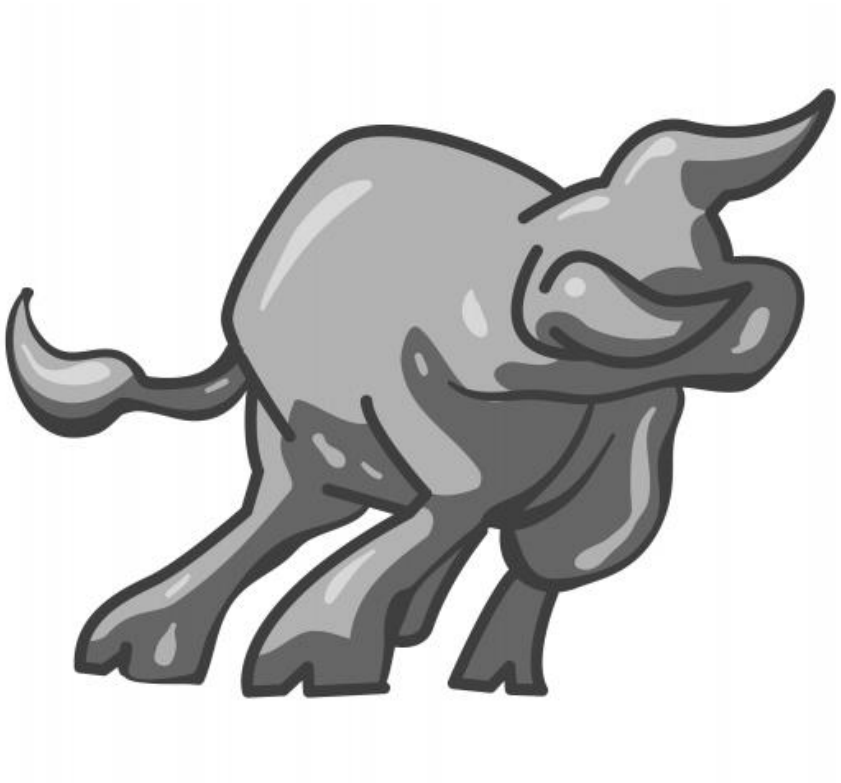
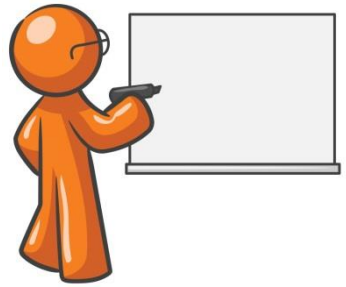
# Azure Synapse Analytics

SQLBI møde nr. 49 - 16. september 2020



ORANGEMAN

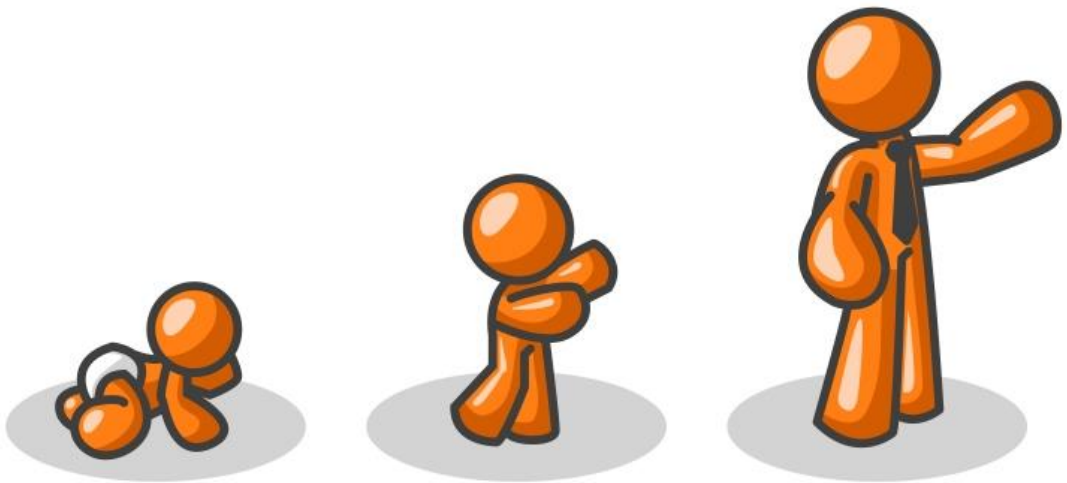
# Agenda



- Modern DW architecture
- Azure Synapse Analytics
- SQL on-demand
- SQL pool (Azure SQL DW)
- MPP architecture
- Data loading with PolyBase
- Apache Spark pool

