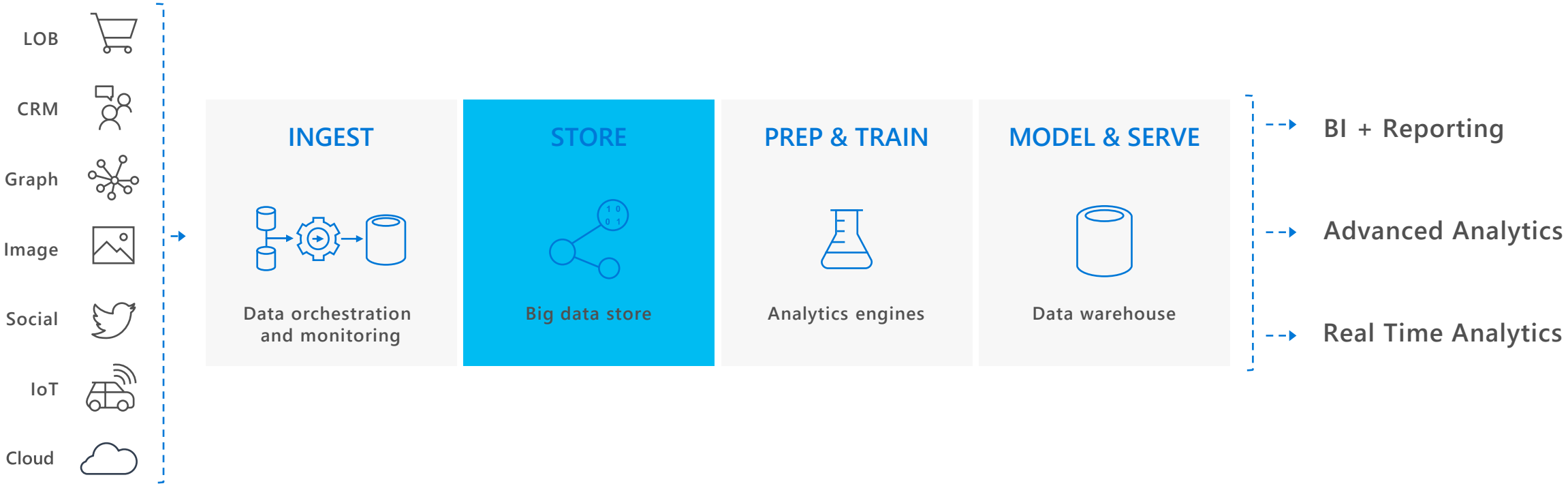
Microsoft

Azure
Data Lake Architekturen & Azure Synapse

Big data & Data warehouse



Defining the data lake

- A data lake is a collection of data, not a platform for data.
  - Hadoop is the preferred platform for data lakes, but not the only one, as there can be relational data lakes as well.
- A data lake handles large volumes of diverse data…
  - Semi- and un-structured data formats, possibly Exabytes of data
- …ingests it quickly…
  - Straight from data source, no wrangling/ETL.
- …and persists it in its original, raw and refined formats.
  - Detailed source data as basis for data engineering/science.

Benefits of a data lake

A data lake is keeping you flexible…

…you can choose and work with whatever tool you like, or you need to work.

A data lake does not do a vendor lock-in…

… you can take your data and move them somewhere else

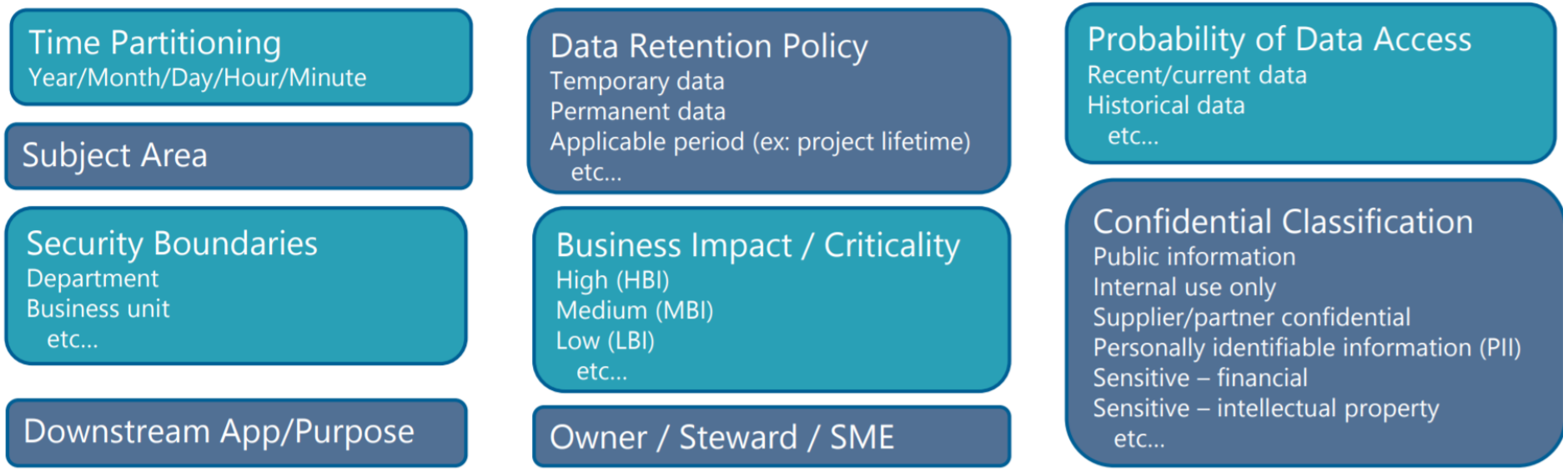
…you are NOT binding yourself to a specific technology or tool…

…and that means for you:

NO MORE MIGRATIONS

1.2 Data Lake Structure – Considerations

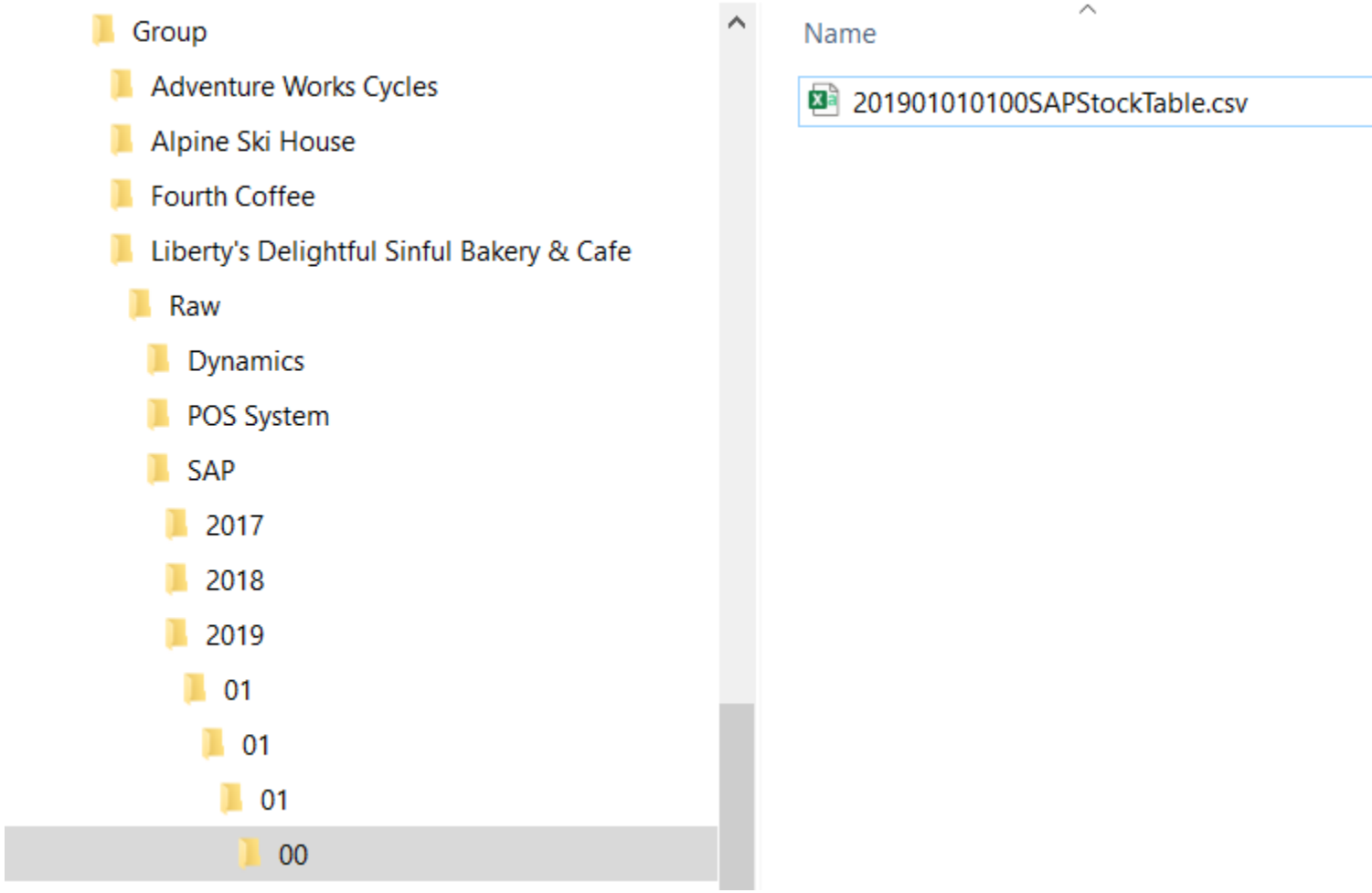
- When organizing the data within a data lake consider:



- These different types of data should be separated into various different zones, for better clarity and enforceability of rules and definitions.

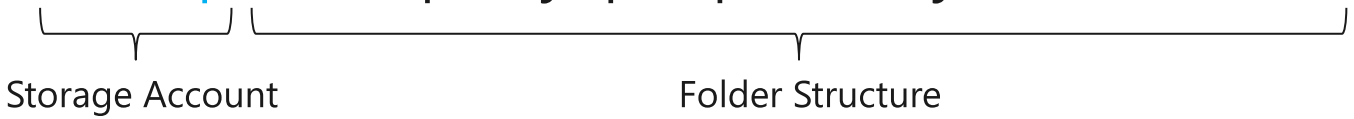
Folder Structure

- Purpose (Raw, Refined etc.)
- System/Origin
- Organisation/Business Unit
- Date
- Sensor Name/ID

