Data Community

SQL Day

PLATINUM SPONSOR

TECHNOLOGY INNOVATION DATA KNOWLEDGE tidk

STRATEGIC PARTNER

Microsoft

GOLD SPONSORS

CLOUDS ON MARS

Future Processing

SPYROSOFT

SILVER SPONSOR

dbWatch DATABASE CONTROL

elitmind think bright

VEEAM

BRONZE SPONSOR

Największa księgarnia IT w Polsce

NOVATECH www.novatech.com.pl

Common Enterprise Analytics Architectures with Azure Data Services and Power BI

Radosław Łebkowski

Technology Solution Professional Data & AI

Microsoft

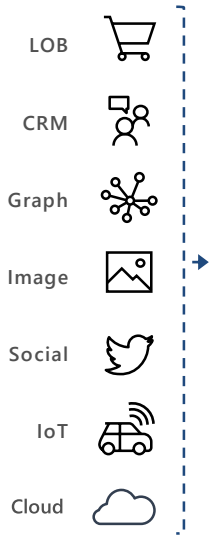
Linkedin

Session goals

- Understand the Microsoft big-data landscape
- Choose the right big-data technology for your needs
- Unlock petabyte scale datasets for interactive analysis in Power BI
- Update on Azure Data Services capabilities

Azure services challenge

Modern Data Warehouse (possible products by four areas)



INGEST

Data orchestration and monitoring

- Azure Data Factory
- SSIS

STORE

Big data store

- Azure Data Lake Storage Gen2
- Blob Storage
- Azure Data Lake Storage Gen1
- SQL Server 2019 Big Data Cluster

PREP

Transform & Clean

- Azure Databricks
- Azure HDInsight
- PolyBase & Stored Procedures
- Power BI Dataflow
- Azure Data Lake Analytics

MODEL & SERVE (& store)

Data warehouse

- Azure SQL Data Warehouse
- Azure Analysis Services
- SQL Database (Single, MI, HyperScale, Serverless)
- SQL Server in a VM
- Cosmos DB
- Power BI Aggregations

- BI + Reporting
- Advanced Analytics
- AI

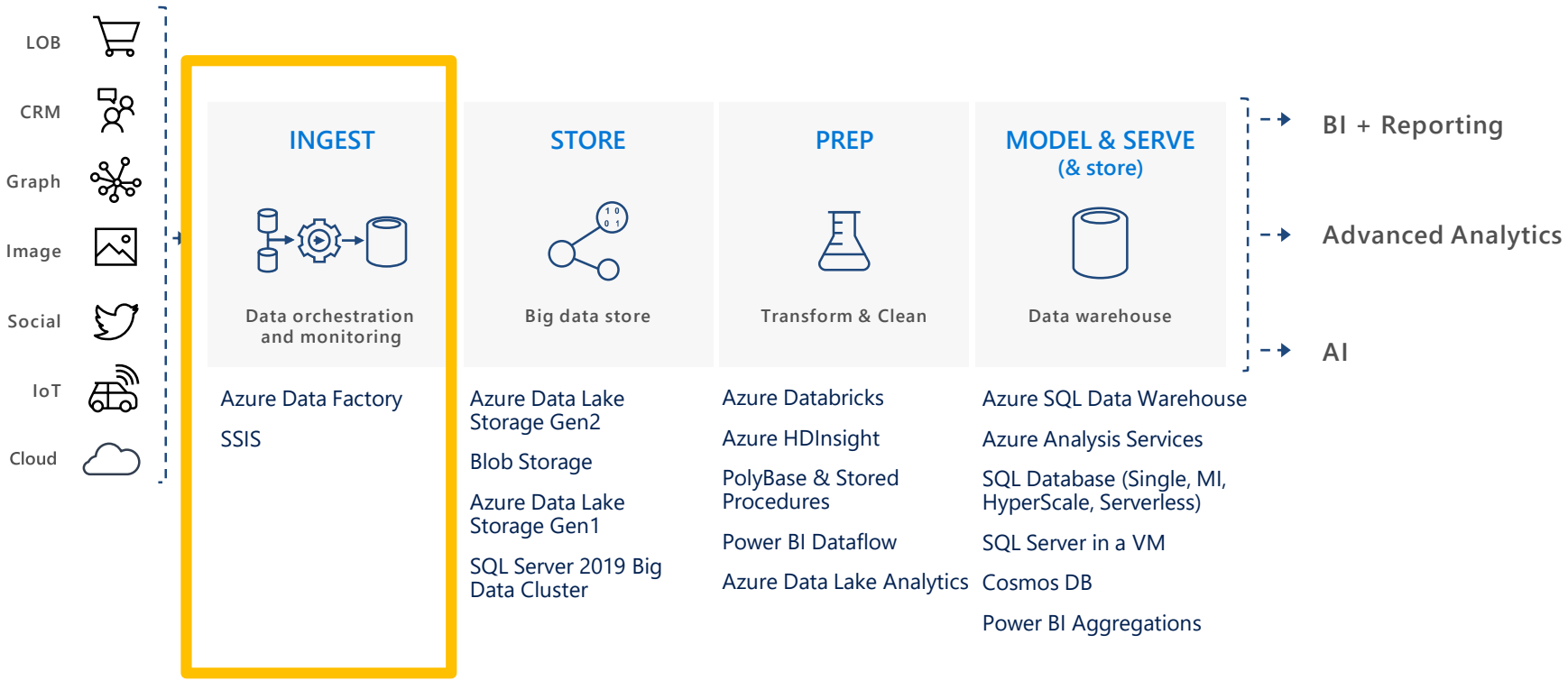
Note: Those products that span more than one area are listed in there primary area

Questions to ask customer

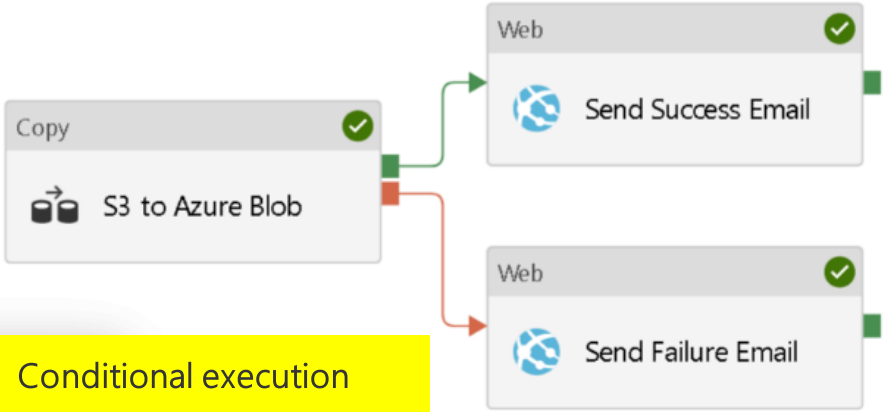
- Can you use the cloud?
- Is this a new solution or a migration?
- Do you want to use Microsoft tools or open source?
- What are your high availability and/or disaster recovery requirements?
- Do you need to master the data (MDM)?
- Are there any security limitations with storing data in the cloud?
- Will you use non-relational data?
- How much data do you need to store (volume)?
- Is this an OLTP or OLAP/DW solution?
- Will you use dashboards and/or ad-hoc queries?
- Will you use batch and/or interactive queries?
- How fast do the operational reports need to run?
- What is the skillset of the developers?
- Will you do predictive analytics?
- How many concurrent users will be accessing the solution at peak-time and on average?
- …

INGEST – DATA ORCHESTRATION AND MONITORING

Modern Data Warehouse (possible products by four areas)

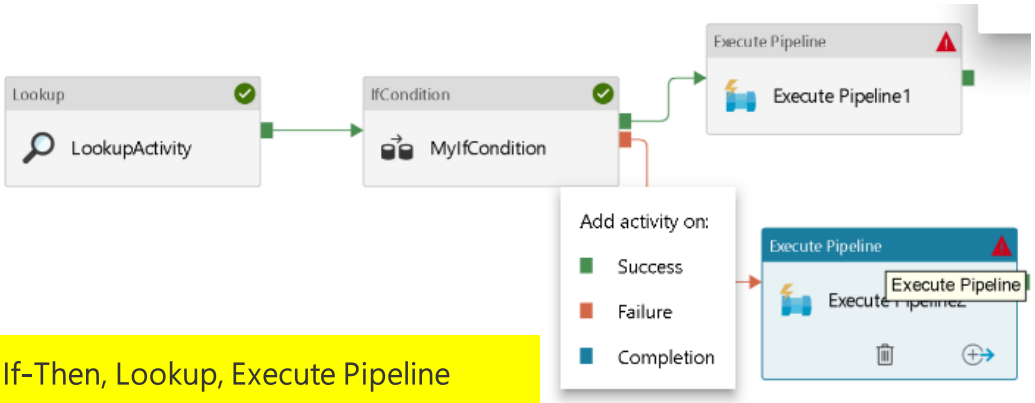


Note: Those products that span more than one area are listed in there primary area

