



Azure Synapse Analytics

What's a Lakehouse?

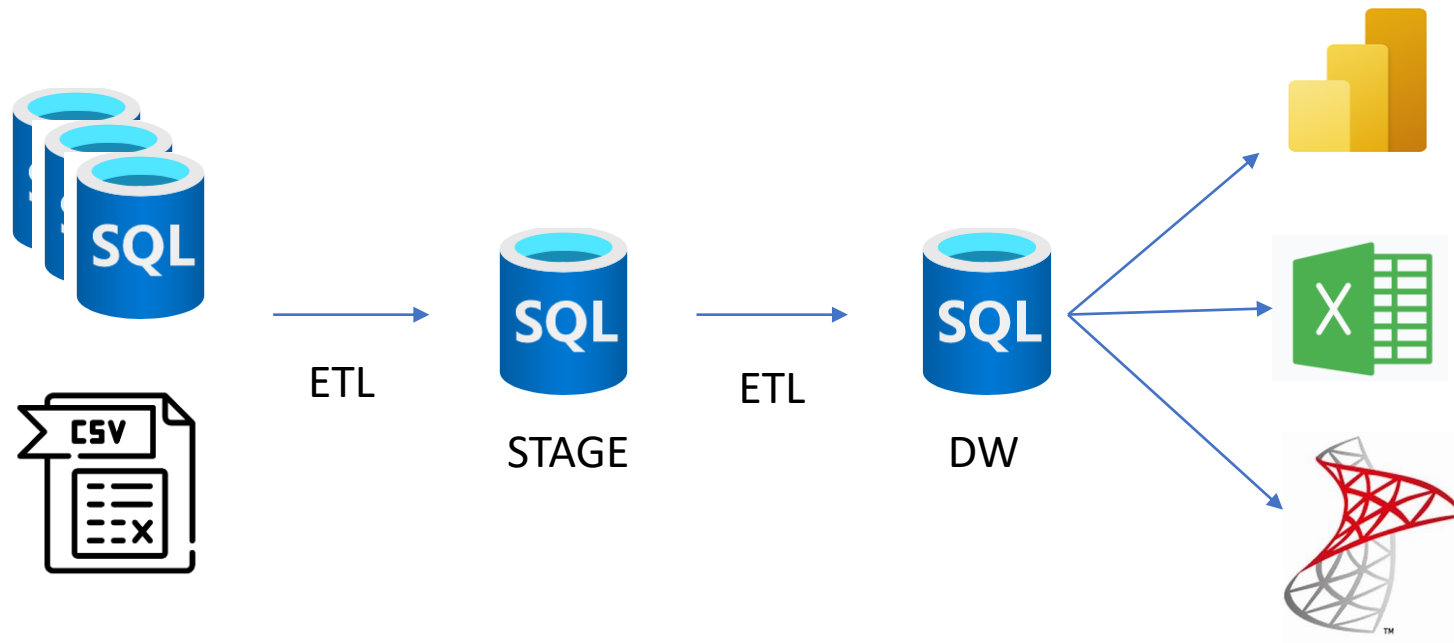
Luke Newport

Technical Architect – Data & AI

Microsoft Technology Center, St. Louis



Legacy On Premises Data Warehouse

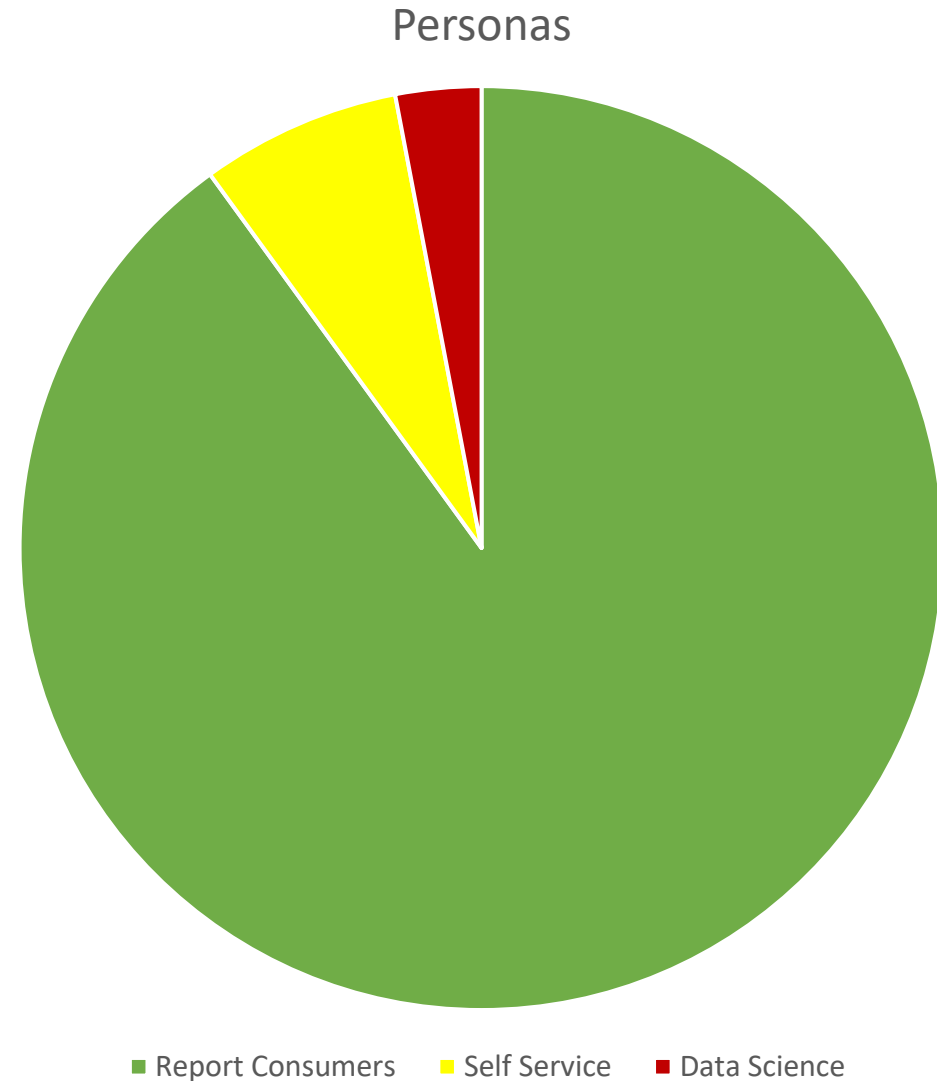


Challenges:

- Scalable – to a point.
- Relational, ideal for:
 - Tabular
 - Columnar
- Not ideal for:
 - JSON
 - Clickstream
 - Unstructured
- Change is time consuming, because of schema on write.
- Difficult to explore (Data Science) and fold back in.

Data Consumers

- Inverse Relationship – Size of Data vs Count of Consumers
- Report Consumers : Published Reports
- Self – Service : Reports + Mash up with DW+ODS
- Data Science : We want it all.
- Traditional DW struggles to serve all three due to volume and velocity.



Modern Data Warehouse (old architecture)



