Oliver Engels | @oengels | https://www.linkedin.com/in/oengels/
Gabi Münster | @SQLMissSunshine | https://www.linkedin.com/in/gabimuenster

From Data Lake to DAX query

DATAmasterminds

POWER BI SENTINEL

Documentation · Data Lineage · Change Tracking · Backup

ROSEN

empowered by technology

SentryOne

HEDDA.IO

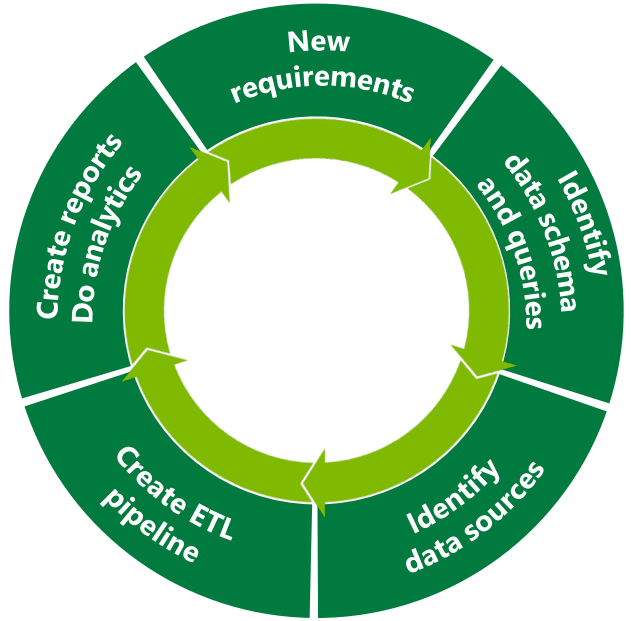
redgate

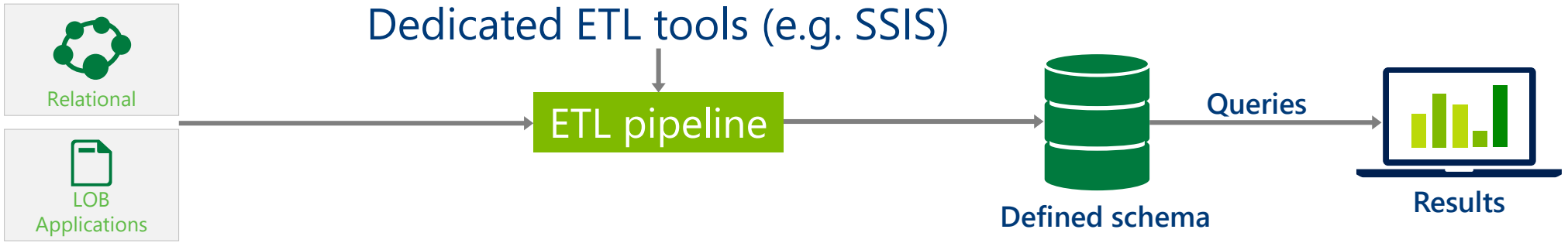
SHOP APOTHEKE
EUROPE

dbWatch
DATABASE CONTROL

Traditional business analytics process

1. Start with end-user requirements to identify desired reports and analysis
2. Define corresponding database schema and queries
3. Identify the required data sources
4. Create a Extract-Transform-Load (ETL) pipeline to extract required data (curation) and transform it to target schema (‘*schema-on-write*’)
5. Create reports. Analyze data

