Architectural blueprints for the Modern Data Warehouse

SQLDay

10th May 2022

ORANGEMAN

SQL Day

14 edycja konferencji SQLDay

9-11 maja 2022, WROCŁAW + ONLINE

Data Community

partner złoty

Future Processing

Infinite DATA

KPMG

partner srebrny

Objectivity

summ-it

UBS

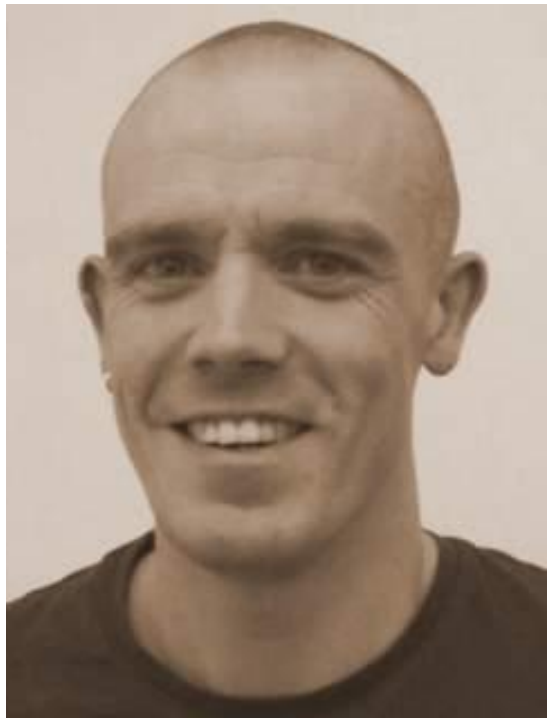
dbWatch DATABASE CONTROL

partner brązowy

elitmind think bright

Just Blindbæk

- Self-employed BI consultant in justB
- Trainer at Orange Man
- Founder
  - Danish Microsoft BI Community (MsBIP.dk)
  - Power BI UG Denmark (PowerBI.dk)
- Strong focus on
  - Azure BI architecture
  - Analysis Services
  - Reporting Services
  - Power BI
- just@justB.dk / blog.justB.dk / @justblindbaek / youtube.com/c/justblindbaek



Microsoft CERTIFIED Trainer

Microsoft CERTIFIED IT Professional

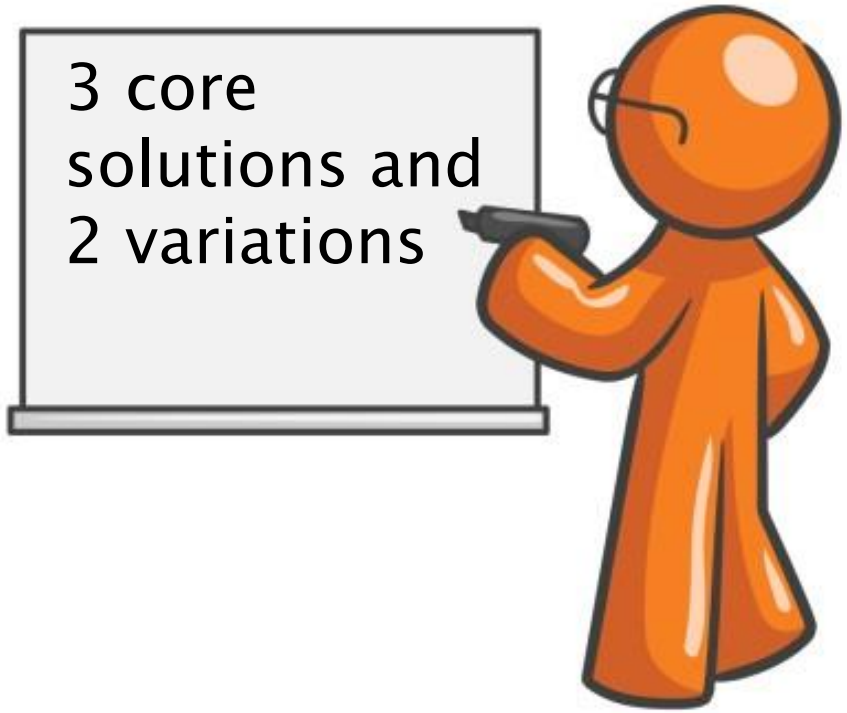
SSAS MAESTRO by Microsoft

MVP Microsoft® Most Valuable Professional

ORANGEMAN

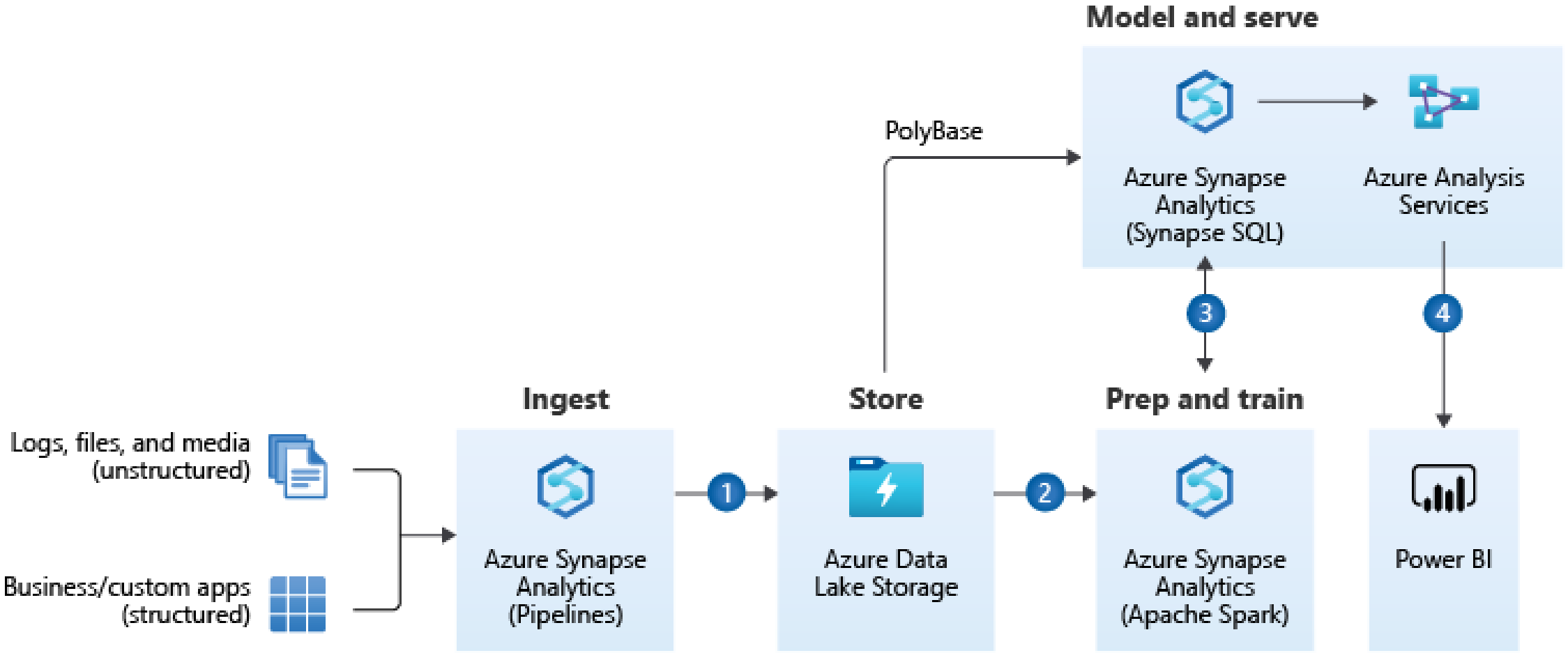
Agenda

- What is a Modern Data Warehouse?
- Traditional Modern Data Warehouse
- Data Lakehouse with Spark
- Self-service with Power BI Dataflows
- Comparison of the three solutions
- Other variations