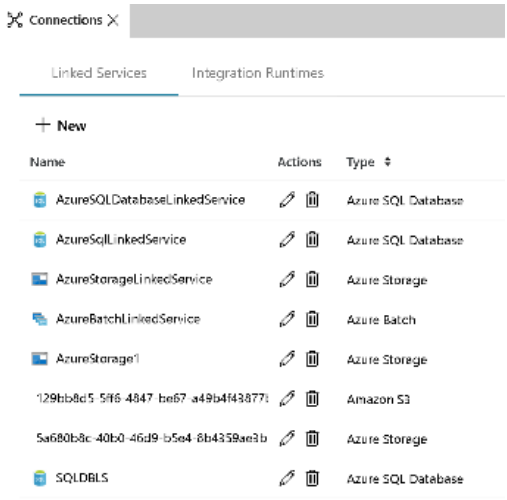


Conditional execution

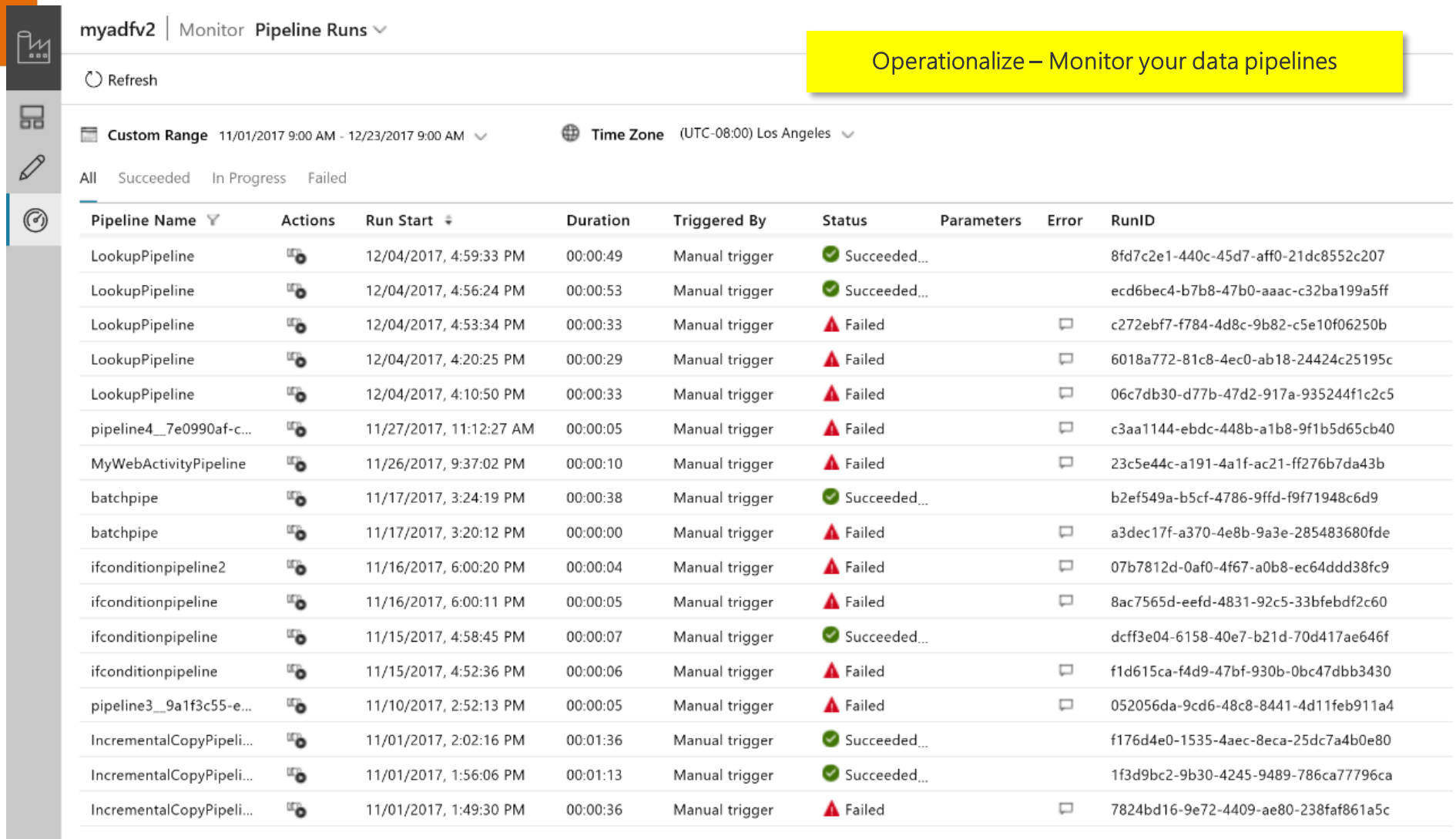
Incremental Delta Data Copy

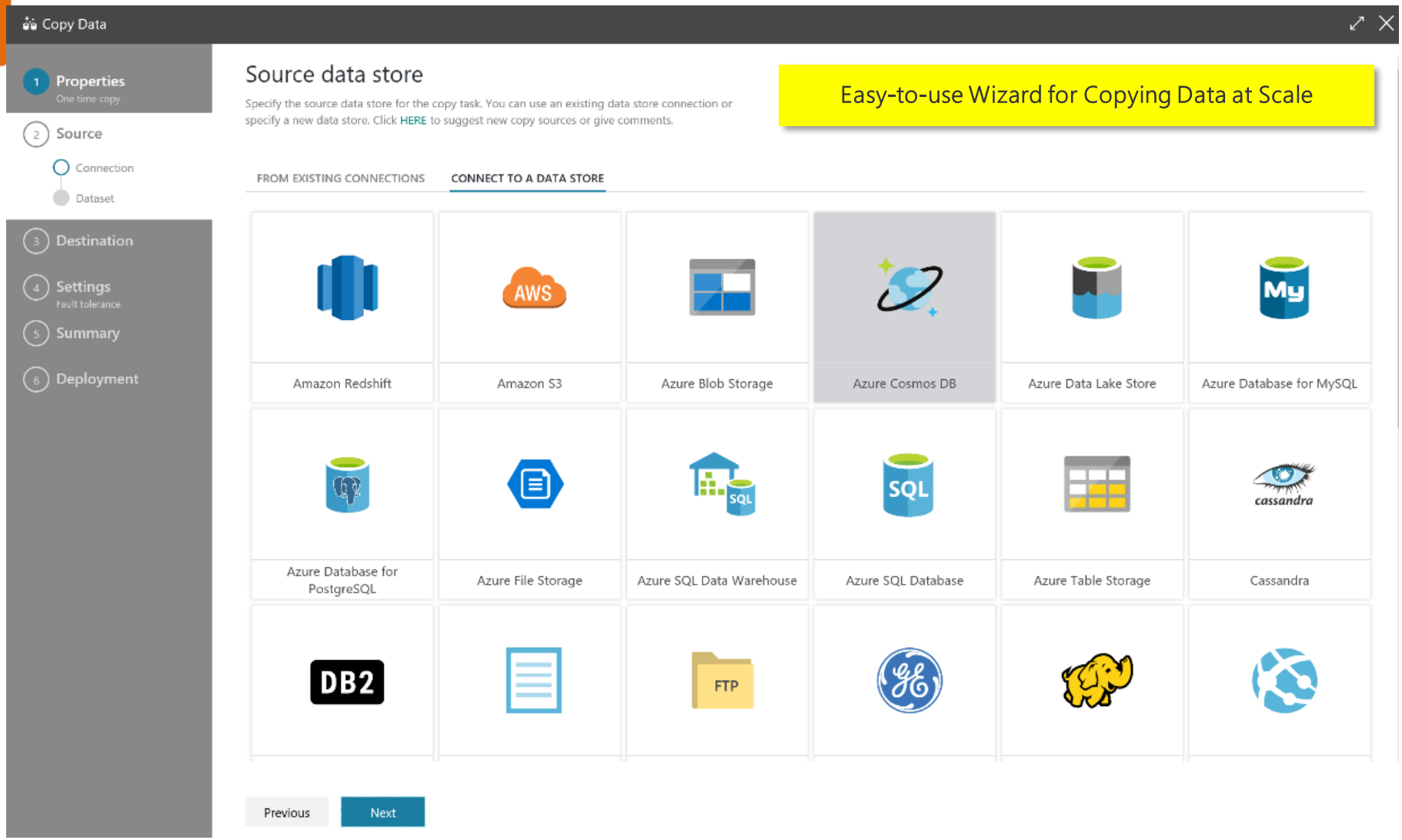


If-Then, Lookup, Execute Pipeline



Connection Managers

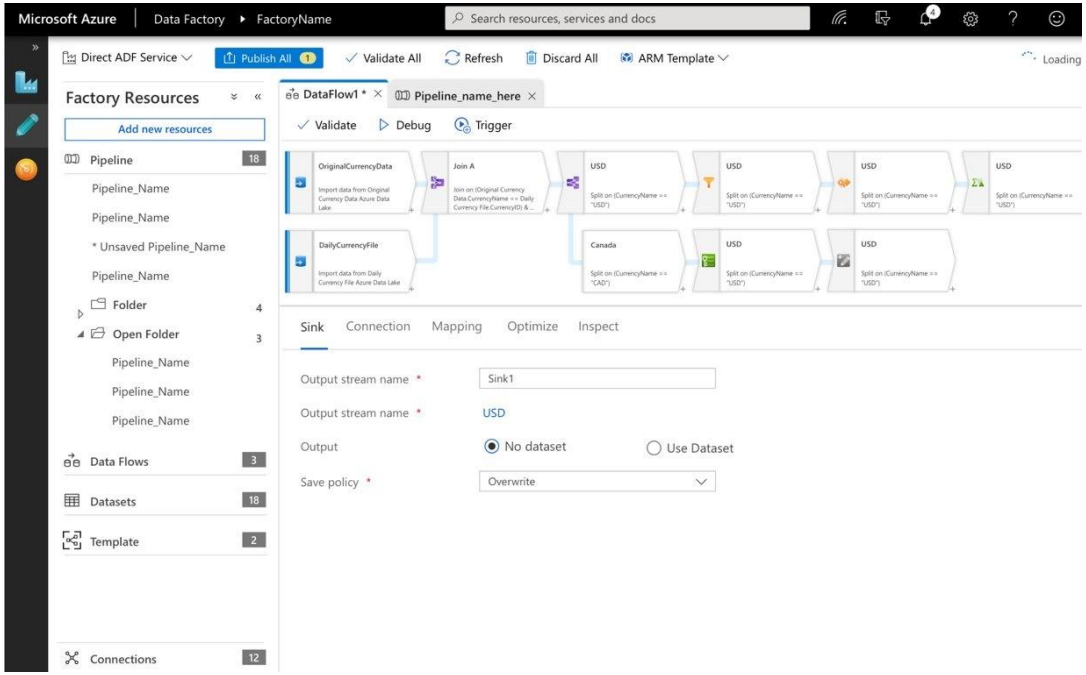

myadfv2 | Monitor Pipeline Runs
Refresh
Operationalize – Monitor your data pipelines
Custom Range 11/01/2017 9:00 AM - 12/23/2017 9:00 AM
Time Zone (UTC-08:00) Los Angeles
All Succeeded In Progress Failed
Pipeline Name Actions Run Start Duration Triggered By Status Parameters Error RunID
LookupPipeline 12/04/2017, 4:59:33 PM 00:00:49 Manual trigger Succeeded... 8fd7c2e1-440c-45d7-aff0-21dc8552c207
LookupPipeline 12/04/2017, 4:56:24 PM 00:00:53 Manual trigger Succeeded... ecd6bec4-b7b8-47b0-aaac-c32ba199a5ff
LookupPipeline 12/04/2017, 4:53:34 PM 00:00:33 Manual trigger Failed c272ebf7-f784-4d8c-9b82-c5e10f06250b
LookupPipeline 12/04/2017, 4:20:25 PM 00:00:29 Manual trigger Failed 6018a772-81c8-4ec0-ab18-24424c25195c
LookupPipeline 12/04/2017, 4:10:50 PM 00:00:33 Manual trigger Failed 06c7db30-d77b-47d2-917a-935244f1c2c5
pipeline4__7e0990af-c... 11/27/2017, 11:12:27 AM 00:00:05 Manual trigger Failed c3aa1144-ebdc-448b-a1b8-9f1b5d65cb40
MyWebActivityPipeline 11/26/2017, 9:37:02 PM 00:00:10 Manual trigger Failed 23c5e44c-a191-4a1f-ac21-ff276b7da43b
batchpipe 11/17/2017, 3:24:19 PM 00:00:38 Manual trigger Succeeded... b2ef549a-b5cf-4786-9ffd-f9f71948c6d9
batchpipe 11/17/2017, 3:20:12 PM 00:00:00 Manual trigger Failed a3dec17f-a370-4e8b-9a3e-285483680fde
ifconditionpipeline2 11/16/2017, 6:00:20 PM 00:00:04 Manual trigger Failed 07b7812d-0af0-4f67-a0b8-ec64ddd38fc9
ifconditionpipeline 11/16/2017, 6:00:11 PM 00:00:05 Manual trigger Failed 8ac7565d-eefd-4831-92c5-33bfebdf2c60
ifconditionpipeline 11/15/2017, 4:58:45 PM 00:00:07 Manual trigger Succeeded... dcff3e04-6158-40e7-b21d-70d417ae646f
ifconditionpipeline 11/15/2017, 4:52:36 PM 00:00:06 Manual trigger Failed f1d615ca-f4d9-47bf-930b-0bc47dbb3430
pipeline3__9a1f3c55-e... 11/10/2017, 2:52:13 PM 00:00:05 Manual trigger Failed 052056da-9cd6-48c8-8441-4d11feb911a4
IncrementalCopyPipeli... 11/01/2017, 2:02:16 PM 00:01:36 Manual trigger Succeeded... f176d4e0-1535-4aec-8eca-25dc7a4b0e80
IncrementalCopyPipeli... 11/01/2017, 1:56:06 PM 00:01:13 Manual trigger Succeeded... 1f3d9bc2-9b30-4245-9489-786ca77796ca
IncrementalCopyPipeli... 11/01/2017, 1:49:30 PM 00:00:36 Manual trigger Failed 7824bd16-9e72-4409-ae80-238faf861a5c


Copy Data
1 Properties
One time copy
2 Source
Connection
Dataset
3 Destination
4 Settings
Fault tolerance
5 Summary
6 Deployment
Source data store
Specify the source data store for the copy task. You can use an existing data store connection or specify a new data store. Click HERE to suggest new copy sources or give comments.
Easy-to-use Wizard for Copying Data at Scale
FROM EXISTING CONNECTIONS
CONNECT TO A DATA STORE
Amazon Redshift
Amazon S3
Azure Blob Storage
Azure Cosmos DB
Azure Data Lake Store
Azure Database for MySQL
Azure Database for PostgreSQL
Azure File Storage
Azure SQL Data Warehouse
Azure SQL Database
Azure Table Storage
Cassandra
DB2
FTP
Previous
Next