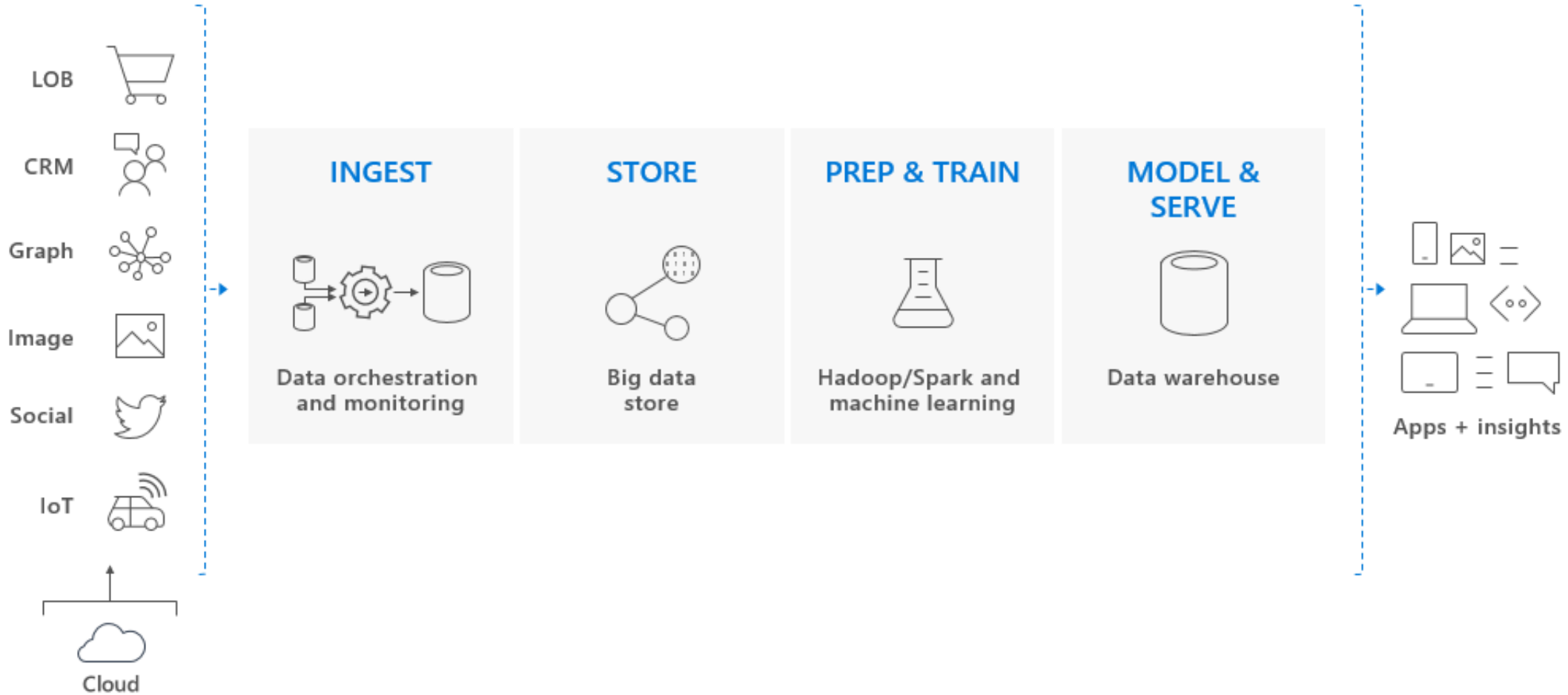
ORANGEMAN

# Modern DW architecture



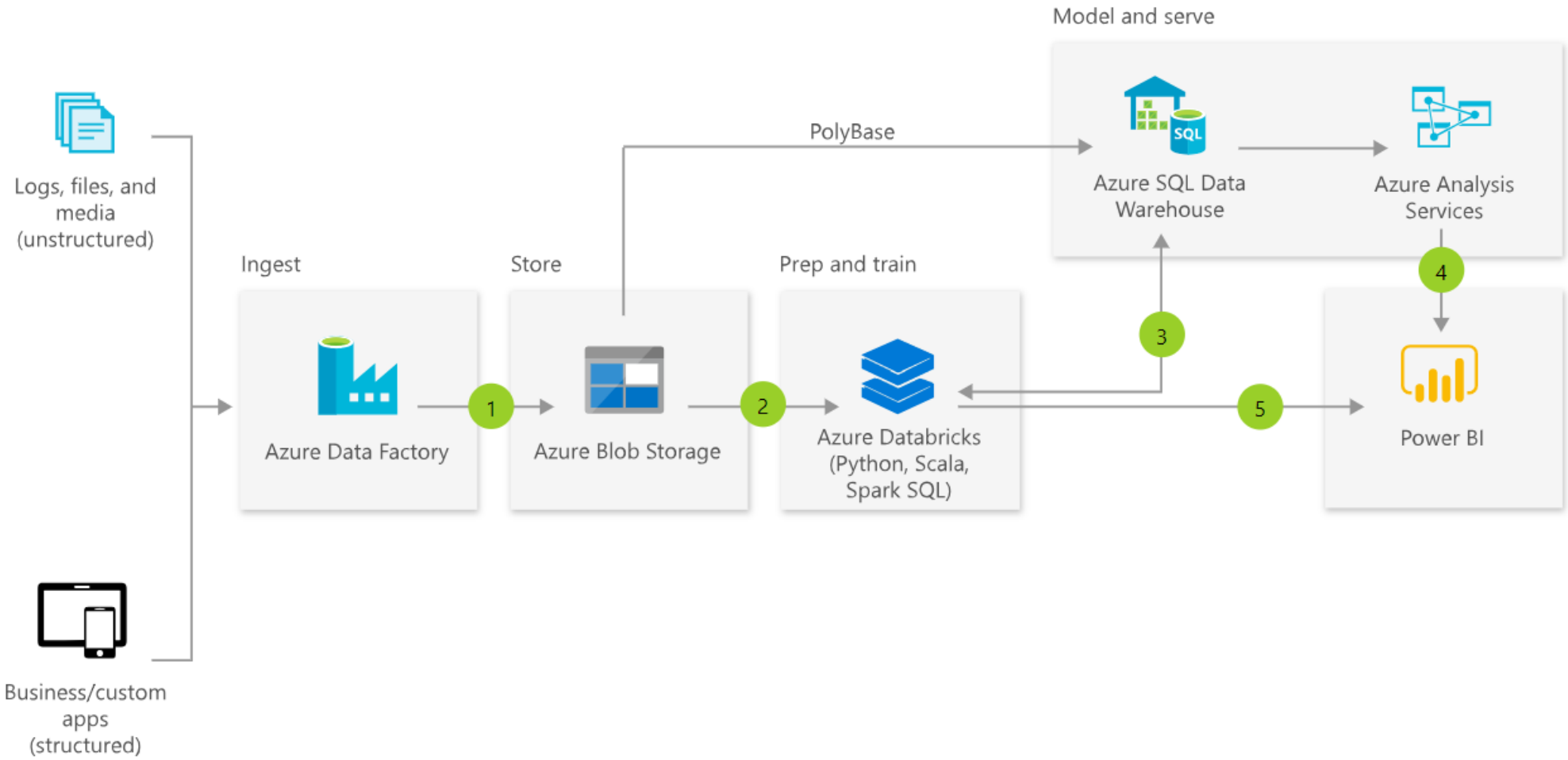
ORANGEMAN

# Key component of big data solution



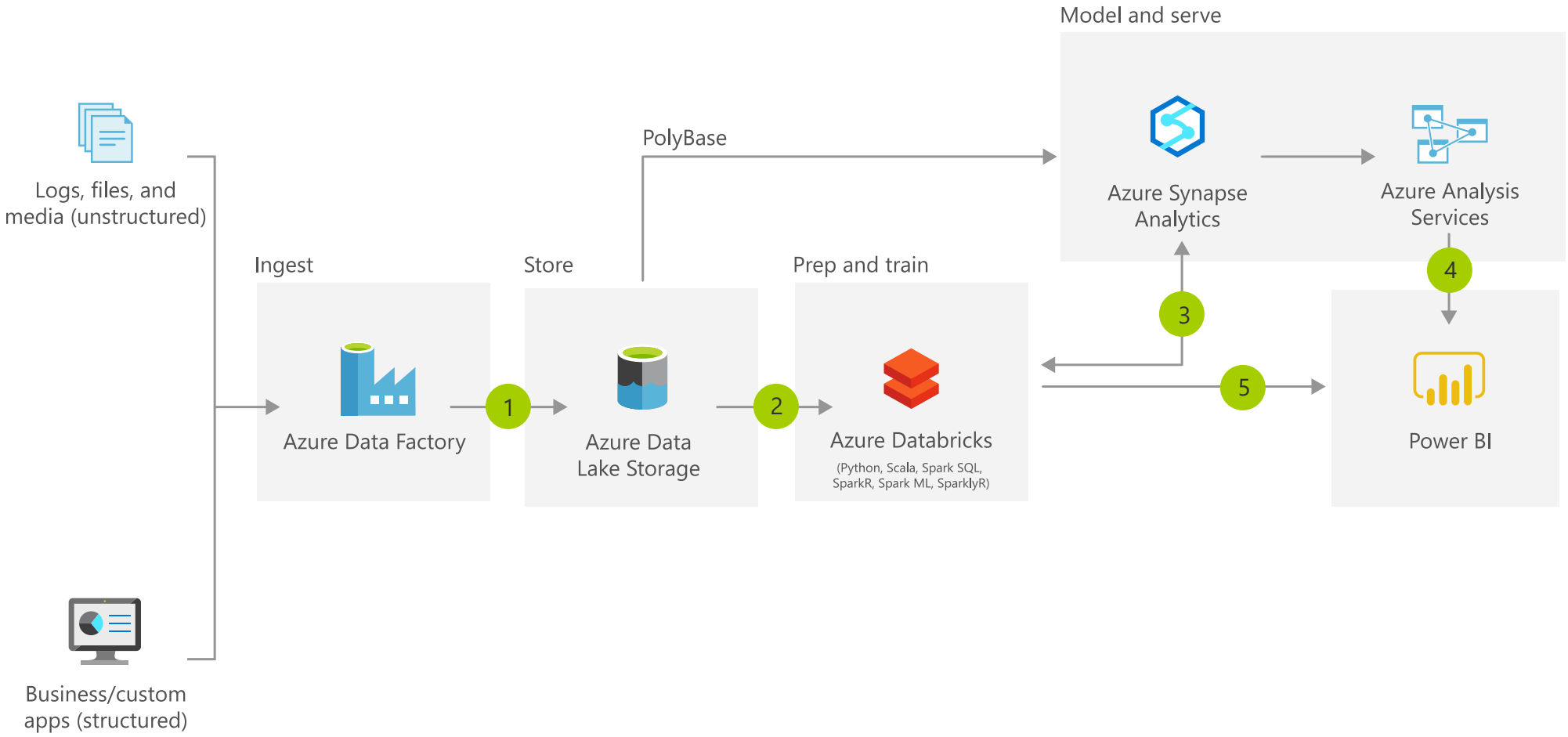

ORANGEMAN

# Modern DW architecture v1



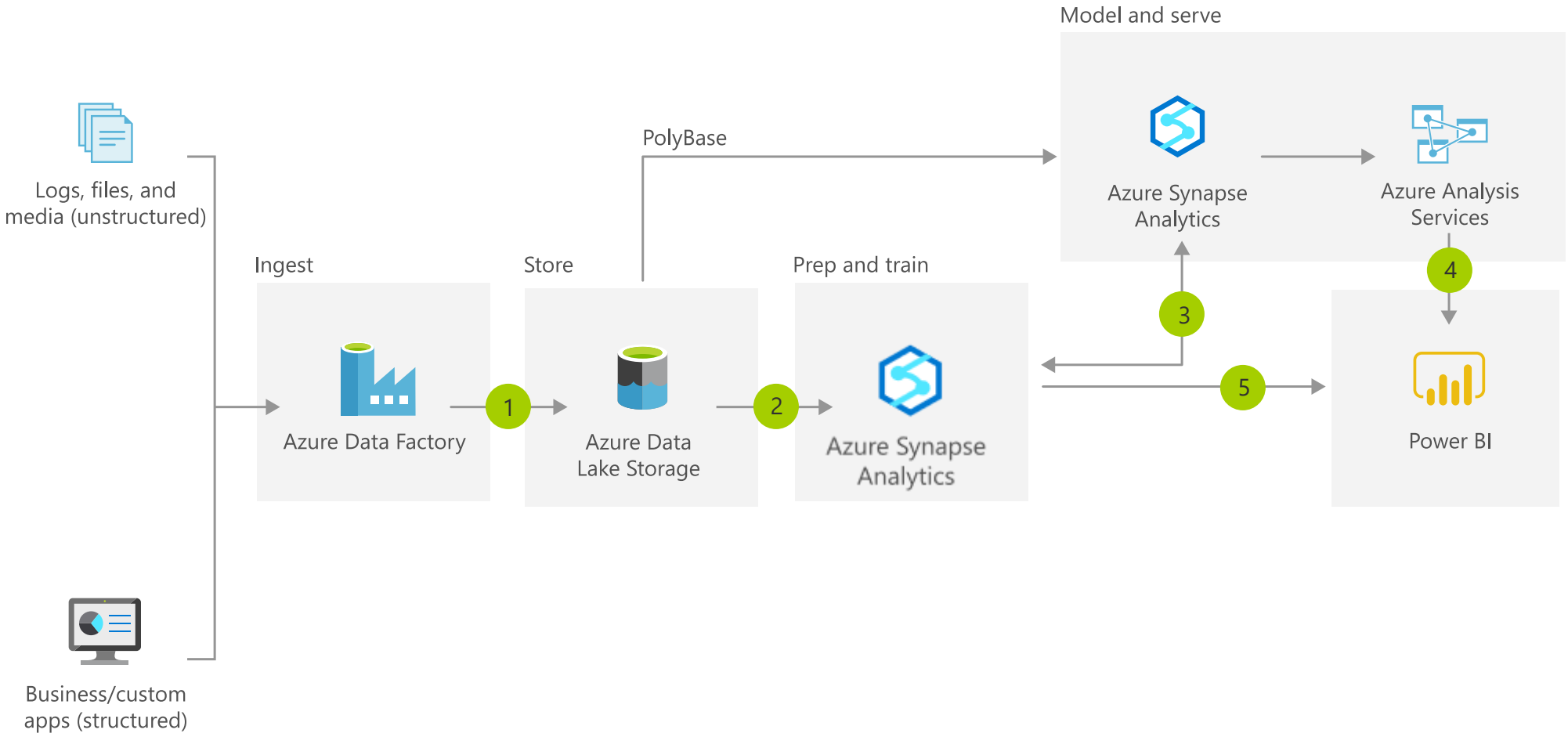

ORANGEMAN

# Modern DW architecture v2




ORANGEMAN

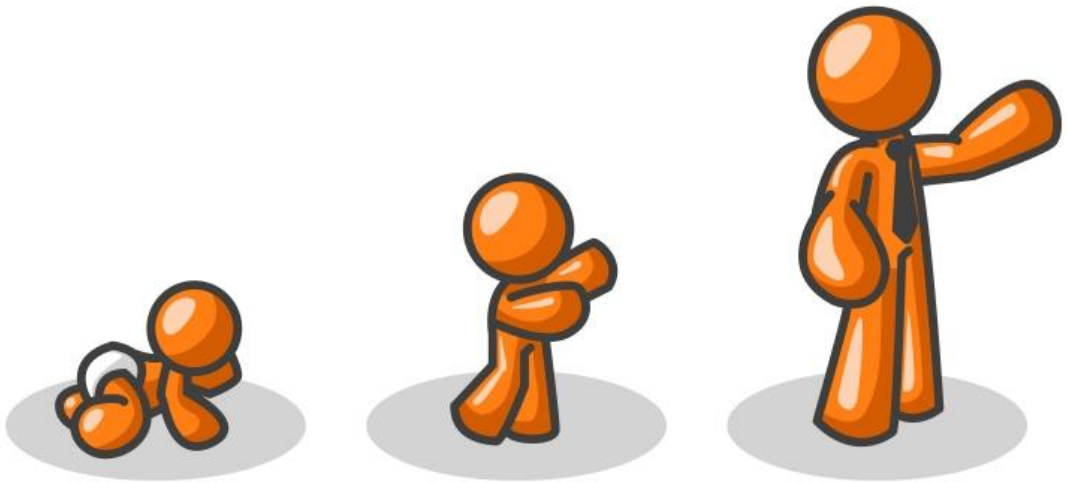
# Modern DW architecture v3 ?




