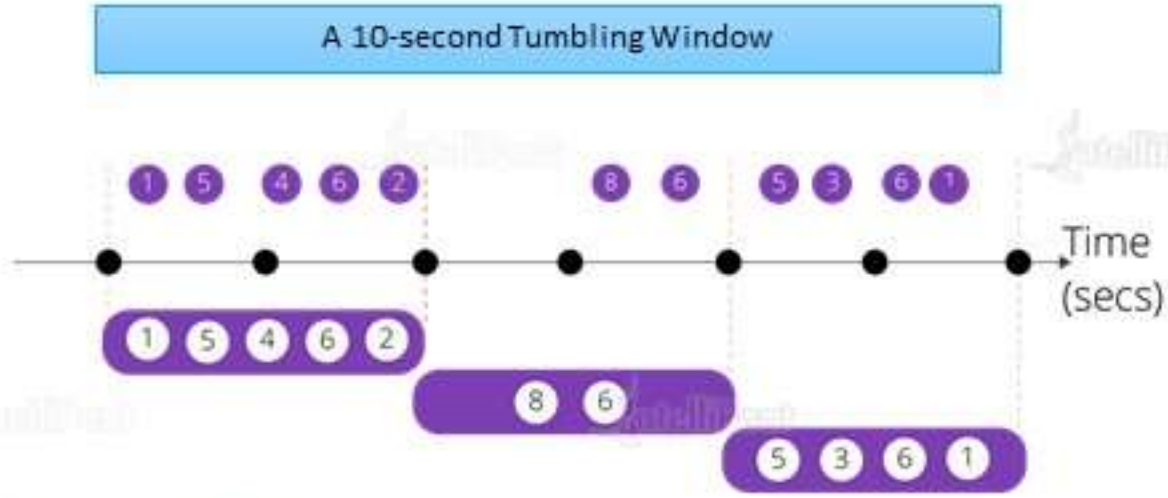


Stream Analytics Windowing Functions

- Tumbling Window
- Hopping Window
- Sliding Window
- Session Window

- These functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below
- Key differentiators of a tumbling window are that they repeat; they do not overlap, and an event cannot belong to more than one tumbling window

Tell me the count of tweets per time zone every 10 seconds




```
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, ThumblingWindow(second, 10)
```

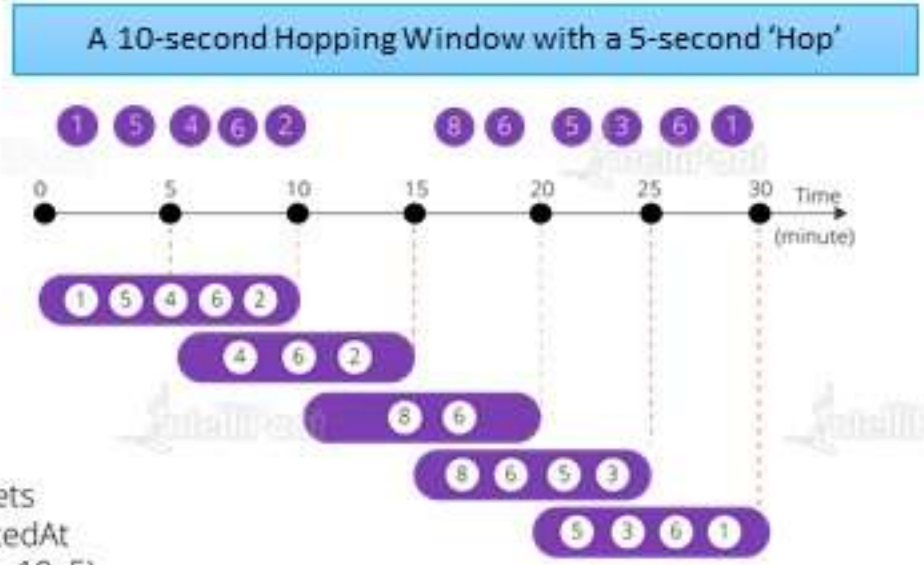
Stream Analytics Windowing Functions

IntelliPaat

- Tumbling Window
- Hopping Window
- Sliding Window
- Session Window

- These functions hop forward in time by a fixed period. It may be easy to think of them as the tumbling windows that can overlap, so events can belong to more than one hopping window result set
- To make a hopping window the same as a tumbling window, specify the hop size to be the same as the window size

