Azure Synapse Analytics

BY: SOUMYA RANJAN NAYAK

PRN: 23070243063

COURSE COORDINATOR: SAHIL SHAH

CONTENT

WHAT IS AZURE SYNAPSE ANALYTICS ?

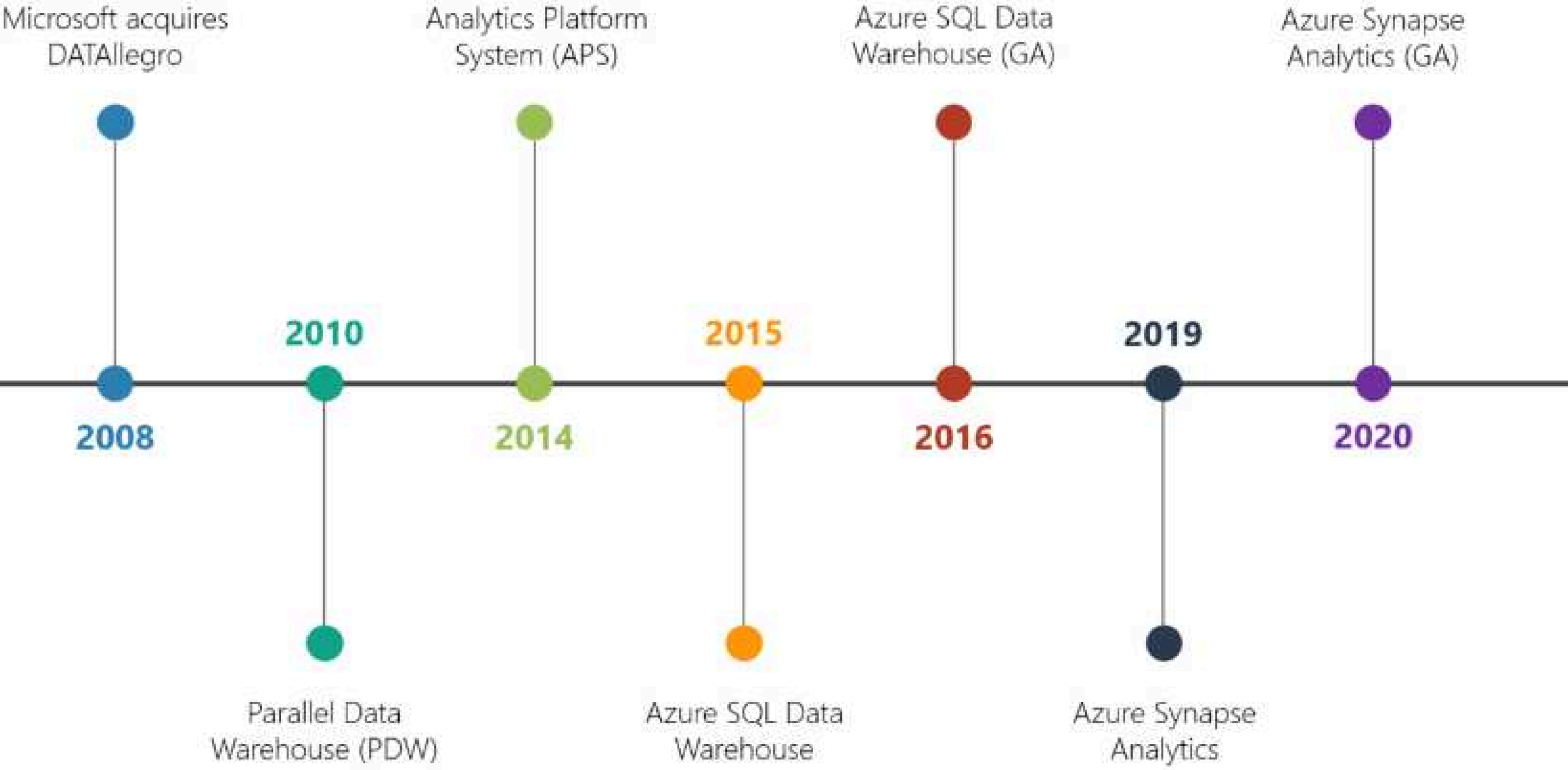
Azure Synapse Analytics is a cloud-based analytics service that enables you to perform big data analytics and data warehousing tasks in a unified, integrated, and secure environment.

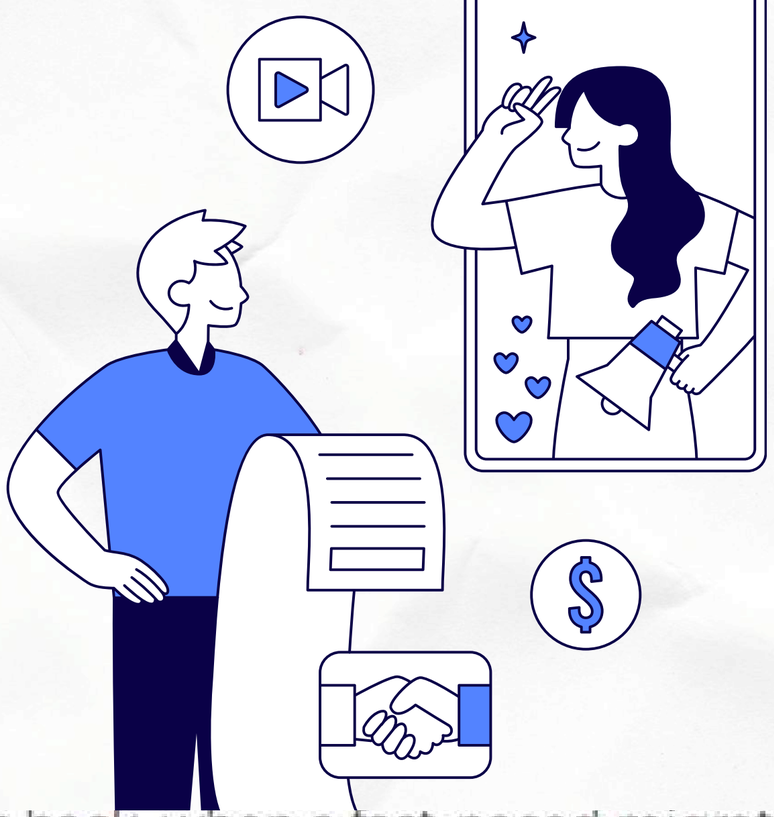
Azure Synapse brings together the best of SQL technologies used in enterprise data warehousing, Spark technologies used for big data, Data Explorer for log and time series analytics, Pipelines for data integration and ETL/ELT, and deep integration with other Azure services such as Power BI, CosmosDB, and AzureML.