ORANGEMAN

# Azure Synapse Analytics

- Preview….



ORANGEMAN

# Azure Synapse Analytics

Limitless analytics service with unmatched time to insight

***Azure Synapse is Azure SQL Data Warehouse evolved****—blending big data, data warehousing, and data integration into a* ***single service*** *for end-to-end analytics at cloud scale.*

ORANGEMAN

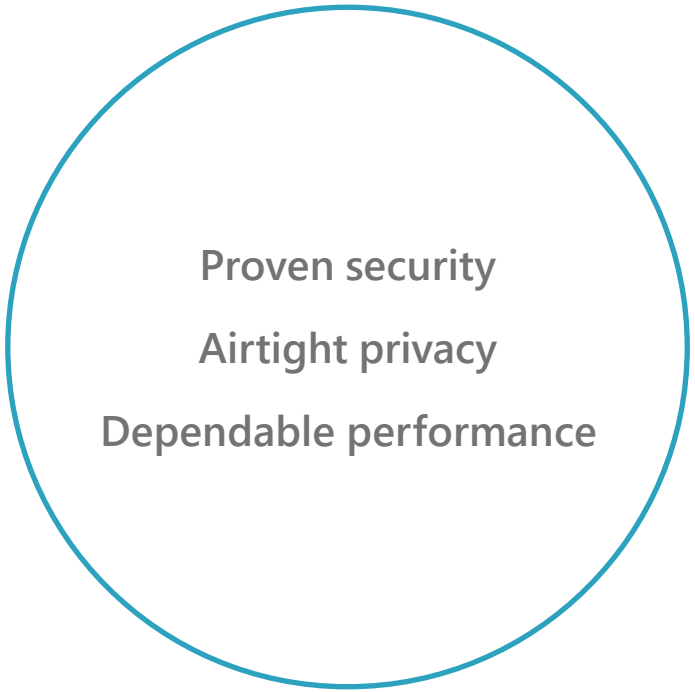
# This is a result of businesses being forced to maintain two critical, yet independent analytics systems

Big Data


Data Lake

OR

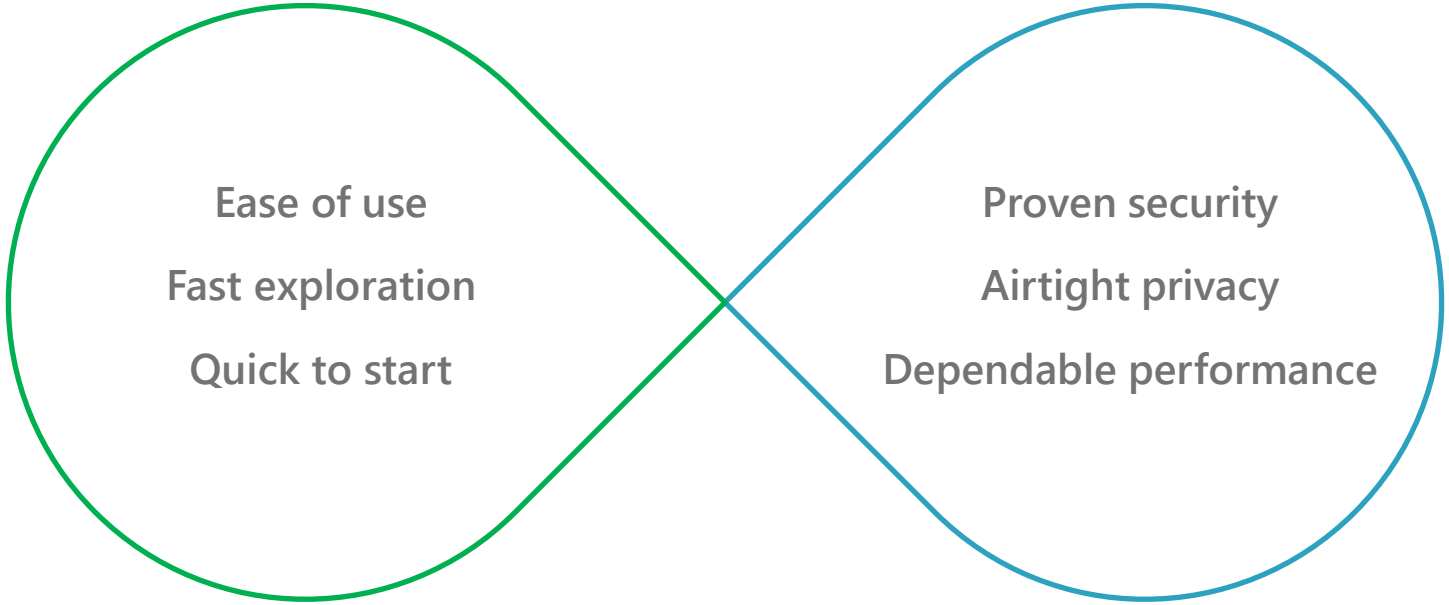
Relational Data




Data Warehouse

ORANGEMAN

# Azure brings these two worlds together, in a single service

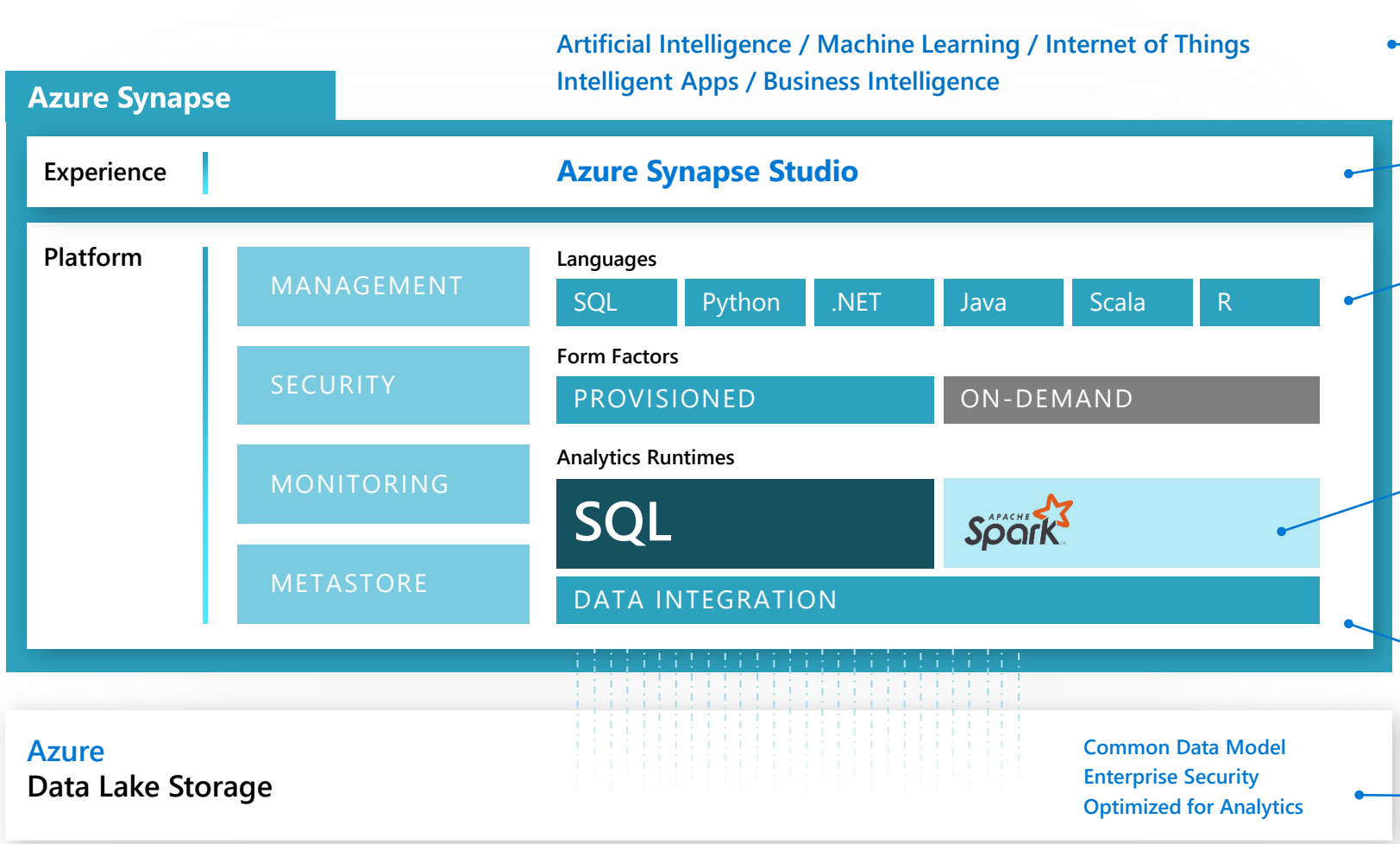



# Welcome to Azure Synapse Analytics

ORANGEMAN

# Azure Synapse Analytics

Limitless analytics service with unmatched time to insight




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 provisioned and serverless on-demand