On-premises data

Oracle, SQL,, Teradata,
fileshares, SAP



Cloud data

Azure, AWS, GCP



SaaS data

Salesforce, Dynamics

INGEST



Azure
Data Factory

PREPARE



Azure
Data Factory



Azure
Databricks

**TRANSFORM
& ENRICH**



Azure
Data Factory



Azure
Databricks

SERVE



Azure
SQL Data
Warehouse

VISUALIZE



Power BI

STORE

Azure Data Lake Storage

Azure Synapse Analytics – *Data Lakehouse*



On-premises data

Oracle, SQL, Teradata,
fileshares, SAP



Cloud data

Azure, AWS, GCP



SaaS data

Salesforce, Dynamics



Azure Synapse Analytics

STORE

Azure Data Lake Storage

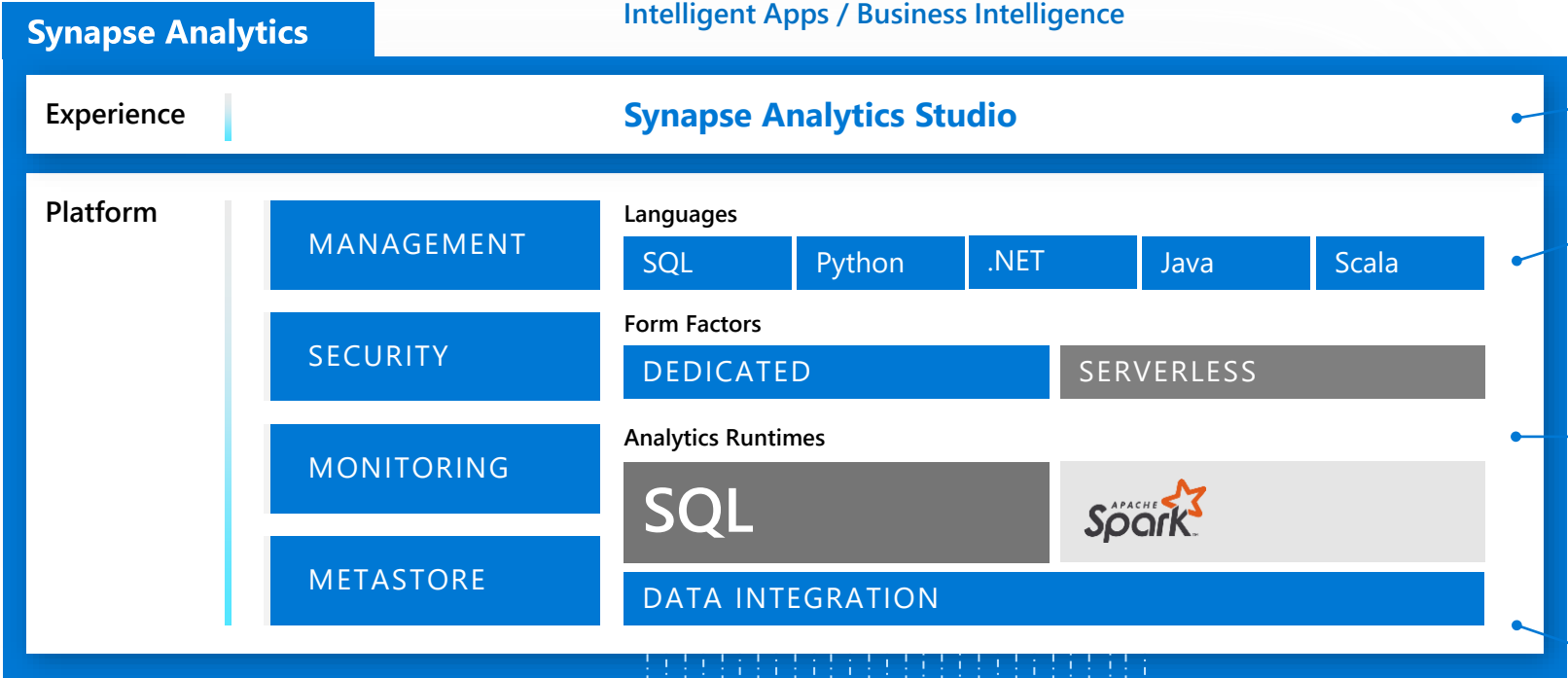
VISUALIZE