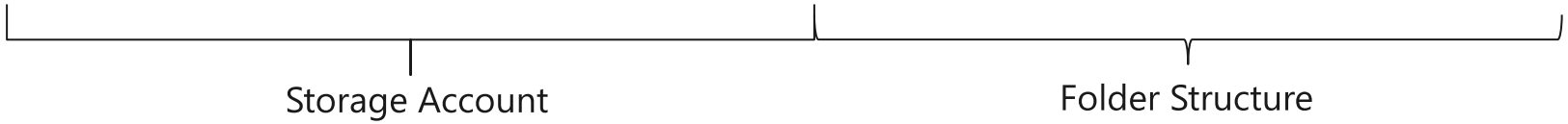


Examples – Storage Accounts

- Group/Company/purpose/System/Date/File

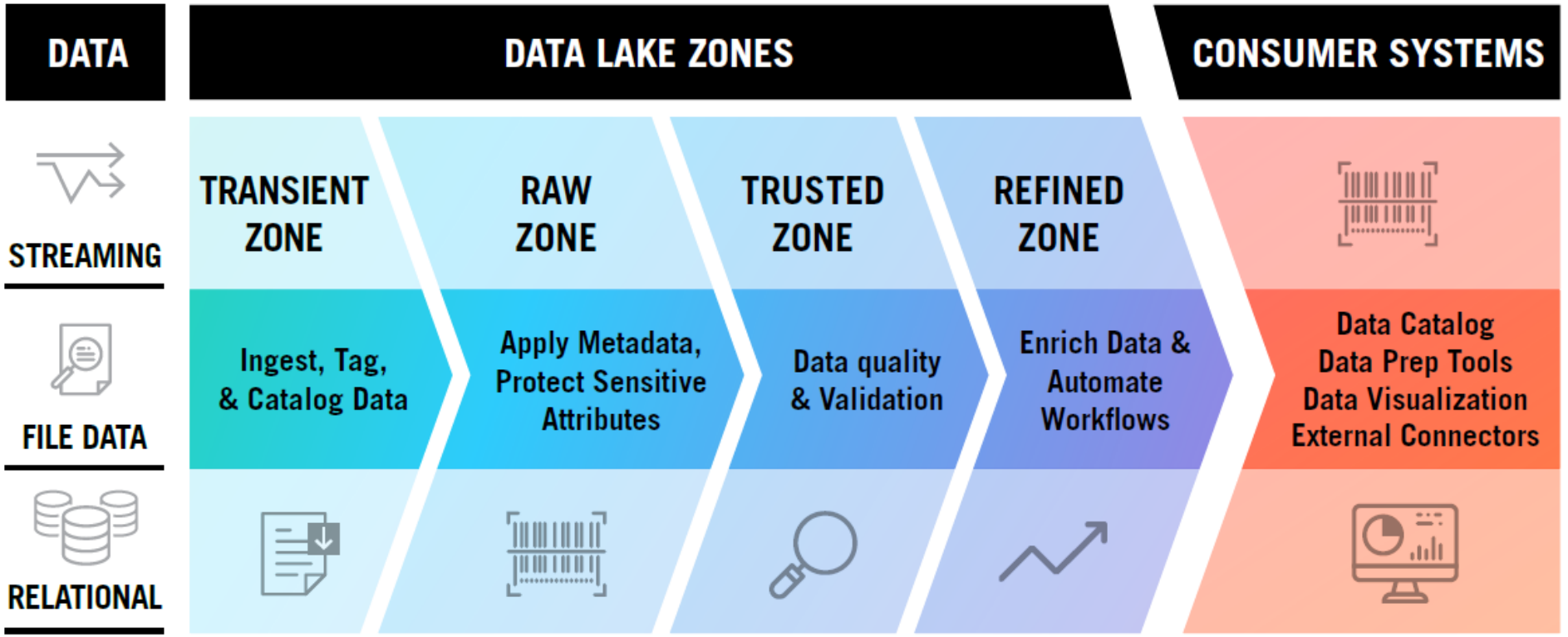


- Group/Company/purpose/System/Date/File

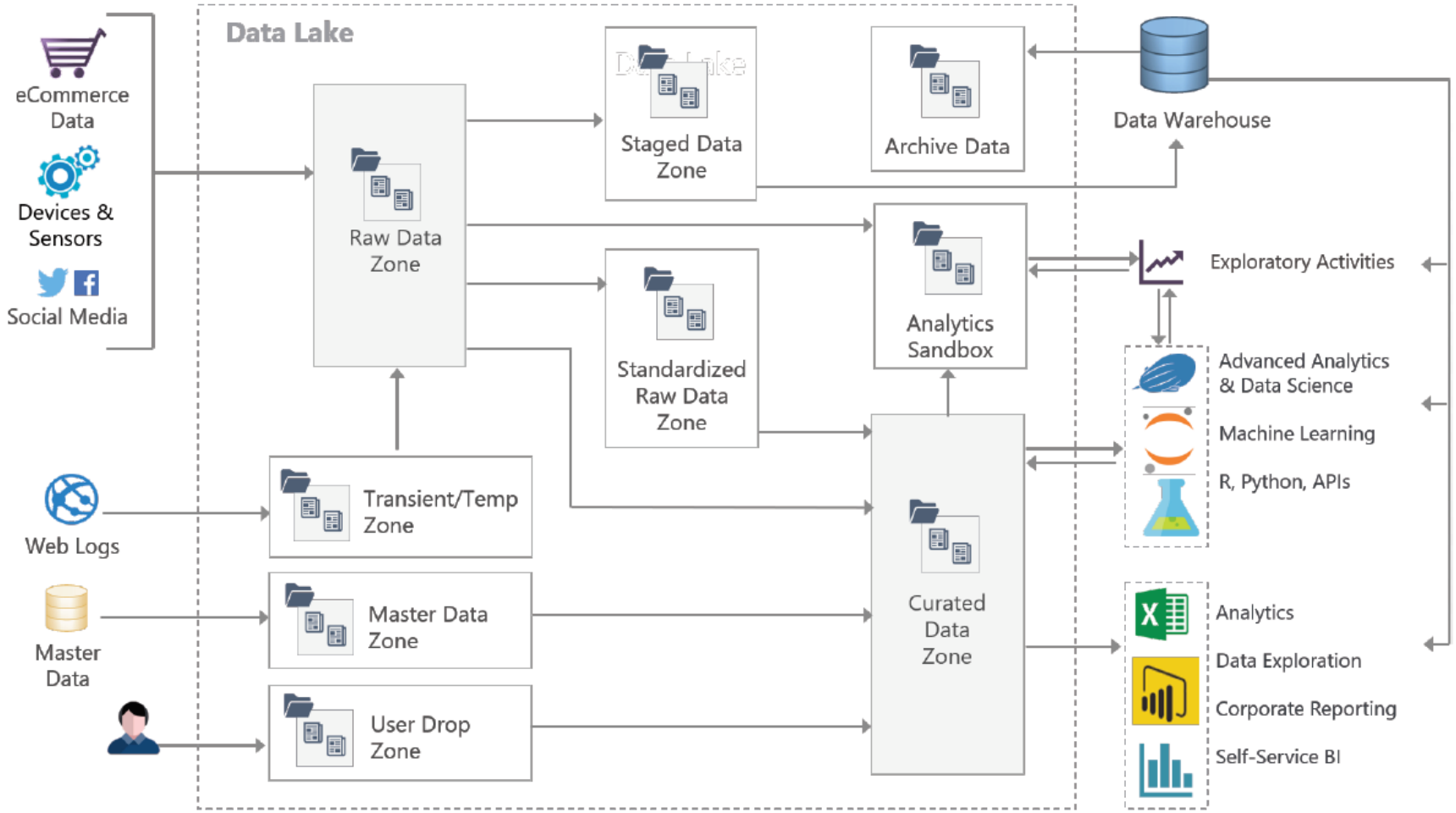


- Group/Company/SecurityMarking/purpose/System/Date/File

Data Lake Structure - Example



Data Lake Structure – more detailed separation example



Data Lake Zones

Last updated: Dec. 30, 2017

Raw Data Zone

- ✓ Exact copy of source data in native format
- ✓ Immutable to change
- ✓ History is retained
- ✓ Data access limited to very few people

Transient/Temp Zone

- ✓ Selectively utilized when data quality validation is required

Master Data Zone

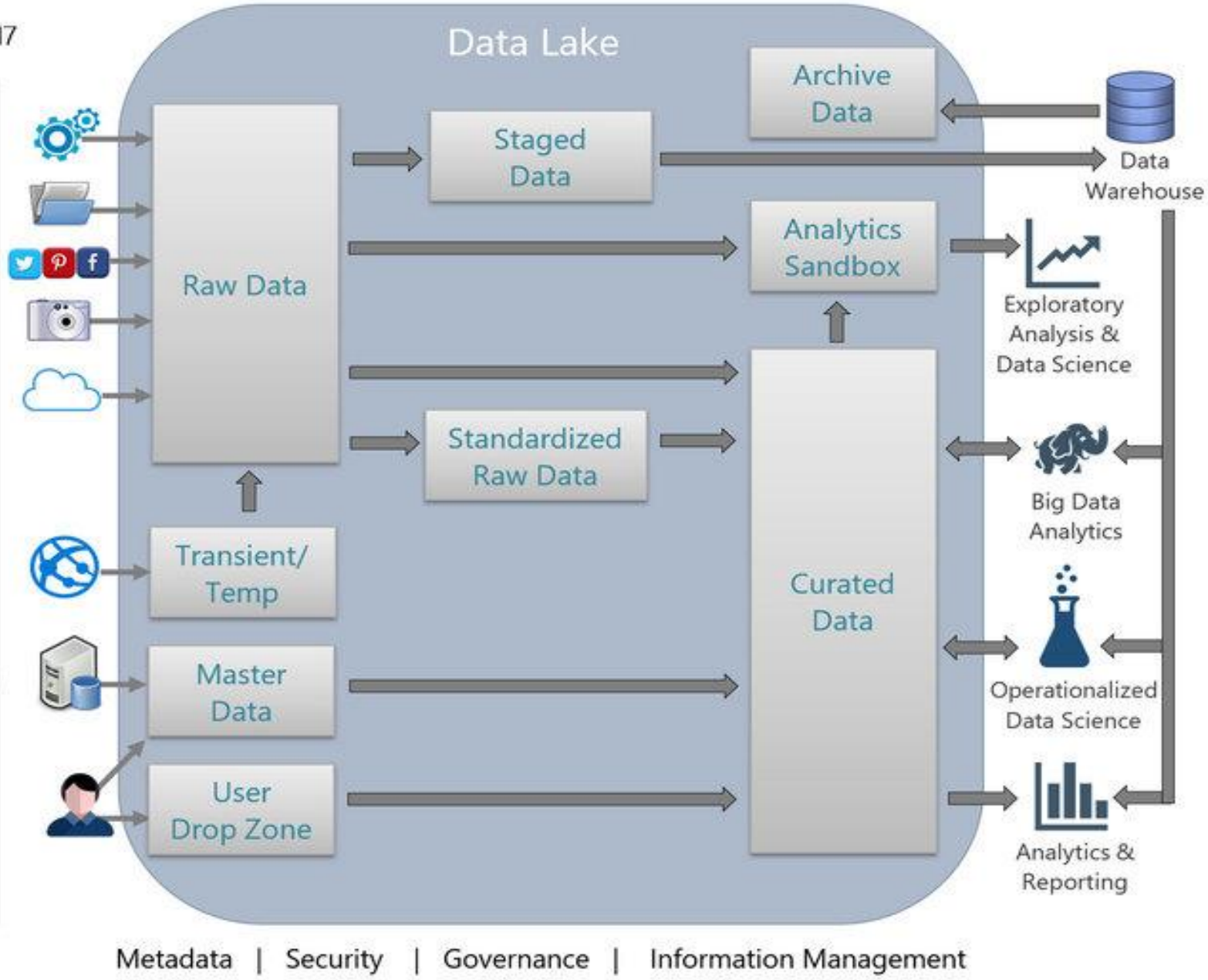
- ✓ Reference data to augment analysis

User Drop Zone

- ✓ Manually generated data to augment analysis

Staged Data Zone

- ✓ Data staged for a specific purpose (ex: to load a data warehouse)



Standardized Raw Data

- ✓ Applicable to data structures which vary in format (ex: multi-structured JSON documents standardized into columns & rows)

Archive Data Zone

- ✓ Active archive with aged data which is available for querying when needed

Analytics Sandbox

- ✓ Workspace for exploratory activities
- ✓ Valuable efforts are productionized to the Curated Data Zone