HISTORY

To attain a better understanding of the Azure Synapse Analytics service, it can help to acknowledge the journey Microsoft's analytics platform has gone through.

- **2008** - Microsoft acquires DATAllegro, a provider of data warehouse appliances.
- **2010** - Microsoft announces SQL Server 2008 R2 Parallel Data Warehouse (aka Project Madison), an enterprise data warehouse appliance based on the massively parallel processing (MPP) technology originally created by DATAllegro.
- **2014** - Microsoft announces the Analytics Platform System (big data in a box), an evolution of the SQL Server Parallel Data Warehouse with the introduction of a dedicated region for Microsoft's Hadoop distribution - HDInsight. Using PolyBase (Microsoft's SQL Server to HDFS bridge technology), customers could join relational data from SQL Server PDW and non-relational data in Hadoop.
- **2015** - Microsoft announces Azure SQL Data Warehouse, an elastic data warehouse in the cloud. Allows customers to dynamically scale (grow and shrink) and pause compute, independent of storage.
- **2016** - Microsoft announces the general availability of Azure SQL Data Warehouse. Offers an availability SLA of 99.9%.
- **2019** - Microsoft announces Azure Synapse Analytics, the next evolution of Azure SQL Data Warehouse. Includes the addition of new preview capabilities such as workload isolation, on-demand query (aka serverless SQL), Azure Synapse Studio, and Apache Spark integration.
- **2020** - Microsoft announces the general availability of Azure Synapse Analytics.
- **2021** - Microsoft announces the public preview of Azure Synapse Data Explorer, complementing the existing Synapse SQL and Apache Spark analytic engines.


Microsoft acquires DATAllegro
2008
2010
Parallel Data Warehouse (PDW)
Analytics Platform System (APS)
2014
2015
Azure SQL Data Warehouse
Azure SQL Data Warehouse (GA)
2016
2019
Azure Synapse Analytics
Azure Synapse Analytics (GA)
2020