Power BI

Azure Synapse Analytics

Artificial Intelligence / Machine Learning / Internet of Things
Intelligent Apps / Business Intelligence



Designed for analytics **workloads at any scale**

SaaS **developer experiences** for code free and code first

Multiple **languages** suited to different analytics workloads

Integrated analytics runtimes available dedicated and serverless
Synapse SQL offering T-SQL for batch, streaming and interactive processing
Apache Spark for big data processing with Python, Scala and .NET

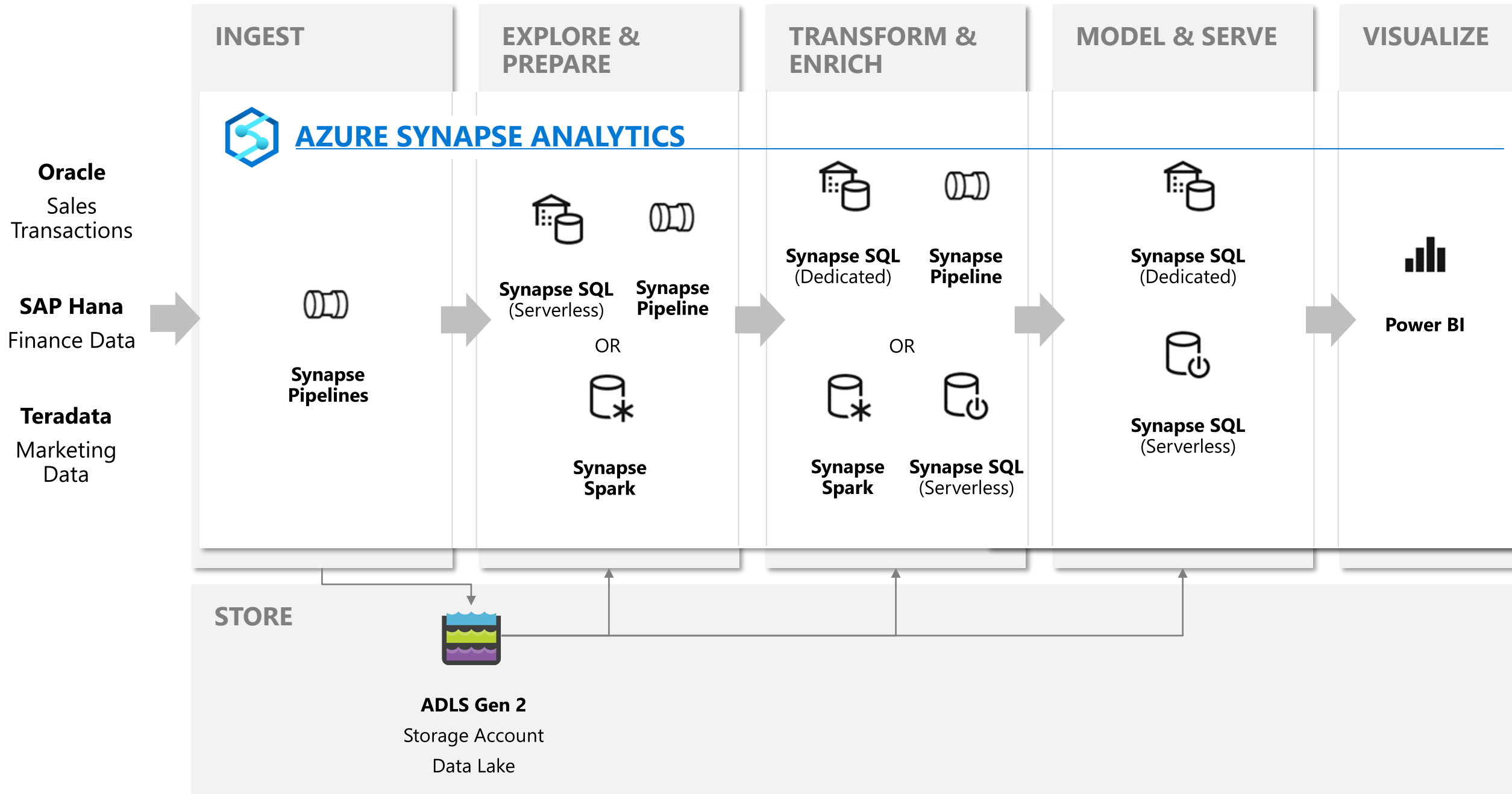
Integrated **platform services** for, management, security, monitoring, and meta-store