Curated Data Zone

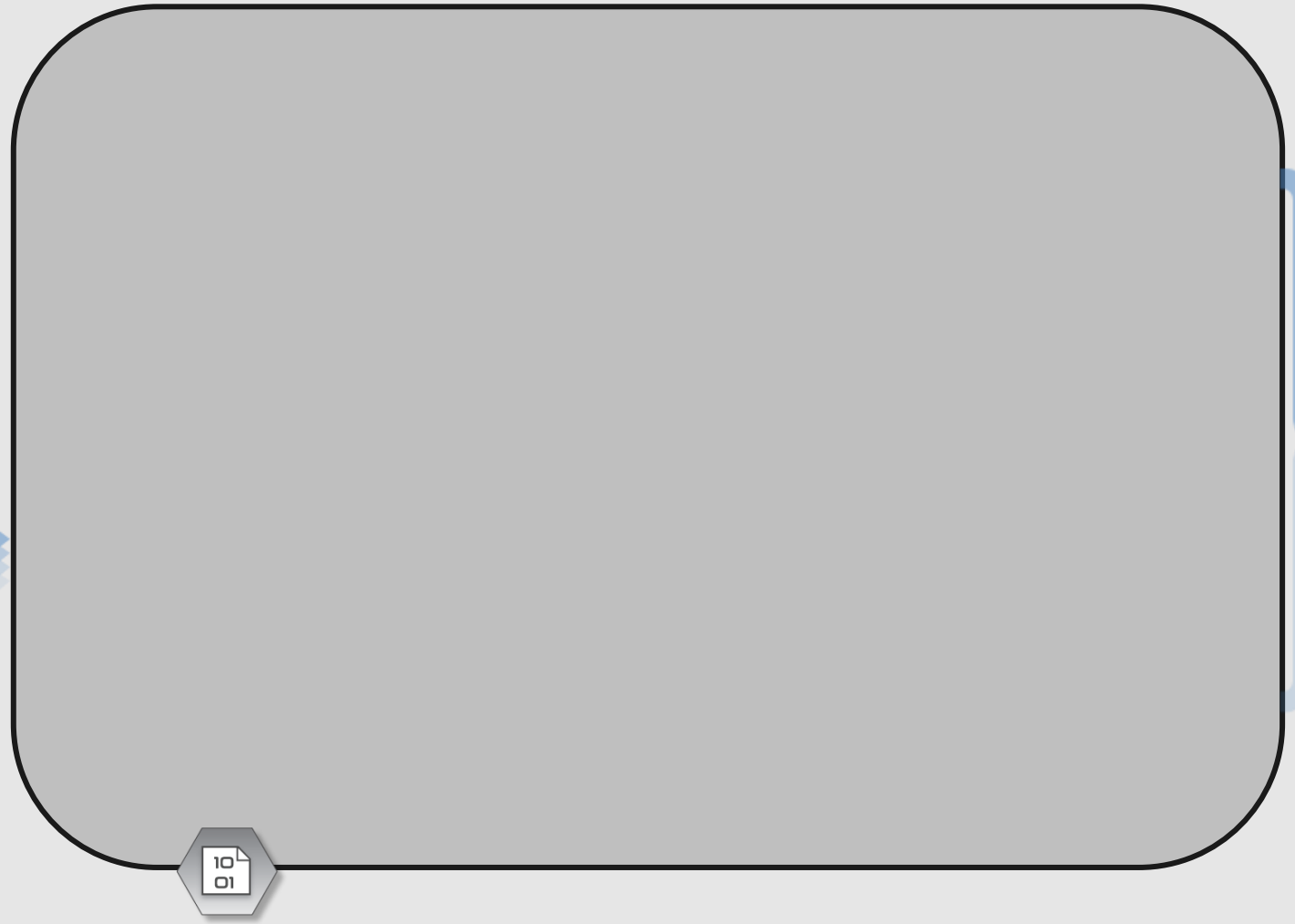
- ✓ Cleansed and transformed
- ✓ Organized optimally for data delivery
- ✓ Supports self-service data access