Every 5 seconds give me the count of tweets over the last 10 seconds



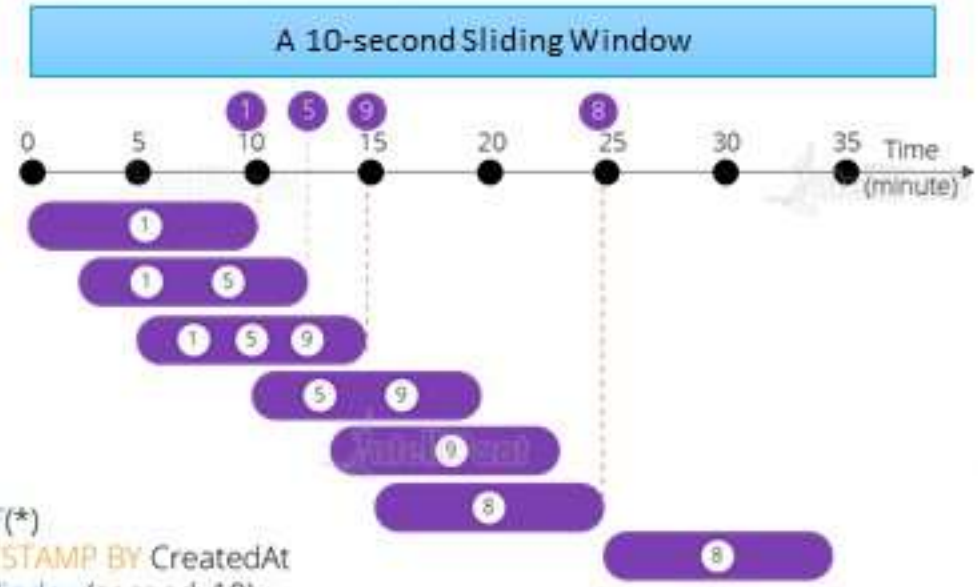

```
SELECT TimeZone, COUNT(*) AS TotalTweets
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY Topic, HoppingWindow(second, 10, 5)
```

Stream Analytics Windowing Functions

- Tumbling Window
- Hopping Window
- **Sliding Window**
- Session Window

- Sliding window functions, unlike tumbling or hopping windows, produce an output only when an event occurs
- Every window has at least one event, and the window continuously moves forward by an € (epsilon)
- Like hopping windows, events can belong to more than one sliding window

Give me the count of tweets for single topic in the last 10 seconds



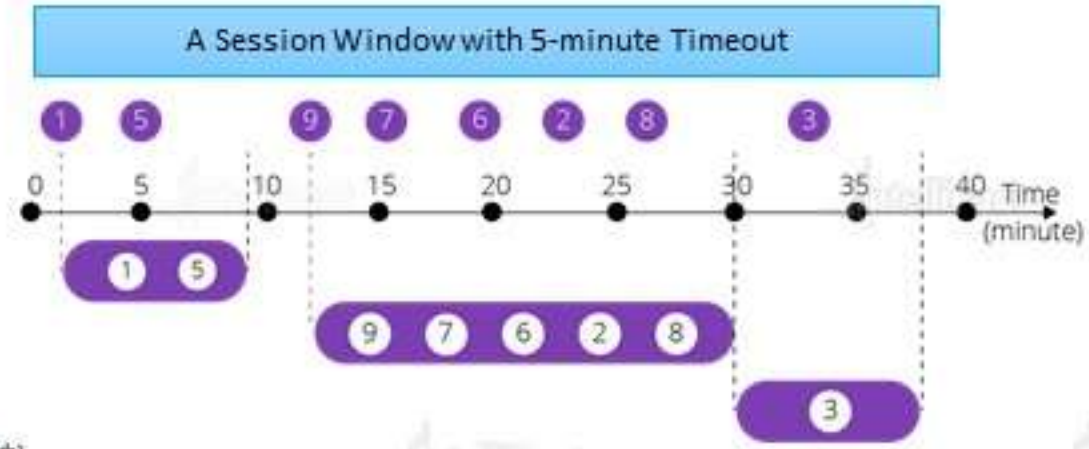

```
SELECT TimeZone, COUNT(*)
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY Topic, SlidingWindow(second, 10)
```

Stream Analytics Windowing Functions

- Tumbling Window
- Hopping Window
- Sliding Window
- **Session Window**

- Session window functions group events that arrive at the same time, filtering periods of time where there is no data
- It has three main parameters: timeout, maximum duration, and partitioning key (optional)

Tell me the count of tweets that occur within 5 minutes to each other




```
SELECT Topic, COUNT(*)
FROM TwitterStream TIMESTAMP BY CreateAt
GROUP BY Topic, SessionWindow(Minute, 5, 10)
```

India: +91-7847955955

US: 1-800-216-8930 (TOLL FREE)

sales@intellipaat.com

24/7 Chat with Our Course Advisor