Data **lake integrated** and Common Data Model aware

Azure
Data Lake Storage

Common Data Model
Enterprise Security
Optimized for Analytics

Modern Data Warehouse (new architecture)



Synapse Studio



Synapse Studio divided into **Activity hubs**.

These organize the tasks needed for building analytics solutions.

The screenshot displays the Synapse Studio interface. On the left, a sidebar contains a list of Activity Hubs: Home, Data, Develop, Integrate, Monitor, and Manage. A red box highlights this sidebar, and a red arrow points from the 'Home' hub to the main workspace. The main workspace is divided into six sections, each representing an Activity Hub:

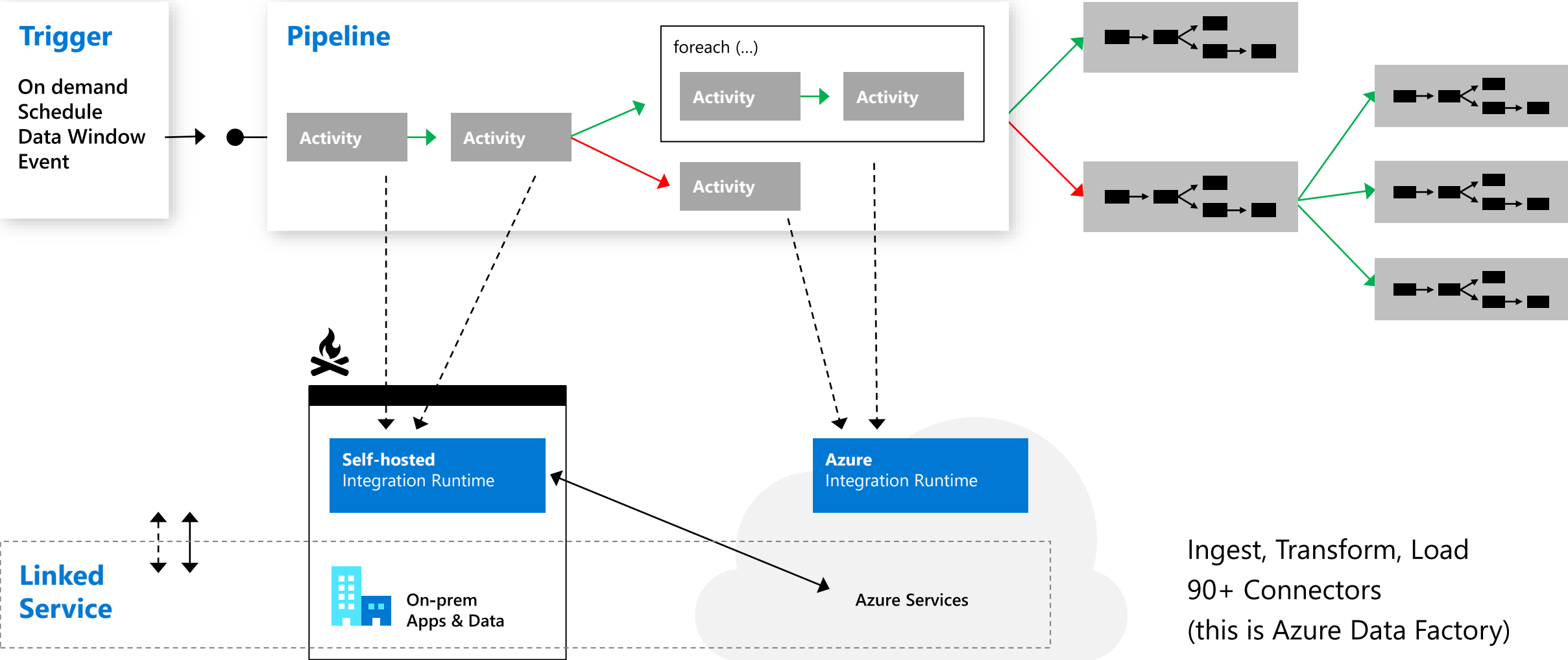
- Home:** Quick-access to common gestures, most-recently used items, and links to tutorials and documentation.
- Data:** Explore structured and unstructured data.
- Develop:** Write code and the define business logic of the pipeline via notebooks, SQL scripts, Data flows, etc.
- Integrate:** Design pipelines that that move and transform data.
- Monitor:** Centralized view of all resource usage and activities in the workspace.
- Manage:** Configure the workspace, pool, linked service, access to artifacts.