Microsoft

Azure
Data Lake one possible Model

One model

-- or how to organize



On-premises data

Oracle, SQL, Teradata, file shares, SAP…

Cloud data

Azure, AWS, GCP…

SaaS data

Salesforce, Dynamics…

Power BI

Cloud

Azure Analysis Services

VNet

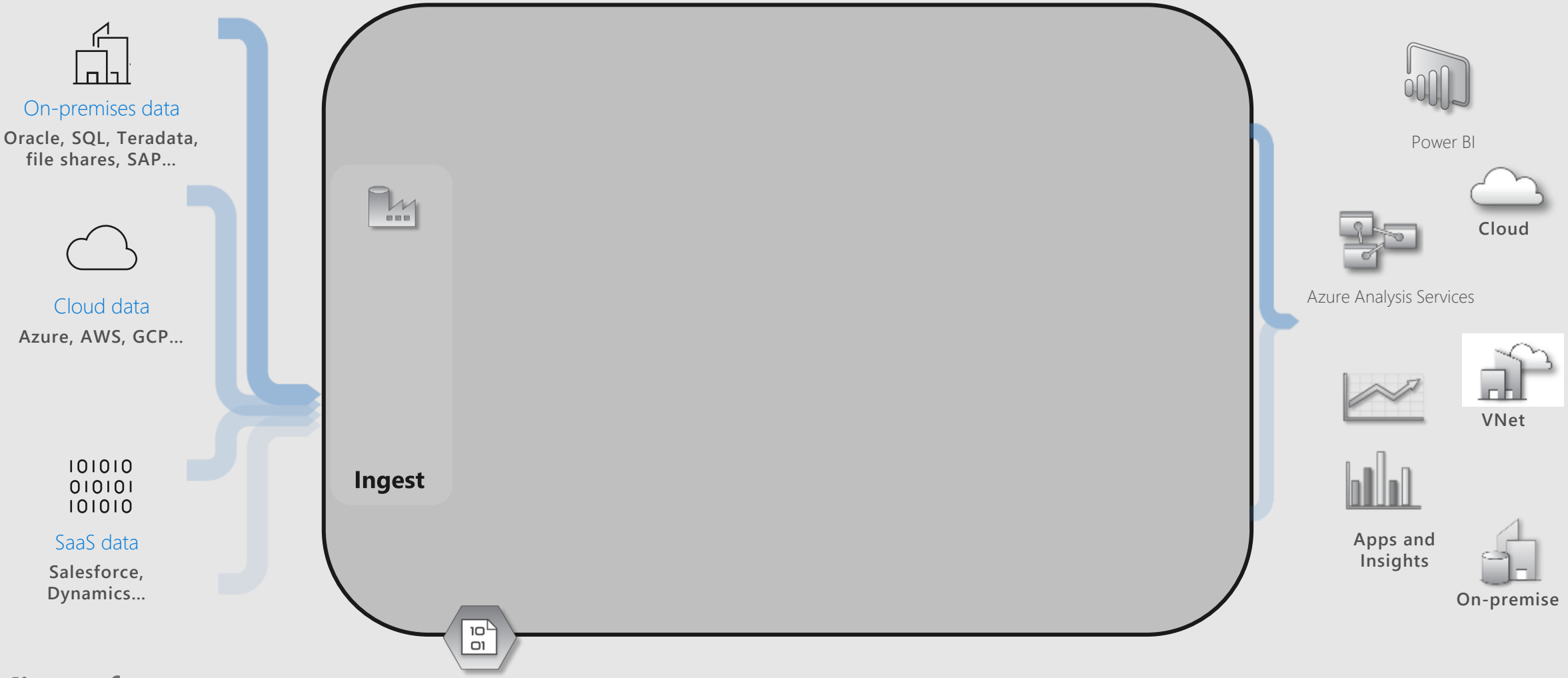
Apps and Insights

On-premise

Microsoft

The Data Bank model

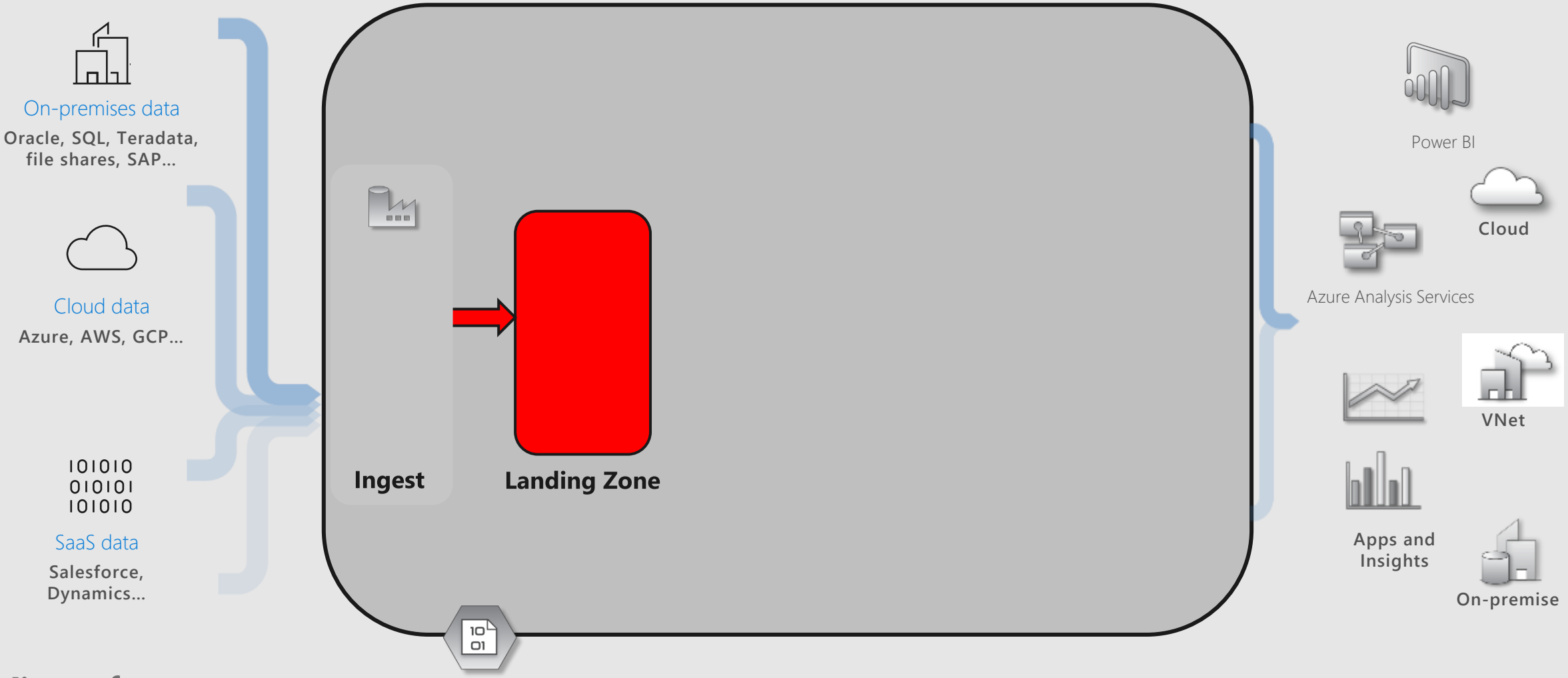
-- or how to organize