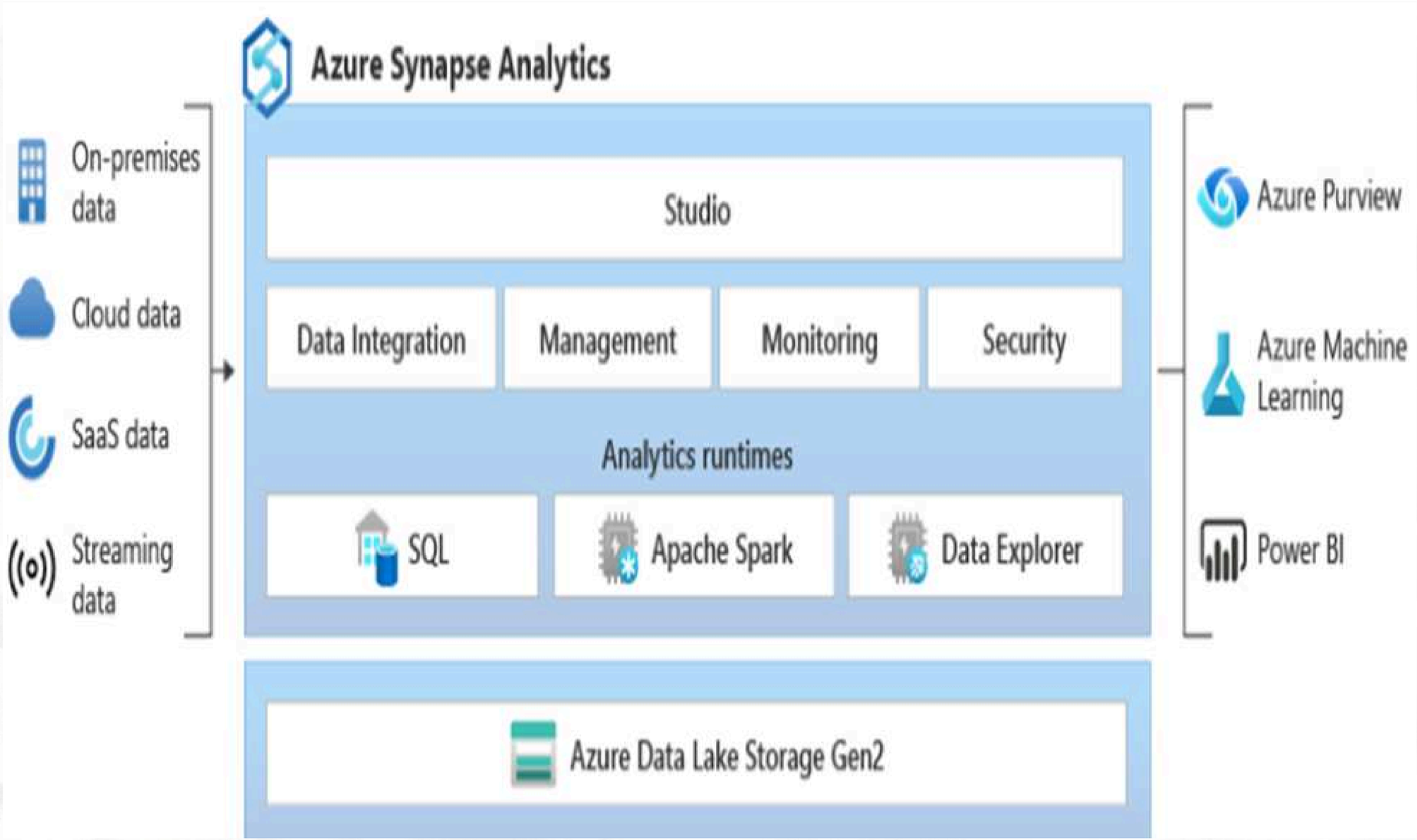


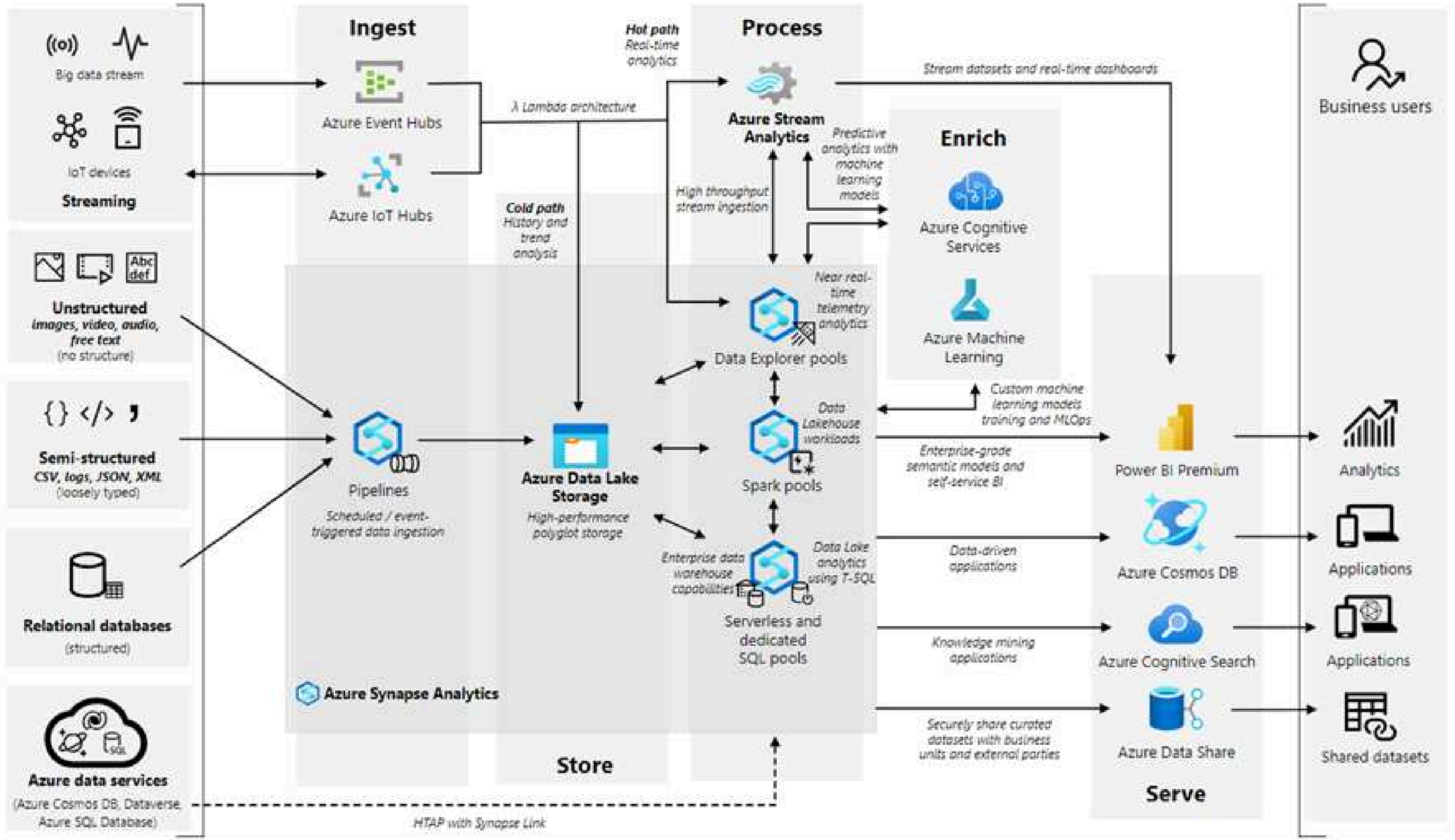
THE NEED

About five years back, when a fast-paced migration to the cloud was happening in the wake of benefits from large-scale computing and data storage, Microsoft noticed a major problem from the customer perspective. There were many widespread applications offering storage, computing, and analysis but something was amiss.

These platforms were built independently and didn't offer a way to connect, leading to wastage of time, effort and cost in learning each of the platforms and reliance on IT. Ironically, these applications were supposed to offer self-service capabilities to business users.