Below the sidebar, there is a 'Recent resources' section with a table listing various resources:

Name
05 Sentiment_Anal
Predict NYCTaxi Tri
001 SQL Pool Secu
005 Predict In-Eng
05 Anomaly_Detec

A 'Show more' link is visible at the bottom of the 'Recent resources' section.

Synapse Integrate



Synapse Data Hub – Linked Storage



Browse Azure Data Lake Storage Gen2 accounts – filesystems, Azure Data Explorer – clusters, Azure Cosmos DB -containers

Linked Cosmos DB
Analytical Store

Linked Azure
Data Explorer

Linked ADLS
Gen2 Account

Container
(filesystem)

The screenshot shows the Microsoft Azure Synapse Analytics interface. The left sidebar displays a list of linked resources under the 'Data' section. The main pane shows the 'rawdata' container with a list of files. The breadcrumb navigation at the top of the main pane shows the file path 'rawdata > taxidata'.

Linked Resources:

- Azure Blob Storage (3)
- Azure Cosmos DB (1)
- Azure Data Explorer (2)
- Azure Data Lake Storage Gen2 (1)
 - wsazuresynapseanalytics (Primary...)
 - default (Primary)
 - rawdata
 - staging
- Integration datasets (24)

File Path: rawdata > taxidata

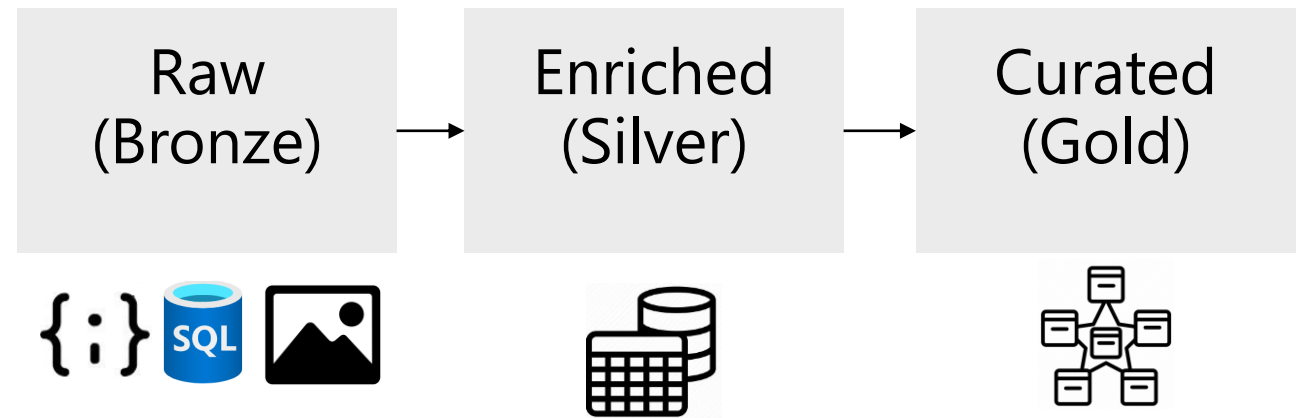
Name	Last Modified	Content Type	Size
part-00000-0300809f-304e-44bc-81bd-bbd63974c3e4-c000.snappy.parq...	8/27/2020, 12:32:19 AM		121.9 MB
part-00000-6b990121-0341-456c-8723-aec72b03f65f-c000.snappy.parqu...	8/27/2020, 12:32:25 AM		535.4 MB
part-00001-0300809f-304e-44bc-81bd-bbd63974c3e4-c000.snappy.parq...	8/27/2020, 12:32:20 AM		124.5 MB
part-00001-6b990121-0341-456c-8723-aec72b03f65f-c000.snappy.parqu...	8/27/2020, 12:32:23 AM		983.7 MB
part-00002-0300809f-304e-44bc-81bd-bbd63974c3e4-c000.snappy.parq...	8/27/2020, 12:32:19 AM		123.7 MB
part-00002-6b990121-0341-456c-8723-aec72b03f65f-c000.snappy.parqu...	8/27/2020, 12:32:21 AM		966.1 MB

Showing 1 to 6 of 6 cached items

Synapse – Azure Data Lake Gen2



- ADLSg2 – polyglot storage
- Posix ACLs
- HDFS endpoint on blob
- AAD Integration
- /.../.../.../*.parquet (delta_log)
- Storage tiering by policy (