**SQL Analytics** offering T-SQL for batch, streaming and interactive processing

**Spark** for big data processing with Python, Scala, R and .NET

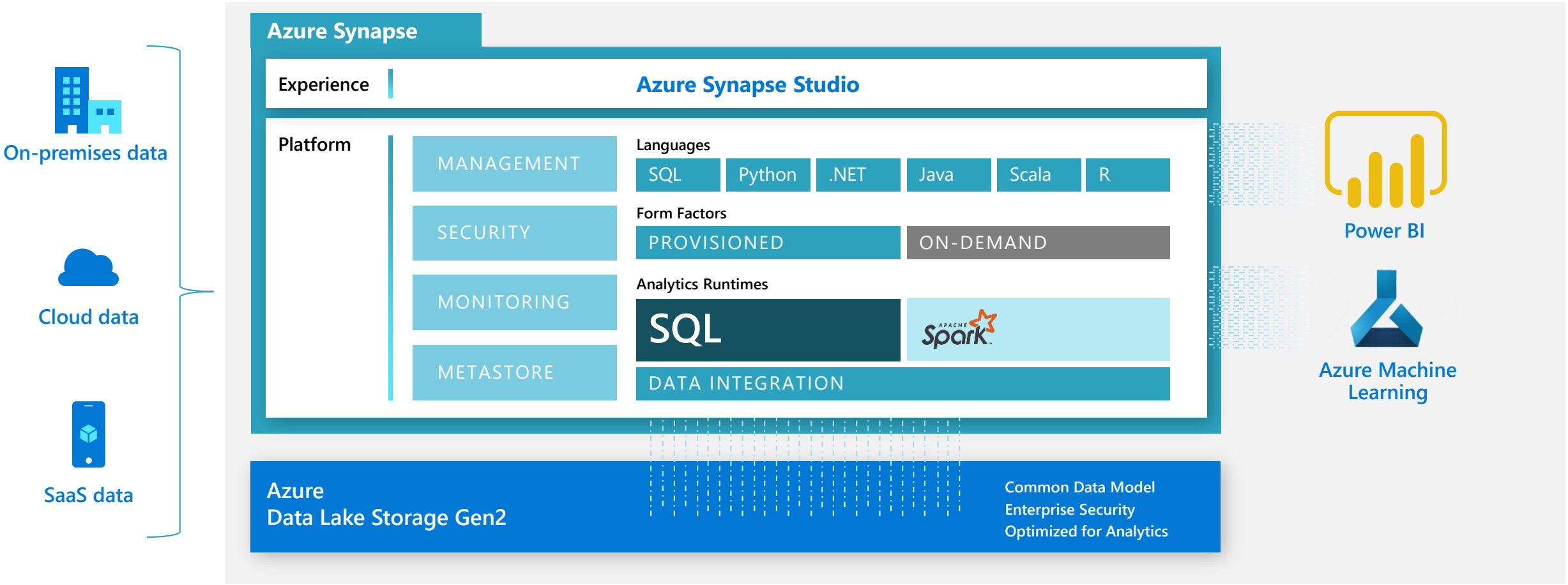
Integrated **platform services** for, management, security, monitoring, and metastore

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

ORANGEMAN

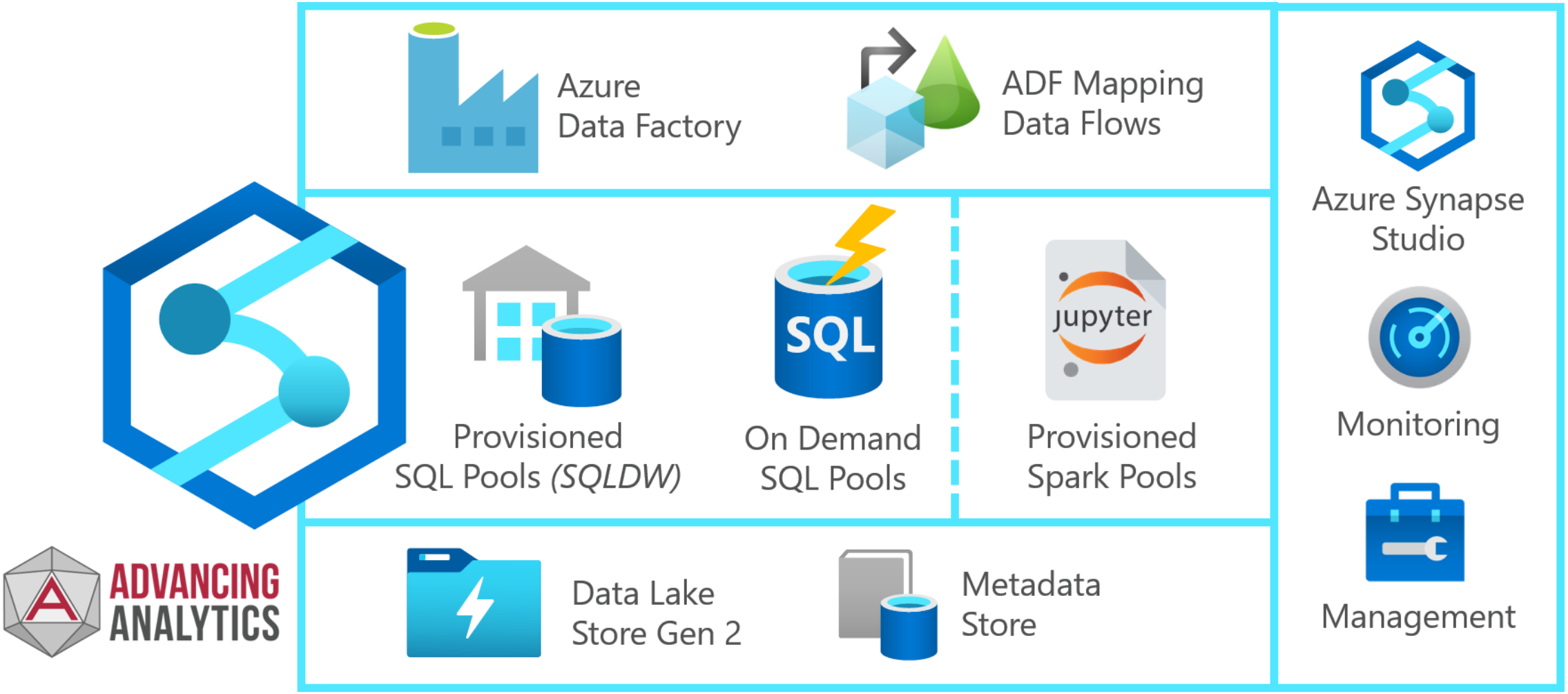
# Analytics using Azure Synapse

Ingest and store data from multiple sources → Prepare your data → Analyze the prepared data → Serve your data with high concurrency support → Build BI dashboards and reports



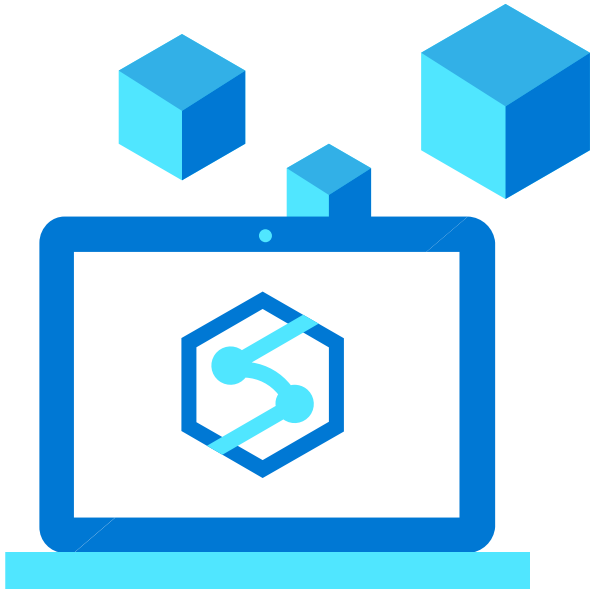

ORANGEMAN

# Analytics using Azure Synapse



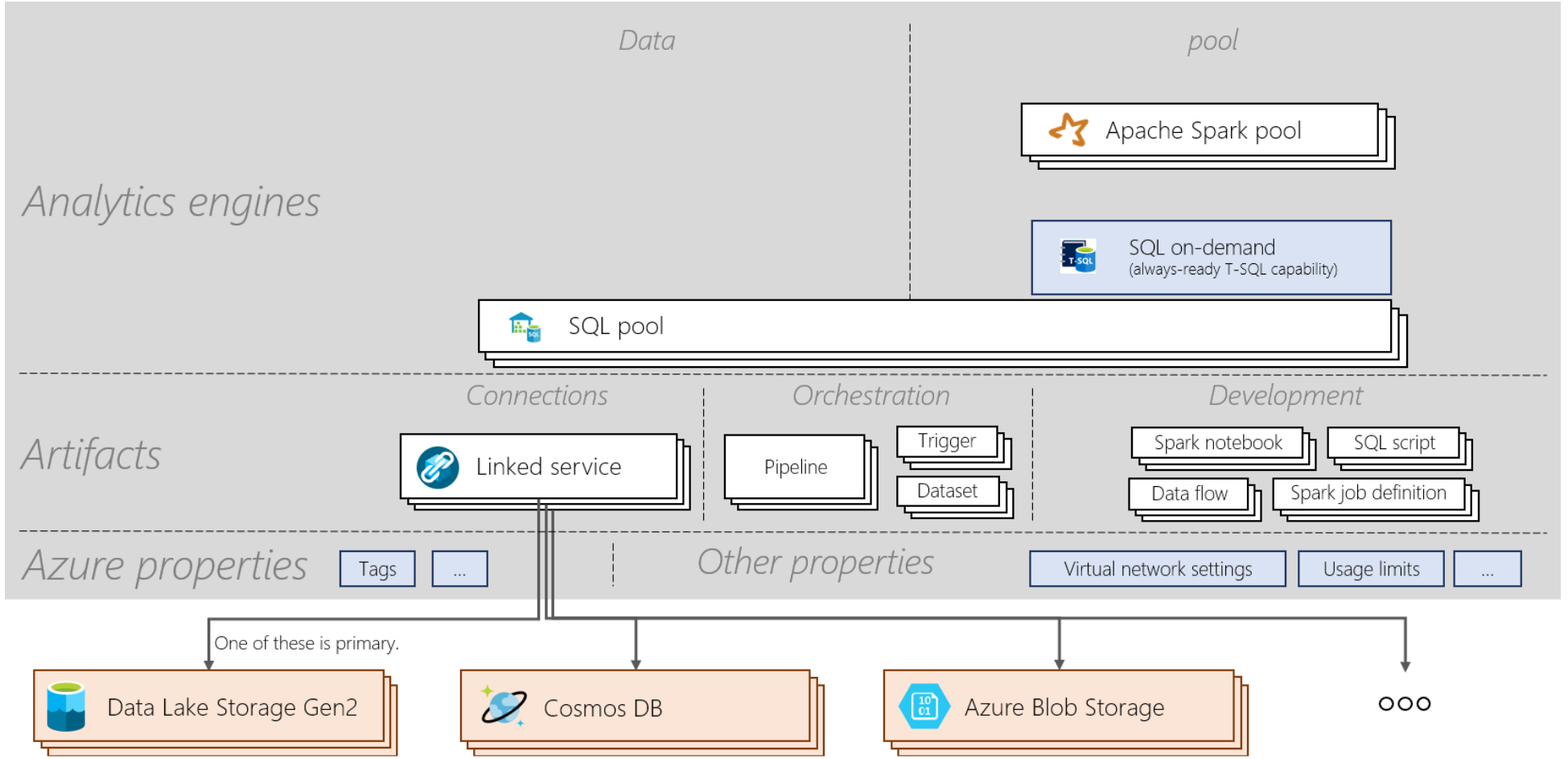
# Synapse Studio

- Makes modern data warehouse solution-building easier than ever!
- Synapse Studio combines ingestion, preparation, analysis, and serving.
- Develop notebooks, SQL scripts, pipelines, Power BI reports, and more… all within one experience.
- Solution-level authoring and monitoring in a single pane of glass.
- Spark and SQL can operate on the same data. No need to duplicate.