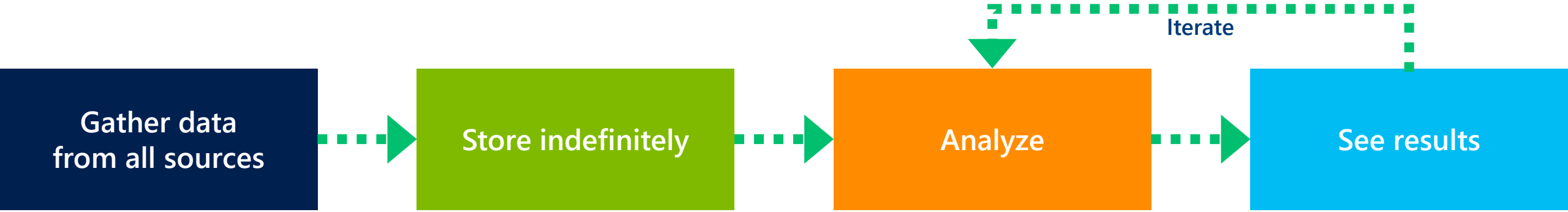




All data not immediately required is discarded or archived

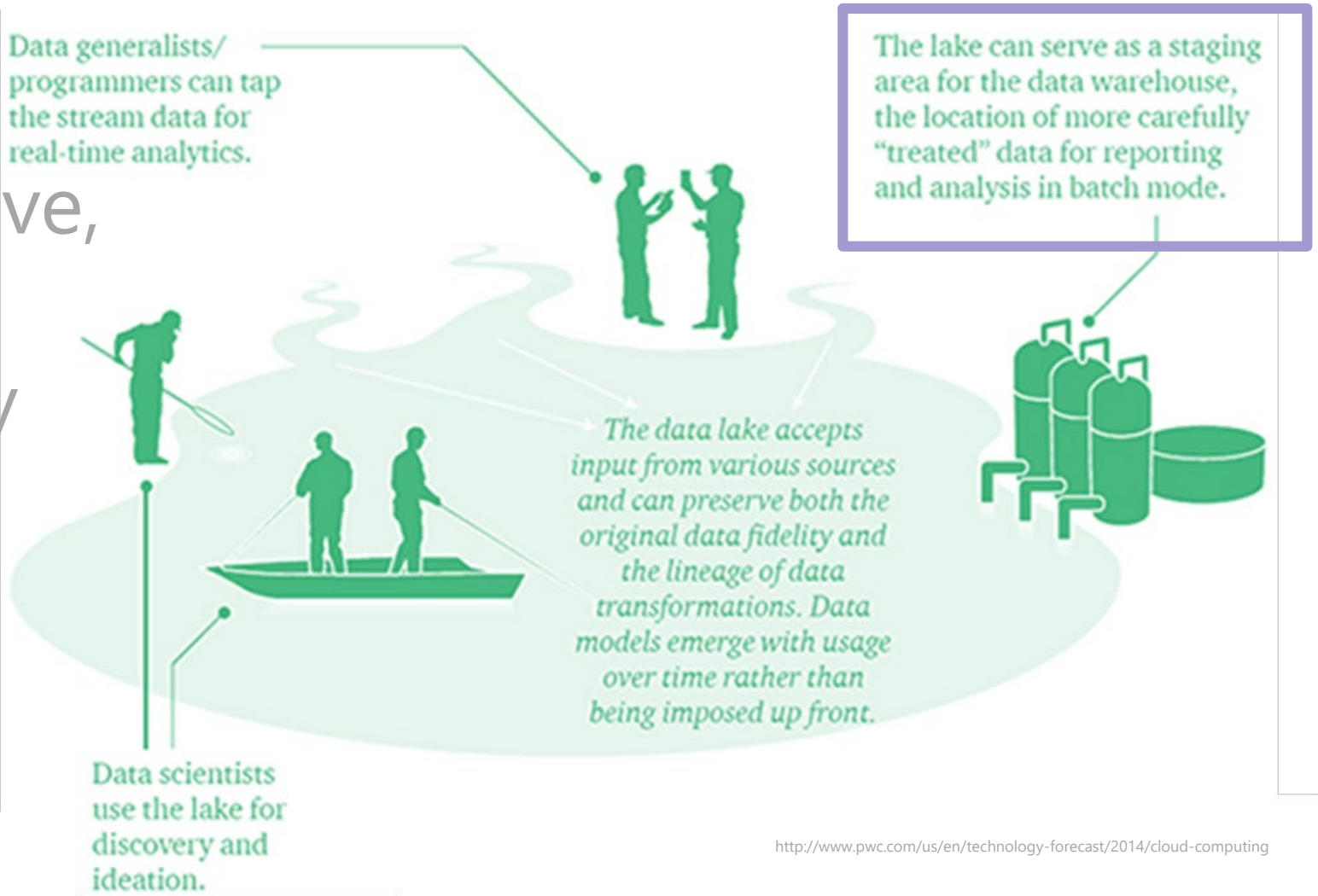
New big data thinking: All data has value

- All data has potential value
- Data hoarding
- No defined schema—stored in native format
- Schema is imposed and transformations are done at query time *(schema-on-read)*.
- Apps and users interpret the data as they see fit



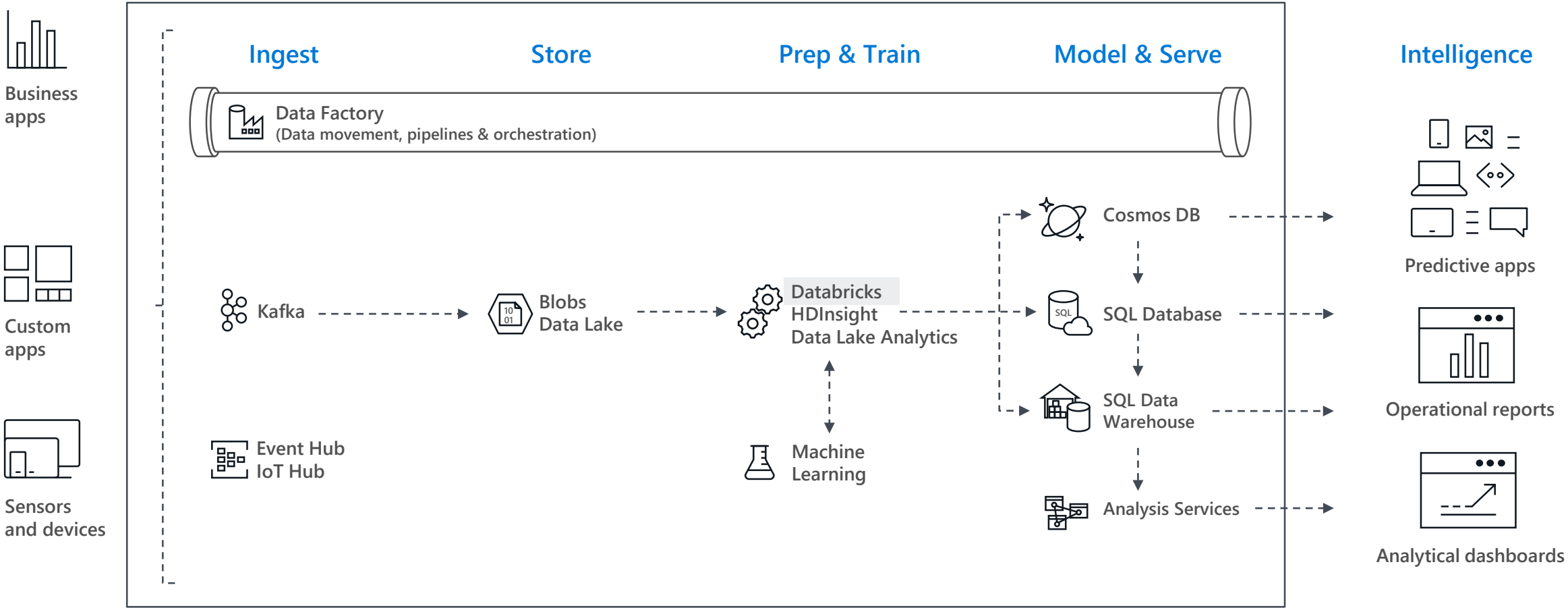
Why data lakes?

A data lake is a massive, easily accessible, centralized repository of large volumes of structured and unstructured data.