ORANGEMAN

Enterprise Data Warehouse in Azure



ORANGEMAN

https://docs.microsoft.com/en-us/azure/architecture/solution-ideas/articles/enterprise-data-warehouse

Enterprise Data Warehouse in Azure

Model and serve

Enterprise Data Warehouse

Azure Synapse Analytics	Tier: Compute Optimized Gen2, Dedicated SQL Pool...	**Monthly: $1,755.90**
Azure Analysis Services	Developer (Hours), 5 Instance(s), 720 Hours	**Monthly: $475.20**
Storage Accounts	Data Lake Storage Gen2, Standard, LRS Redundancy...	**Monthly: $71.88**

Estimated upfront cost $0.00

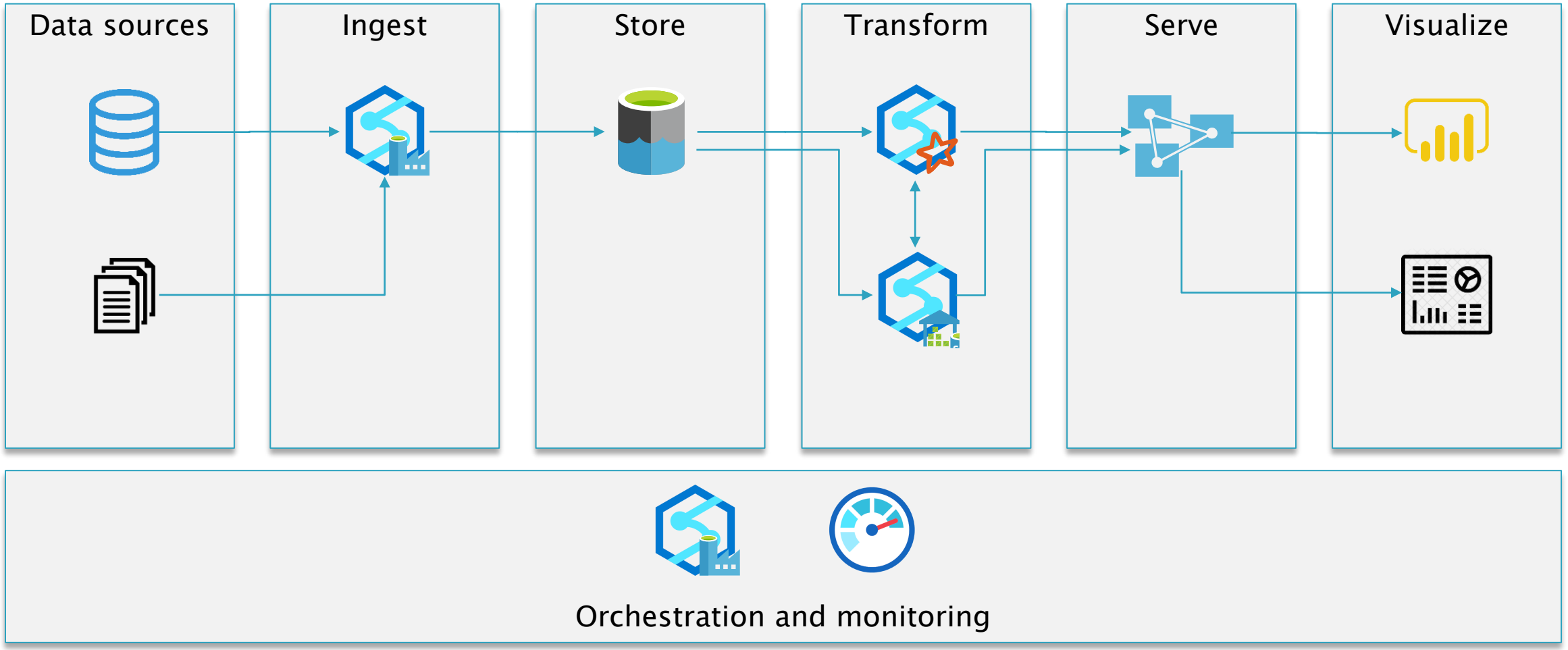
Estimated monthly cost $2,302.98

(structured) (Pipelines) (Apache Spark)

ORANGEMAN

https://docs.microsoft.com/en-us/azure/architecture/solution-ideas/articles/enterprise-data-warehouse