ORANGEMAN

# Azure Synapse Analytics Achitecture



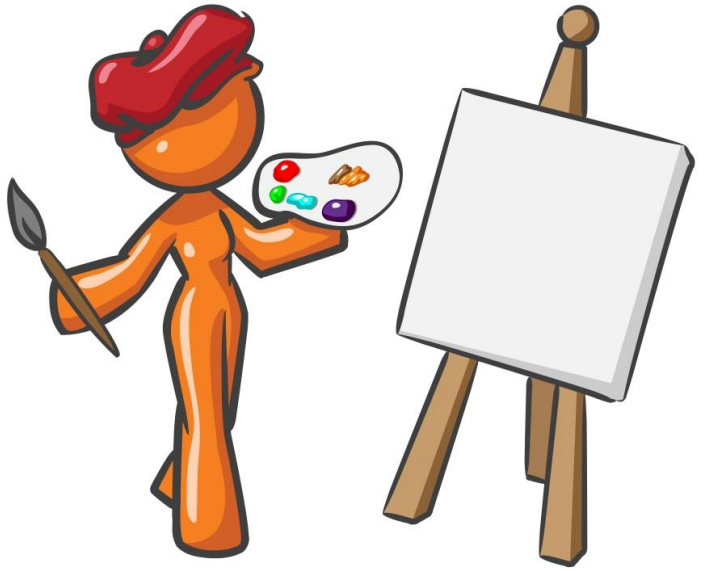

ORANGEMAN

# Azure Synapse Analytics Terminology

- **Synapse workspace:** A securable collaboration boundary. Has an associated ADLS Gen2 account and file system (for temporary data).
- **SQL on-demand:** Distributed data processing system that lets you run T-SQL queries over data in data lake. It is serverless.
- **SQL pool (SQL DW):** 0-to-N SQL provisioned resources with their corresponding databases
- **Apache Spark pool:** 0-to-N Spark provisioned resources with their corresponding databases.
- **Data Integration (Data Factory):** Ingest data between various sources and orchestrate activities running within or outside a workspace

ORANGEMAN

# Demo

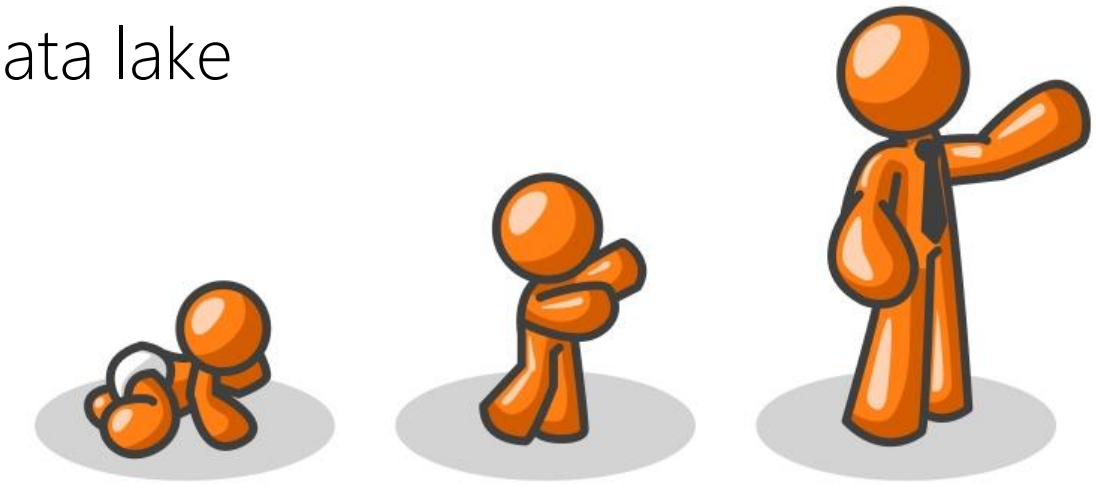


1. Create a Workspace and use existing Data Lake
2. Setup firewall rules
3. Set security on Data Lake
   - Storage Blob Reader or Contributor role
4. Launch Synapse Studio
   - Browse storage account
   - Query PARQUET file with the use of SQL on-demand

Quickstart: https://docs.microsoft.com/en-us/azure/synapse-analytics/quickstart-create-workspace and https://docs.microsoft.com/en-us/azure/synapse-analytics/quickstart-synapse-studio

# SQL on-demand in Synapse

- Query service over the data in your data lake



ORANGEMAN

# SQL on-demand in Synapse

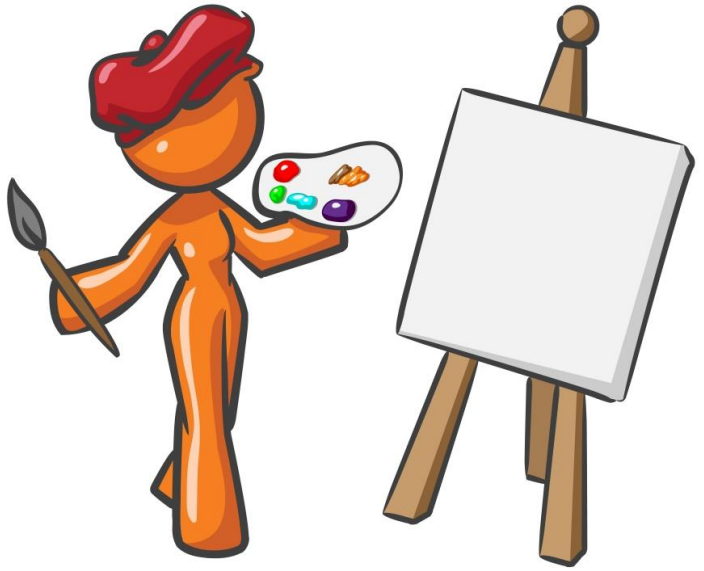
- Allows you to query files in your Azure storage accounts
- No local storage, only metadata objects are stored in databases
- Analyze your Big Data in seconds to minutes
- Supported T-SQL
  - Full SELECT surface area is supported, including a majority of SQL functions
  - CETAS - CREATE EXTERNAL TABLE AS SELECT
  - DDL statements related to views and security only
- There is no charge for resources reserved, you are only being charged for the data scanned by queries you run, hence this model is a true pay-per-use model. ($5 per TB of data processed)

# SQL on-demand scenarios

- **Basic discovery and exploration** - Quickly reason about the data in various formats (Parquet, CSV, JSON) in your data lake, so you can plan how to extract insights from it.
- **Logical data warehouse** – Provide a relational abstraction on top of raw or disparate data without relocating and transforming data, allowing always up-to-date view of your data.
- **Data transformation** - Simple, scalable, and performant way to transform data in the lake using T-SQL, so it can be fed to BI and other tools, or loaded into a relational data store (Synapse SQL databases, Azure SQL Database, etc.).

# Demo

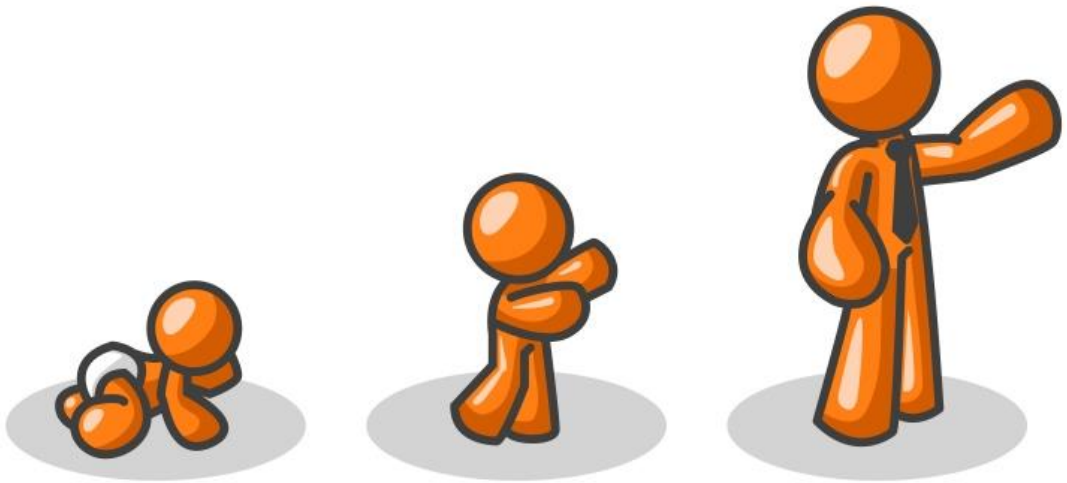
1. Create a Database
2. Create a data source
3. Query files
4. Create an external table (CETAS)



Quickstart: https://docs.microsoft.com/en-us/azure/synapse-analytics/quickstart-sql-on-demand
and https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/create-external-table-as-select