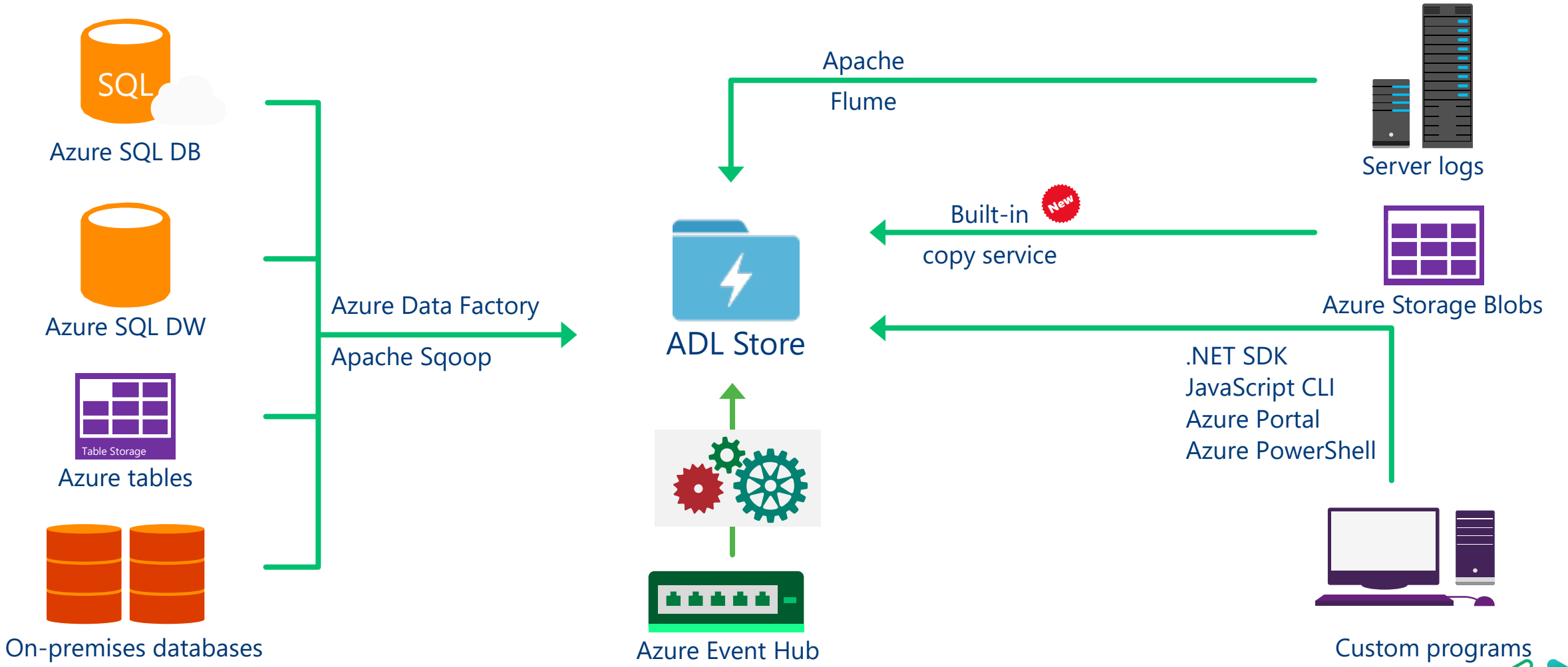
http://www.pwc.com/us/en/technology-forecast/2014/cloud-computing