Microsoft

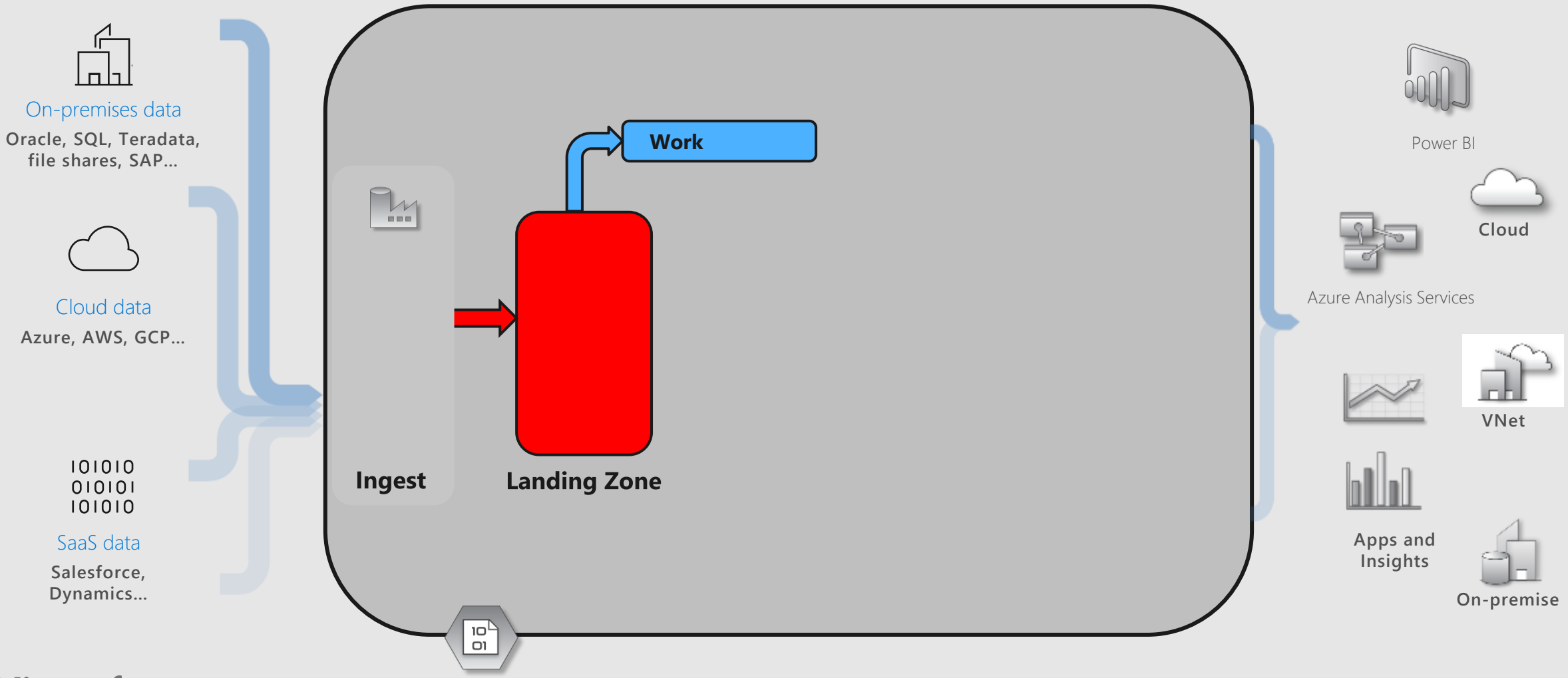
One model

-- or how to organize



Microsoft

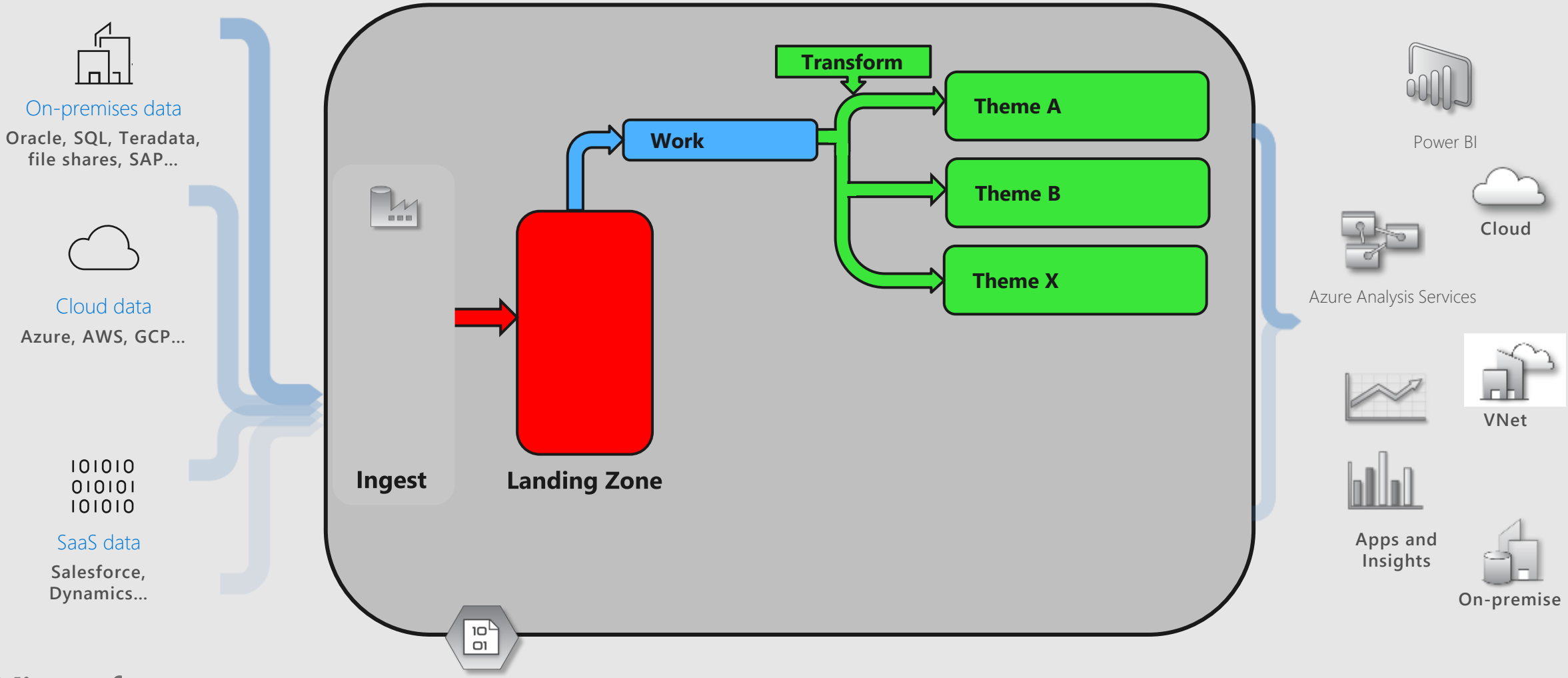
One model



Microsoft

One model

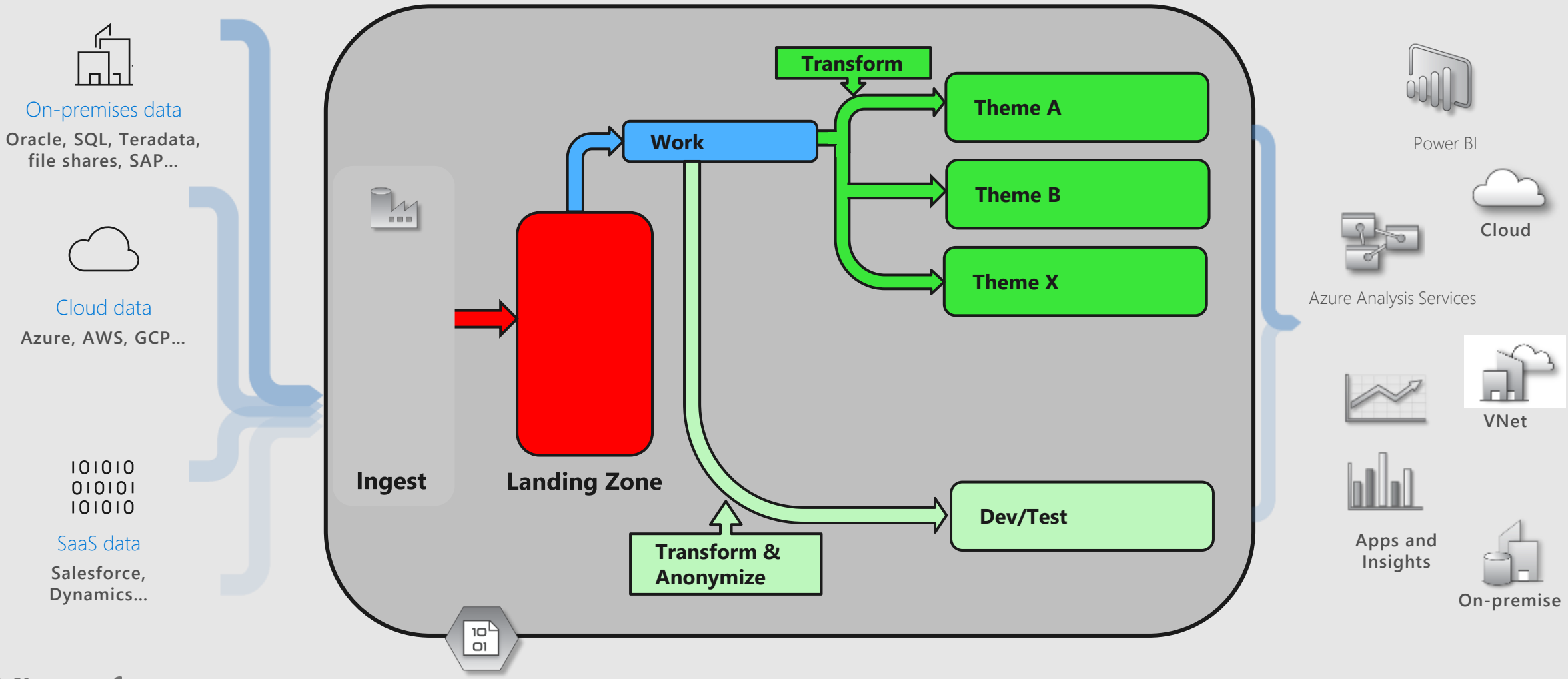
-- or how to organize



Microsoft

One model

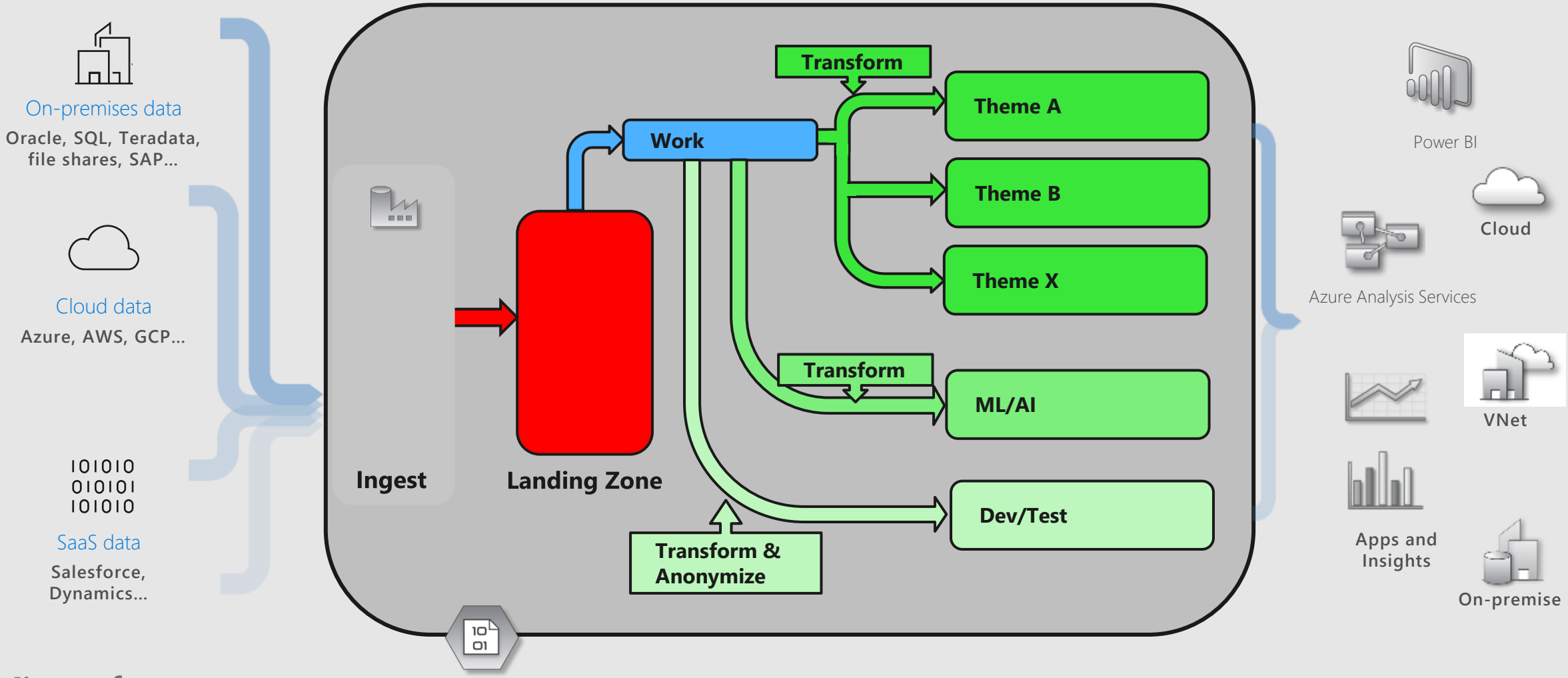
-- or how to organize



Microsoft

One model

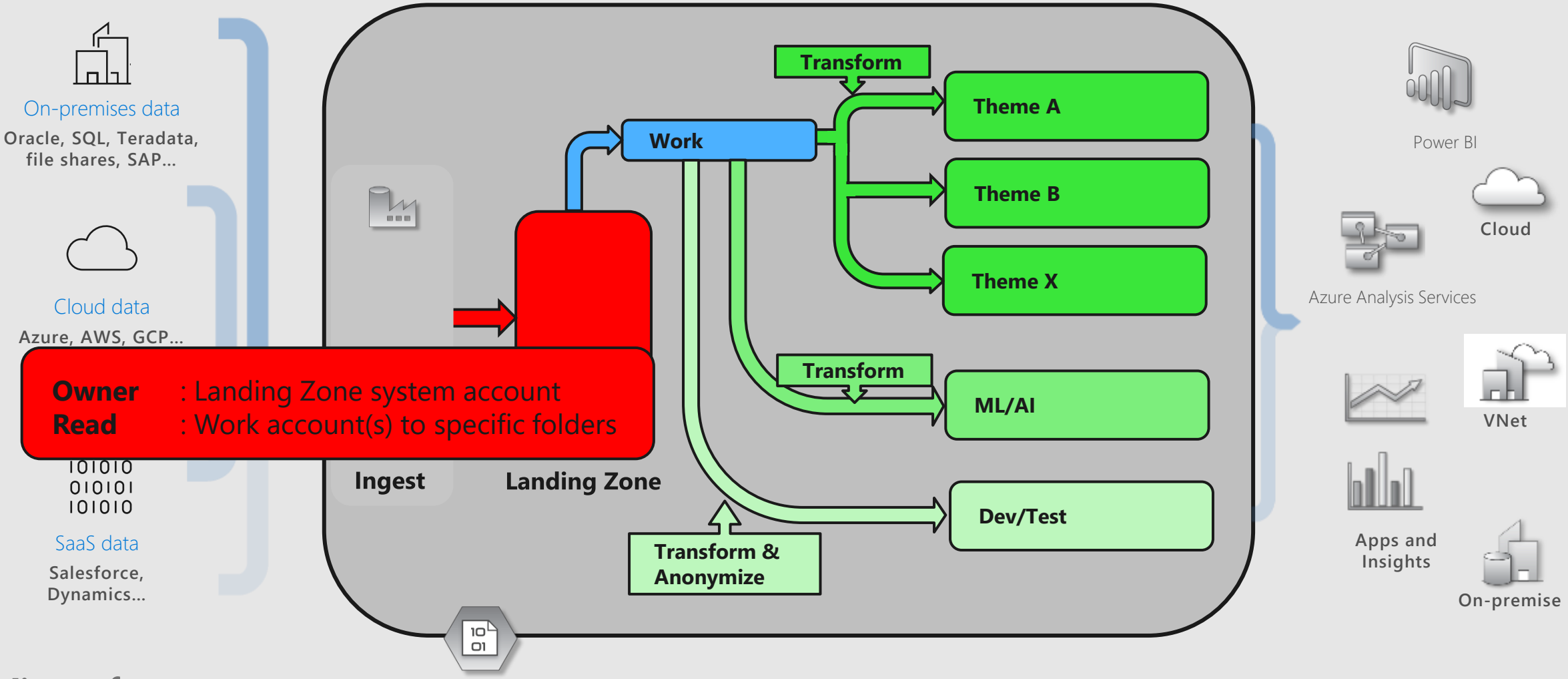
-- or how to organize