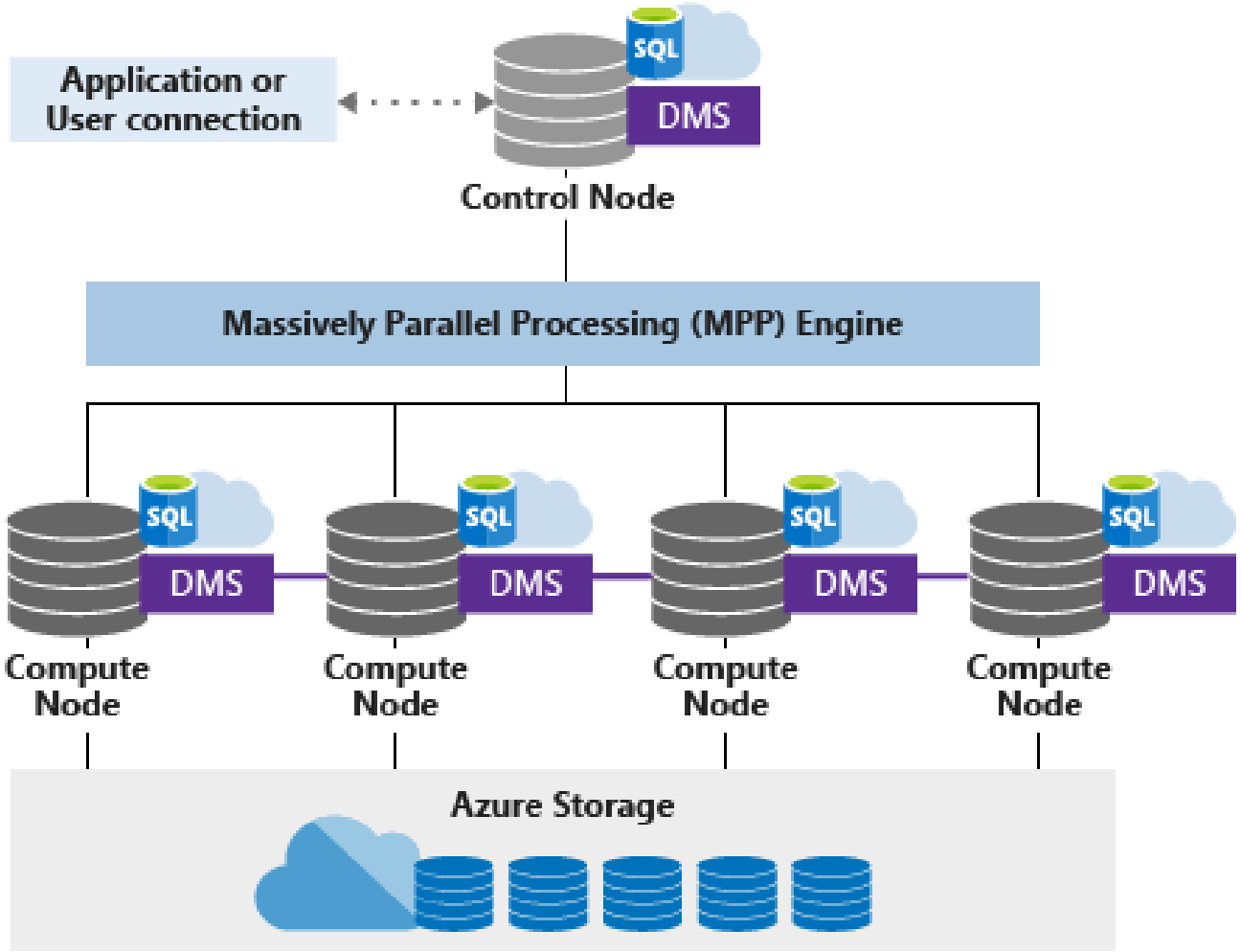
# SQL pool in Synapse

- Azure SQL Data Warehouse



ORANGEMAN

# MPP architecture components




ORANGEMAN

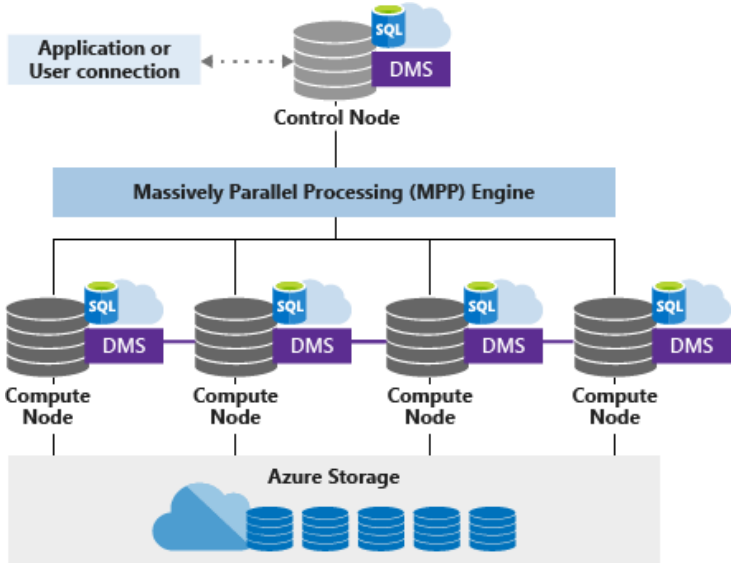
# Decoupled storage and compute

- Independently size compute power irrespective of your storage needs.
- Grow or shrink compute power without moving data.
- Pause compute capacity while leaving data intact, so you only pay for storage.
- Resume compute capacity during operational hours.


ORANGEMAN

# MPP architecture components

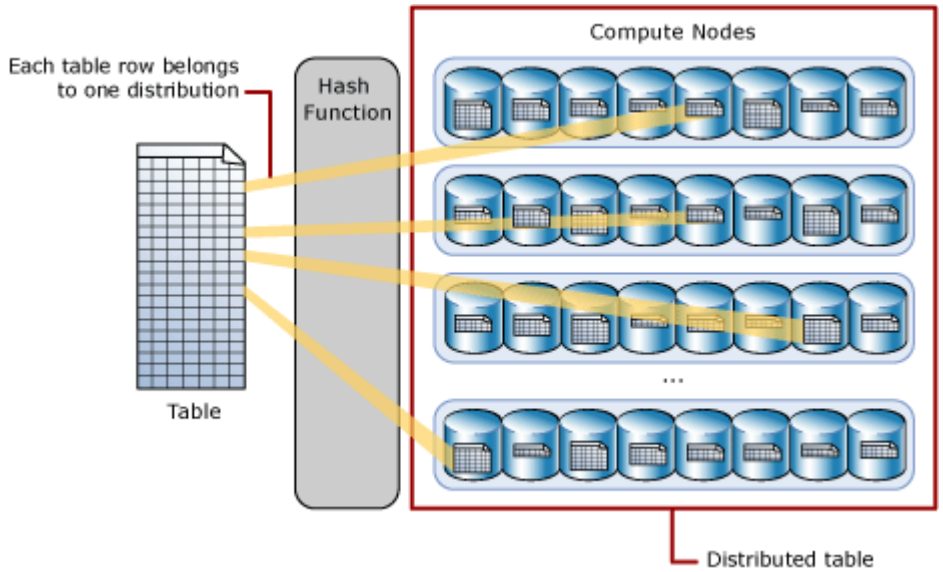
- Azure storage
  - Charges separately for your storage consumption
  - Sharded into distributions
- Control node
  - The brain of the data warehouse
  - Optimize and coordinate parallel queries
- Compute nodes
  - Provide the computational power
  - Ranges from 1 to 60 - determined by the service level
- Data Movement Service
  - Ensures the right data gets to the right location.

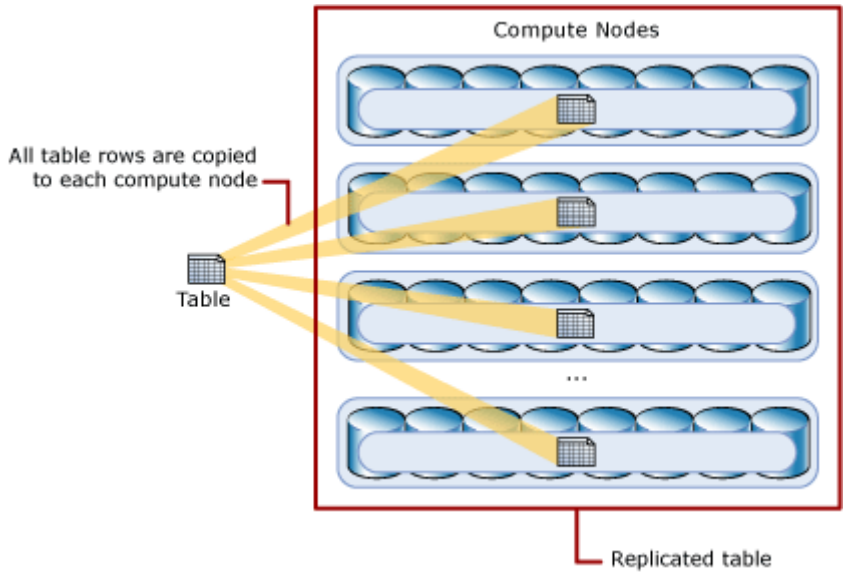



ORANGEMAN

# Distributions

- Hash-distributed tables
- Round-robin distributed tables
- Replicated Tables




ORANGEMAN

# Enterprise Grade Cloud Data Warehouse

| On-demand | Secure | Scalable |
| --- | --- | --- |
| Automated provisioning in seconds | Virtual Networks | 4,000+ core compute scale |
| Re-size compute in minutes | Firewalls | Infinite data storage |
| Create multiple instances | Azure Active Directory, MFA | Compute scale independent of storage |
| Pause and Resume warehouses | Transparent Data Encryption | Adaptive caching |
| Consume compute and storage independently | Object level security | Instant data movement |
| | Azure Key Vault | 128 active concurrent queries, 1024 connections per cluster |
| | Auditing and Threat Detection | |
| | Vulnerability Assessment | |
| | Compliance | |

ORANGEMAN

# Enterprise Grade Cloud Data Warehouse

| Provisioning | Administration | Tuning | Backup / Restore |
| --- | --- | --- | --- |
| Provision and load through integrated portal | Automatic columnar index optimization | Multi-level cost-based query optimization | Automated backup snapshots every 8 hours |
| Portal and programmatic provisioning | Auto-create statistics | Graphical execution plans | Geo-redundant backups every 24 hours |
| Provision in 33 global regions | Automatic Intelligent Insights | Resource utilization/contention | User-defined restore points |
| | One-click scale-up during peak demand | Data distribution/skew | |
| | Maintenance Windows | | |

ORANGEMAN

# SQL Data Warehouse Gen2

A next-generation data warehouse solution that extends Microsoft's lead in cloud data warehouse price/performance with better performance, greater scale, and higher query concurrency.

You can upgrade your existing data warehouse to Gen2 without any additional costs – giving you even greater price/performance.