BIG DATA & ADVANCED ANALYTICS AT A GLANCE



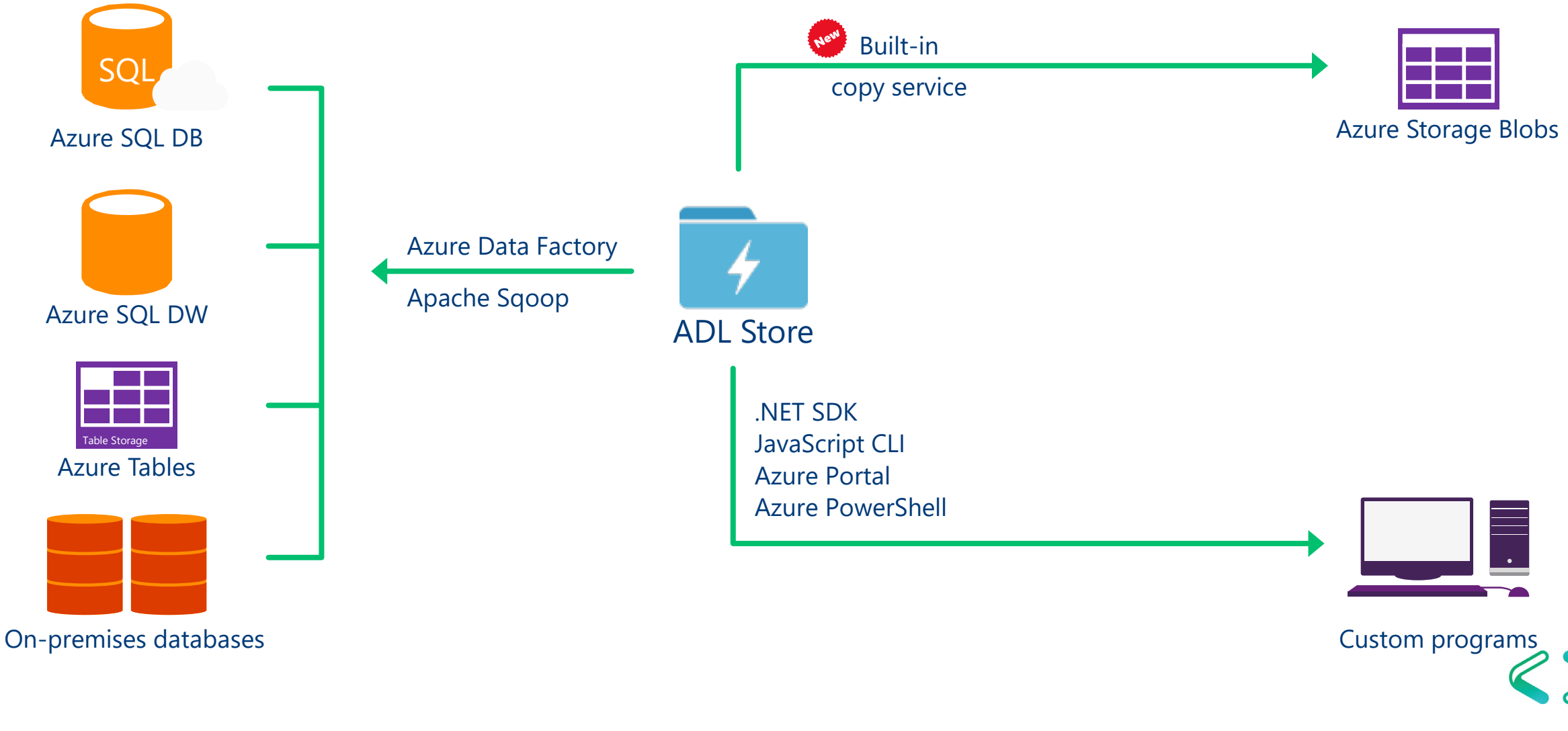
ADL Store: Ingress

Data can be ingested into Azure Data Lake Store from a variety of sources



ADL Store: Egress

Data can be exported from Azure Data Lake Store into numerous targets/sinks



Azure Data Factory & Databricks

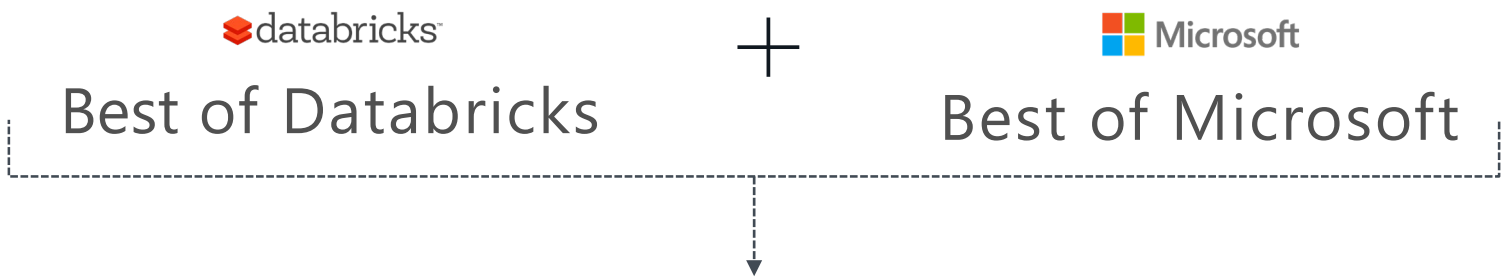
Ingest, prepare, and transform using Azure Databricks and Data Factory

Databricks Notebook Activity in Azure Data Factory

Run any notebook on an existing Databricks Cluster

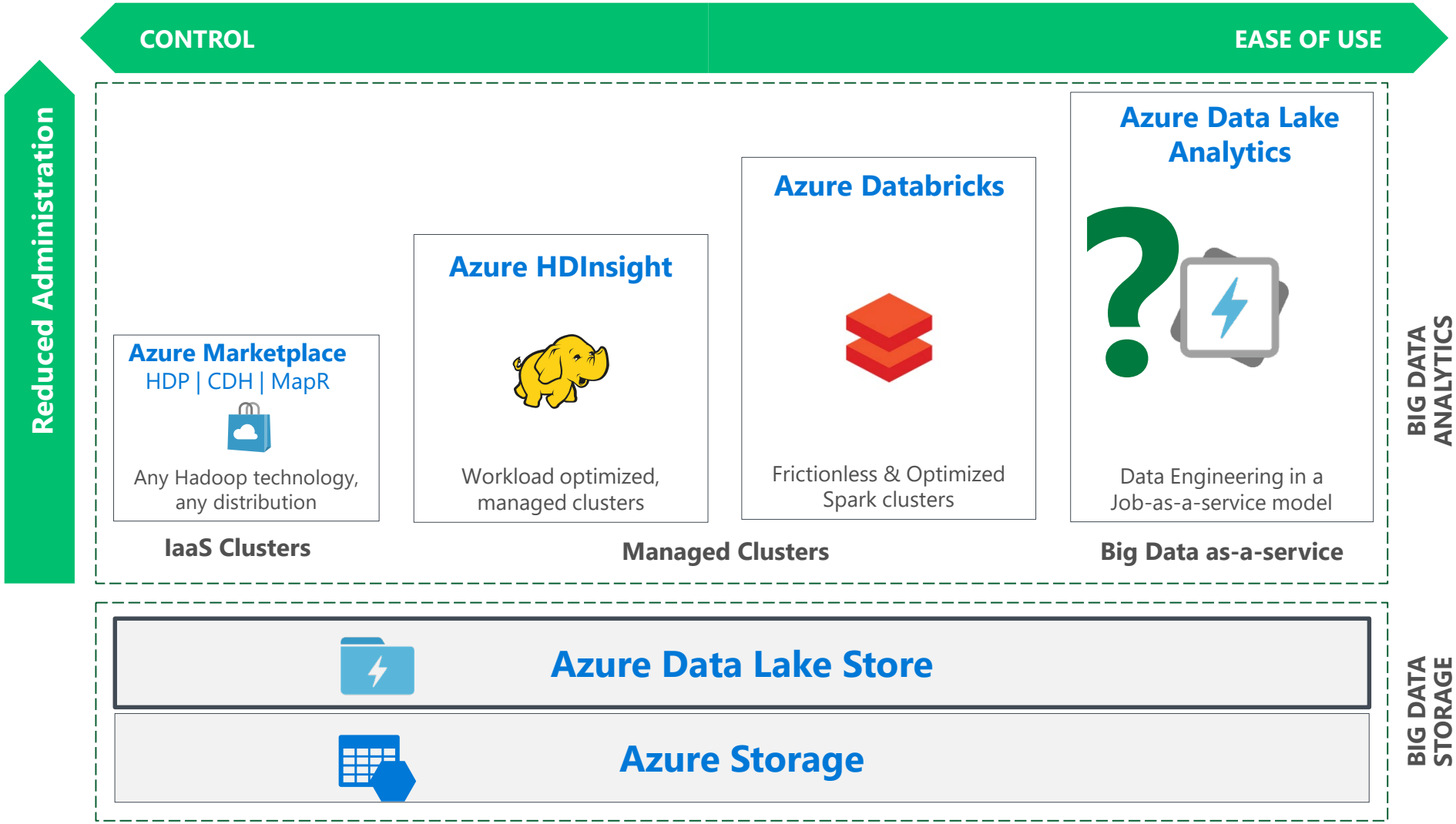
What is Azure Databricks?