Microsoft

One model

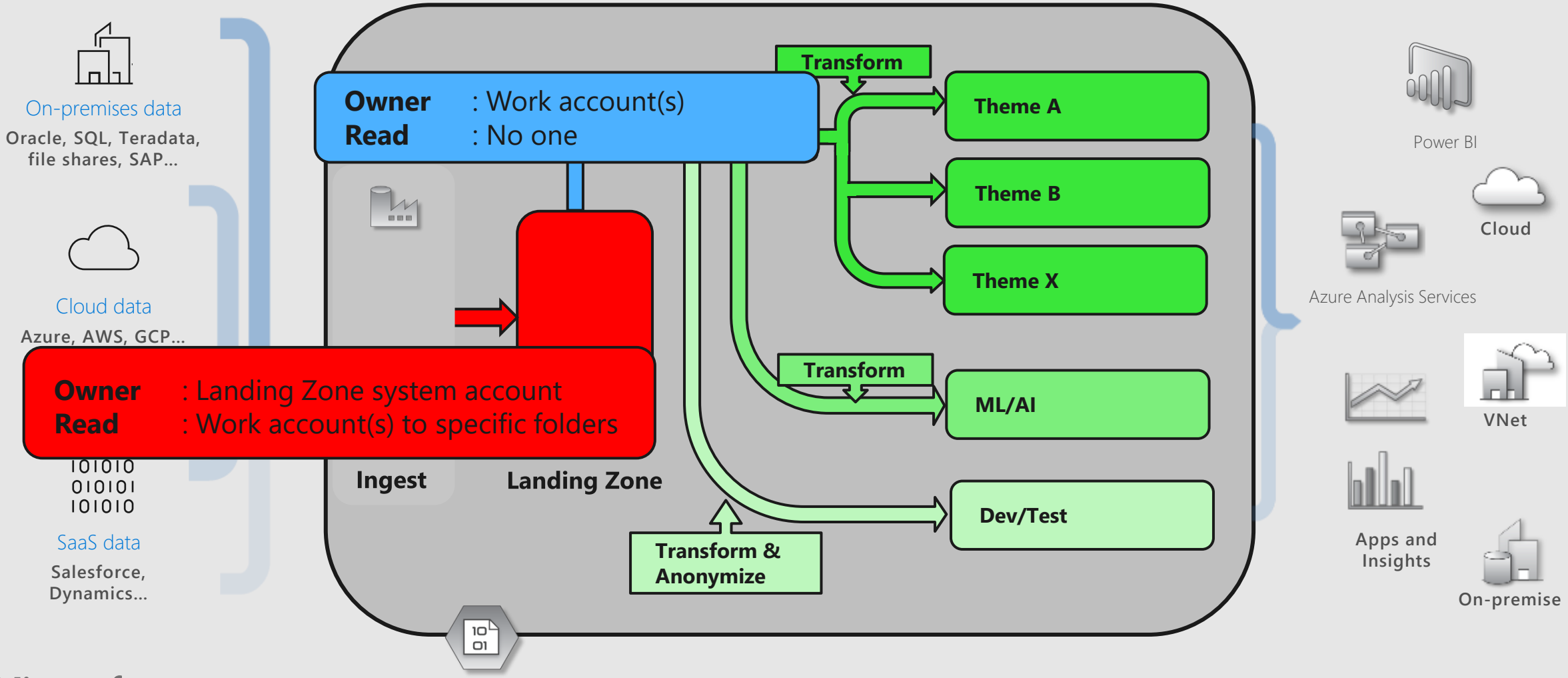
-- or how to organize



Microsoft

One model

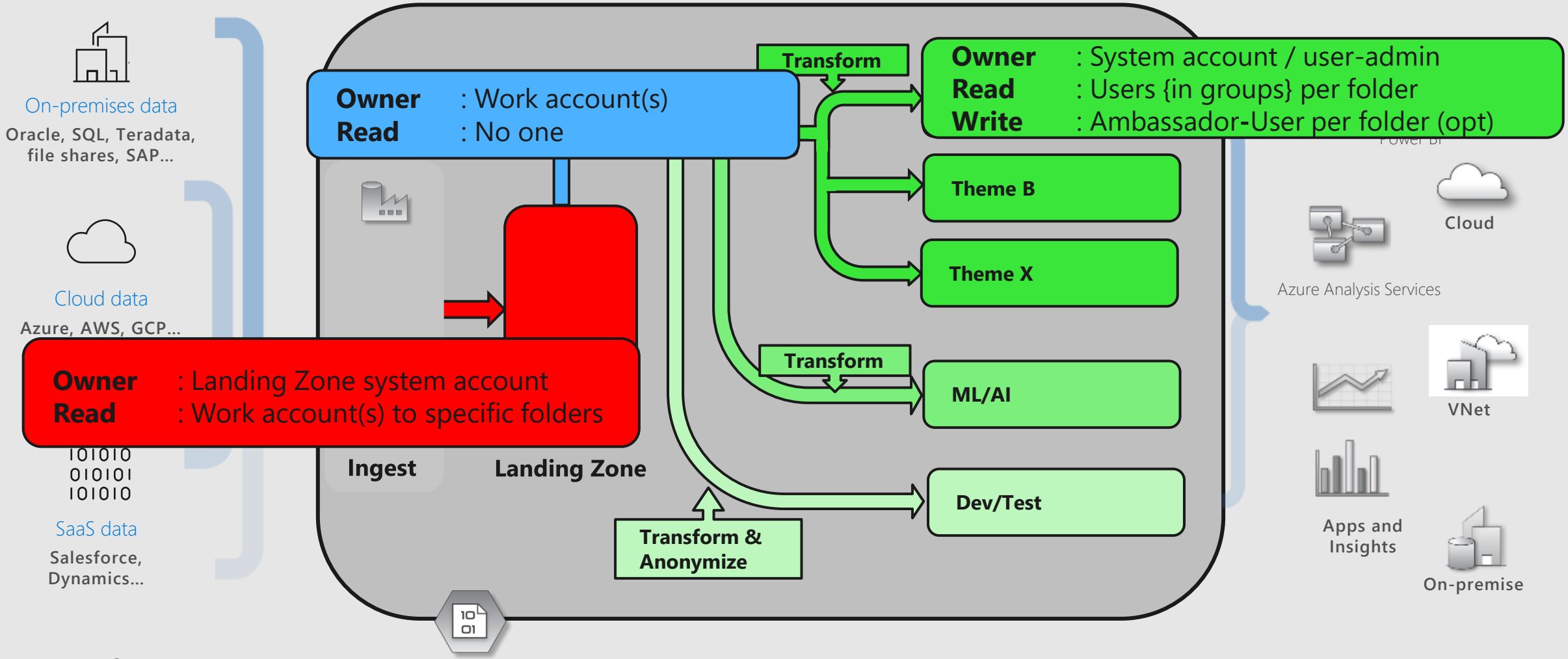
-- or how to organize



Microsoft

One model

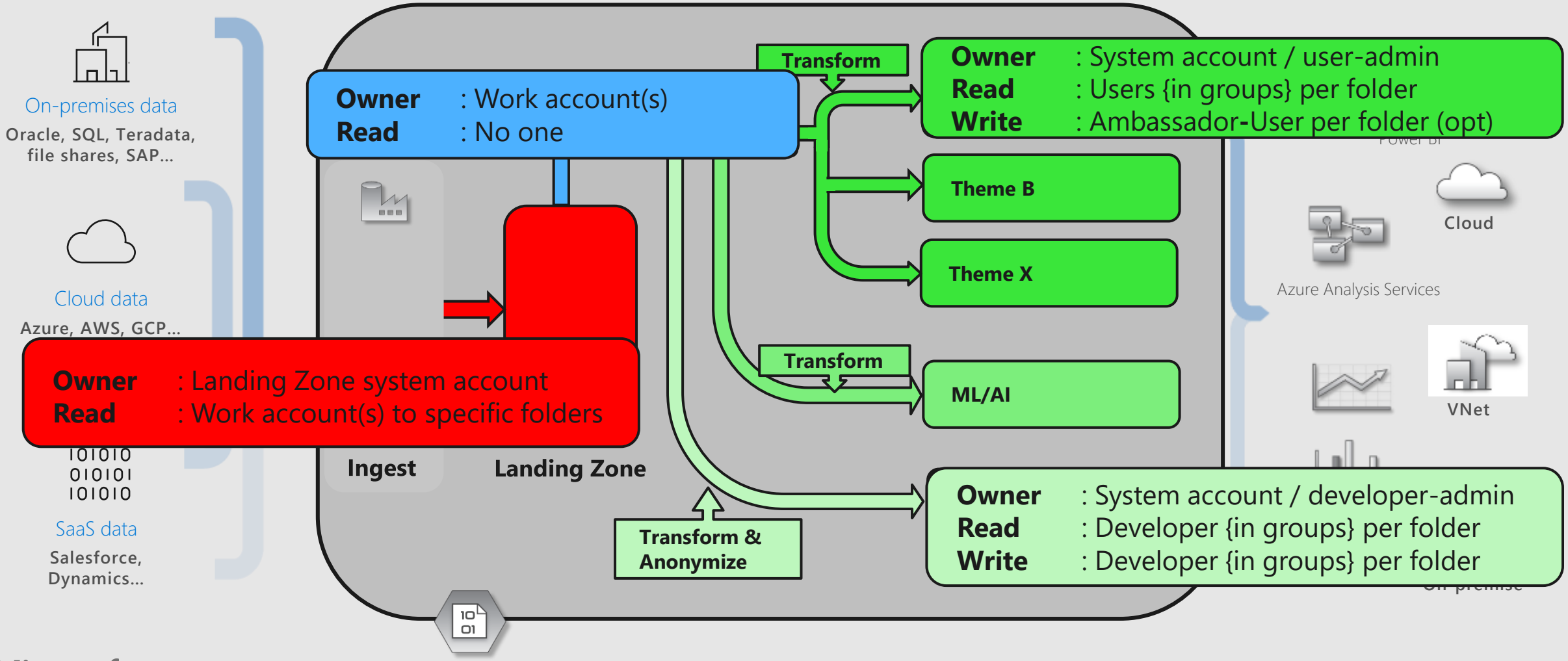
-- or how to organize



Microsoft

One model

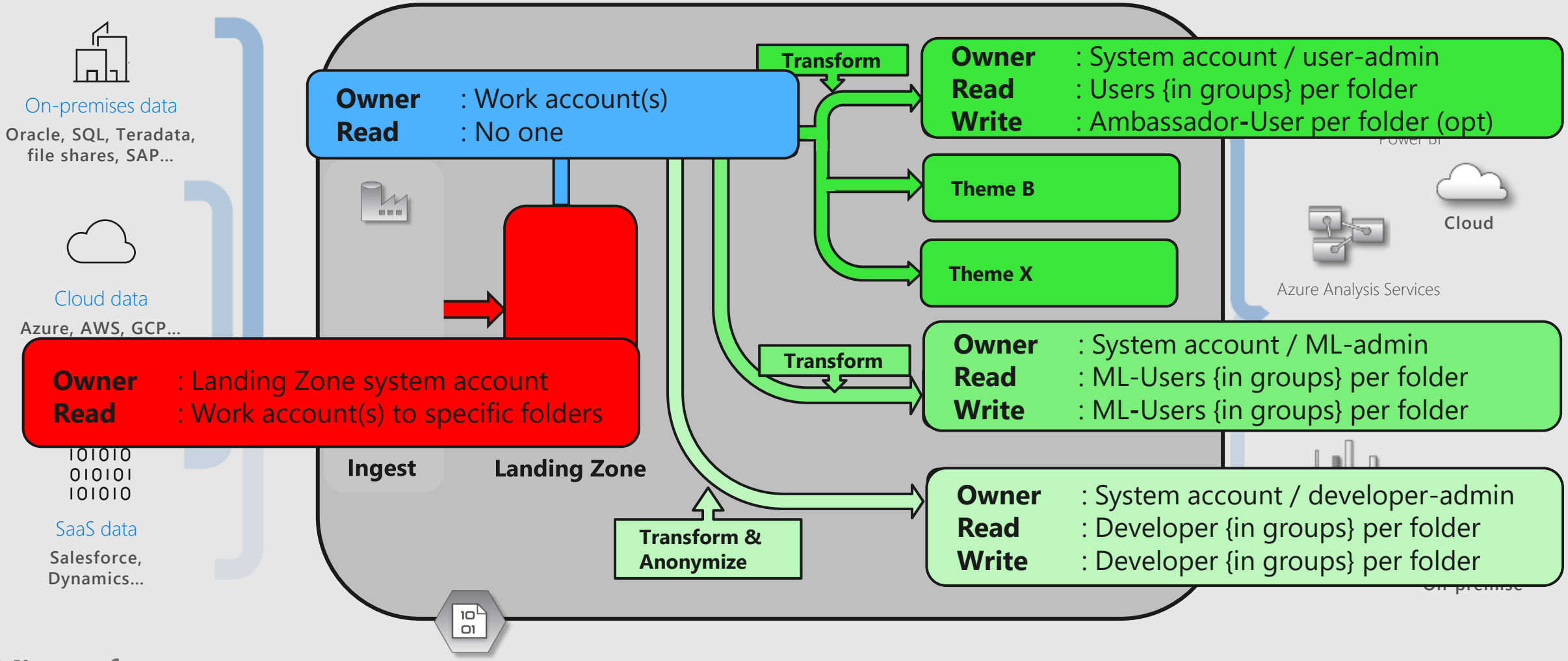
-- or how to organize



Microsoft

One model

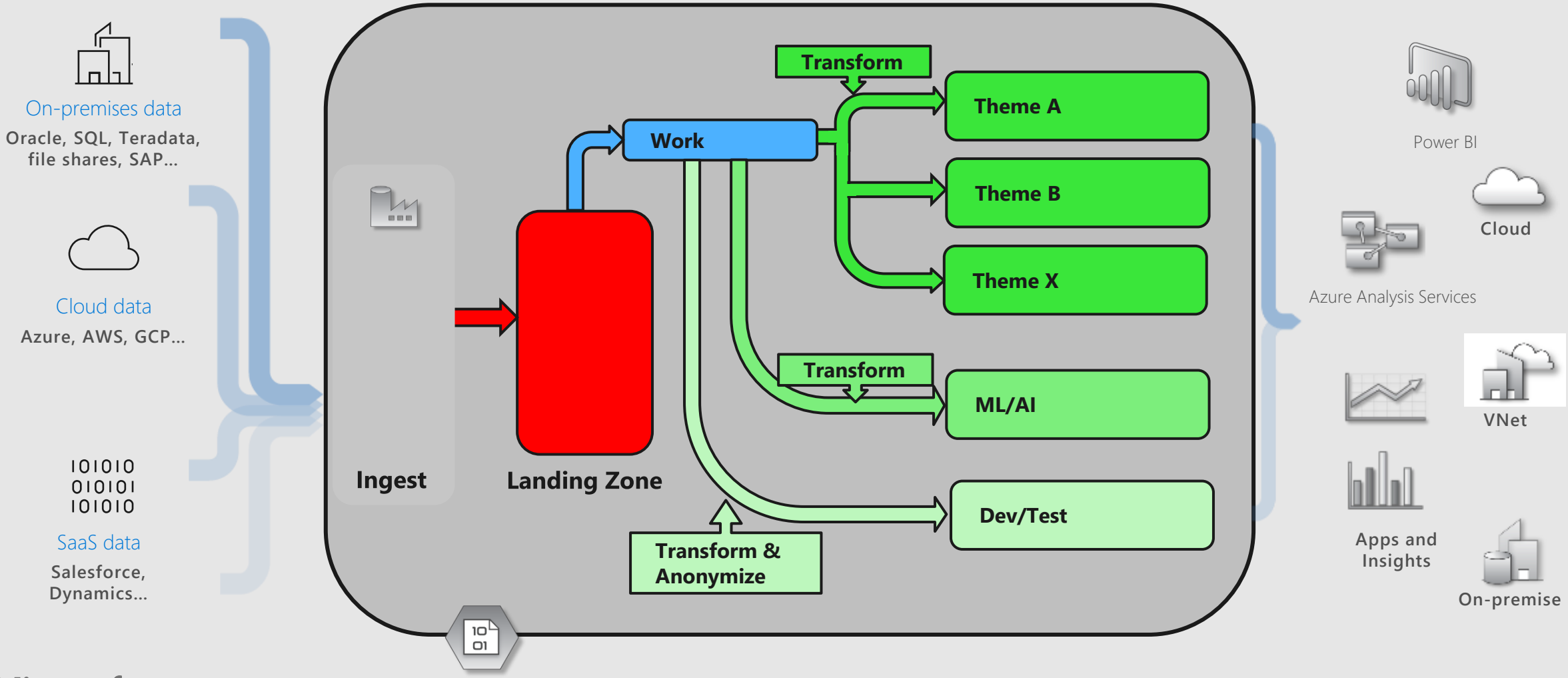