## 5x improvement in performance

Better performance for your data and analytics workloads

## 4x concurrent queries

Now up to 128 concurrent queries for expanding your data warehouse

## 5x scalability

Now up to 30,000 cDWU giving you more compute capabilities for your data assets

## Retains all elastic functionality

Pause, Resume and Scale operations

ORANGEMAN

# SQL Data Warehouse SKU recommendations

| SQL DW SKU | DWU | Active Dataset (TB) | Max Concurrency |
|---|---|---|---|
| Compute Optimized Gen2 tier | 30000c | >= 90 | 128 |
| | 15000c | 45 | 128 |
| | 10000c | 30 | 128 |
| | 7500c | 22.5 | 128 |
| | 6000c | 18 | 128 |
| | 5000c | 15 | 64 |
| | 3000c | 9 | 64 |
| | 2500c | 7.5 | 48 |
| | 2000c | 6 | 48 |
| | 1500c | 4.5 | 32 |
| | 1000c | 3 | 32 |
| Compute Optimized Gen1 tier | 600 | NA | 24 |
| | 500 | NA | 20 |
| | 400 | NA | 16 |
| | 300 | NA | 12 |
| | 200 | NA | 8 |
| | 100 | NA | 4 |

ORANGEMAN

# Azure SQL Data Warehouse Gen2 pricing

| | DWU | PRICE |
|---|---|---|
| DW100c | 100 | $1.51/hour |
| DW500c | 500 | $7.55/hour |
| DW1000c | 1000 | $15.10/hour |
| DW2000c | 2000 | $30.20/hour |
| DW3000c | 3000 | $45.30/hour |
| DW5000c | 5000 | $75.50/hour |
| DW10000c | 10000 | $151/hour |
| DW15000c | 15000 | $226.50/hour |
| DW30000c | 30000 | $453/hour |

| SQL Server 2019 Enterprise | |
|---|---|
| 2 cores | $28,512 |
| 4 cores | $57,024 |
| 8 cores | $114,048 |
| 16 cores | $228,096 |

**Price example**

DW500c * 5 hours * 365

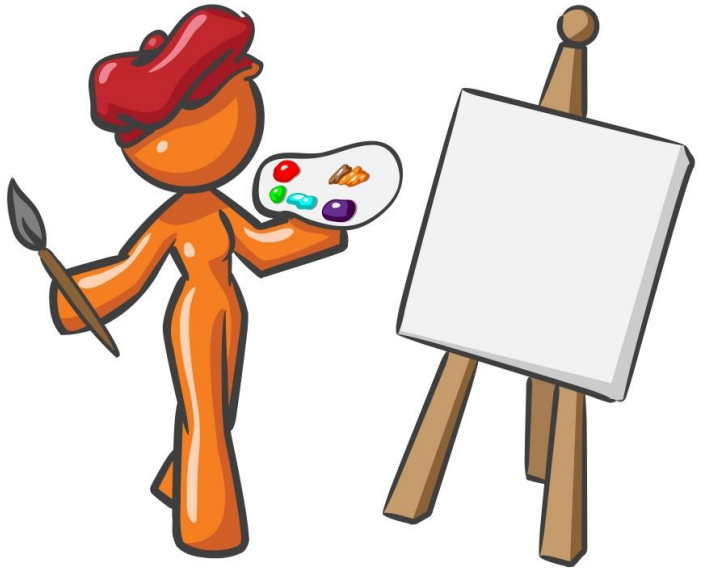
7.55 * 5 * 365 = 13,778.75

- Data storage is charged at the rate of $148.68/1 TB/month ($0.21/1 TB/hour)
- Your data warehouse is copied to geo-redundant storage for disaster recovery. Disaster recovery storage is billed starting at $0.12/GB/month.

ORANGEMAN

# Demo

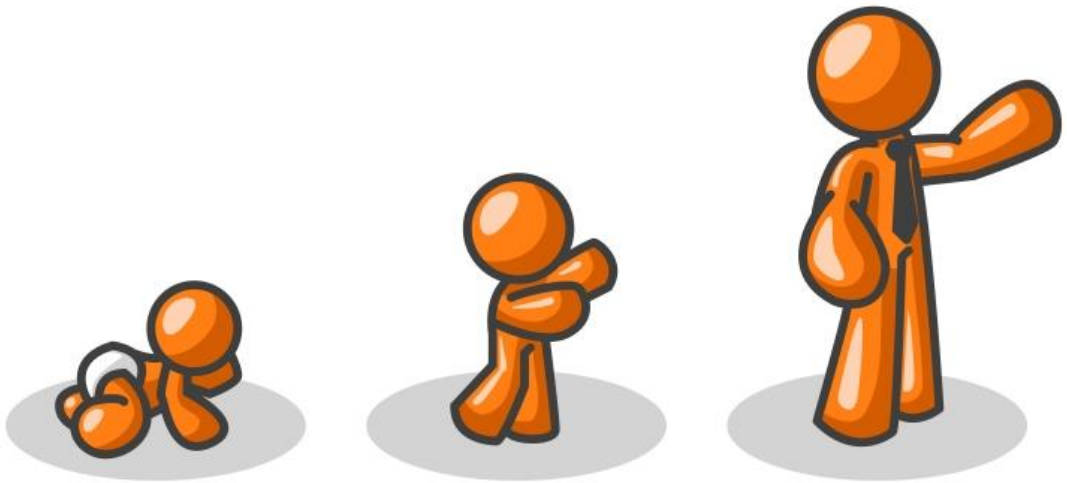
1. Create a Synapse SQL pool
2. Bulk Load data with the wizard
3. Connect to the SQL pool in SSMS as server admin
4. Scale compute in portal



Quickstart: https://docs.microsoft.com/en-us/azure/synapse-analytics/quickstart-create-sql-pool-portal
and https://docs.microsoft.com/en-us/azure/synapse-analytics/quickstart-load-studio-sql-pool

# PolyBase

- Data loading technique



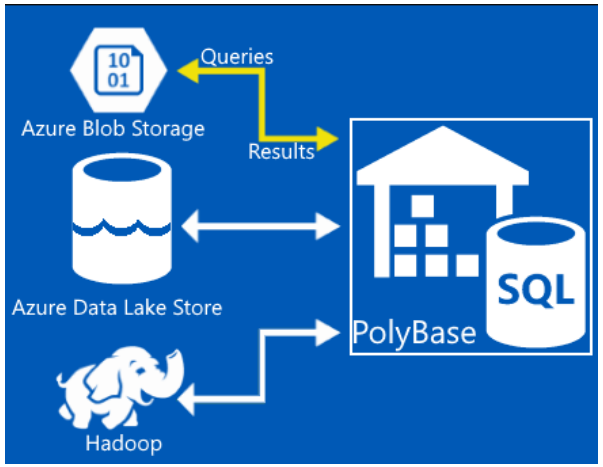
ORANGEMAN

# PolyBase

- Best practice data loading technique from Azure Data Lake Store to Azure SQL DW (Synapse Analytics)
- Massively Parallel Processing Architecture

- External data source
- External tables

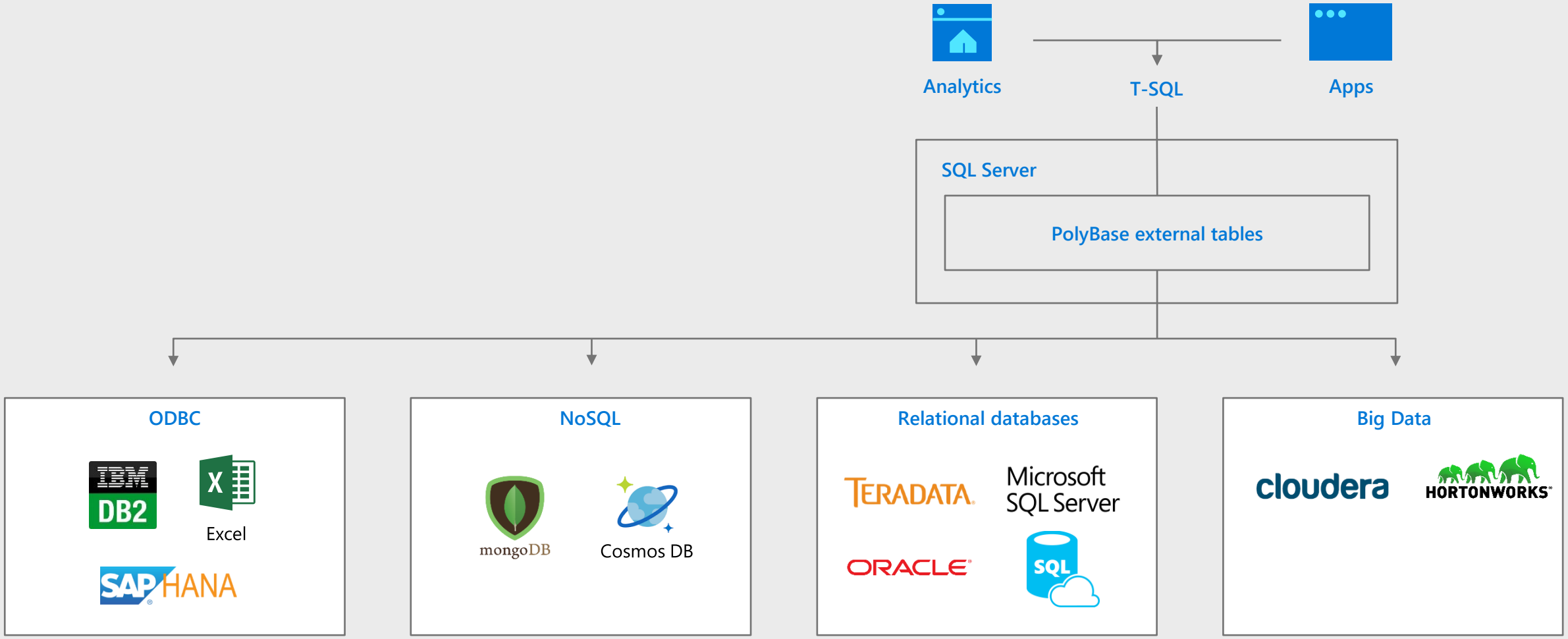
- CTAS – Create Table As Select
- COPY




ORANGEMAN

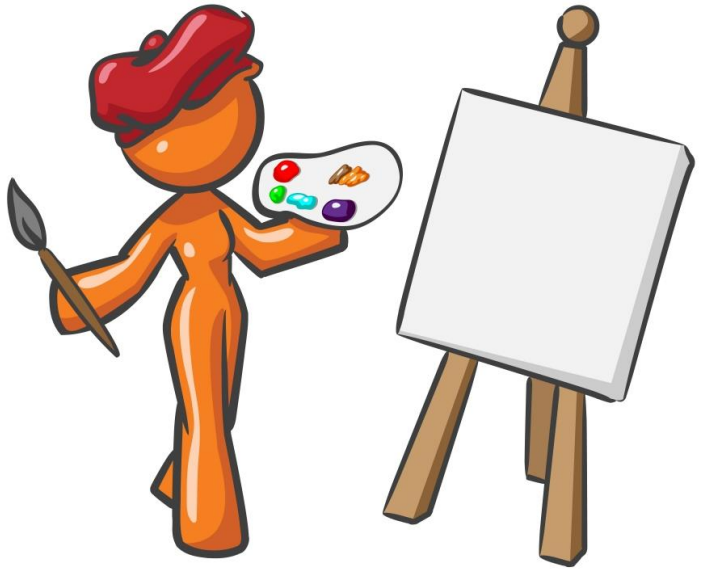
# SQL Server is the hub for integrating data