A fast, easy and collaborative Apache® Spark™ based analytics platform optimized for Azure

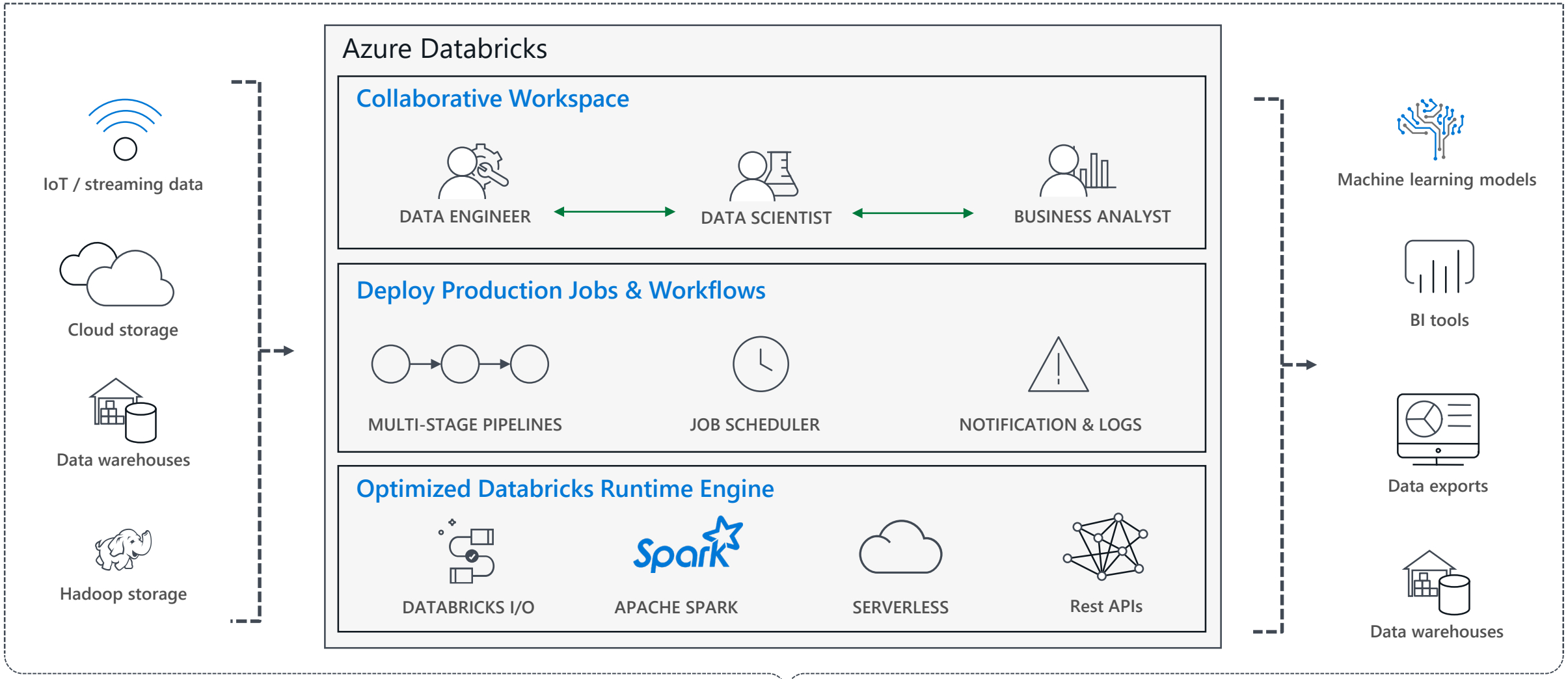


- Designed in collaboration with the founders of Apache Spark
- One-click set up; streamlined workflows
- Interactive workspace that enables collaboration between data scientists, data engineers, and business analysts.
- Native integration with Azure services (Power BI, SQL DW, Cosmos DB, Blob Storage)
- Enterprise-grade Azure security (Active Directory integration, compliance, enterprise-grade SLAs)

KNOWING THE VARIOUS BIG DATA SOLUTIONS



Azure Databricks



Enhance Productivity

Build on secure & trusted cloud

Scale without limits

Collaborative Workspace

GET STARTED IN SECONDS

Single click to launch your new Spark environment

INTERACTIVE EXPLORATION

Explore data using interactive notebooks with support for multiple programming languages including R, Python, Scala, and SQL

COLLABORATION

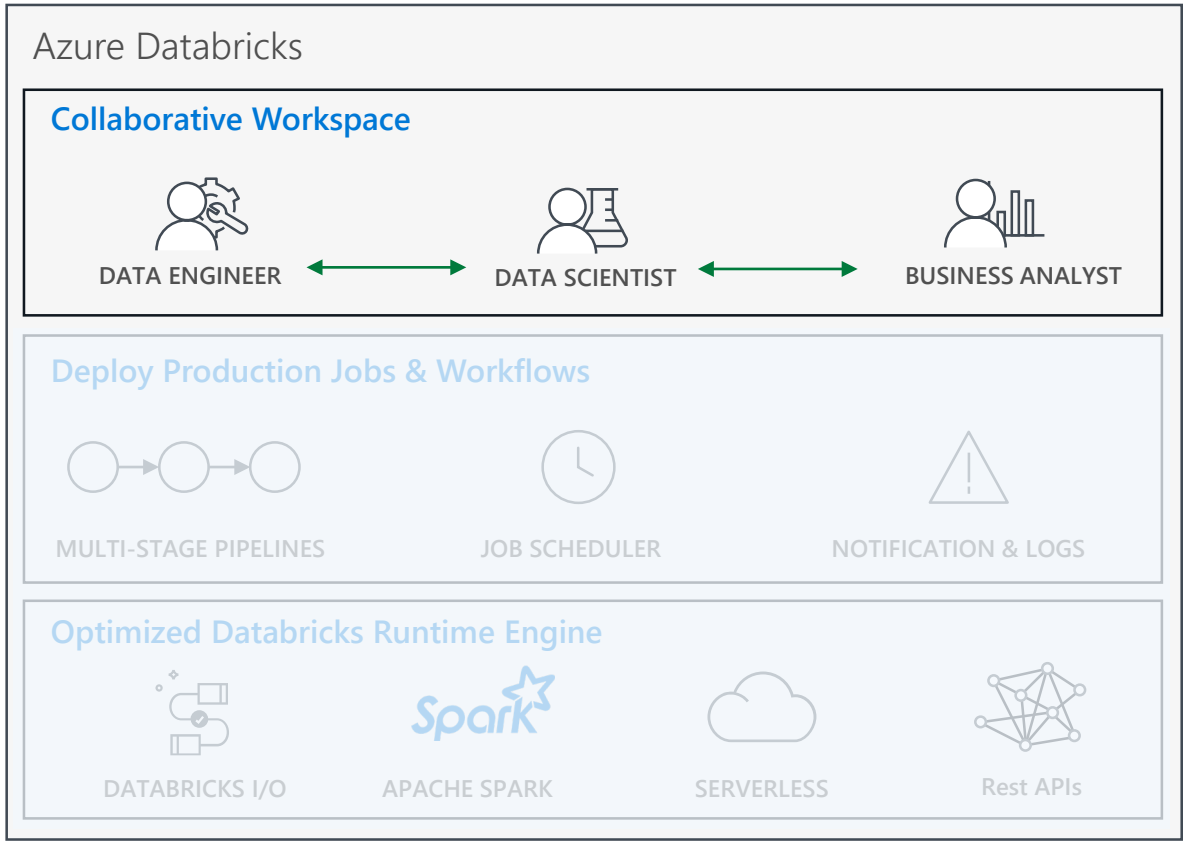
Work on the same notebook in real-time while tracking changes with detailed revision history, GitHub, or Bitbucket