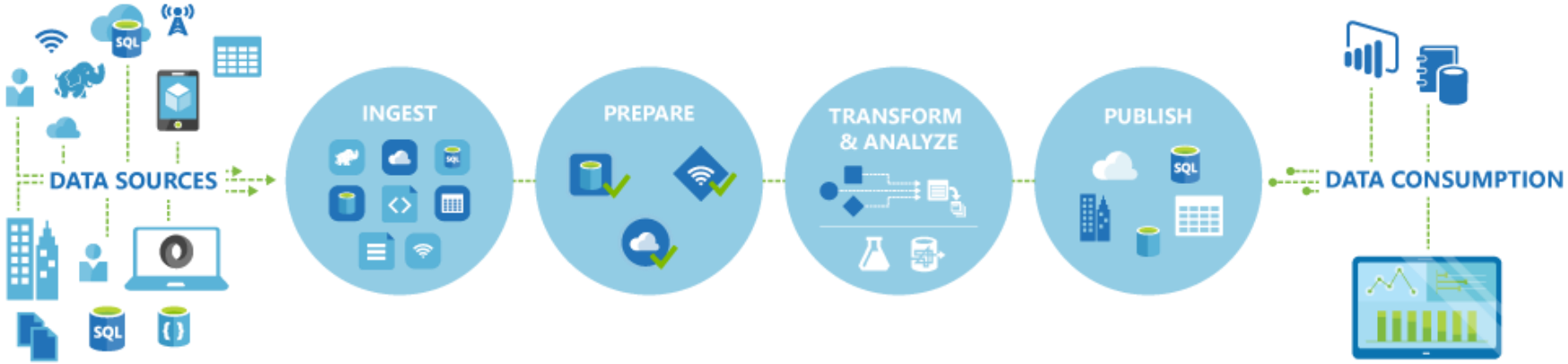
Enterprise Data Warehouse in Azure



ORANGEMAN

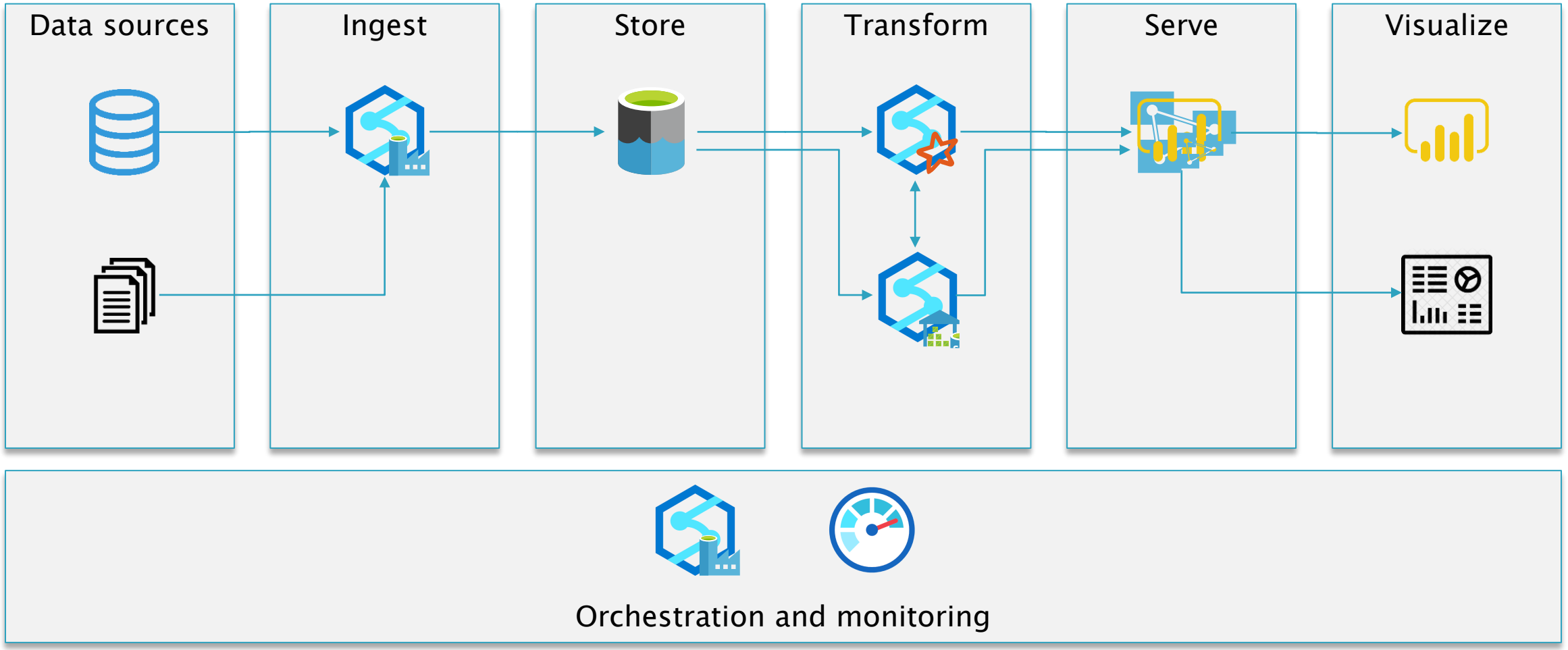
Azure Data Factory in Synapse

- Hybrid data integration at scale
- 80+ connectors provided to Cloud & Hybrid
- Create, plan, administer and monitor data pipelines
- Execute activities (copy, transform or orchestration)
- Parameters, control flow and triggers



ORANGEMAN

Enterprise Data Warehouse in Azure

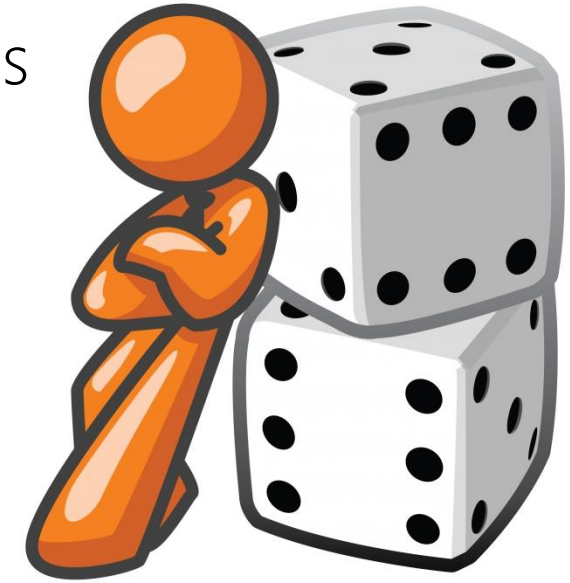


ORANGEMAN

Analysis Services in Power BI

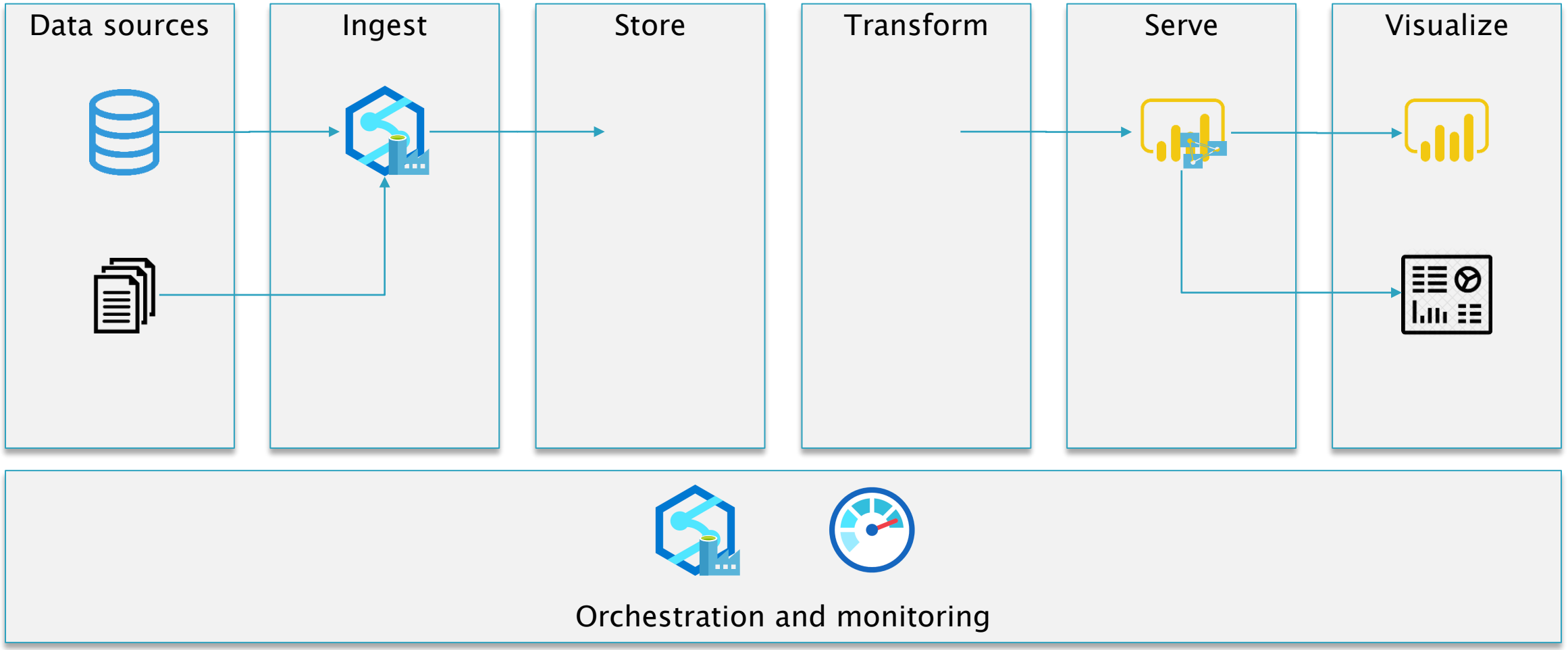
- Analysis Services engine inside
- XMLA endpoint to manage the Semantic data models
- Deploy models directly to a “premium” workspace from Tabular Editor or SQL Server Data Tools.