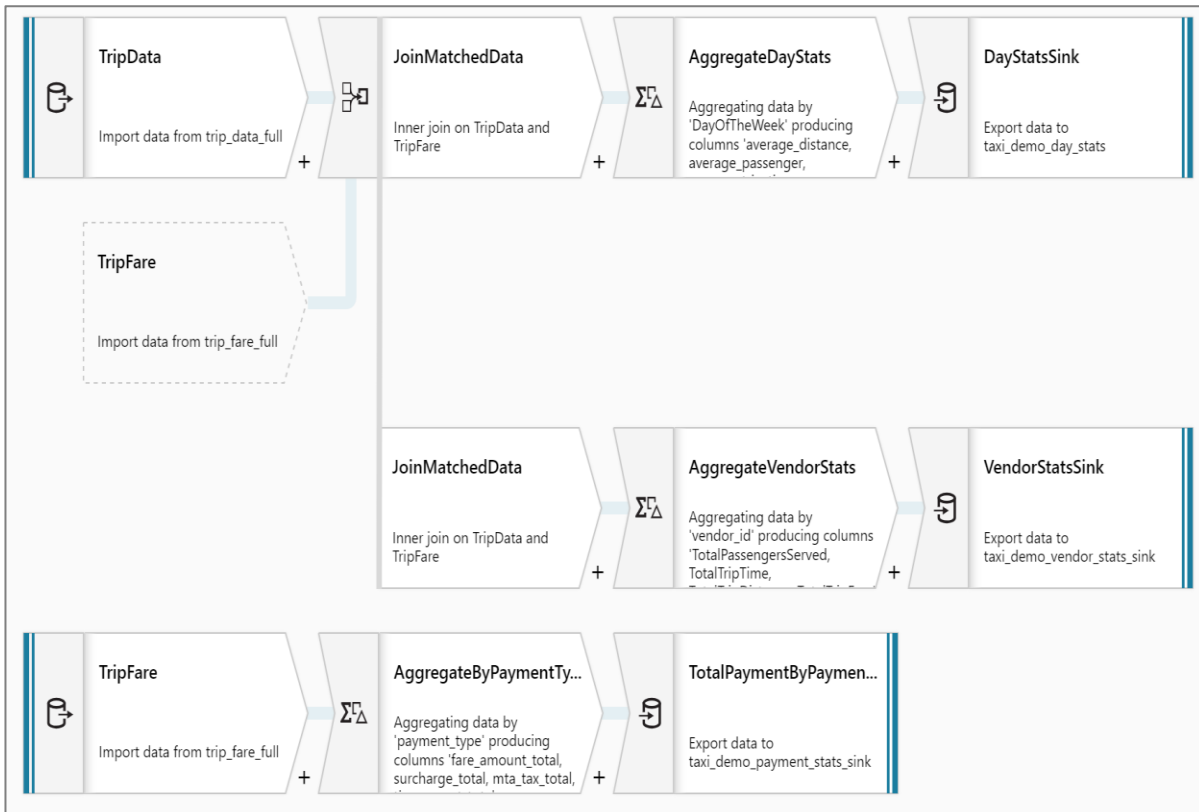


Synapse - Prep & Transform Data – Low Code



Data Flows

Code free data transformation @scale



Power Query

Code free data preparation @scale

The screenshot shows the Microsoft Azure Synapse Analytics Power Query interface. The table displayed has the following columns and data:

	storeid	productCode	quantity	1.2_quantity	advertising	price	weekStarting	id
1	2	surface.go	105	9.265	1	159	6/15/2017	d0bd47a7-2ad6-4f0a-b1de-ed1386cae5ea
2	2	surface.go	80	8.987	0	269	7/27/2017	64cc74c2-c7da-4e12-af64-c95bd429934
3	2	surface.go	68	8.832	1	209	8/3/2017	9a2d164b-5e44-44d7-9837-c9ae6566c99
4	2	surface.go	28	7.966	0	209	8/10/2017	b8cd9987-1d5a-44f1-9346-719d73b1f7f0
5	2	surface.go	16	7.378	0	209	8/24/2017	ac0ec099-e102-4bf6-9775-983b151dc0d3
6	2	surface.go	253	10.14	1	189	8/31/2017	3d22c002-b04c-4092-4bcb-b3b0f01c9d8f
7	2	surface.go	107	9.283	0	189	9/7/2017	b5e19699-d584-449e-9c98-c19d5288bc7b
8	2	surface.go	66	8.803	0	189	9/14/2017	e99a5838-b8f1-413a-a23d-4d000c4b5282
9	2	surface.go	65	8.794	0	179	9/21/2017	c3278682-16c0-4832-b76f-5d4a3e7a5d0
10	2	surface.go	17	7.455	0	269	10/12/2017	440190c1-b2ed-46f4-a5e0-8d5899d41dc
11	2	surface.go	337	10.428	1	124	10/19/2017	5294983-5044-4068-e925-e778d938a627
12	2	surface.go	19	7.56	0	159	10/26/2017	0a3f799c-07a3-450f-a6bc-516d552bf1ad
13	2	surface.go	89	9.101	1	159	11/2/2017	d0eab000-e45e-474f-864e-284980ba156
14	2	surface.go	113	9.341	0	129	11/9/2017	932a3b9f-2f54-4e0b-bc4d-c02180bc7f60
15	2	surface.go	284	10.255	0	99	11/16/2017	638edd94-5873-4b37-a078-5ac145eb549e
16	2	surface.go	171	9.75	1	159	11/23/2017	c1dc14fb-29a5-490a-9191-34e070f946c
17	2	surface.go	265	10.187	1	249	11/30/2017	6fb16536-dd0c-49cc-a7a6-9e7a23ce793f
18	2	surface.go	63	8.754	0	269	12/7/2017	241dd46b-fb0e-48c9-92e4-6ad658b9319e
19	2	surface.go	263	10.18	1	139	12/14/2017	25da81a1-e423-4a0b-9142-deac1177f793
20	2	surface.go	8	6.798	0	269	12/21/2017	eed8291c-4f02-419f-8aad-6c69abede24
21	2	surface.go	14	7.294	0	269	12/28/2017	8b964b5a-3eee-4132-9071-4299c568587a
22	2	surface.go	13	7.203	0	269	1/4/2018	be1e79fa-311c-4e46-a215-6ac770ffab4
23	2	surface.go	1176	11.623	1	99	1/11/2018	896b1fa5-fc91-4955-b09c-9c01a5c74e9
24	2	surface.go	18	7.526	0	269	1/18/2018	90cd5b43-0f6e-4f05-b4c5-69e09784b0c
25	2	surface.go	55	8.625	0	269	1/25/2018	7a33a43e-79a0-4b13-b306-6bac3f7b5a87
26	2	surface.go	320	10.375	1	149	2/1/2018	6821e07f-568f-4b2a-8484-5e6959d05155
27	2	surface.go	43	8.378	0	149	2/8/2018	5d33c961-852f-4efe-944d-2a793598e8fe