-- or how to organize



Microsoft

One model

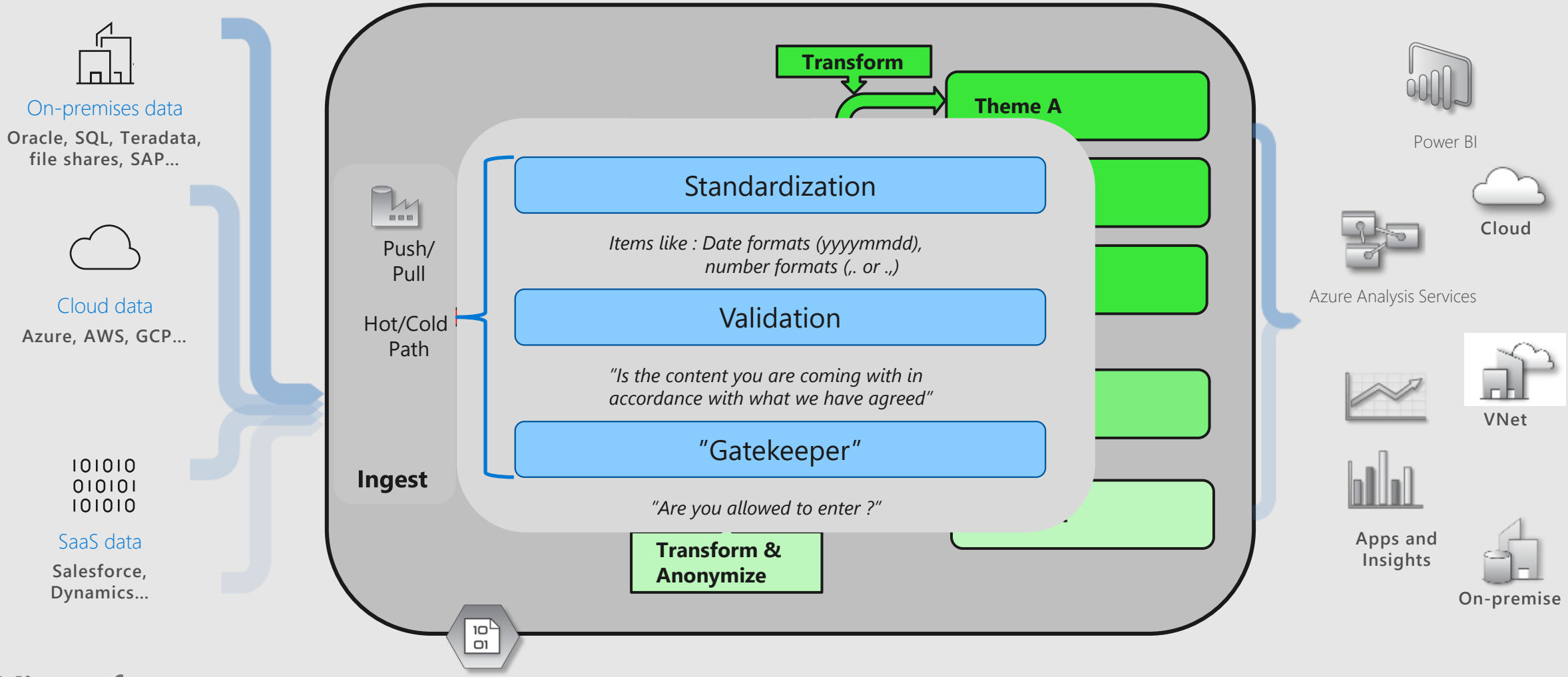
-- or how to organize



Microsoft

One model

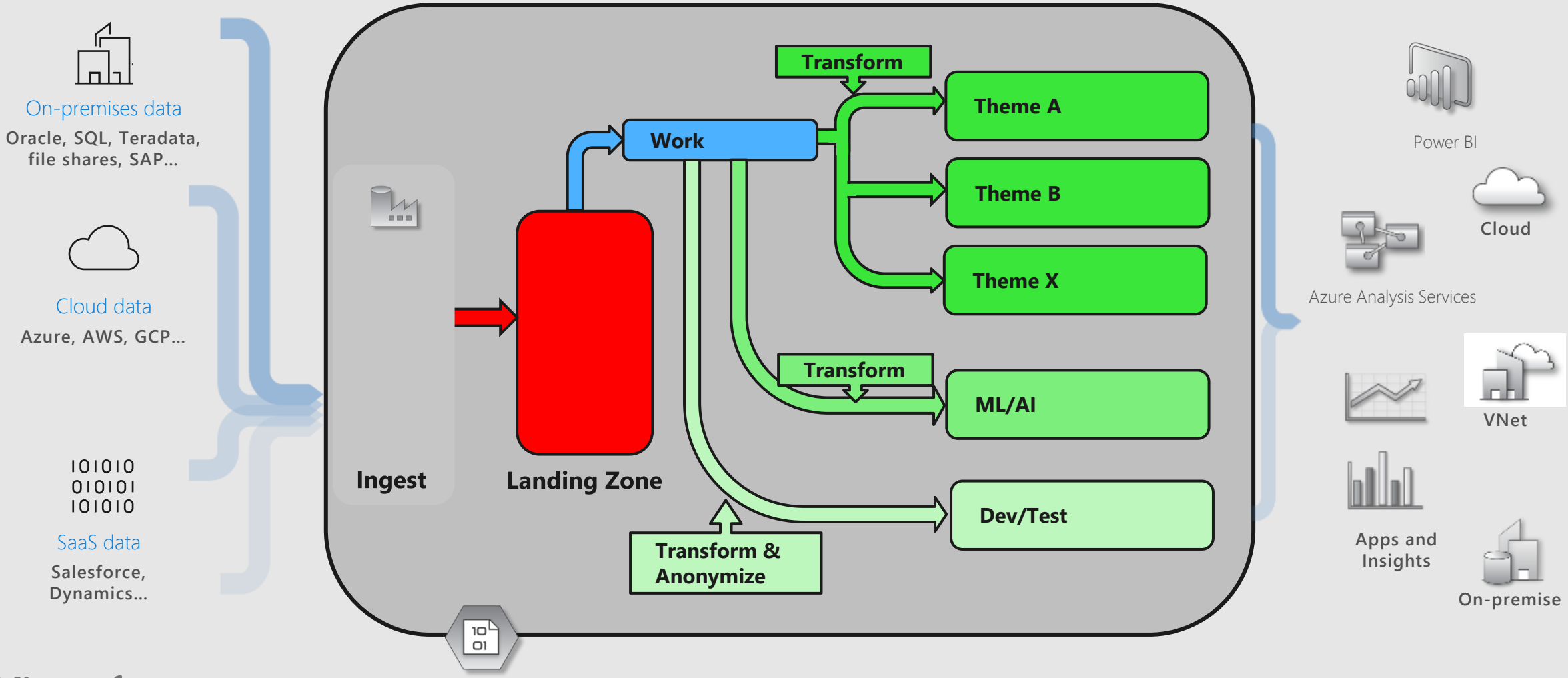
-- or how to organize



Microsoft

One model

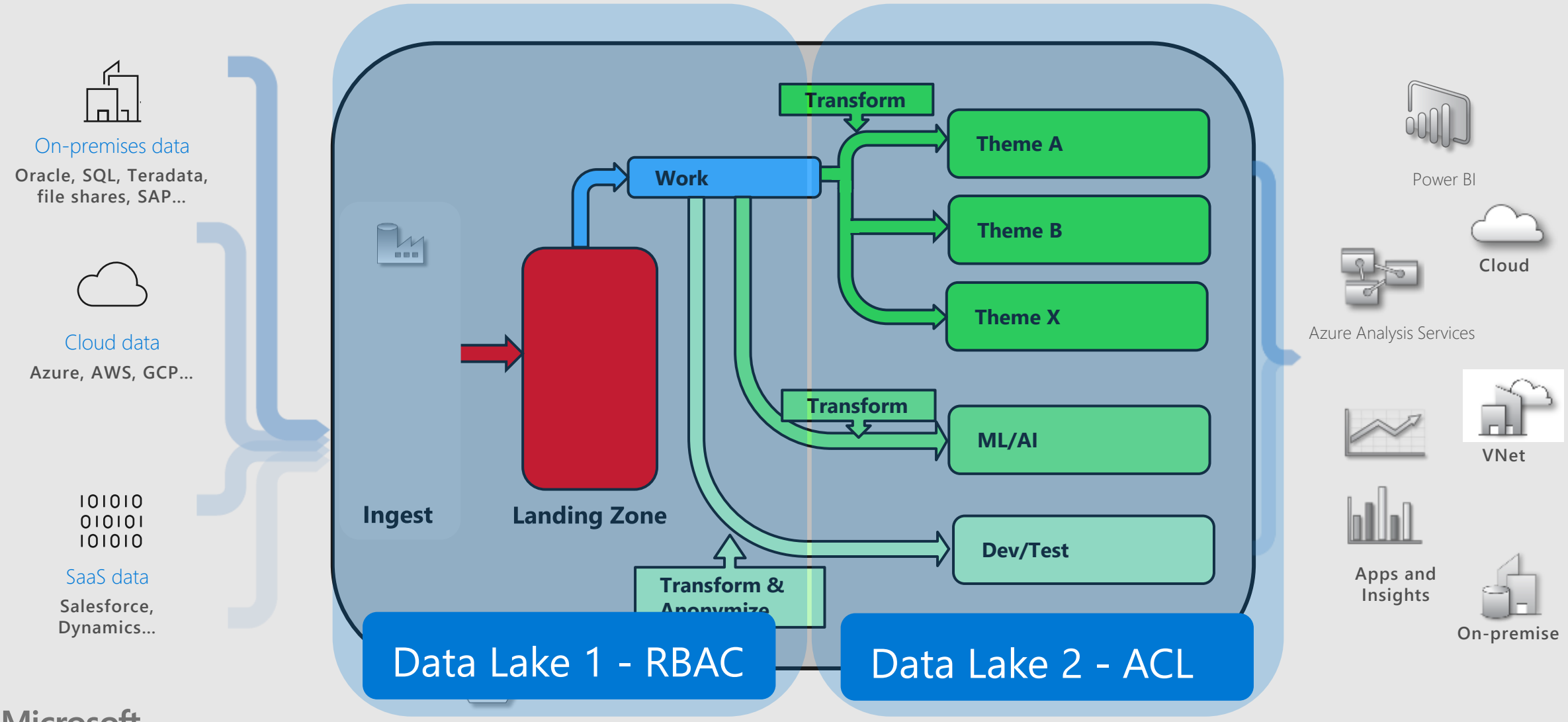
-- or how to organize



Microsoft

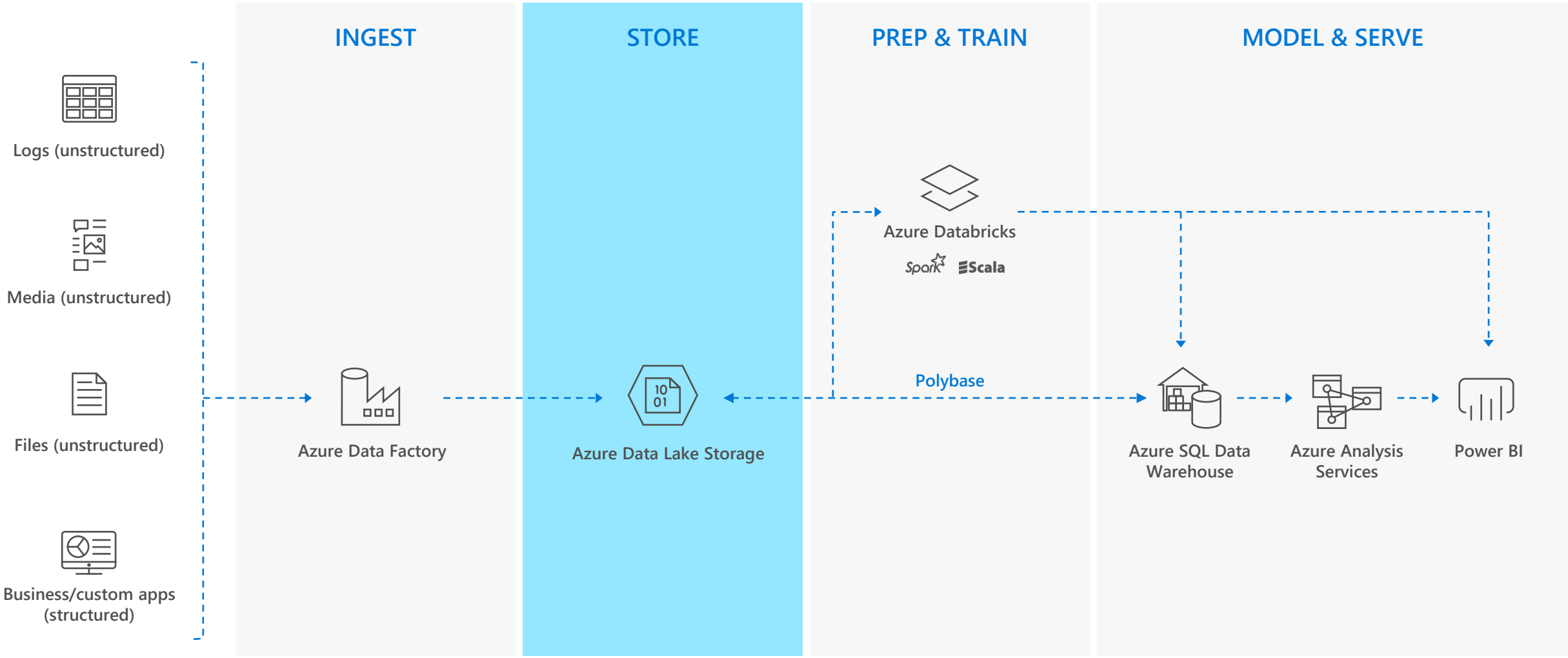
One model

-- or how to organize



Microsoft

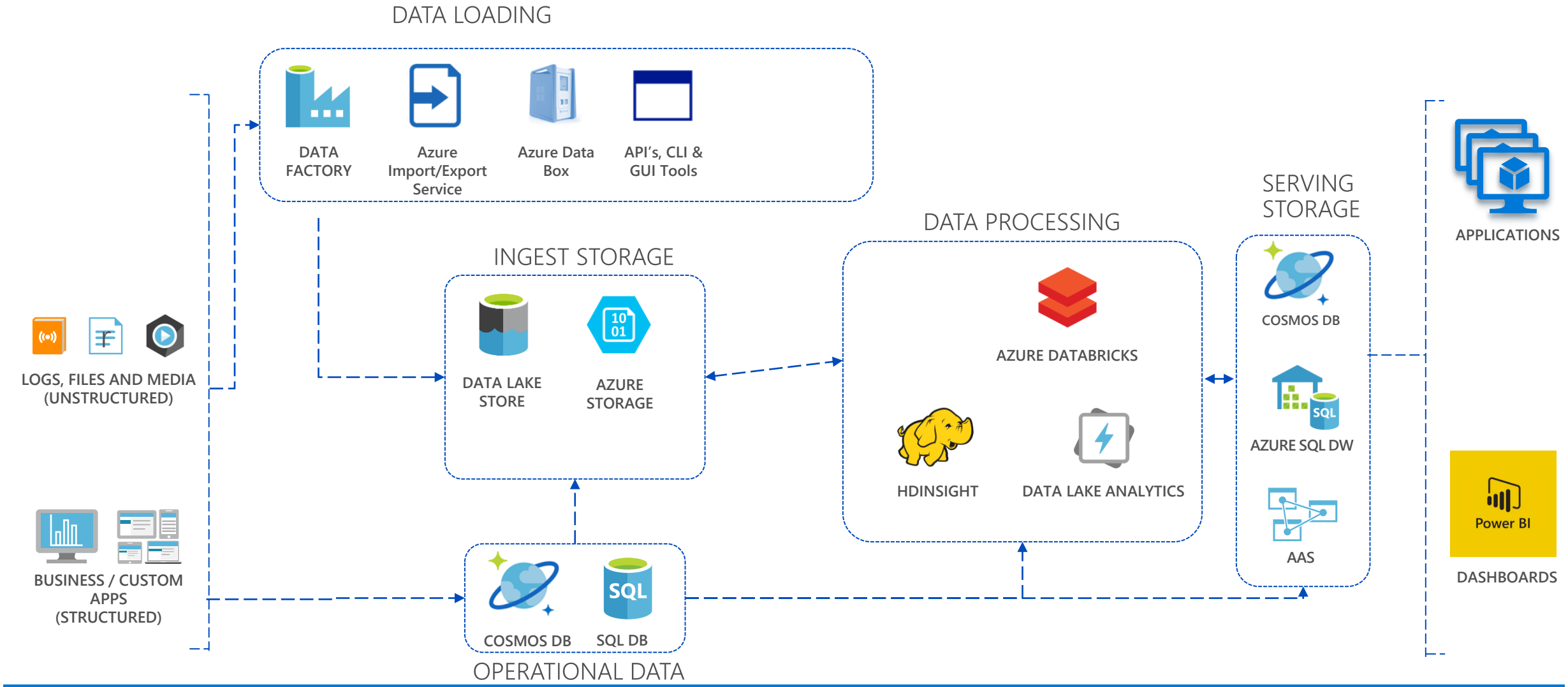