VISUALIZATIONS

Visualize insights through a wide assortment of point-and-click visualizations. Or use powerful scriptable options like matplotlib, ggplot, and D3

DASHBOARDS

Rich integration with PowerBI to discover and share your insights in powerful new ways



Deploy Production Jobs & Workflows

JOBS SCHEDULER

Execute jobs for production pipelines on a specific schedule

NOTEBOOK WORKFLOWS

Create multi-stage pipelines with the control structures of the source programming language

RUN NOTEBOOKS AS JOBS

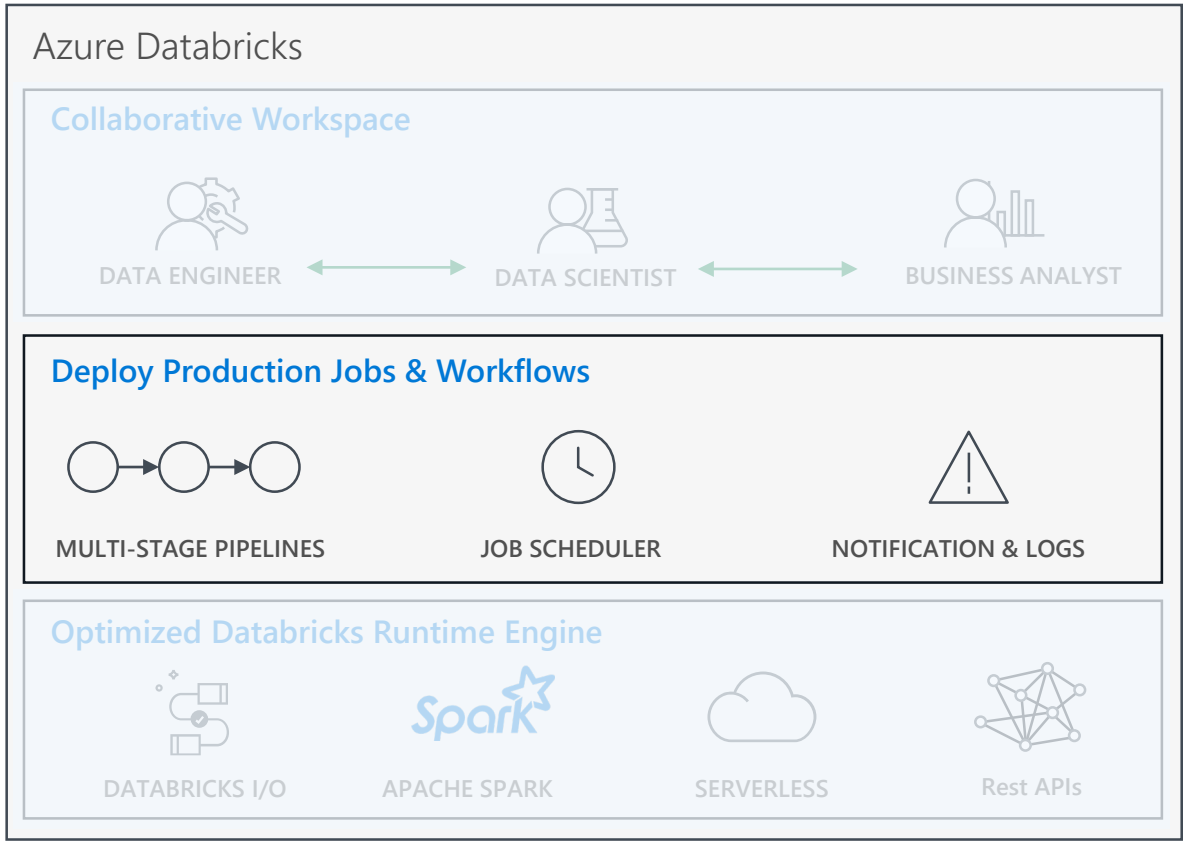
Turn notebooks or JARs into resilient Spark jobs with a click or an API call

NOTIFICATIONS AND LOGS

Set up alerts and quickly access audit logs for easy monitoring and troubleshooting

INTEGRATE NATIVELY WITH AZURE SERVICES

Deep integration with Azure SQL Data Warehouse, Cosmos DB, Azure Data Lake Store, Azure Blob Storage, and Azure Event Hub



Optimized Databricks Runtime Engine

OPTIMIZED I/O PERFORMANCE

The Databricks I/O module (DBIO) takes processing speeds to the next level — significantly improving the performance of Spark in the cloud

FULLY-MANAGED PLATFORM ON AZURE

Reap the benefits of a fully managed service and remove the complexity of big data and machine learning