ADF Mapping Data Flow

Data Flow is a new feature of Azure Data Factory that allows you to build data transformations in a visual user interface

- Transform Data, At Scale, in the Cloud, Zero-Code
  - Cloud-first, scale-out ELT
  - Code-free dataflow pipelines
- Serverless scale-out transformation execution engine
- Maximum Productivity for Data Engineers
  - Does NOT require understanding of Spark / Scala / Python / Java
- Resilient Data Transformation Flows
  - Built for big data scenarios with unstructured data requirements
  - Operationalize with Data Factory scheduling, control flow and monitoring



Ingest – Data Orchestration and Monitoring

Product: Azure Data Factory (ADF)

Overview: With Mapping Data Flow, can now transform data, so ETL tool. Copy Data tool to easily copy from source to destination. Power Query support this semester

Use cases: Any new project, converting SSIS packages

How to use: PaaS

Watch out for: Row-by-row ETL can be slower, data needs to be moved to Databricks, limited by compute size of Databricks. Mapping Data Flow in public preview

Area also used for: Prep

Ingest – Data Orchestration and Monitoring

Product: SSIS

Overview: Very popular product, used for on-prem ETL for many years

Use cases: Too big of an effort to migrate existing packages, skillset, staying on-prem

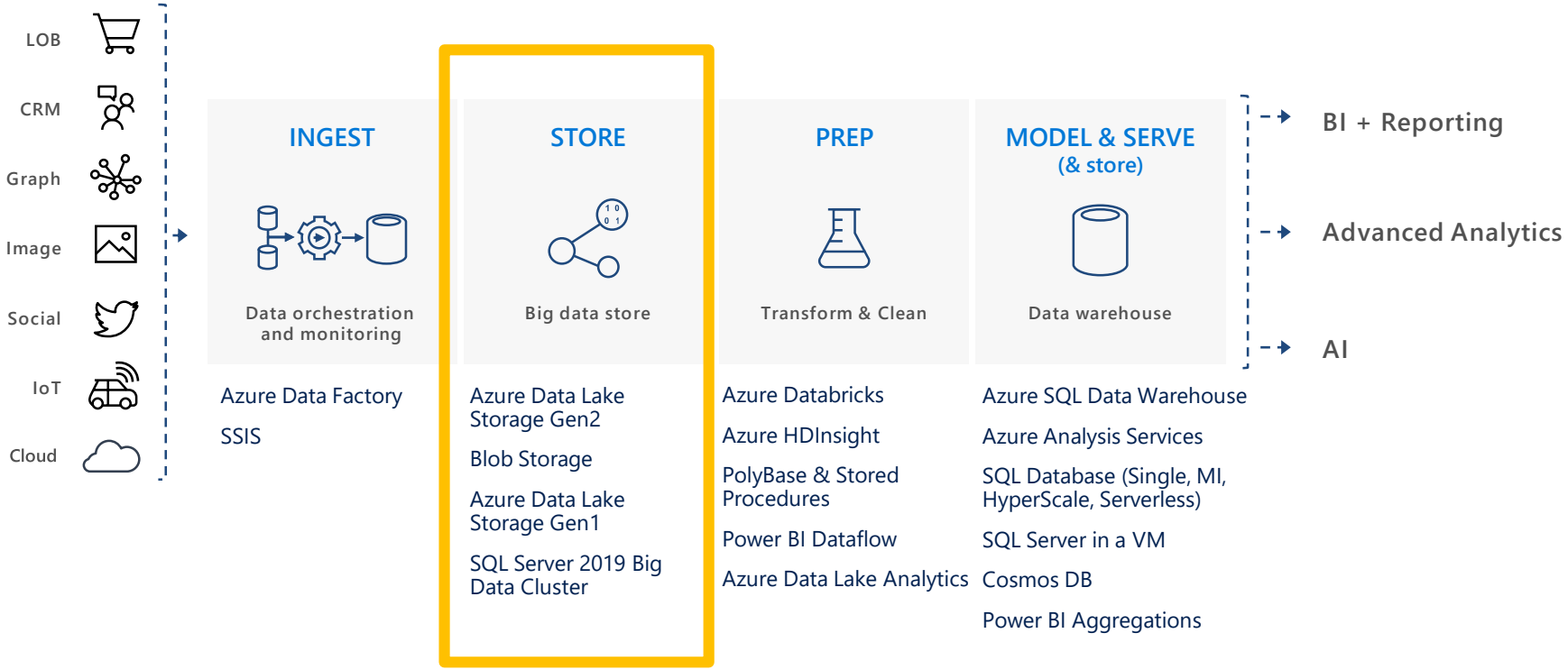
How to use: Visual Studio, change destination adapter and deploy to IR in ADF

Watch out for: Row-by-row ETL can be slower, data needs to be moved to IR, limited by compute size of IR

Area also used for: Prep

STORE – BIG DATA STORE

Modern Data Warehouse (possible products by four areas)

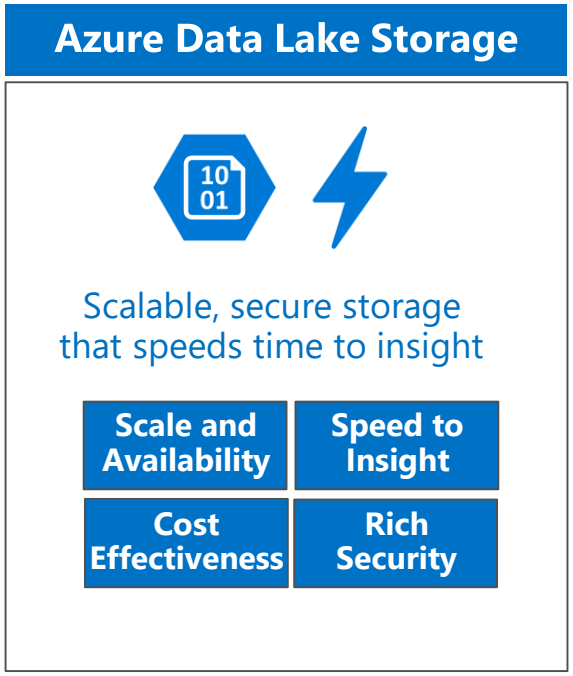


Note: Those products that span more than one area are listed in there primary area

Azure Data Lake Storage Gen2

Brings together the best of Azure Data Lake Store and Blob Storage

- Hadoop compatible file system interface for Azure Blob Storage
- Fine grained file and folder permissions (ACLs)
- Atomic file system operations
- Full support for all Blob features (AAD Integration, Zone Redundant and RA-Geo Redundant Storage)
- Pricing at Blob Storage levels
- Available in all 50 Azure regions (at GA)



Upgrade path for existing ADLS Customers

databricks

HORTONWORKS

Strong Partner Support

hadoop

APACHE Spark

Optimized for performance with Spark and Hadoop analytics engines

Store – Big Data Store

Product: Azure Data Lake Storage Gen2 (ADLS Gen2)

Overview: GA Feb 7th Combines best features of blob storage and ADLS Gen1

Use cases: Any new project. Convert Blob and Gen1 over time

How to use: PaaS

Watch out for: Not all features are available yet (soft delete, snapshots, object level storage tiers and lifecycle management). Some products may not support it yet. Blob Storage APIs and Azure Data Lake Gen2 APIs aren't interoperable with each other yet. 5TB file size limit

Area also used for: None

https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-known-issues

Store – Big Data Store

Product: Blob Storage

Overview: Original storage, most popular

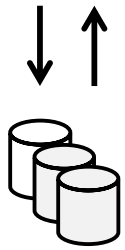
Use cases: Don't use for new projects unless need feature not available in ADLS Gen2 yet, or for non-analytical use cases that only need object storage rather than hierarchical storage (i.e. video, images, backup files). Don't migrate to Gen2 if current data does not need features of ADLS Gen2

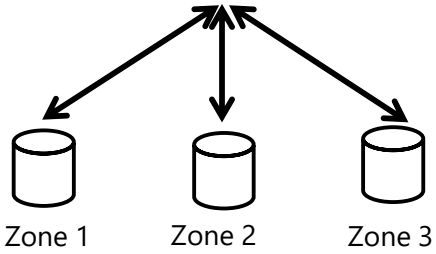
How to use: PaaS

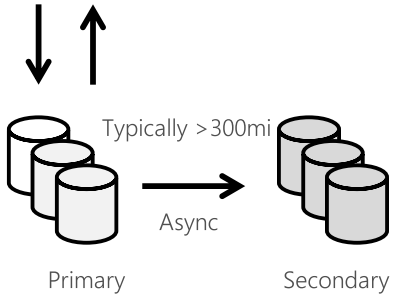
Watch out for: Account limit: 2PB for US and Europe, 500TB for all other regions including UK; File size limit: 4.75TB

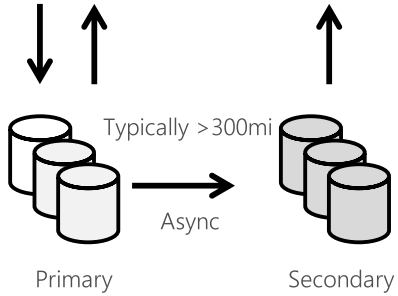
Area also used for: None

Azure Storage Replication Options









LRS

Multiple replicas across a datacenter

Protect against disk, node, rack failures

Write is ack'd when all replicas are committed

Superior to dual-parity RAID

11 9s of durability

SLA: 99.9%

ZRS

Replicas across 3 Zones

Protect against disk, node, rack and zone failures

Synchronous writes to all 3 zones

12 9s of durability

Available in 8 regions

SLA: 99.9%

GRS

Multiple replicas across each of 2 regions

Protects against major regional disasters

Asynchronous to secondary

16 9s of durability

SLA: 99.9%

RA-GRS

GRS + Read access to secondary

Separate secondary endpoint

RPO delay to secondary can be queried

SLA: 99.99% (read), 99.9% (write)

Store – Big Data Store

Product: Azure Data Lake Storage Gen1 (ADLS Gen1)

Overview: Originally for better performance over Blob storage

Use cases: Don't use for new projects. Convert to ADLS Gen2 to save money and get more features

How to use: PaaS

Watch out for: Will not have any new features

Area also used for: None

Store – Big Data Store

Product: SQL Server 2019 Big Data Cluster

Overview: Combines together the SQL Server database engine, Spark, and HDFS (including ADLS Gen2) into a unified data platform deployed as containers on Kubernetes. Also uses PolyBase to access many types of data sources

Use cases: Hybrid cloud. Data virtualization, data lake, and AI platform. Read/write non-relational data in HDFS, reads other sources. Scale-out compute via Spark and SQL. Query relational and non-relational together. Can be MPP-like option in future (needs updateable distributed tables and replicated dimensional tables)