MODERN DATA WAREHOUSE



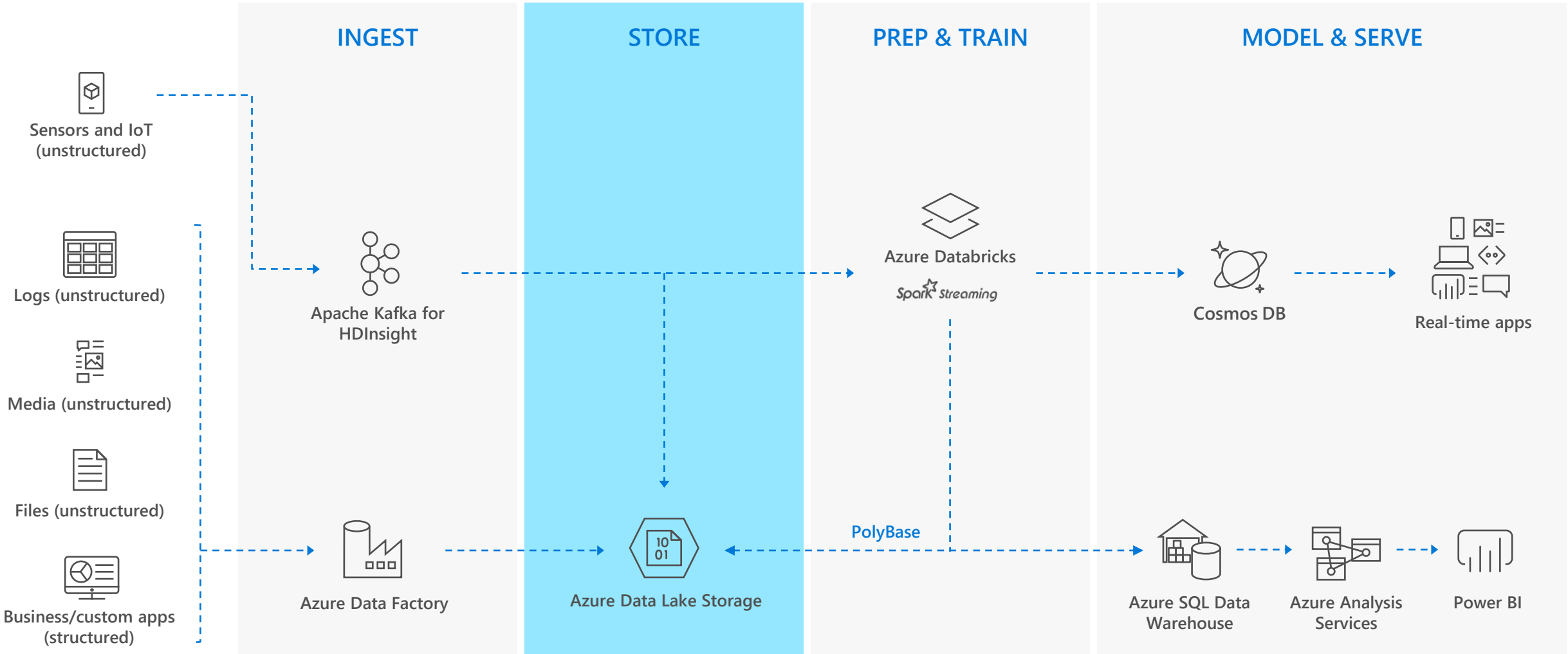
Enterprise Data Catalog

Data Warehousing Pattern IN Azure

Loading and preparing data for analysis with a data warehouse



REAL-TIME ANALYTICS



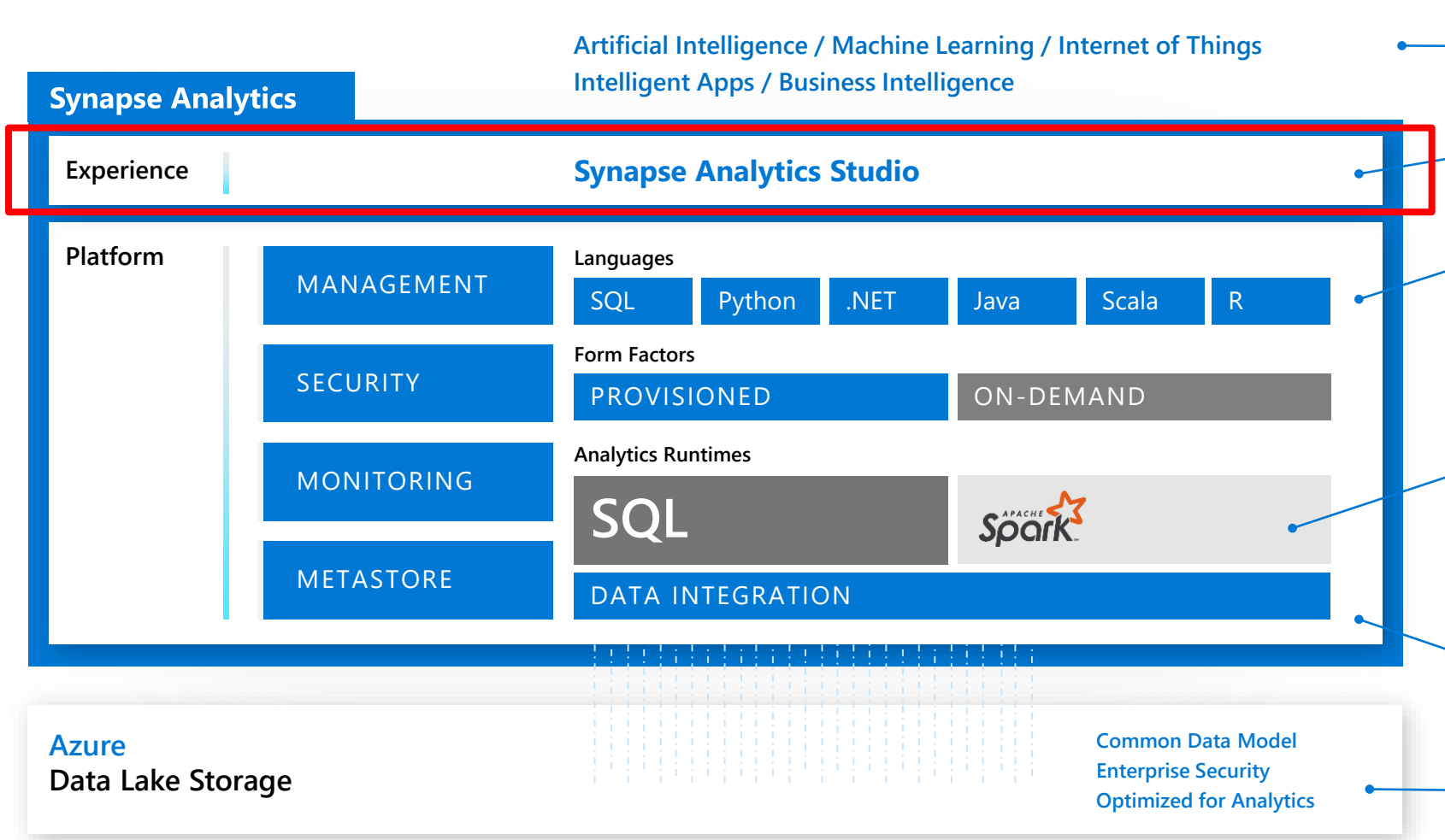
Enterprise Data Catalog

Microsoft

Azure Synapse Analytics

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

Demo Azure Synapse Analytics