SERVERLESS INFRASTRUCTURE

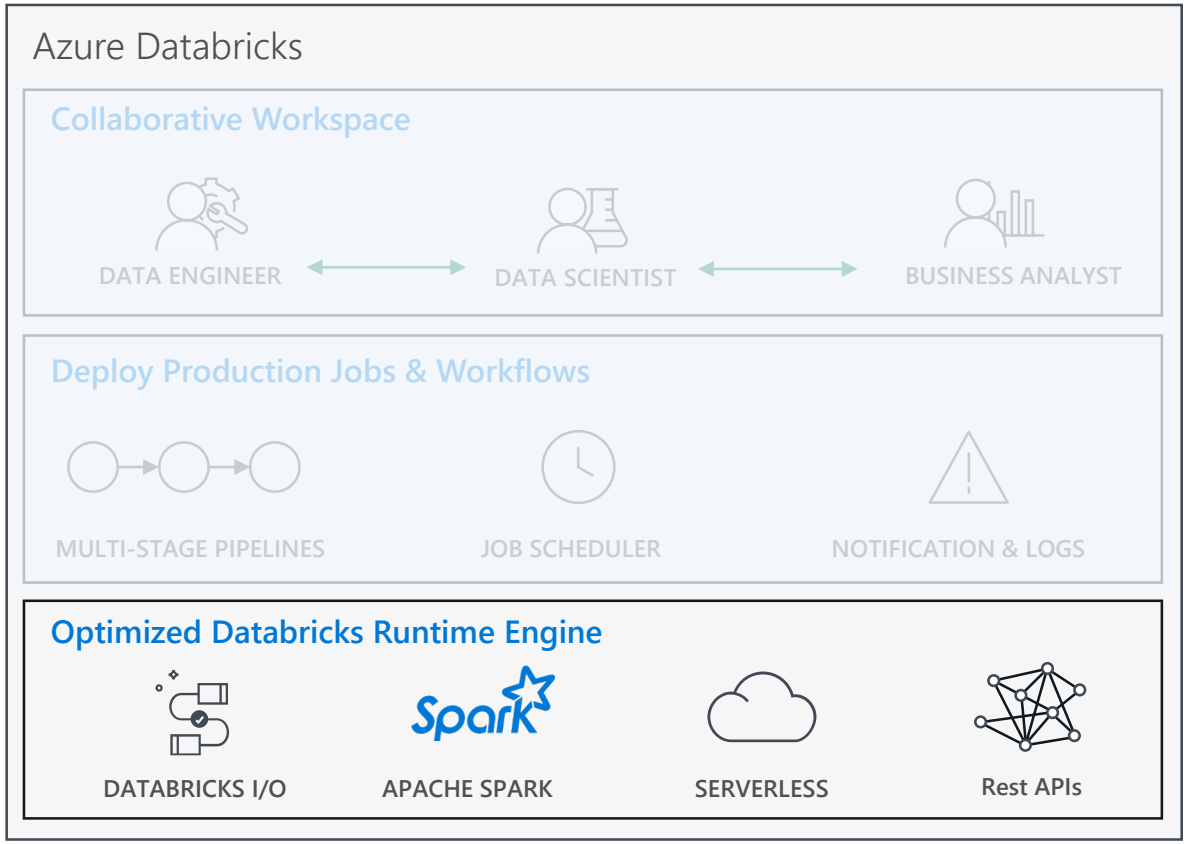
Databricks' serverless and highly elastic cloud service is designed to remove operational complexity while ensuring reliability and cost efficiency at scale

OPERATE AT MASSIVE SCALE

Without limits globally

SUPPORT FOR GPU ENABLED VMS

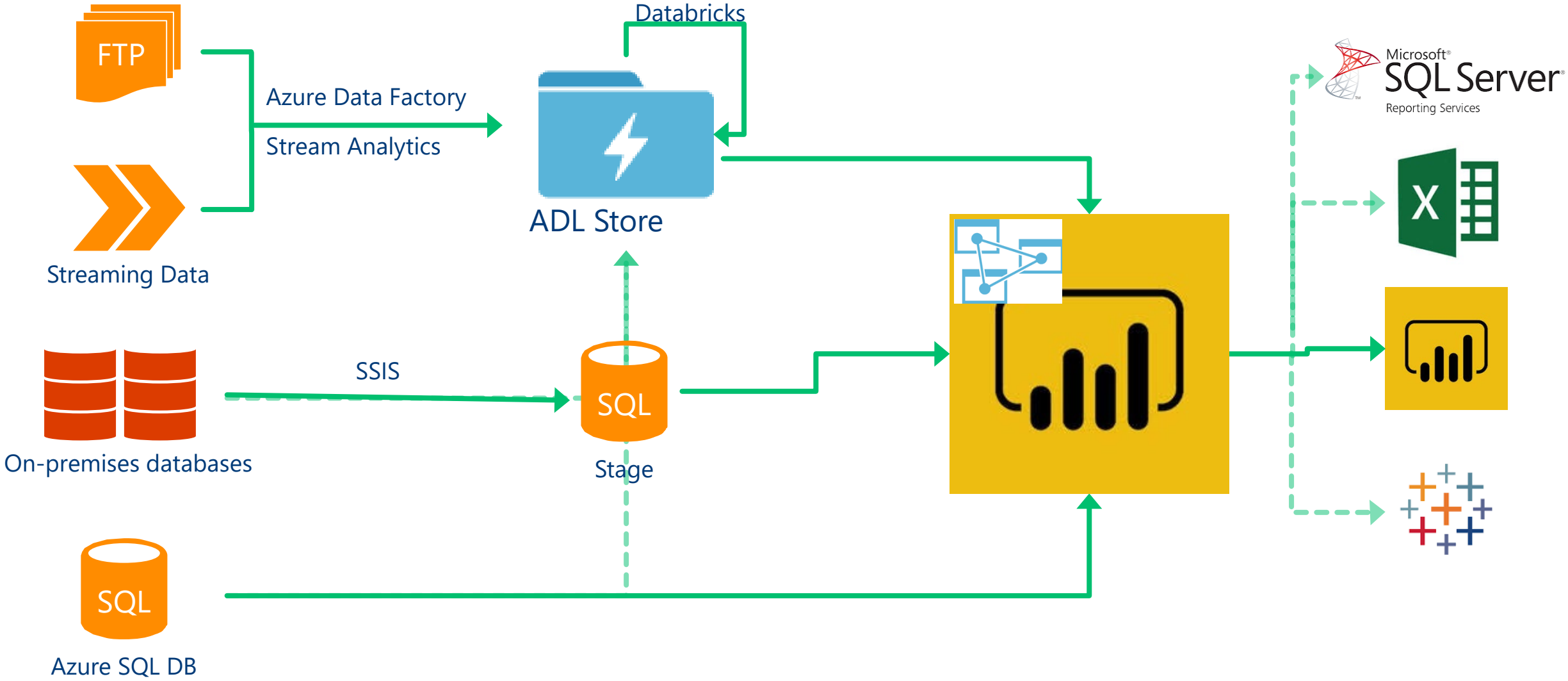
Specialized compute for your deep learning needs