Azure Synapse x Apache Spark – Code First



- **Apache Spark 3.2 derivation**
 - Linux Foundation Delta Lake 1.1 support
 - .Net Core 3.0 support
 - Python 3.8 + Anacondas support
- **Tightly coupled to other Azure Synapse services**
 - Integrated security and sign on
 - Integrated Metadata
 - Integrated and simplified provisioning
 - Integrated UX including interact based notebooks
 - Fast load of Synapse SQL (provisioned) pools
- **Core scenarios**
 - Data Prep/Data Engineering/ETL
 - Machine Learning via Spark ML and Azure ML integration
 - Extensible through library management
- **Efficient resource utilization**
 - Fast Start
 - Auto scale (up and down)
 - Auto pause
 - Min cluster size of 3 nodes
- **Multi Language Support**
 - .Net (C#), PySpark, Scala, Spark SQL, Java

Synapse Dedicated SQL Pools

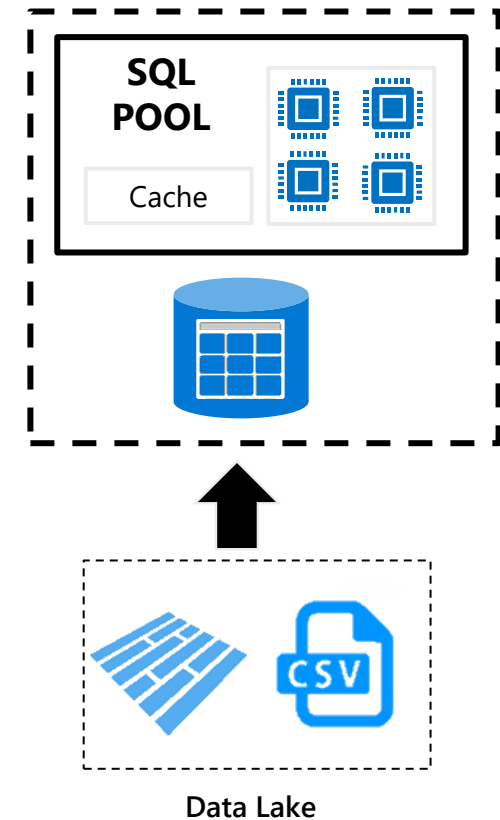


Advanced storage system

- MPP – evolution of Azure SQL DW
- ColumnStore Indexes
- Table partitions
- Distributed tables
- Isolation modes
- Materialized Views
- Nonclustered Indexes
- Result-set caching