ARCHITECTURE

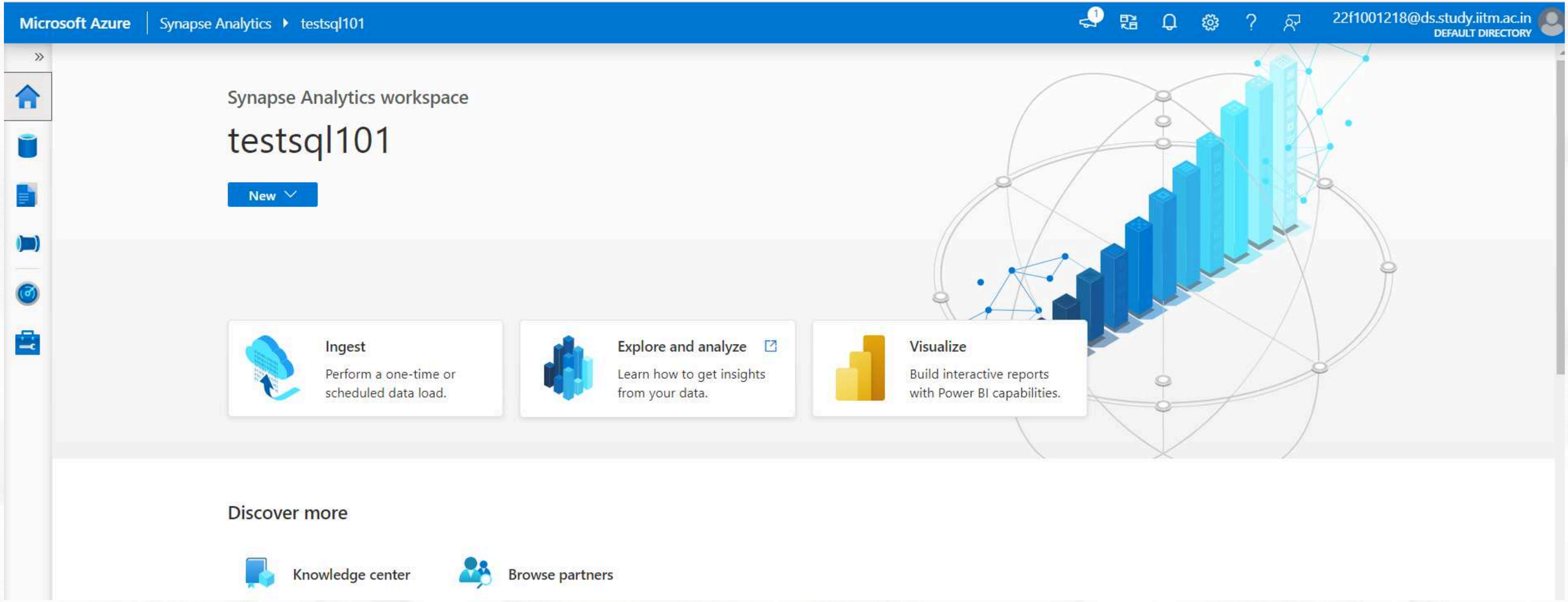



Ingest
Process
Enrich
Serve
Store
Big data stream
IoT devices
Streaming
Unstructured
images, video, audio, free text
(no structure)
Semi-structured
CSV, logs, JSON, XML
(loosely typed)
Relational databases
(structured)
Azure data services
(Azure Cosmos DB, Dataverse, Azure SQL Database)
Azure Event Hubs
Azure IoT Hubs
λ Lambda architecture
Hot path
Real-time analytics
Cold path
History and trend analysis
Azure Stream Analytics
High throughput stream ingestion
Predictive analytics with machine learning models
Stream datasets and real-time dashboards
Azure Cognitive Services
Azure Machine Learning
Near real-time telemetry analytics
Data Explorer pools
Custom machine learning models training and MLOps
Pipelines
Scheduled / event-triggered data ingestion
Azure Data Lake Storage
High-performance polyglot storage
Data Lakehouse workloads
Spark pools
Enterprise-grade semantic models and self-service BI
Power BI Premium
Data-driven applications
Azure Cosmos DB
Enterprise data warehouse capabilities
Data Lake analytics using T-SQL
Serverless and dedicated SQL pools
Knowledge mining applications
Azure Cognitive Search
Securely share curated datasets with business units and external parties
Azure Data Share
HTAP with Synapse Link
Azure Synapse Analytics
Business users
Analytics
Applications
Applications
Shared datasets
Discover & govern
Azure Purview
Platform
Microsoft Entra ID
Microsoft Cost Management
Azure Key Vault
Azure Monitor
Microsoft Defender for Cloud
Azure DevOps & GitHub
Azure Policy

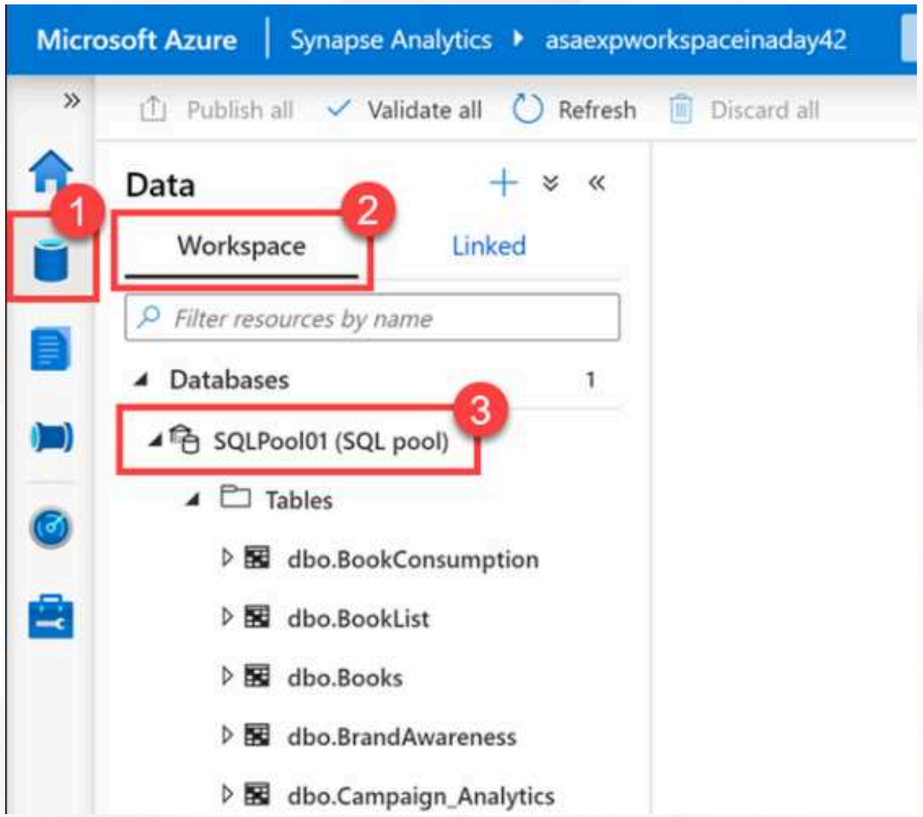
KEY COMPONENTS

- **Workspace:** A central hub for all Synapse Analytics resources, like SQL and Apache Spark pools, data flows, linked services, and pipelines.
- **SQL Pool:** Ideal for large-scale data warehousing, employing a parallel processing setup for speedy data analysis.
- **Apache Spark Pool:** Handles parallel data processing tasks efficiently, supporting various languages like SQL, Python, R, and Scala.
- **Data Flow:** A visual ETL tool simplifying data transformation tasks with a drag-and-drop interface.
- **Linked Services:** Secure connections to external data sources, enabling seamless data transfer.
- **Pipelines:** Automates ETL workflows, allowing scheduling, monitoring, and management of data processing tasks.
- **Integration with Power BI:** Enhances analytics capabilities by leveraging Power BI's visualization and business intelligence features for creating interactive dashboards and reports.