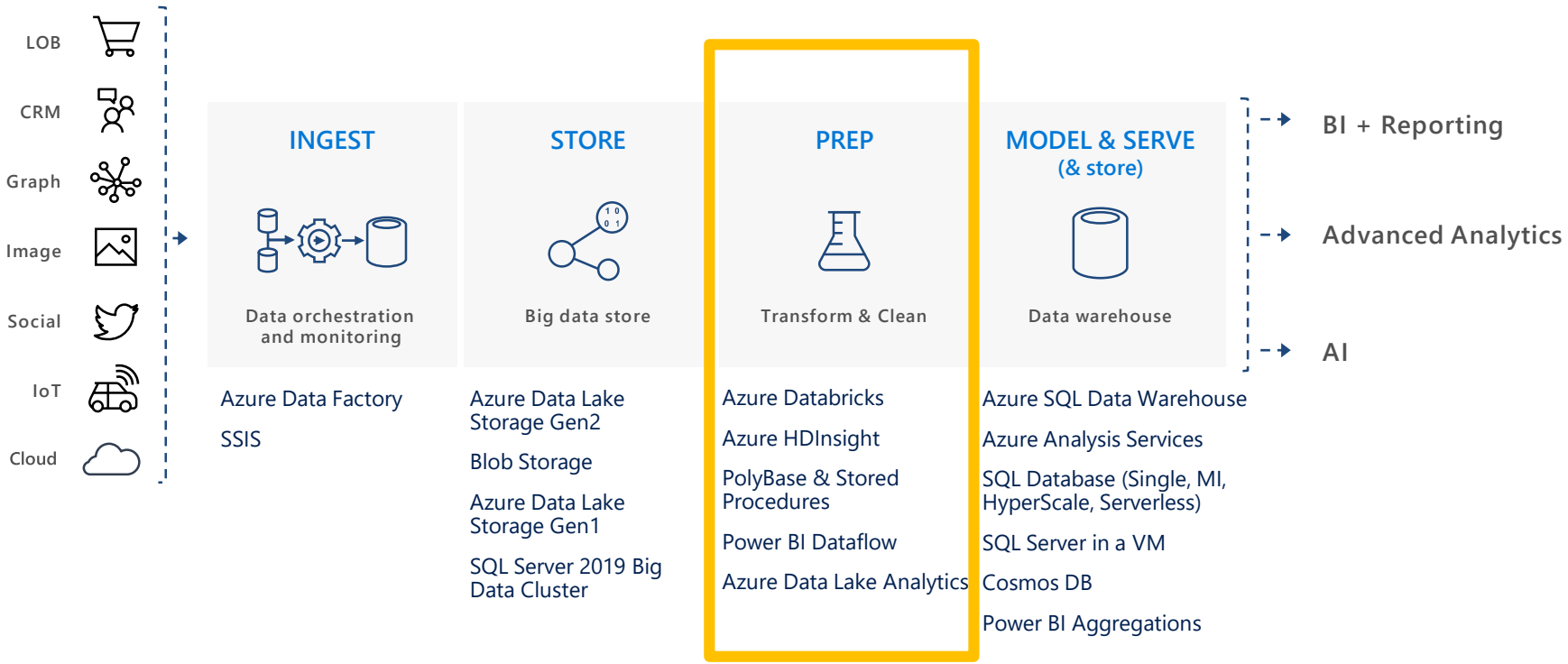
How to use: IaaS

Watch out for: Look to use PaaS solutions first (SQL DW). SQL Server 2019 in community technology preview

Area also used for: Prep

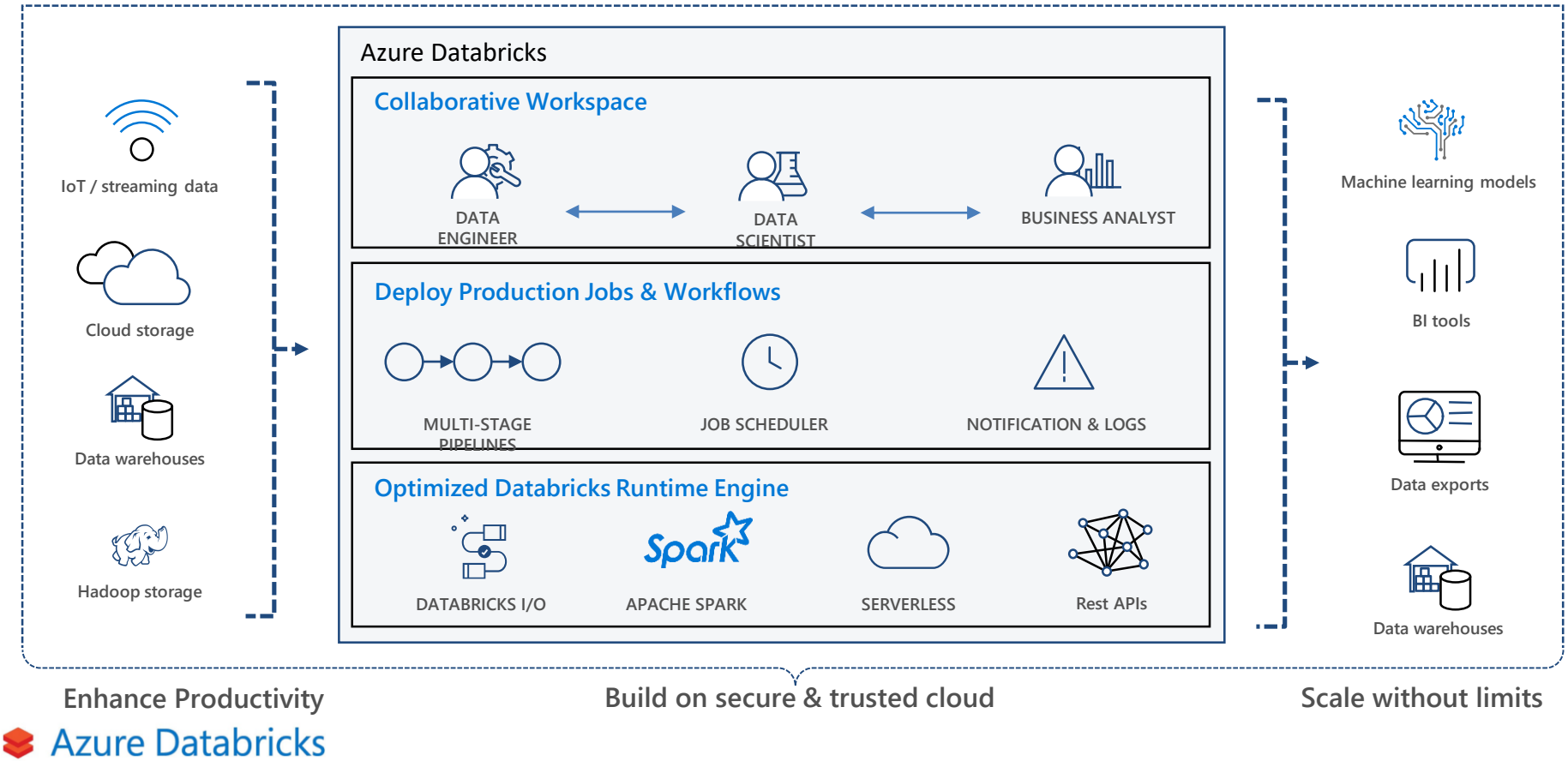
PREP – TRANSFORM AND CLEAN

Modern Data Warehouse (possible products by four areas)



Note: Those products that span more than one area are listed in there primary area

Azure Databricks



Azure Databricks key audiences & benefits

Data scientist

Integrated workspace

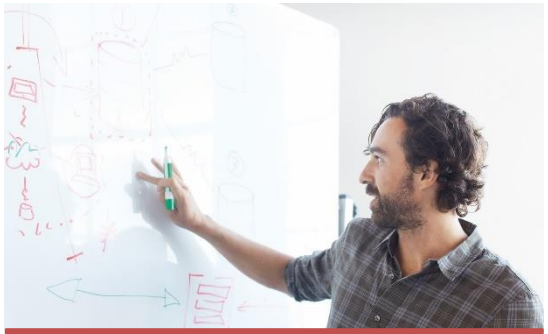
Easy data exploration

Collaborative experience

Interactive dashboards

Faster insights

- Best Spark & serverless
- Databricks managed Spark



Data engineer

Improved ETL performance

- Zero management clusters, serverless

Easy to schedule jobs

Automated workflows

Enhanced monitoring & troubleshooting

- Automated alerts & easy access to logs

Zero Management Spark

Cluster democratization (serverless)

CDO, VP of analytics

Fast, collaborative analytics platform accelerating time to market

No dev-ops required

Enterprise grade security

- Encryption
- End-to-end auditing
- Role-based control
- Compliance

Unified analytics platform

Azure Databricks

Prep – Transform and Clean

Product: Azure Databricks

Overview: Tool for curating and processing massive amounts of data and developing, training and deploying models on that data, and managing the whole workflow process throughout the project

Use cases: Comfortable with Spark and notebooks, integration with ADLS, SQL DW, PBI, etc, need auto-scaling and auto-termination, need fast Spark

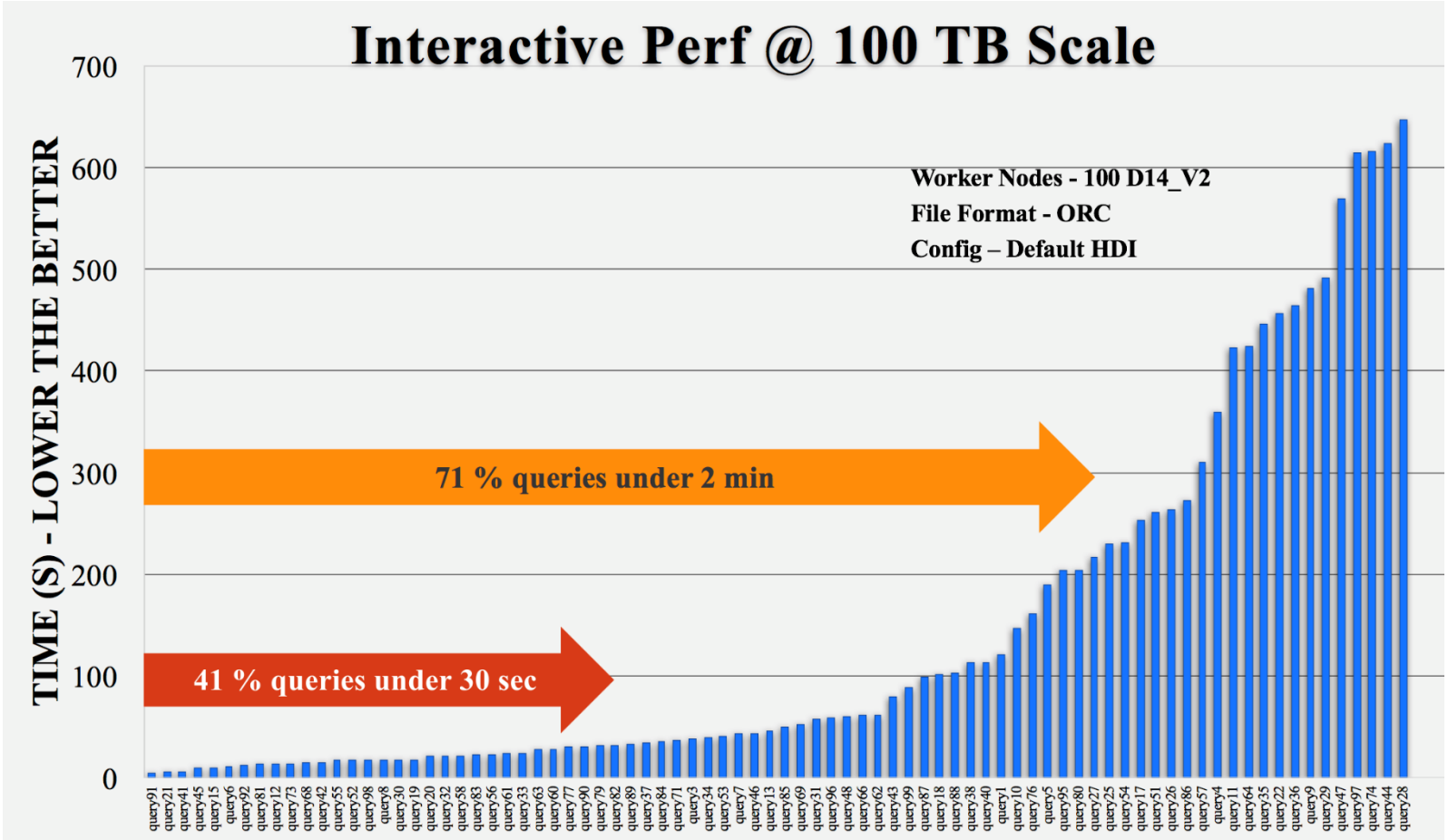
How to use: PaaS

Watch out for: Avoid if you don't like to write code – for data engineer/scientist, steep learning curve