- Requires dedicated capacity when deploying the models
- Recommended license options
  - Power BI Embedded capacity
  - Power BI Premium Per User



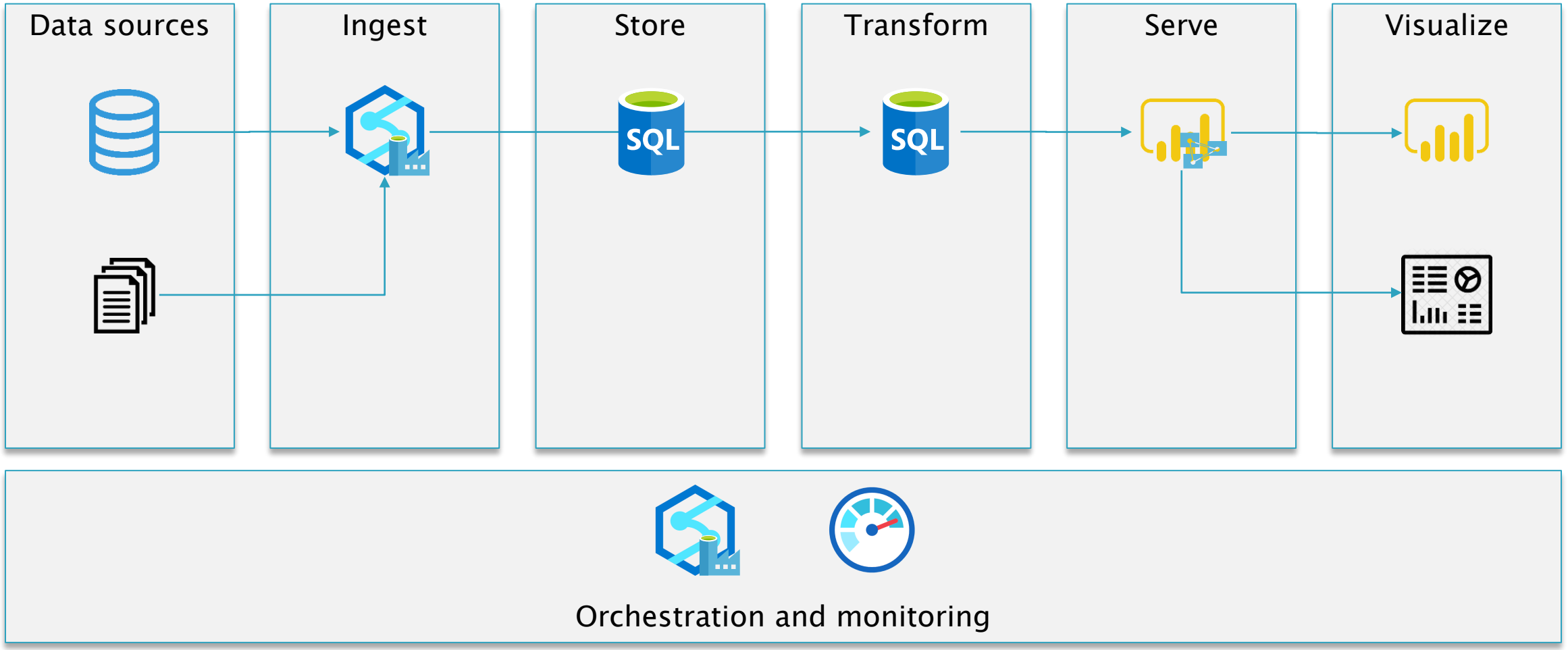
ORANGEMAN

Where to Store and Transform?