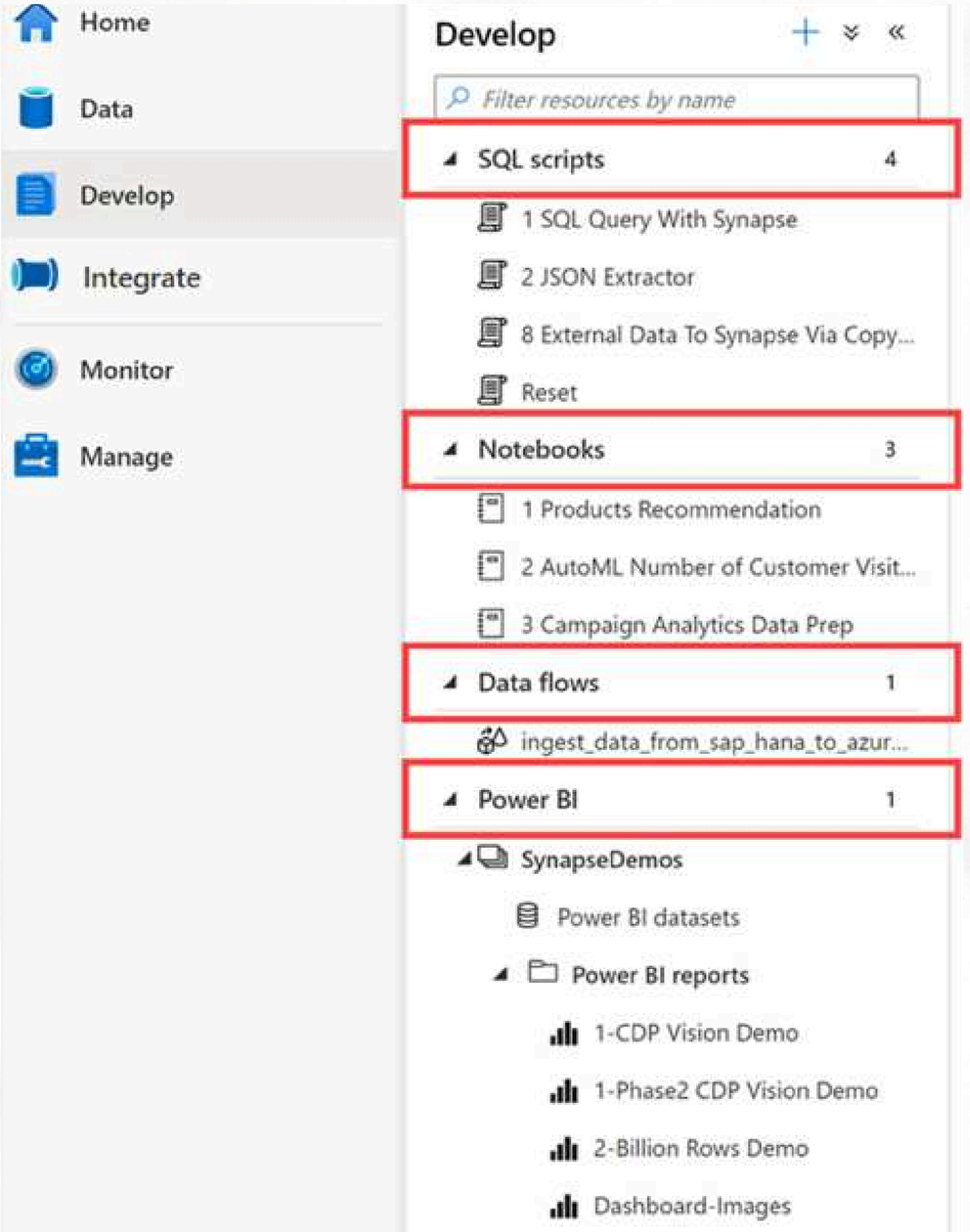
SYNAPSE ANALYTICS WORKSPACE



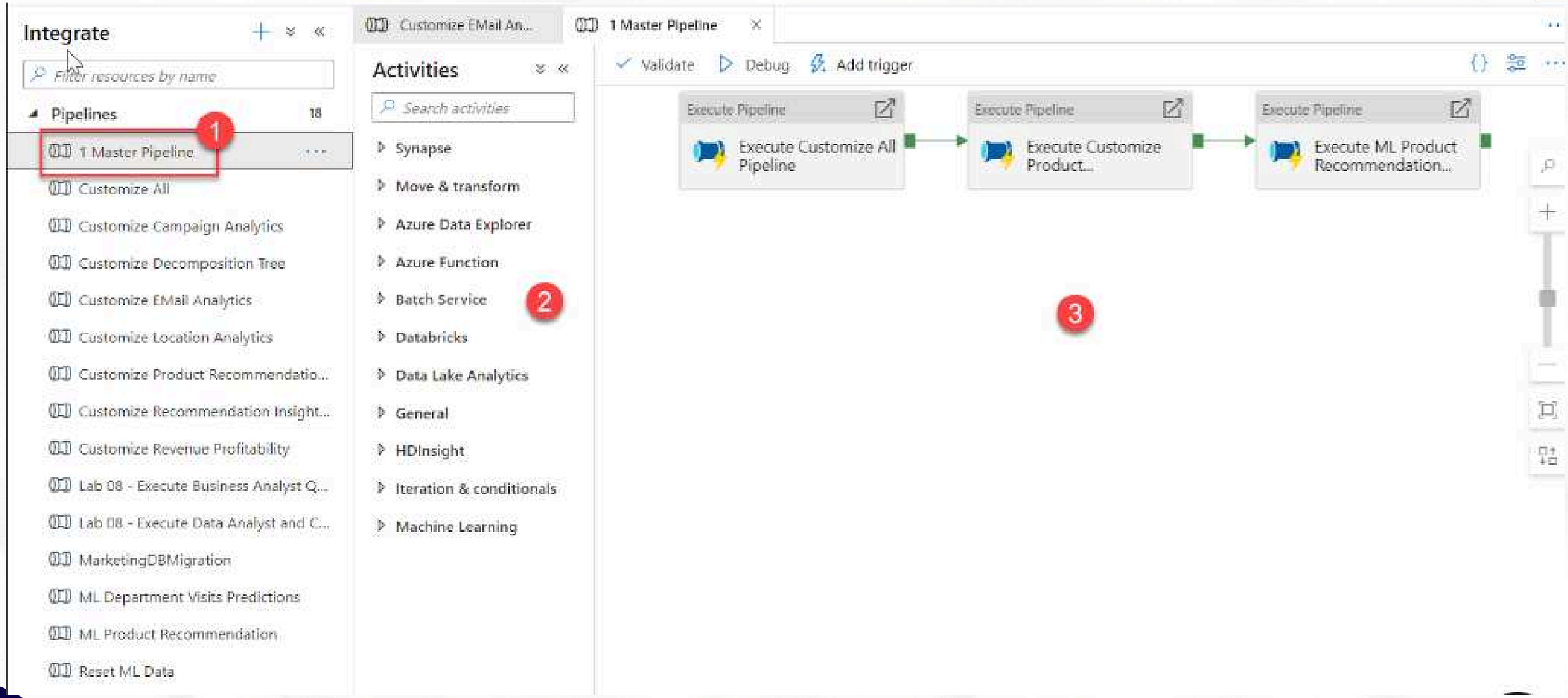
DATA HUB



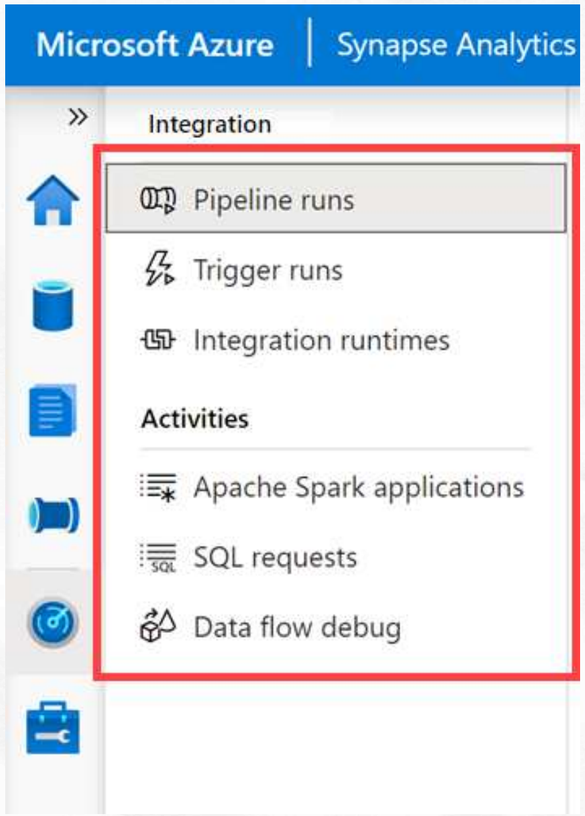
DEVELOP HUB

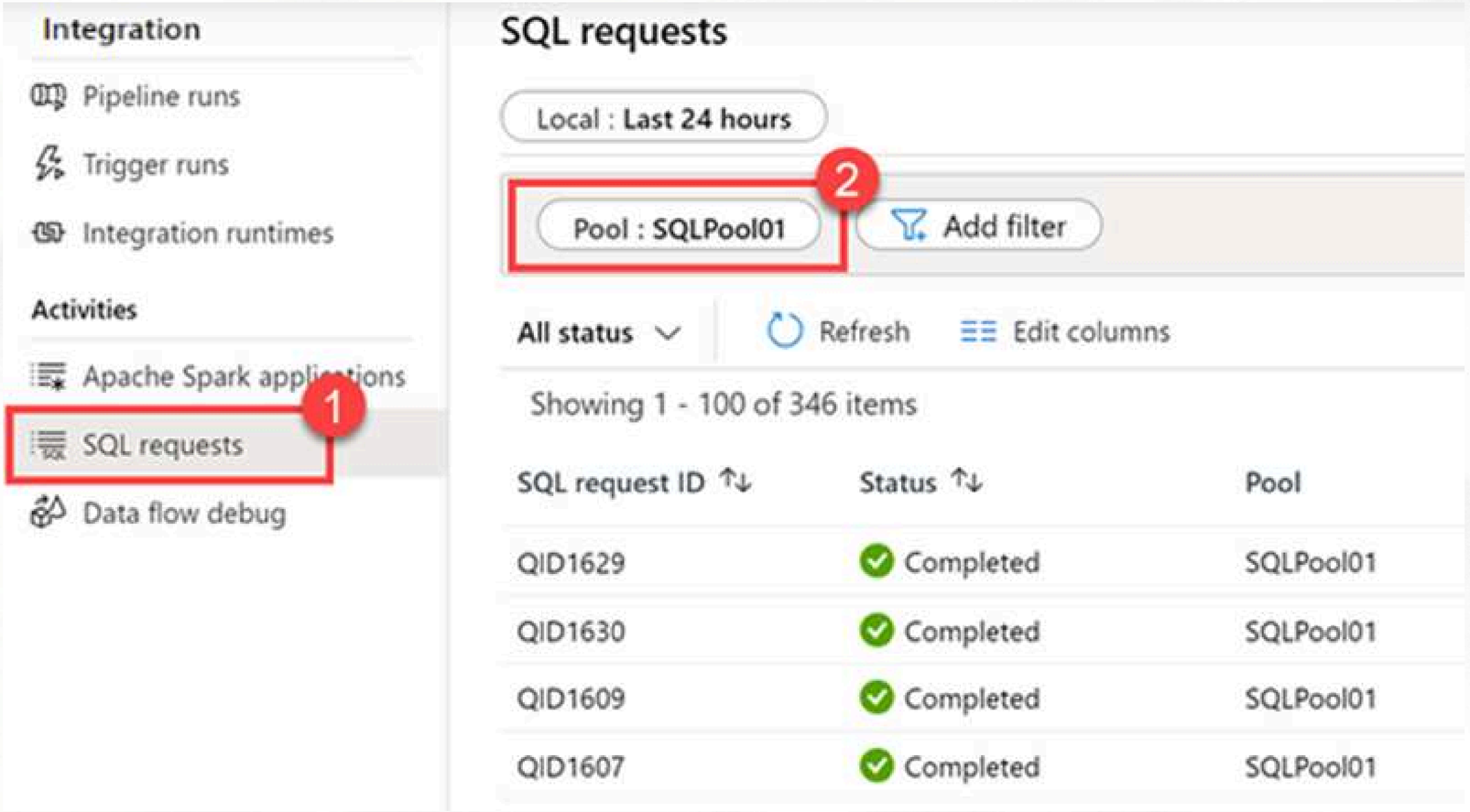


INTEGRATION HUB

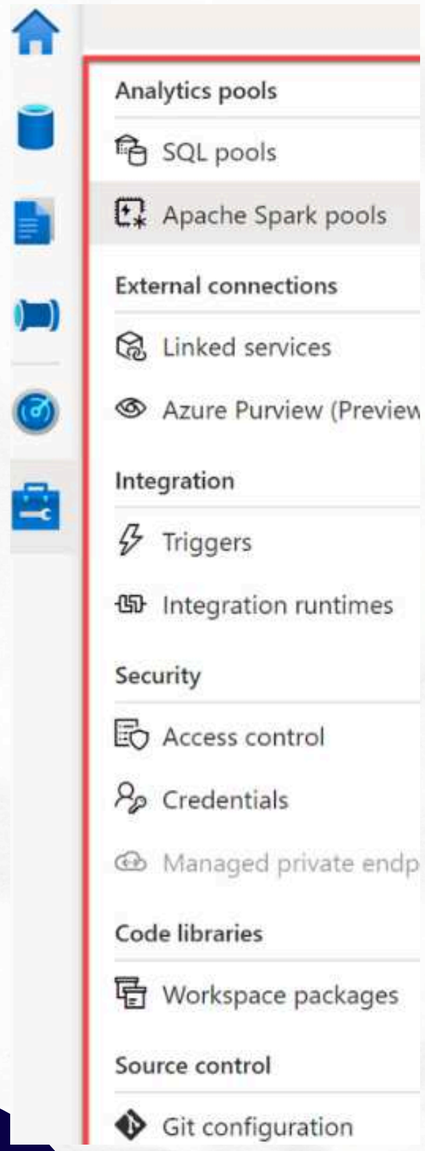


MONITOR HUB


Microsoft Azure
Synapse Analytics
Integration
Pipeline runs
Trigger runs
Integration runtimes
Activities
Apache Spark applications
SQL requests
Data flow debug


Integration
Pipeline runs
Trigger runs
Integration runtimes
Activities
Apache Spark applications
SQL requests
Data flow debug
1
SQL requests
Local : Last 24 hours
2
Pool : SQLPool01
Add filter
All status
Refresh
Edit columns
Showing 1 - 100 of 346 items
SQL request ID
Status
Pool