Area also used for: Ingest, Model & Serve

HDInsight Clusters

- Support Java and Python by default
- Customizable through script actions


Interactive Perf @ 100 TB Scale
TIME (S) - LOWER THE BETTER
700
600
500
400
300
200
100
0
Worker Nodes - 100 D14_V2
File Format - ORC
Config – Default HDI
71 % queries under 2 min
41 % queries under 30 sec

Prep – Transform and Clean

Product: Azure HDInsight (HDI)

Overview: Deploys and provisions Apache Hadoop clusters in the Azure cloud. Hortonworks under the covers

Use cases: Databricks is the preferred product over HDI, unless the customer has a mature Hadoop ecosystem already established or if want to use other Hadoop tools that are available 24/7

How to use: PaaS

Watch out for: No integration with SQL DW, always running and incurring costs

Area also used for: Ingest, Model & Serve

Prep – Transform and Clean

Product: PolyBase & Stored Procedures (within SQL DW)

Overview: Process T-SQL queries that copy raw data from data lake (ADLS or Blob storage) via an external table into SQL DW, then clean via stored procedures

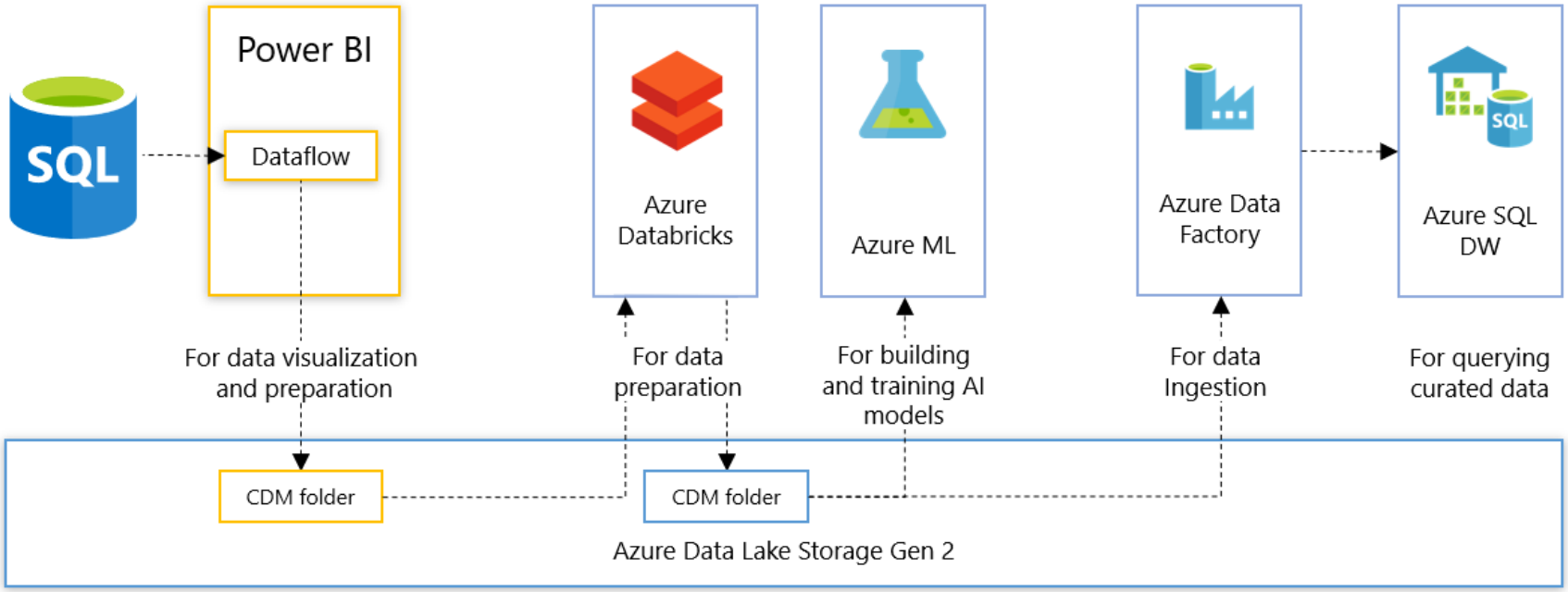
Use cases: ELT. Stick with T-SQL and don't want to deal with Spark or Hive or other more-difficult technologies

How to use: Via external tables in SQL DW or SQL Server 2016+ (IaaS)

Watch out for: Cleaning data in SQL DW can affect user queries, increase storage space, more expensive, not have clean data in data lake, no pushdown queries

Area also used for: Ingest

Power BI Dataflows



Business analysts

Low/no code

Data scientists

Data engineers

Medium to high code

Prep – Transform and Clean

Product: Power BI Dataflows

Overview: Integrates data lake and data prep technology (Power Query) directly into Power BI Service, independent of PBI reports. Self-service data prep

Use cases: Individual solution or for small workloads. For Data Analysts and Business Analysts. Can transform data that lands in the data lake and can then be used as part of an enterprise solution

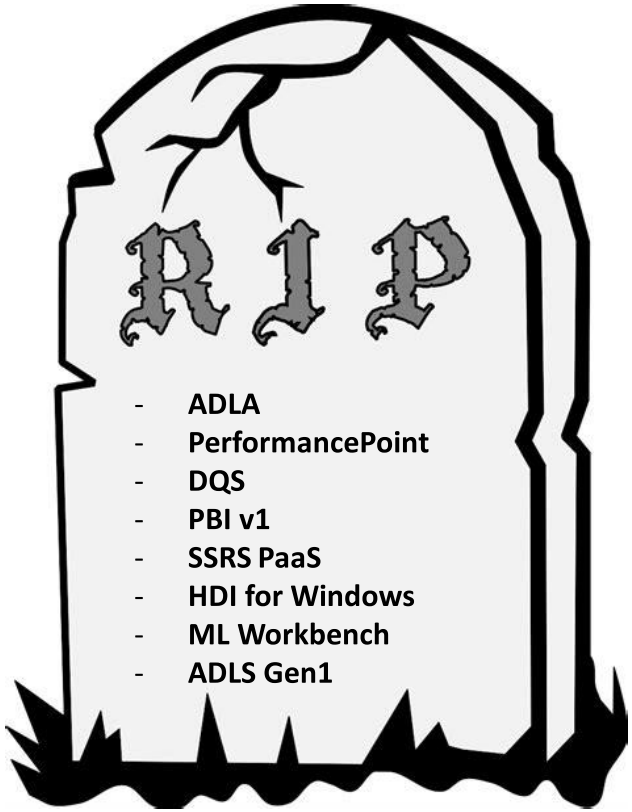
How to use: Power BI Service

Watch out for: Don't use this to replace a data warehouse or ADF. Does not work for streaming data or Direct Query. In public preview

Area also used for: Ingest, Store

Prep – Transform and Clean

Product: Azure Data Lake Analytics (ADLA)



Prep – Transform and Clean

Product: Azure Data Lake Analytics (ADLA)

Overview: Dynamically provisions resources so you can run queries on petabytes of data. Query-as-a-service using U-SQL

Use cases: Do not use for new projects. Use for transforming large amounts of data in a data lake or replacing long-running monthly batch processing with shorter running distributed processes. Predictable performance with no startup time

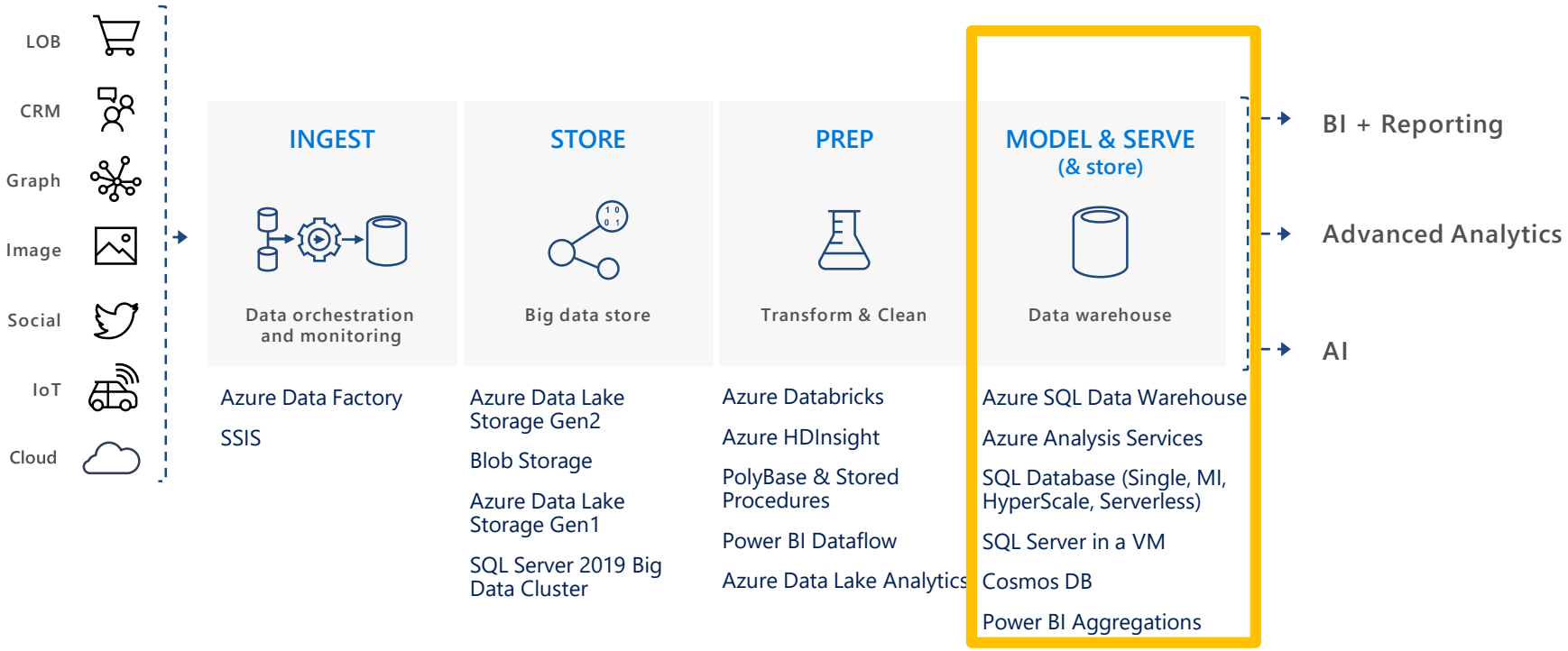
How to use: PaaS

Watch out for: No ADLS Gen2 support. Does not support interactive queries, persistence, or indexing