## Easily combine across relational and non-relational data stores



# Extra demos

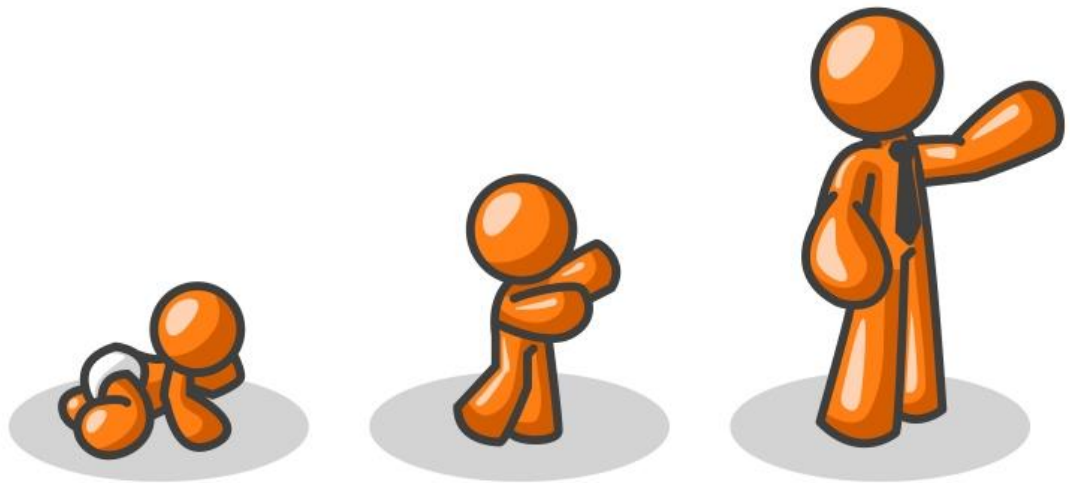
1. Configure access to external Data Lake
2. Create the External Tables
3. Load the data with CTAS or COPY



Quickstart:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-load-from-azure-data-lake-store

ORANGEMAN

# Apache Spark pool

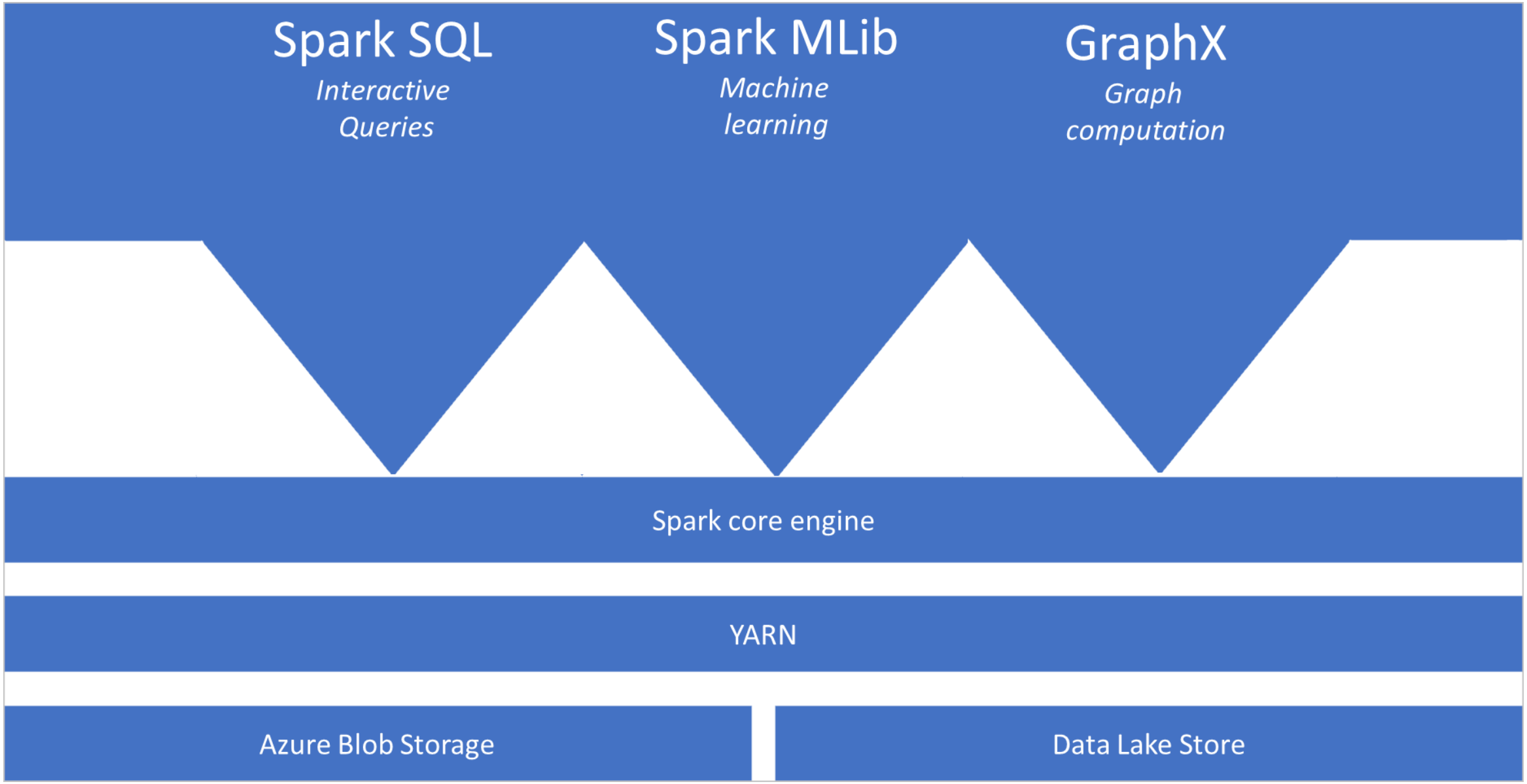


ORANGEMAN

# Industry-standard Apache Spark

- The most popular open source big data engine used for data preparation, data engineering, ETL, and machine learning
- ML models with SparkML algorithms and AzureML integration for Apache Spark 2.4
- Simplified resource model that frees you from having to worry about managing clusters.
- Fast Spark start-up and aggressive autoscaling.
- Built-in support for .NET for Spark allowing you to reuse your C# expertise and existing .NET code within a Spark application
- Price: $0.18 per vCore-hour

ORANGEMAN


Spark SQL
Interactive Queries
Spark MLib
Machine learning
GraphX
Graph computation
Spark core engine
YARN
Azure Blob Storage
Data Lake Store

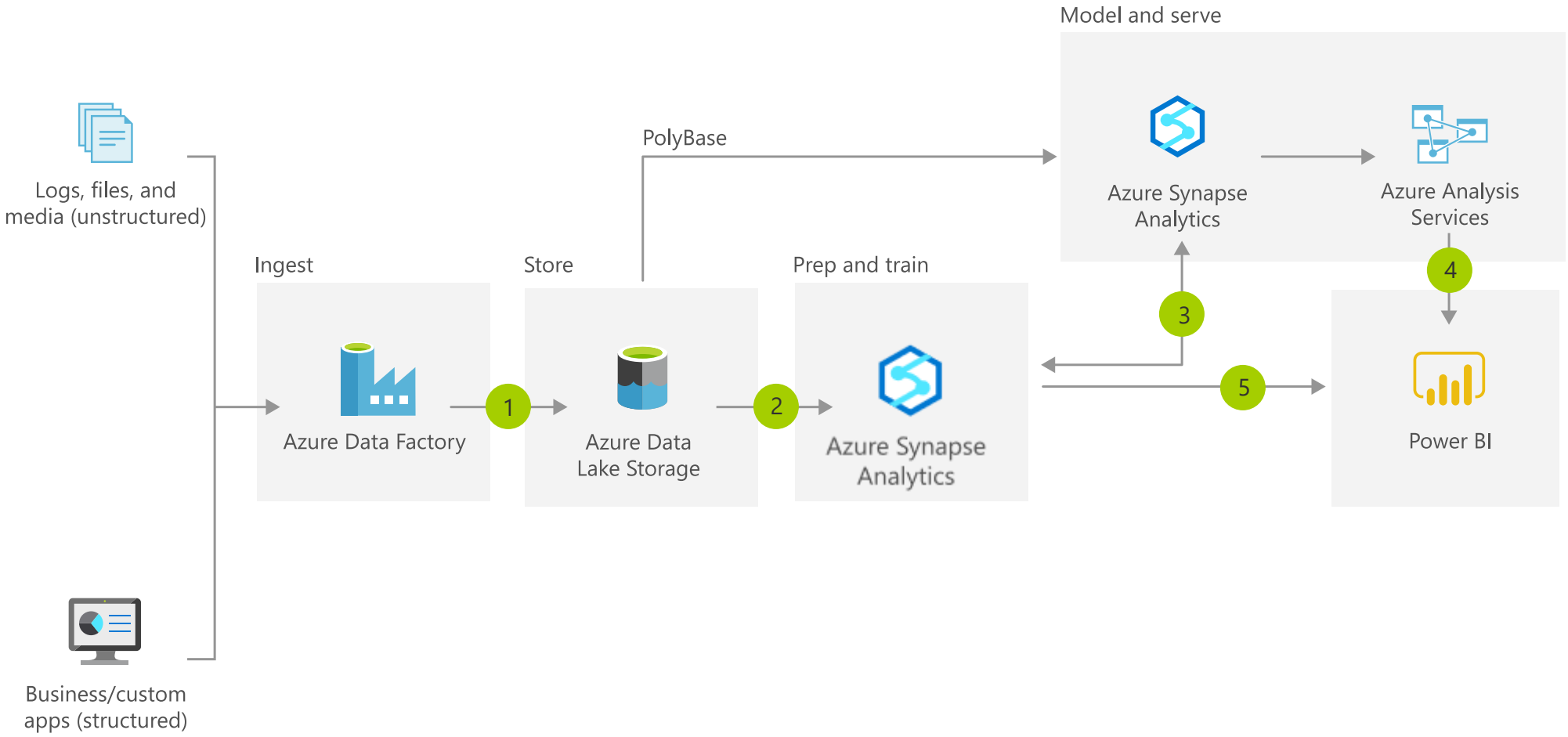
ORANGEMAN

# Demo

1. Create an Apache Spark pool
2. Create a notebook
   - New notebook
   - Add code
   - Run cell
   - View job execution

Quickstart:
https://docs.microsoft.com/en-us/azure/synapse-analytics/quickstart-create-apache-spark-pool-studio and https://docs.microsoft.com/en-us/azure/synapse-analytics/quickstart-apache-spark-notebook

# Modern DW architecture v3 ?




ORANGEMAN