QID1629
Completed
SQLPool01
QID1630
Completed
SQLPool01
QID1609
Completed
SQLPool01
QID1607
Completed
SQLPool01

MANAGE HUB



1. SQL POOLS: MANAGE PROVISIONED SQL POOLS AND ON-DEMAND SQL SERVERLESS POOLS. YOU CAN ADD, PAUSE, OR SCALE POOLS TO OPTIMIZE COSTS.
2. APACHE SPARK POOLS: CONFIGURE AUTO-PAUSE AND AUTO-SCALE SETTINGS FOR SPARK POOLS. PROVISION NEW POOLS AS NEEDED.
3. LINKED SERVICES: MANAGE CONNECTIONS TO EXTERNAL RESOURCES SUCH AS DATA LAKE STORAGE, AZURE KEY VAULT, POWER BI, AND SYNAPSE ANALYTICS. EASILY ADD NEW LINKED SERVICES.
4. AZURE PURVIEW (PREVIEW): INTEGRATES WITH AZURE PURVIEW FOR DATA GOVERNANCE AND LINEAGE WITHIN SYNAPSE ANALYTICS.
5. TRIGGERS: CENTRAL LOCATION FOR CREATING OR REMOVING PIPELINE TRIGGERS. TRIGGERS CAN ALSO BE ADDED DIRECTLY FROM PIPELINES.
6. INTEGRATION RUNTIMES: LISTS COMPUTE INFRASTRUCTURE FOR DATA INTEGRATION CAPABILITIES. MODIFY PARAMETERS AND MANAGE RUNTIMES EFFICIENTLY.
7. ACCESS CONTROL: ADD OR REMOVE USERS TO SECURITY GROUPS LIKE WORKSPACE ADMIN, SQL ADMIN, AND APACHE SPARK ADMIN.
8. CREDENTIALS: STORES AUTHENTICATION INFORMATION USED BY SYNAPSE ANALYTICS.
9. MANAGED PRIVATE ENDPOINTS: MANAGE PRIVATE ENDPOINTS FOR SECURE CONNECTIONS WITHIN A VIRTUAL NETWORK TO AZURE SERVICES OR PRIVATE LINK SERVICES.
10. WORKSPACE PACKAGES: INCLUDES CUSTOM CODE OR SPECIFIC VERSIONS OF OPEN-SOURCE LIBRARIES FOR USE IN APACHE SPARK POOLS.
11. GIT CONFIGURATION: CONNECT WORKSPACE TO A GIT REPOSITORY FOR ENABLING SOURCE CONTROL.