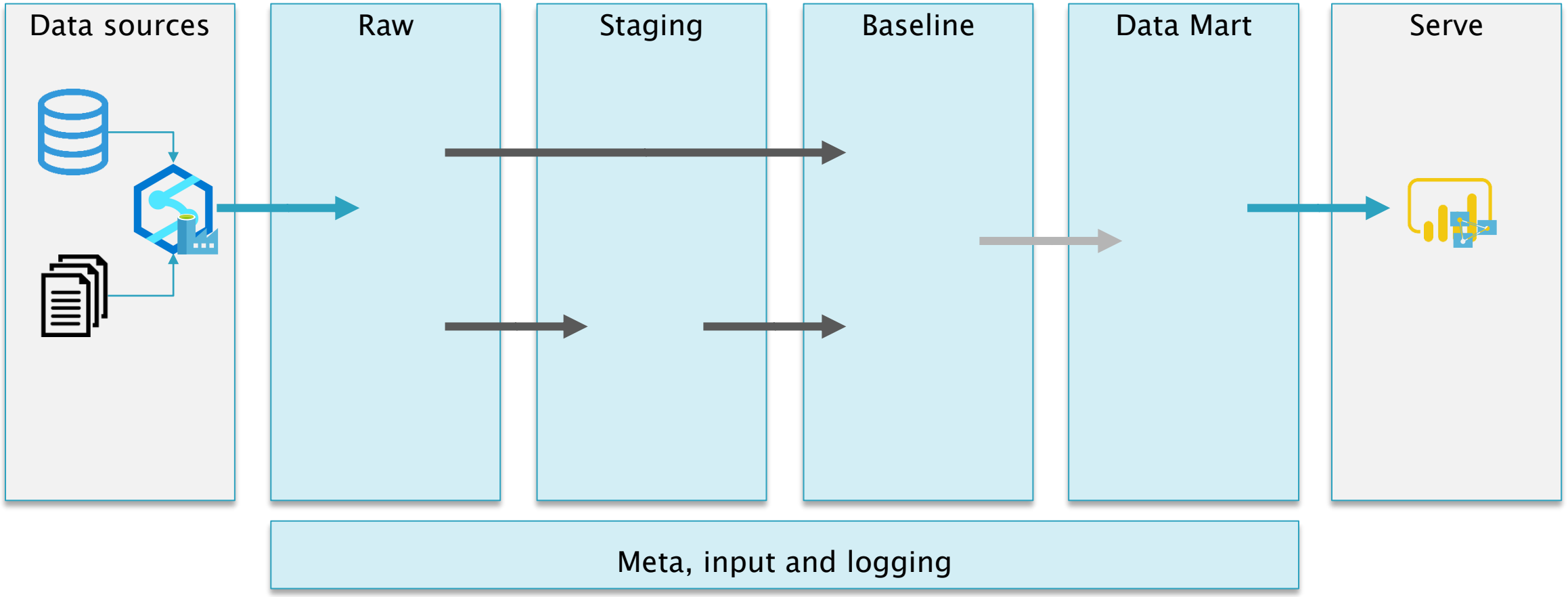
ORANGEMAN

(1) Traditional Modern Data Warehouse



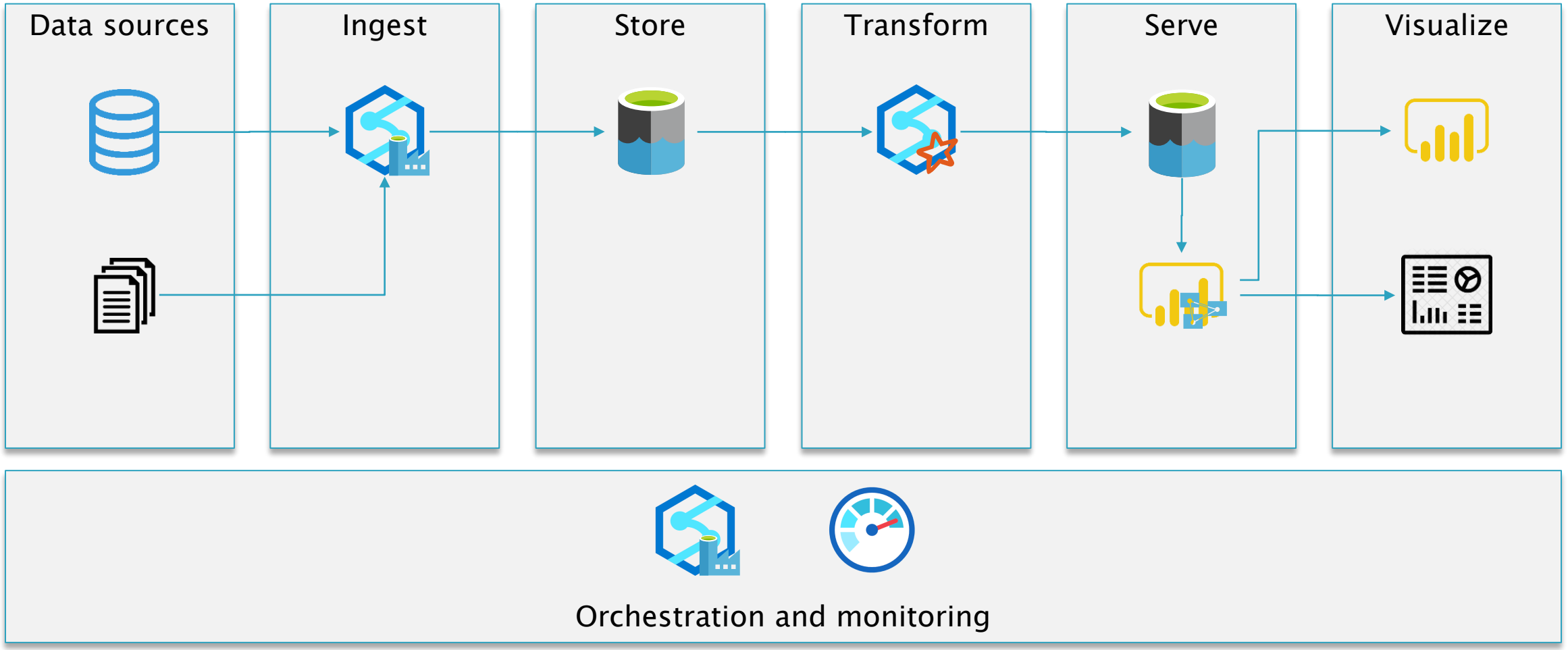
ORANGEMAN

Transform "framework" in SQL Database



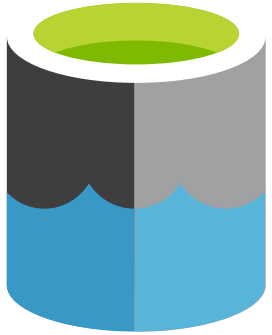
ORANGEMAN

(2) Data Lakehouse with Spark

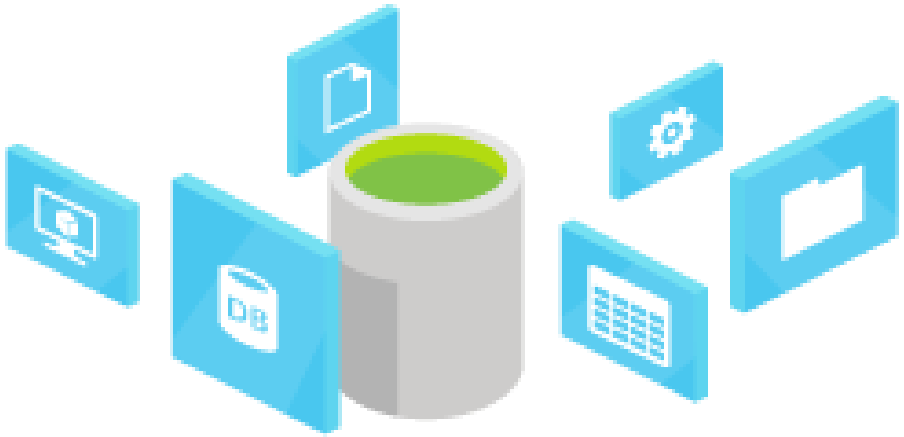


ORANGEMAN

Azure Data Lake Store



- Storage Account with “hierarchical namespace” enabled
- Unlimited cheap, cheap, cheap storage
- Can handle structured, semi-structured, and unstructured data
- Apache Hadoop file system compatible (HDFS)

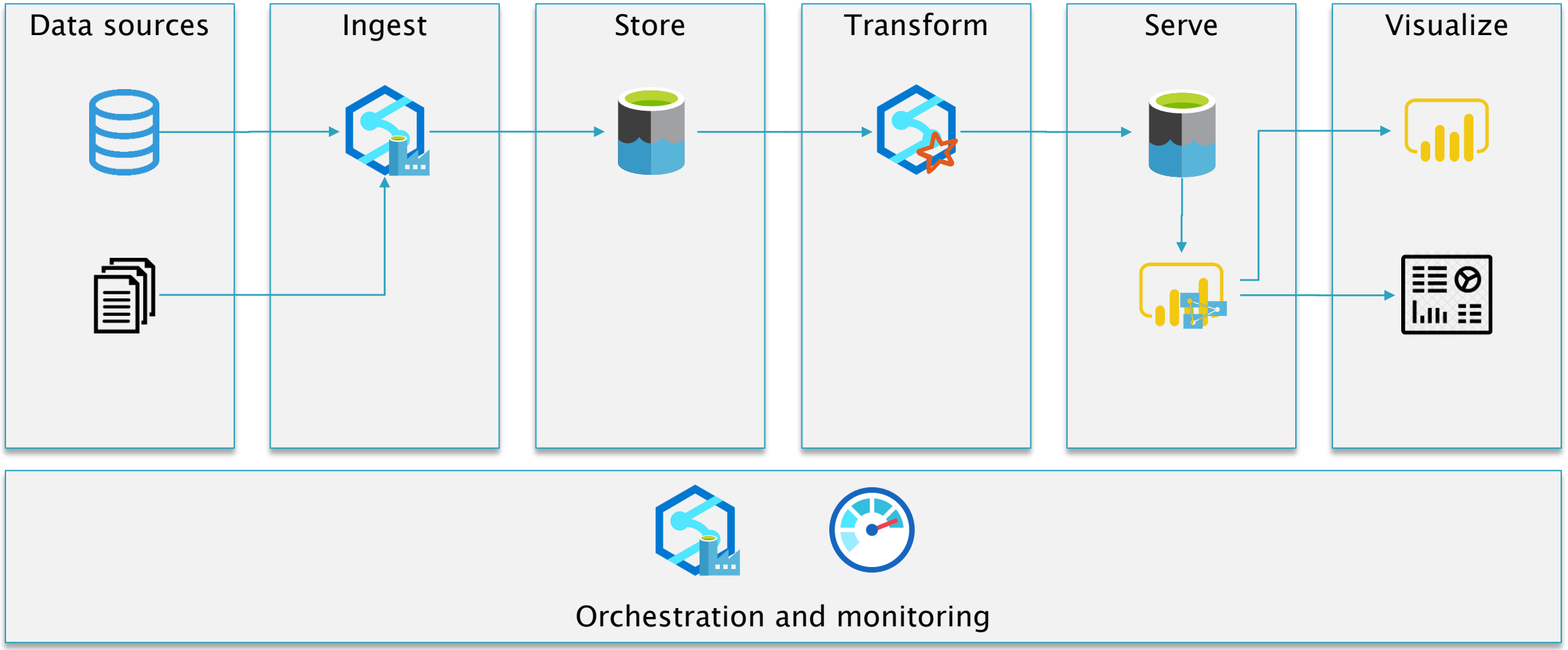


Example of date partitioning at year/month level:

- Raw Data
  - Sales
    - Salesforce
      - CustomerContacts
        - 2018
          - 201812
            - CustContact_2018_12_01.csv
            - CustContact_2018_12_02.csv
            - CustContact_2018_12_03.csv
            - ...
      - CustomerActivity
        - 2018
          - 201812
            - CustActivity_2018_12_01.csv
            - CustActivity_2018_12_02.csv
            - CustActivity_2018_12_03.csv
            - ...