Area also used for: None

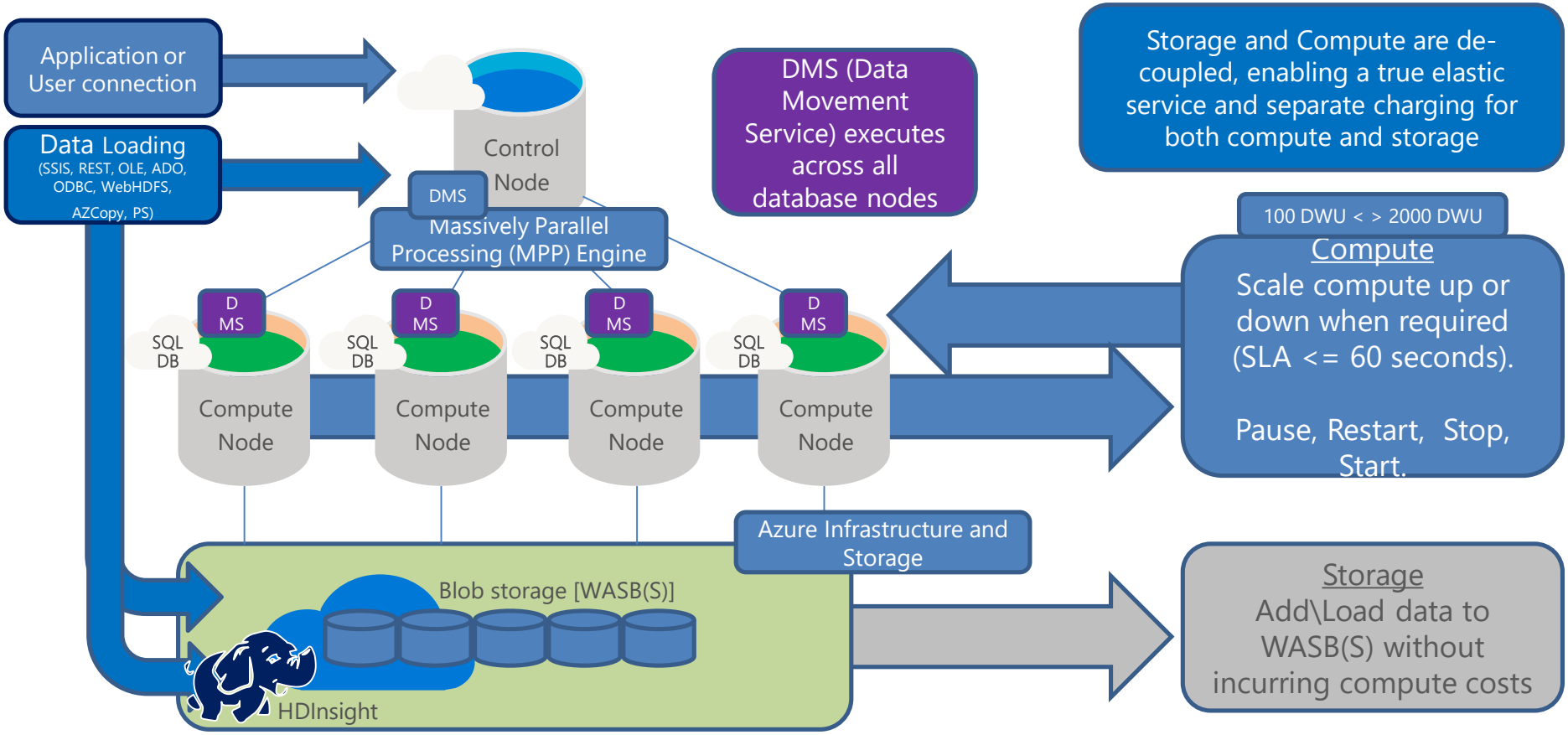
MODEL & SERVE – DATA WAREHOUSE

Modern Data Warehouse (possible products by four areas)



Note: Those products that span more than one area are listed in there primary area

Azure SQL Data Warehouse Architecture



Model & Serve – Data Warehouse

Product: Azure SQL Data Warehouse (SQL DW)

Overview: SQL-based, fully-managed, petabyte-scale cloud data warehouse. Can scale compute and storage independently allowing you to burst compute, and can be paused

Use cases: MPP technology that shines when used for ad-hoc queries and operational reports in relational format (queries run 20-100x faster)

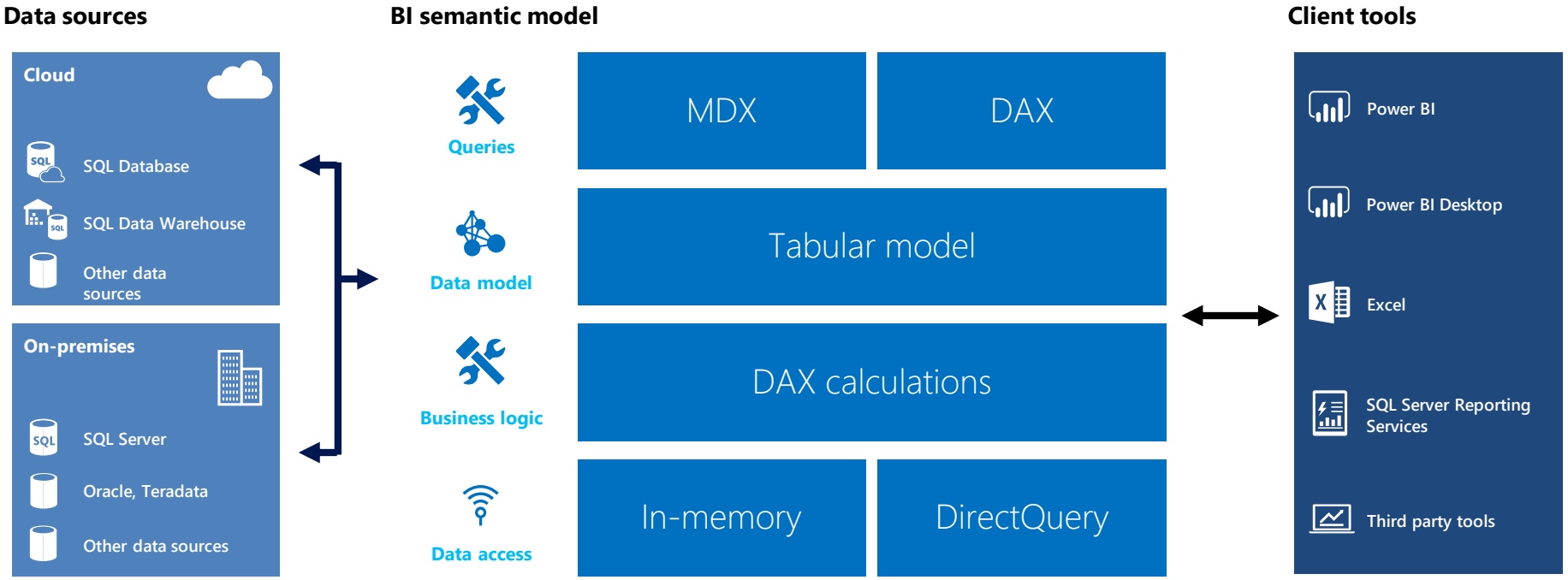
How to use: PaaS

Watch out for: Don't use if need high concurrency, no geo-replication, no cross-database queries. Requires data to be copied from ADLS into SQL DW but this can be done quickly using PolyBase

Area also used for: Prep

Proven analytics engine

Azure Analysis Services based on SQL Server technology



Azure Analysis Services Cubes

Reasons to report off cubes instead of the data warehouse:

- Semantic layer
- Handle many concurrent users
- Aggregating data for performance
- Multidimensional analysis
- No joins or relationships
- Hierarchies, KPI's
- Row-level Security
- Advanced time-calculations
- Slowly Changing Dimensions (SCD)
- Required for some reporting tools

Model & Serve – Data Warehouse

Product: Azure Analysis Services (AAS)

Overview: Tabular models of aggregated data (OLAP)

Use cases: Queries in milliseconds (dashboards), high concurrency, semantic layer . Can do vertical scale-out for high availability and high concurrency. Built-in hierarchies and KPI's and Advanced time-calculations

How to use: PaaS

Watch out for: Does not support multidimensional cubes, time to process the cube, not real-time, slower performance for ad-hoc queries Area also used for: None

Model & Serve – Data Warehouse

Product: SQL Database. Programming model: Instance (MI), Database (Single); Service tiers: General Purpose, Business Critical, Hyperscale; Compute tier: Serverless; Resource grouping concept: Elastic Pools

Overview: Database-as-a-service in different flavors: MI (near 100% compatibility, great for on-prem migrations), Single (additional DTU model, less compatibility, lower price of entry), Hyperscale (up to 100TB database size, higher performance), Serverless (price benefits for bursty workloads, compute range and auto-pause)

Use cases: Migrating on-prem SQL Server or any new projects that need a relational database. Mostly for OLTP but can be used for smaller data warehouses

How to use: PaaS

Watch out for: Simple/Bulk-logged recovery mode not supported which affects data loading. Database size limits: Singleton: DTU Basic tiers (2GB), Standard tiers (1TB), Premium tier (4TB); vCore General Purpose tier (4TB), Business Critical tier (4TB); Managed Instance: General Purpose tier (8TB), Business Critical tier (4TB); Hyperscale in public preview (GP only), Hyperscale MI in private preview (GP only), optimized for OLTP; Serverless in private preview

Area also used for: Prep

Model & Serve – Data Warehouse

Product: SQL Server in a VM

Overview: SQL Server in a VM

Use cases: Need control over / access to the operating system, have to run the app or agents side-by-side with the DB, need to use older version of SQL Server, SSRS, DW in the 4TB-50TB range

How to use: IaaS. Provision SQL Server image from Azure Marketplace

Watch out for: Max IOPS and database size depends on managed disks used

Area also used for: Prep

Model & Serve – Data Warehouse

Product: Cosmos DB

Overview: A globally distributed, multi-model (key-value, graph, and document) database service. It fits into the NoSQL camp by having a non-relational model (supporting schema-on-read and JSON documents)

Use cases: Works really well for large-scale OLTP solutions. Spark to Cosmos DB connector for DW aggregations. Use for data lake to have one datastore for both operational and analytical queries

How to use: PaaS

Watch out for: Data lake - cost and having to convert all files to JSON. DW – speed of query joins and group by, Spark SQL not 100% SQL compatible

Area also used for: Store, Prep

Model & Serve – Data Warehouse

Product: Power BI with Aggregations

Overview: Tabular models of aggregated data

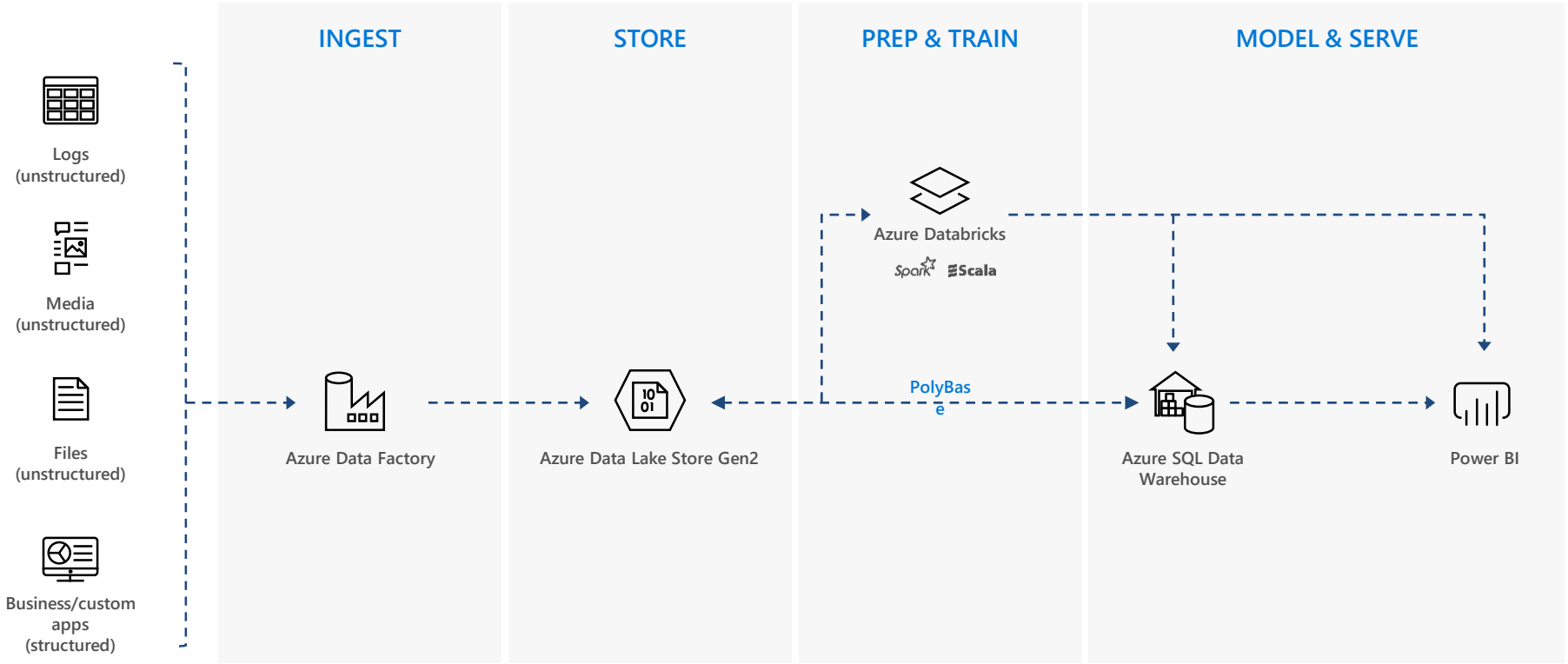
Use cases: Replaces AAS

How to use: Power BI Desktop and Power BI Service

Watch out for: Aggregations in preview

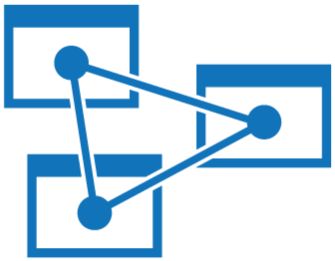
Area also used for: None

Modern Data Warehouse



Microsoft Azure also supports other Big Data services like Azure HDInsight to allow customers to tailor the above architecture to meet their unique needs.