DEMO

Transform data with Databricks

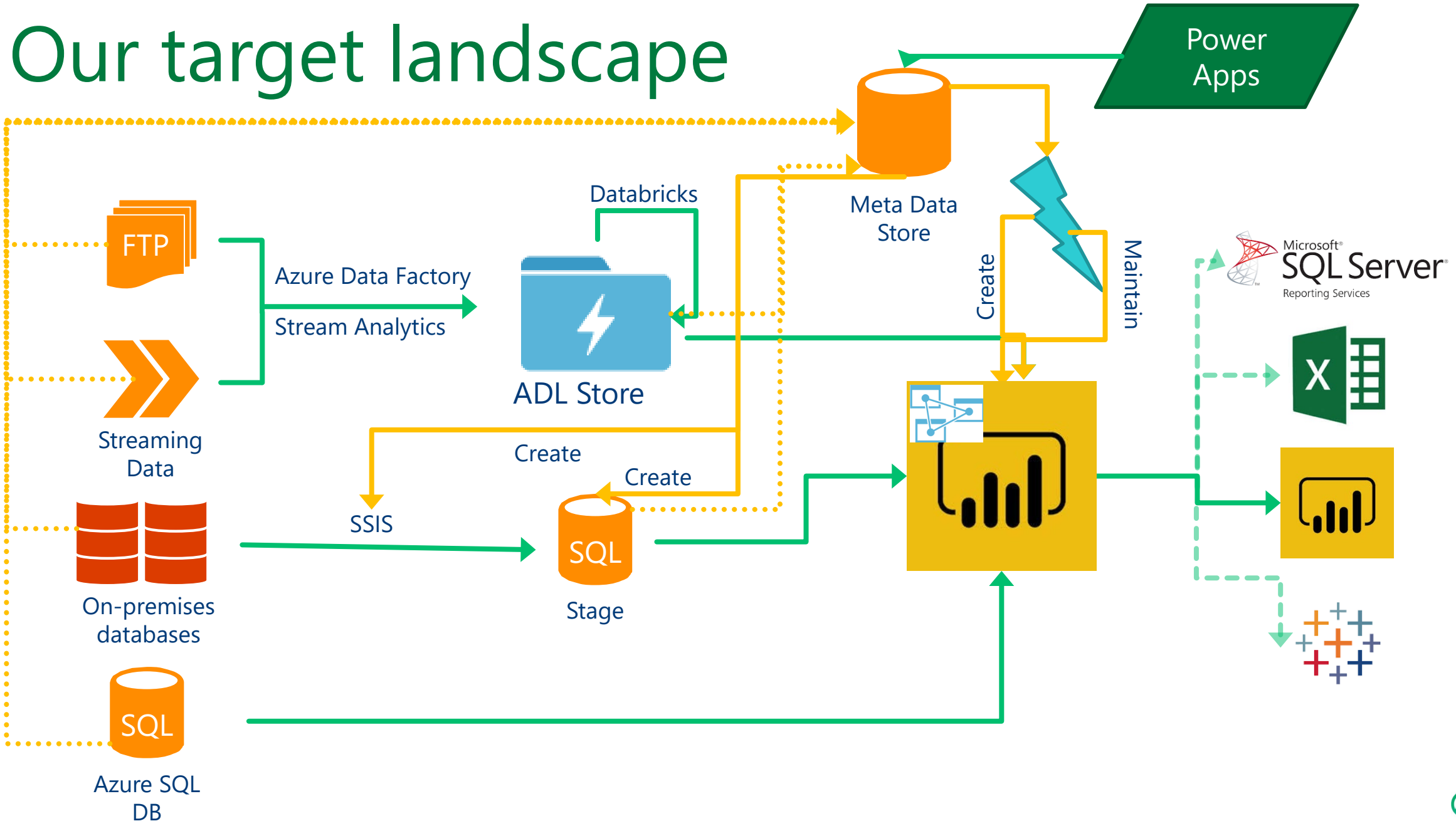
Our target landscape



DEMO

Create Power BI Model

Our target landscape



Spark .NET

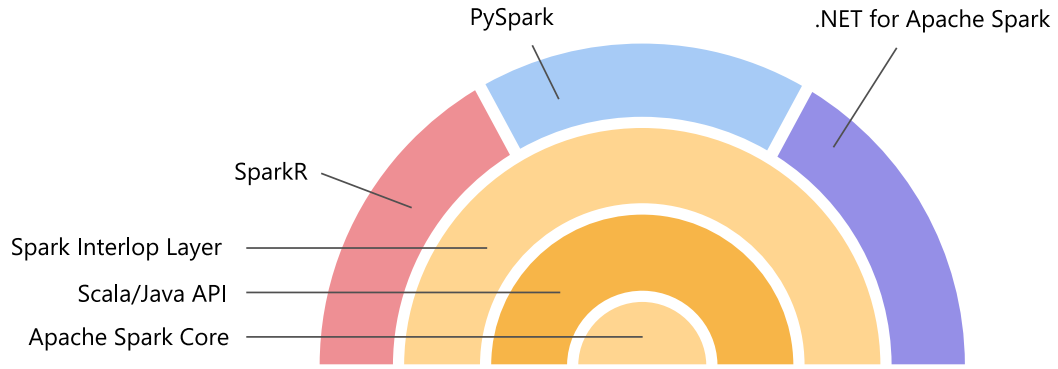
free, open-source, and cross-platform big data analytics framework

written on the Spark interop layer

compliant with .NET Standard

available by default in Azure HDInsight

can be installed in Azure Databricks, Azure Kubernetes Service, AWS Databricks, AWS EMR, and more




```
// Create a Spark session
var spark = SparkSession
    .Builder()
    .AppName("word_count_sample")
    .GetOrCreate();

// Create a DataFrame
DataFrame dataFrame = spark.Read().Text("input.txt");

// Manipulate and view data
var words = dataFrame.Select(Split(dataFrame["value"], " ").Alias("words"));

words.Select(Explode(words["words"])
    .Alias("word"))
    .GroupBy("word")
    .Count()
    .Show();
```