Complete SQL object model

- Tables
- Views
- Stored procedures
- Functions
- **It's SQL Server at scale**



Serverless SQL Pools

Overview

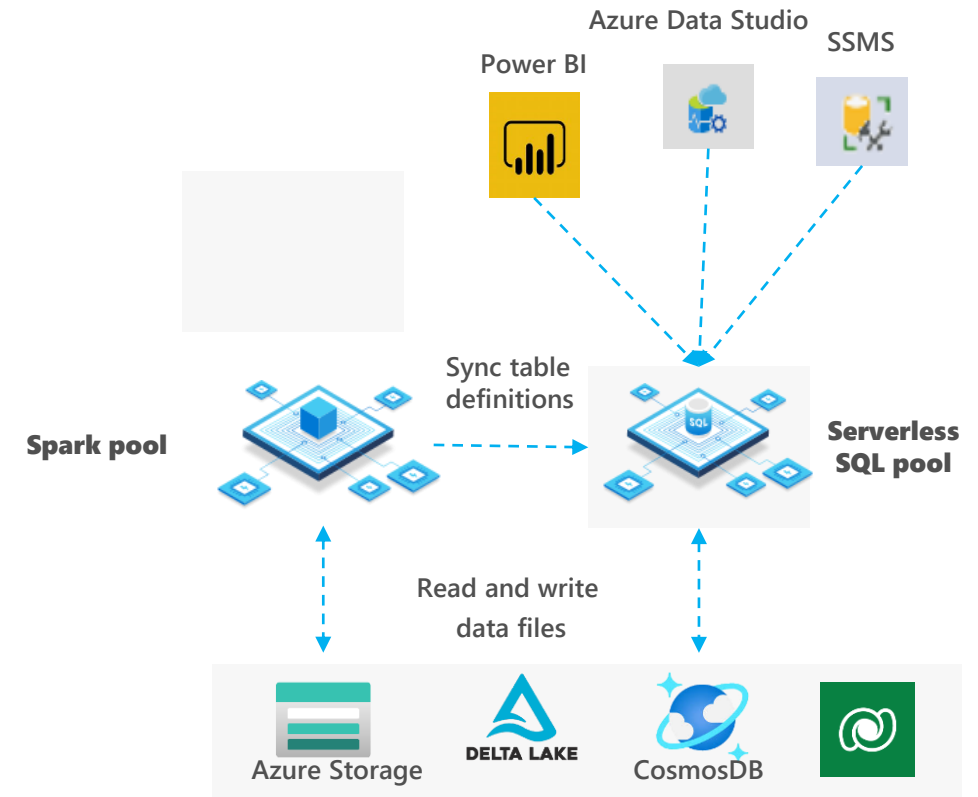
- An interactive query service that enables you to use standard T-SQL queries over Data Lake, CosmosDB, Dataverse.

Benefits

- Use T-SQL language
- Supports any tool or library that uses T-SQL to query data
- Automatically synchronize tables from Spark pool
- Queries multiple storages (Lake, CosmosDB)

Serverless experience

- Auto Scale & Manage
- Pay-per-use model
- Fast (roughly DWU 2000 per workspace)
- Automatic schema inference



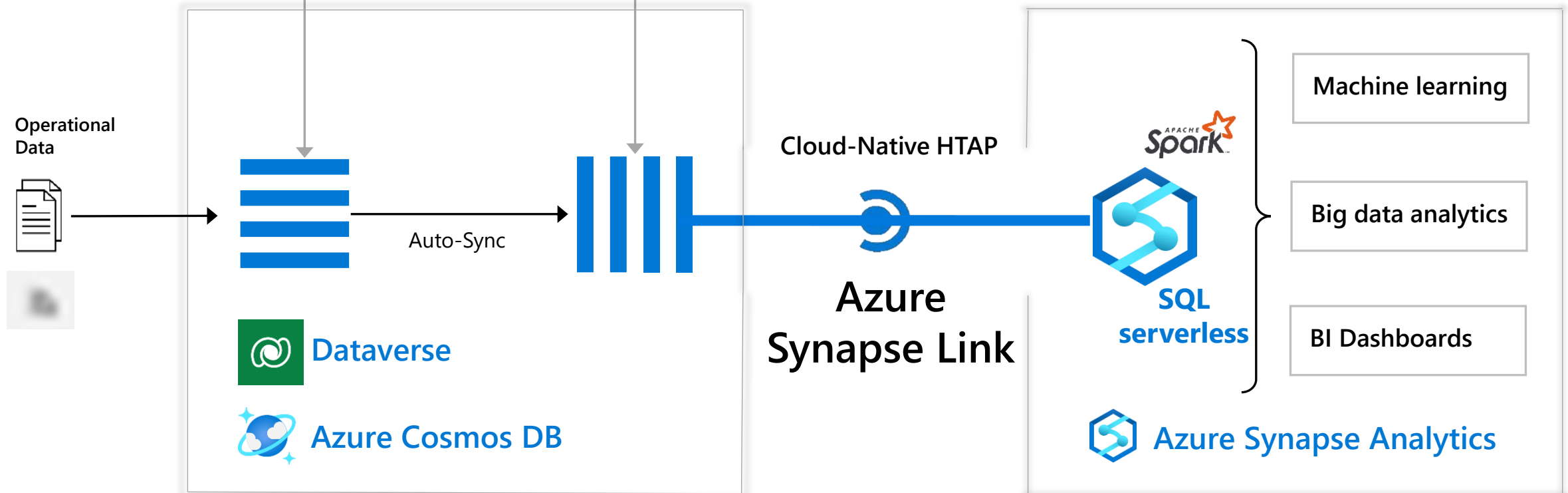
Synapse Link – near real-time analytics

Transactional Store

Row store optimized for transactional operations

Analytical Store

Column store optimized for analytical queries



Demos!

A walk through Synapse ingestion, exploration, transformation, serving and more!

Repo: github.com/lnewport (pdf of deck will be here)

The Big Picture