FEATURES

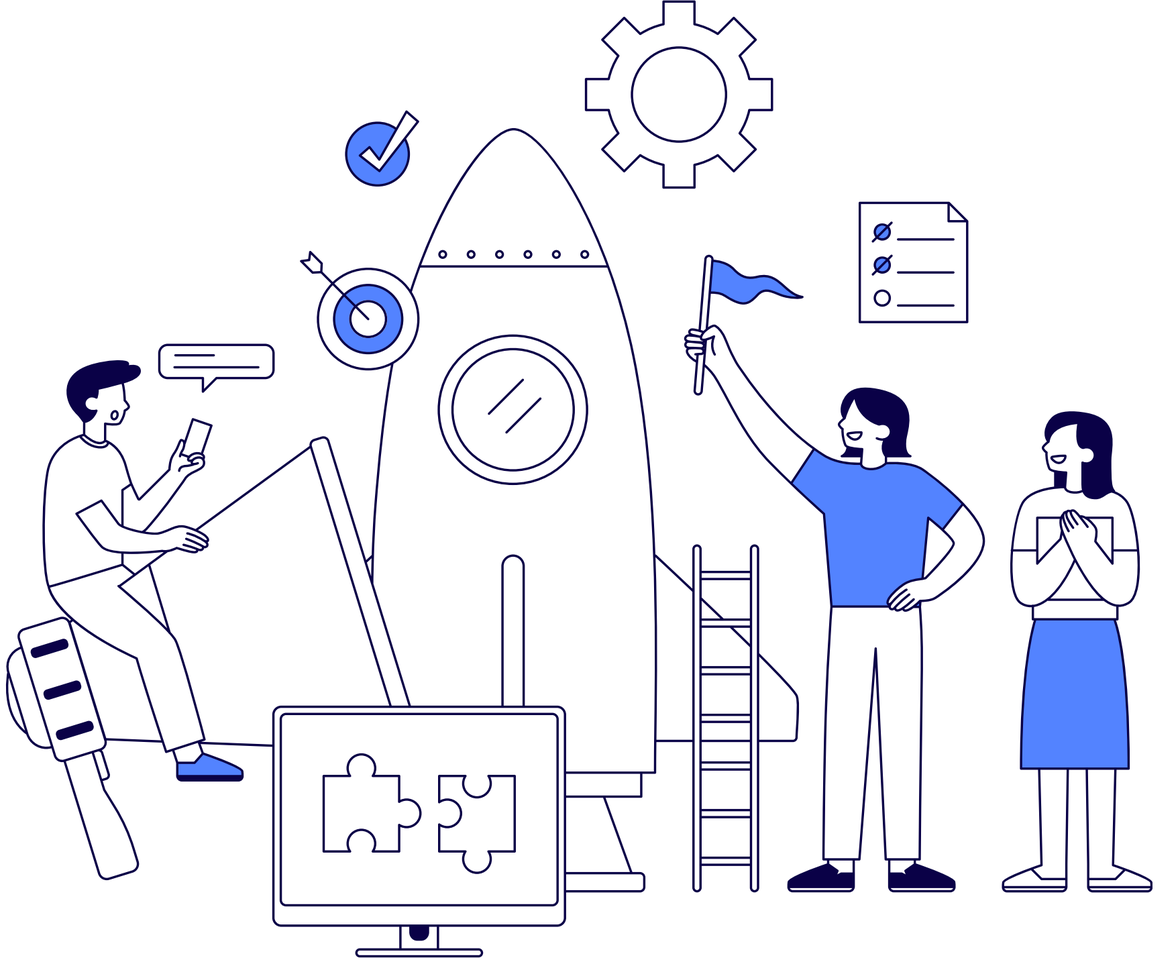
Where it fits?

The most common business use-cases for Azure Synapse Analytics are:

- *Data Warehouse*: Ability to integrate with various data platforms and services.
- *Descriptive/Diagnostic Analytics*: Use T-SQL queries against the Synapse database to perform data exploration and discovery.
- *Realtime Analytics*: Azure Synapse Link enables integration with disparate operational data sources to implement real-time analytics solutions.
- *Advanced Analytics*: Uses Azure Databricks to support decision-making by leveraging Azure Databricks.
- *Reporting & Visualization*: Integrate with PowerBI to empower and enhance business decision-making.

Companies Currently Using Azure Synapse

COMPANY NAME	WEBSITE	HQ ADDRESS	CITY	STATE	ZIP	COUNTRY
Programmers.io	programmers.io	8951 Cypress Wate...	Coppell	TX	75019-4...	US
LTIMindtree	ltimindtree.com	Global Village RVCE...	Bengaluru	KA	560002	IN
Pepsi	pepsi.com	700 Anderson Hill Rd	Purchase	NY	10577	US
Unilever	unilever.com	Unilever House 100...	London	England	EC4Y 0...	GB
Amentum	amentum.com	20501 Seneca Mea...	Germantown	MD	20876	US
Combined Insurance	combinedinsurance....	8750 Bryn Mawr Ave	Chicago	IL	60631	US



CERTIFICATIONS

EXAM DP-203

HANDS ON PROJECT FOR DATA ENGINEERS USING ALL THE SERVICES AVAILABLE IN AZURE SYNAPSE ANALYTICS (DP-203, DP-500)

INGOUDE
COMPANY

THANK YOU

Name: Soumya Ranjan Nayak

PRN: 23070243063

Hands-on Azure Synapse Analytics

```
SELECT
    TOP 100 *
FROM
    OPENROWSET(
        BULK
'https://testingsql101.blob.core.windows.net/testingfile101/used_cars_data.csv
',
        FORMAT = 'CSV',
        HEADER_ROW = TRUE,
        PARSER_VERSION = '2.0'
    ) AS [result]

SELECT
    Location, COUNT(*) as total
FROM
    OPENROWSET(
        BULK
'https://testingsql101.blob.core.windows.net/testingfile101/used_cars_data.csv
',
        FORMAT = 'CSV',
        HEADER_ROW = TRUE,
        PARSER_VERSION = '2.0'
    ) AS [result]
    group by Location


SELECT
    S.No., max(price)
FROM
    OPENROWSET(
        BULK
'https://testingsql101.blob.core.windows.net/testingfile101/used_cars_data.csv
',
        FORMAT = 'CSV',
        HEADER_ROW = TRUE,
        PARSER_VERSION = '2.0'
    ) AS [result]
    group by Location
```