Enterprise BI



Azure
Analysis Services

All BI users

Power BI
Premium

Self-service BI users

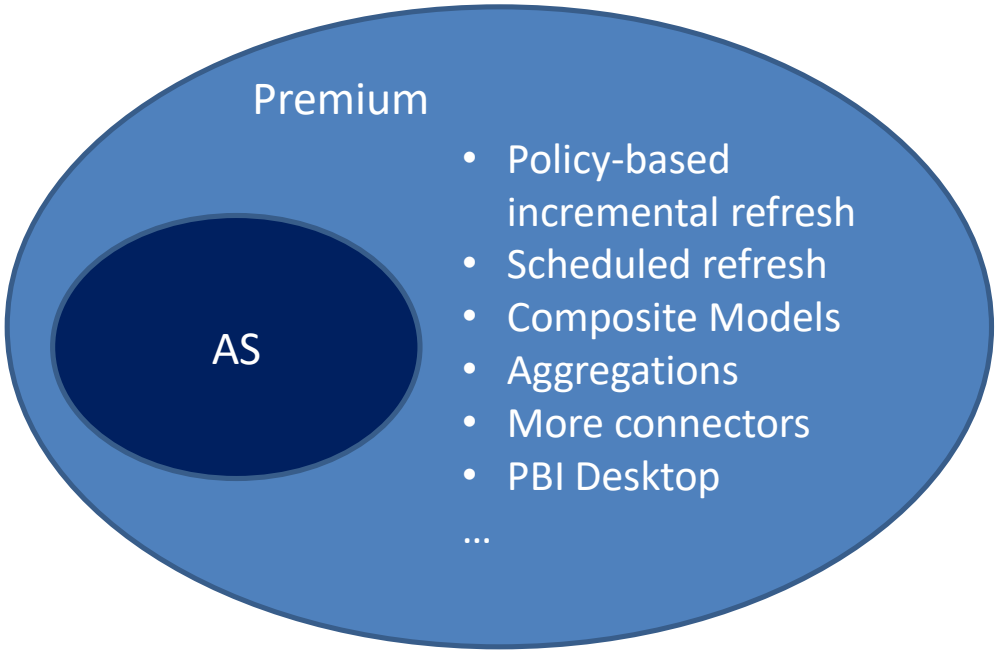
Power BI

Azure AS vs Power BI

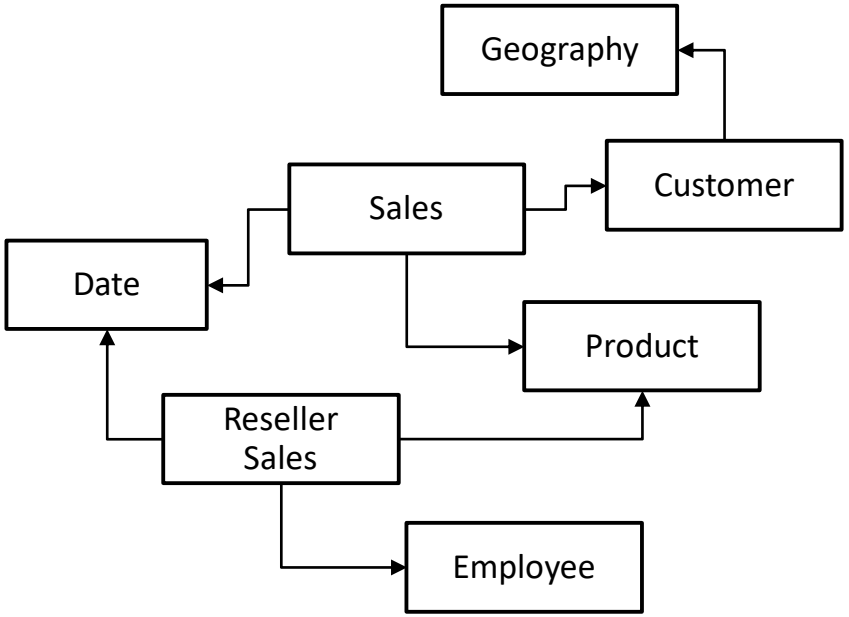
Power BI: a unified platform for self-service and enterprise BI

Power BI

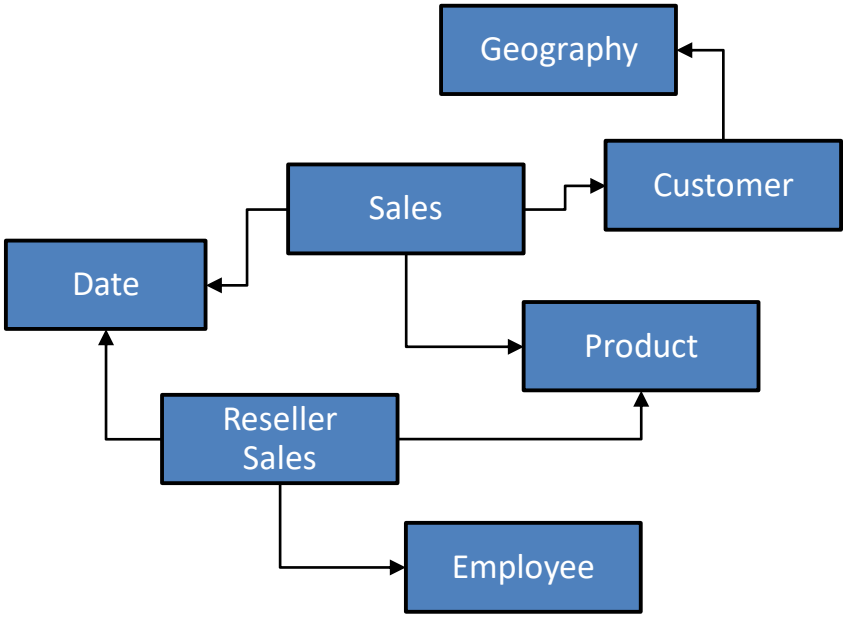
SQL Server Analysis Services (Tabular)