ORANGEMAN

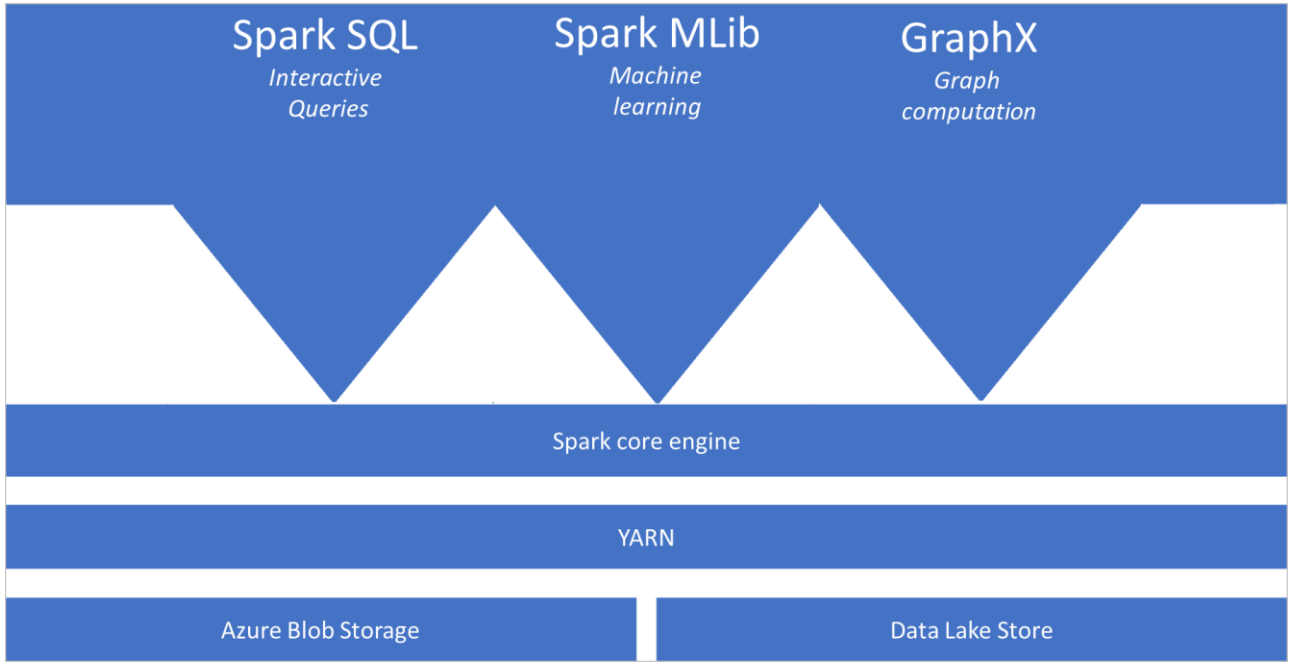
Data Lakehouse with Spark



ORANGEMAN

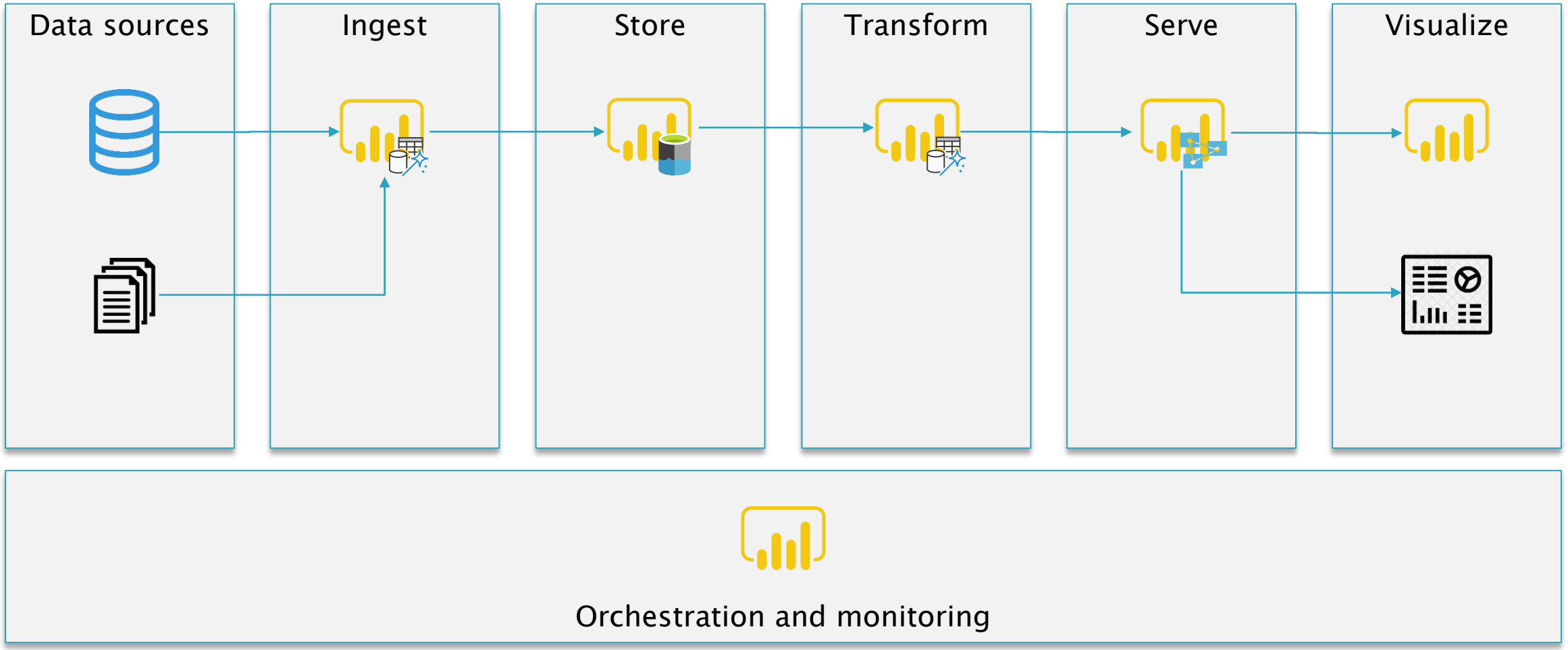
Spark in Synapse

- Industry-standard Apache Spark in-memory engine
- The most popular open-source big data engine used for data preparation, data engineering, ETL, and machine learning
- Highly extensible with support for C#, Scala and PySpark alongside Spark SQL, GraphX, Streaming and Machine Learning Library (Mllib)
- Fast start-up and aggressive autoscaling



ORANGEMAN

(3) Self-service with Dataflows in Power BI

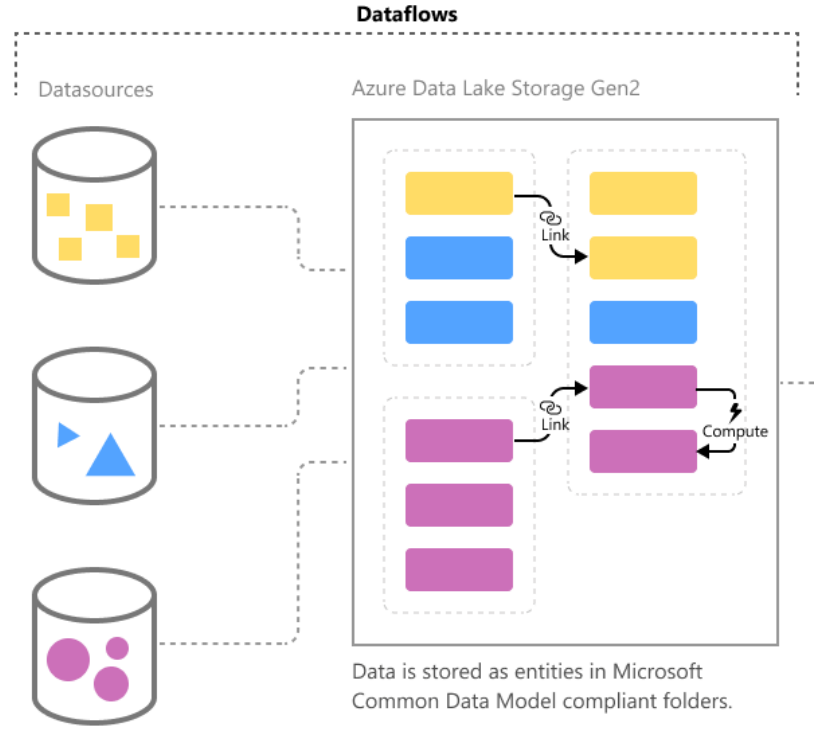


ORANGEMAN

Dataflows in Power BI

- Self-service data prep
- Power Query in the cloud
- Wide range of cloud and on-premise sources
- Link entities between dataflows
- Automatic triggers refresh processes for dependent entities in destination dataflows

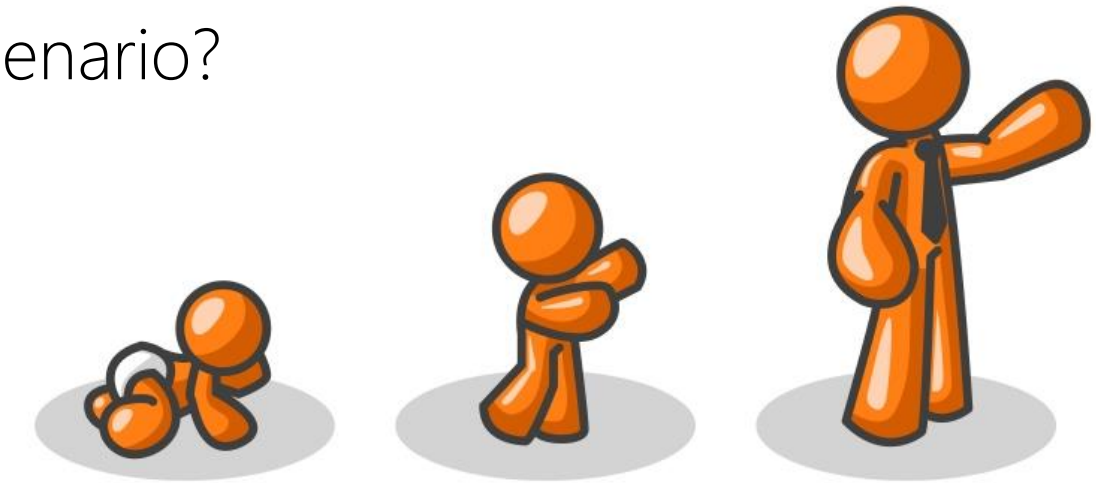
- Bring your own Data Lake storage, enabling you to use other Azure services to read/write the underlying data



ORANGEMAN

Comparison of the solutions

- The conclusion – what's best in your senario?



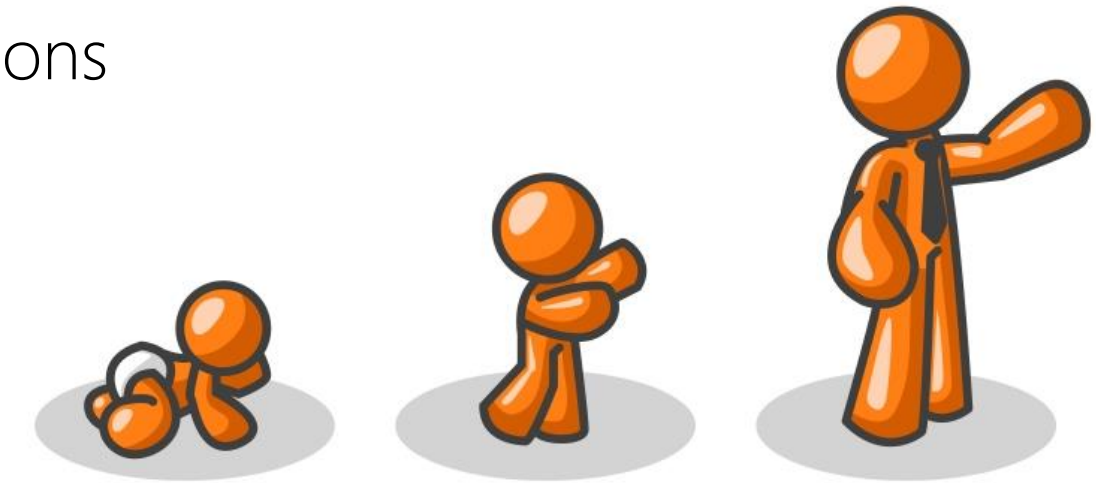
ORANGEMAN

Comparison of the solutions

	Traditional Database	Data Lakehouse	Power BI Dataflows
Transformation language	SQL	SQL or PySpark	M or GUI
Developer persona	Pro	Pro	Citizen
Portability to other services	High	High	Low
Scalability in data volume	Medium	High	Low
Pricing model	Compute	Compute	Free?
Monthly cost	Low	Low	Very low
Extensibility with AI, ML etc.	Low	High	Low
Data processing	Schema on write	Schema on read	Schema on write
DevOps possibilities	High	High	Very low