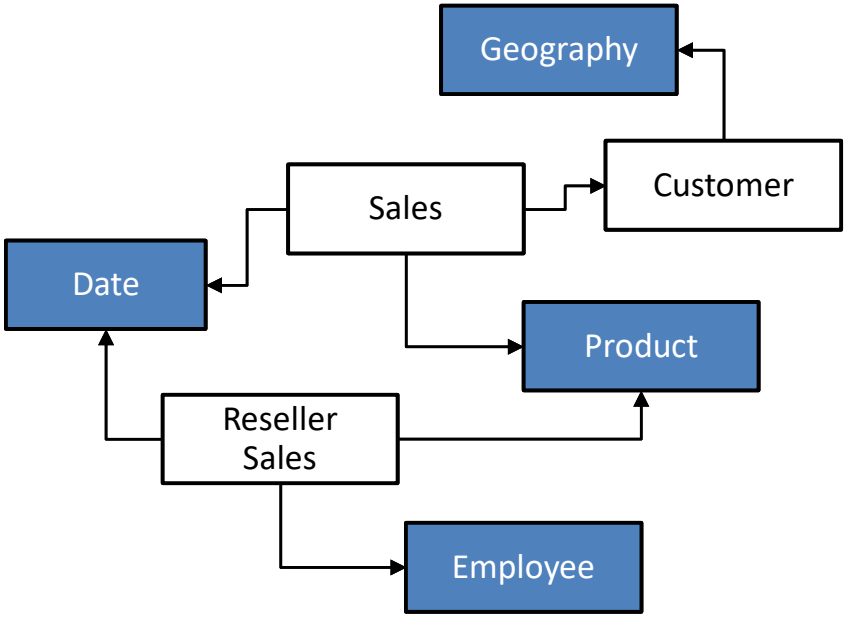
DirectQuery



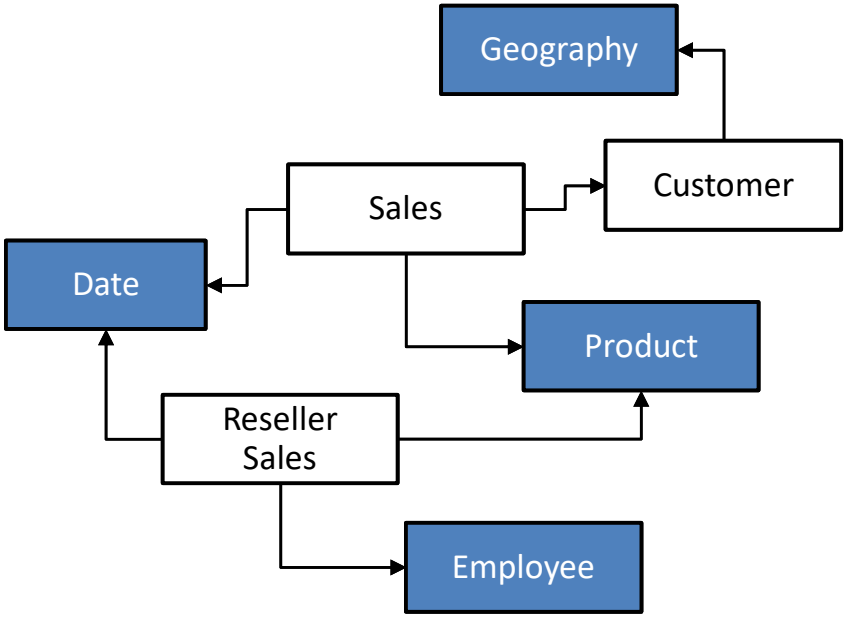
Import



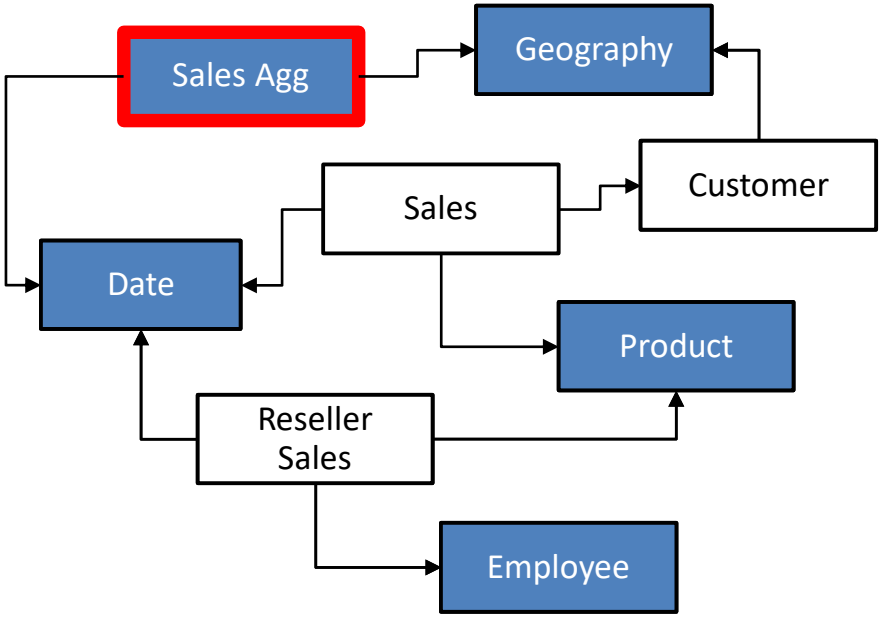
DirectQuery & Import



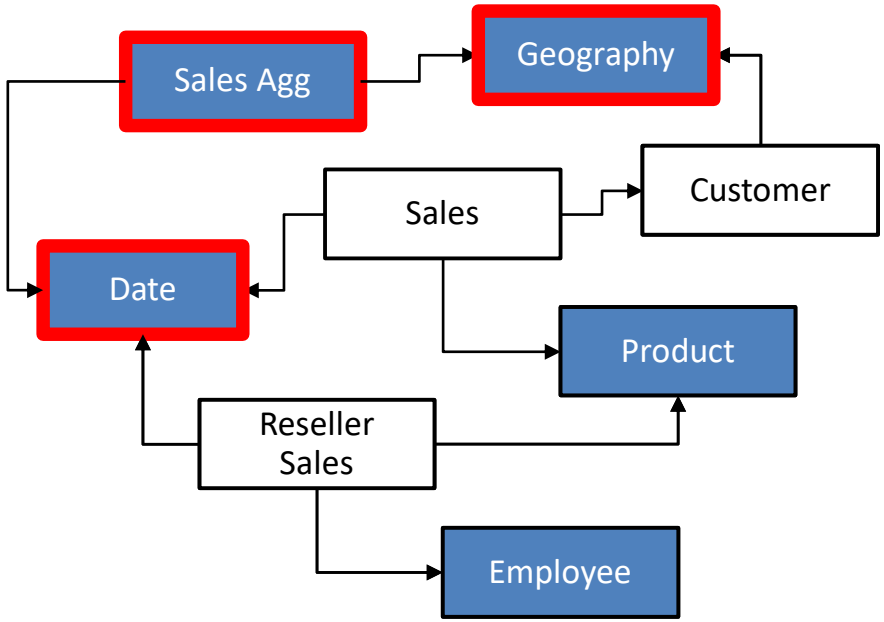
DirectQuery & Import



Aggregations



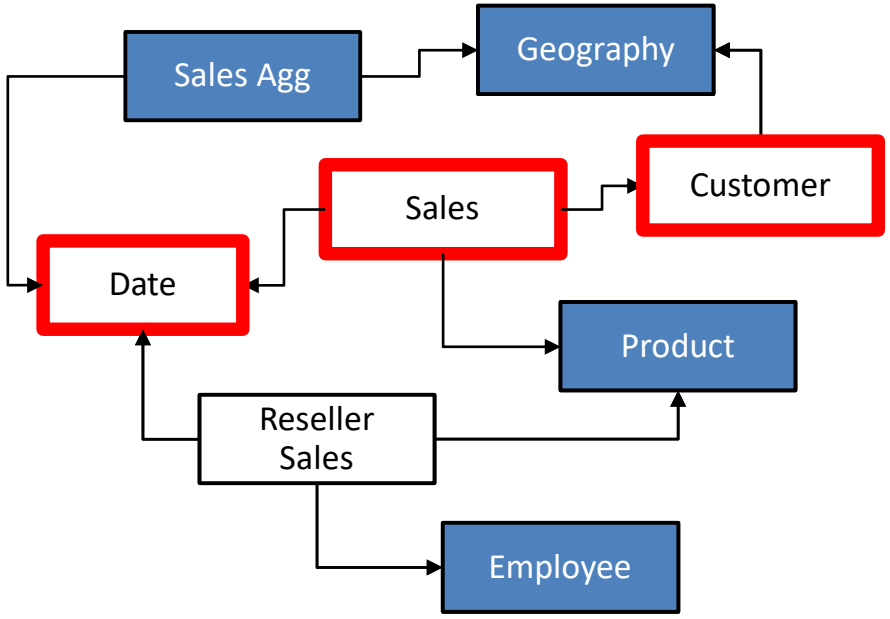
Aggregations




```
SummarizeColumns(
        Date[Year],
        Geography[City],
        "Sales", Sum(Sales[Amount])
)
```

Hits in-memory cache

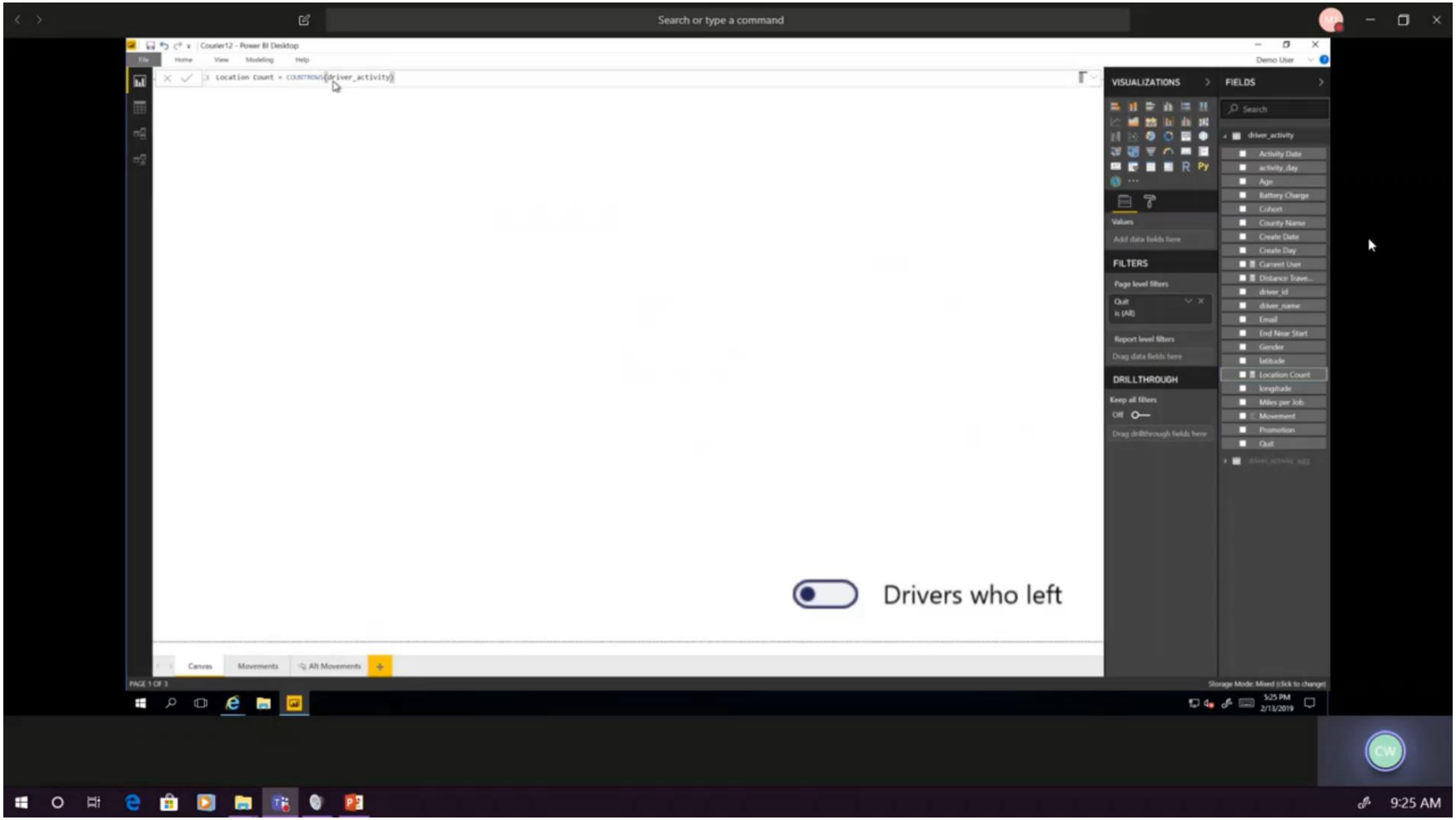
Aggregations




```
SummarizeColumns(
        Date[Year],
        Customer[Name],
        "Sales", Sum(Sales[Amount])
)
```

DirectQuery

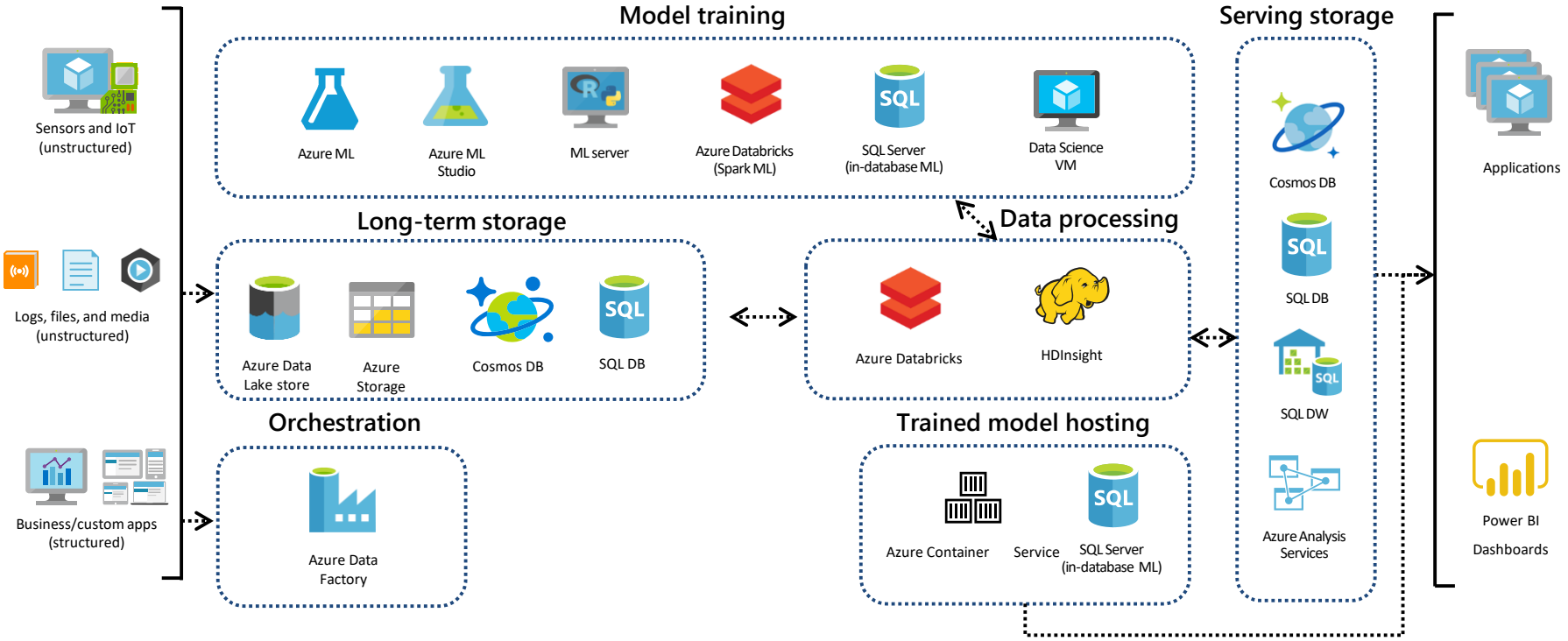
Power BI Aggregations demo



Trillion row demo: https://aka.ms/TrillionRowDemo

Advanced analytics/Big data pattern in Azure

Data collection and understanding, modeling, and deployment



Dziękuję za uwagę

Data Community

SQL Day

PLATINUM SPONSOR

TECHNOLOGY INNOVATION DATA KNOWLEDGE tidk

STRATEGIC PARTNER

Microsoft

GOLD SPONSORS

CLOUDS ON MARS

Future Processing

SPYROSOFT

SILVER SPONSOR

dbWatch DATABASE CONTROL

elitmind think bright

VEEAM

BRONZE SPONSOR

Największa księgarnia IT w Polsce

NOVATECH www.novatech.com.pl