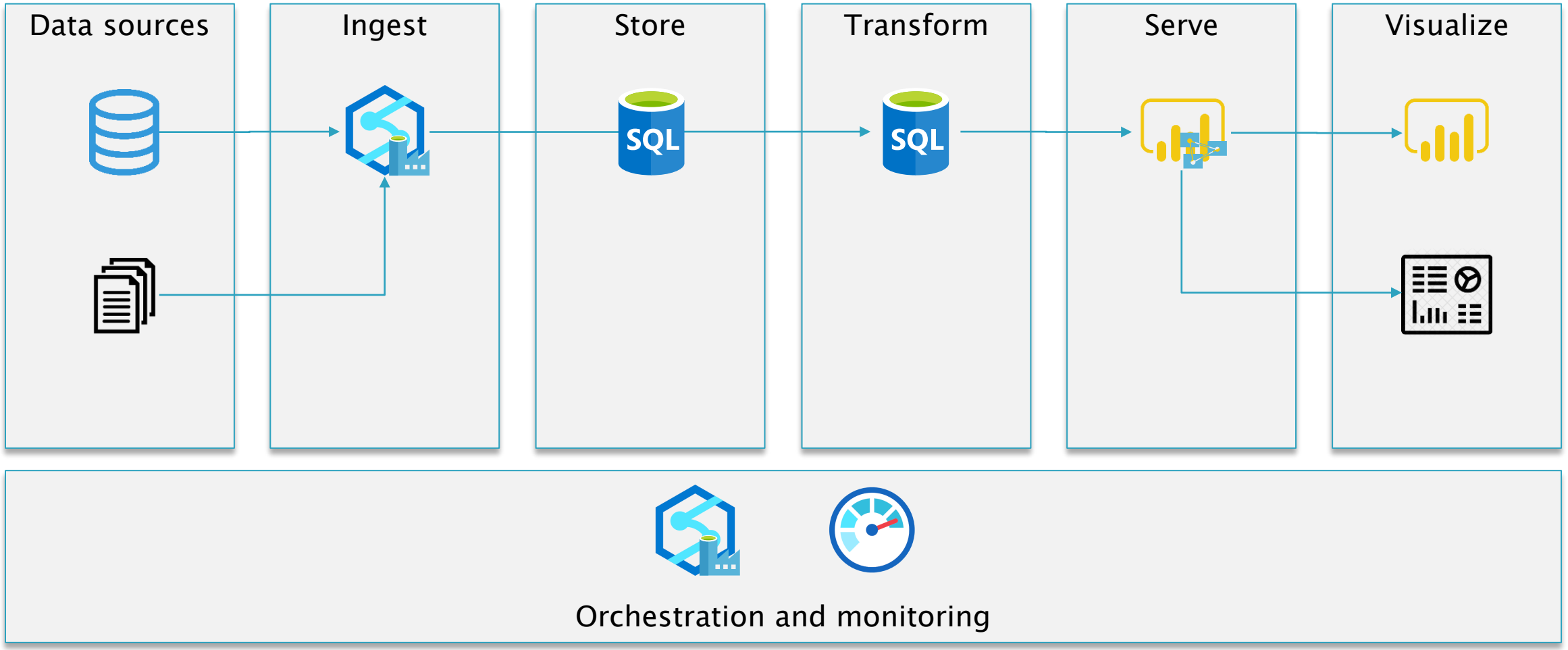
Other variations

- Mix'n'match and combining the solutions



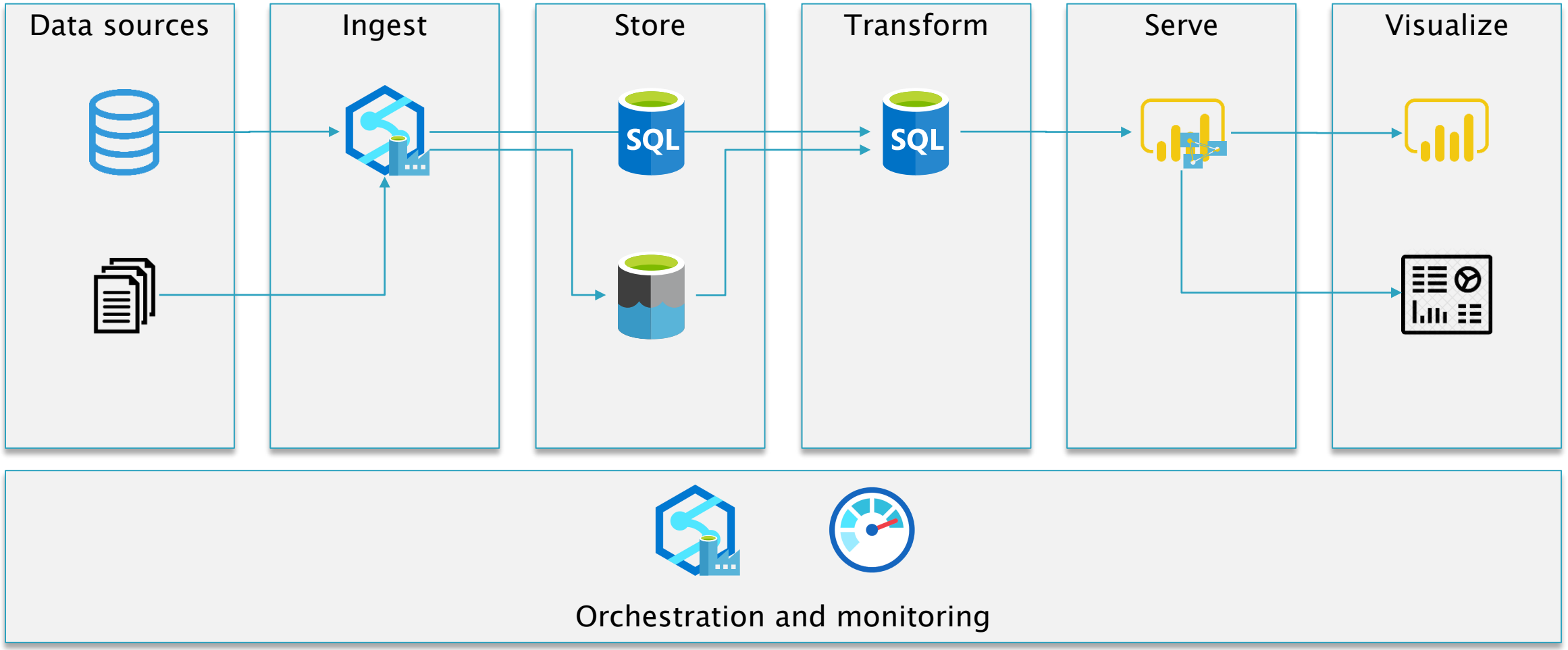
ORANGEMAN

Traditional Modern Data Warehouse



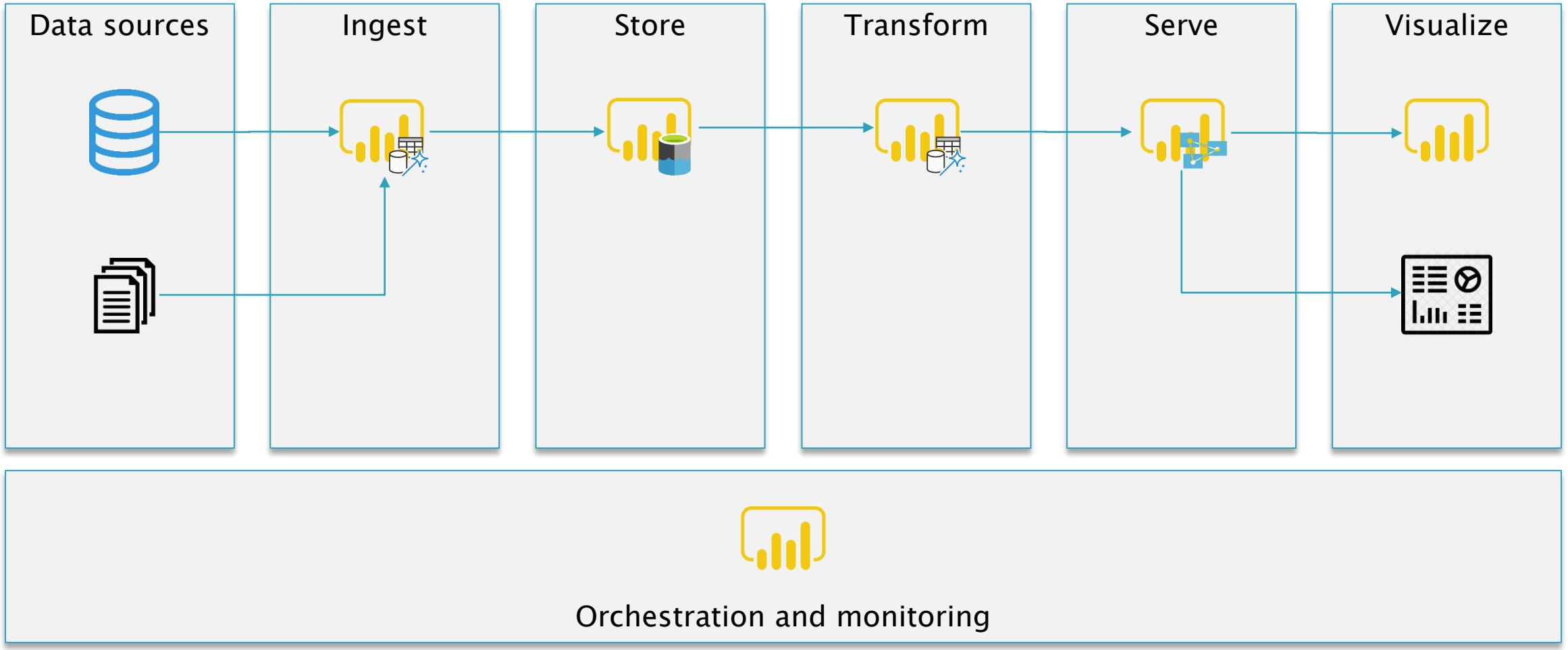
ORANGEMAN

Traditional Modern Data Warehouse with Lake



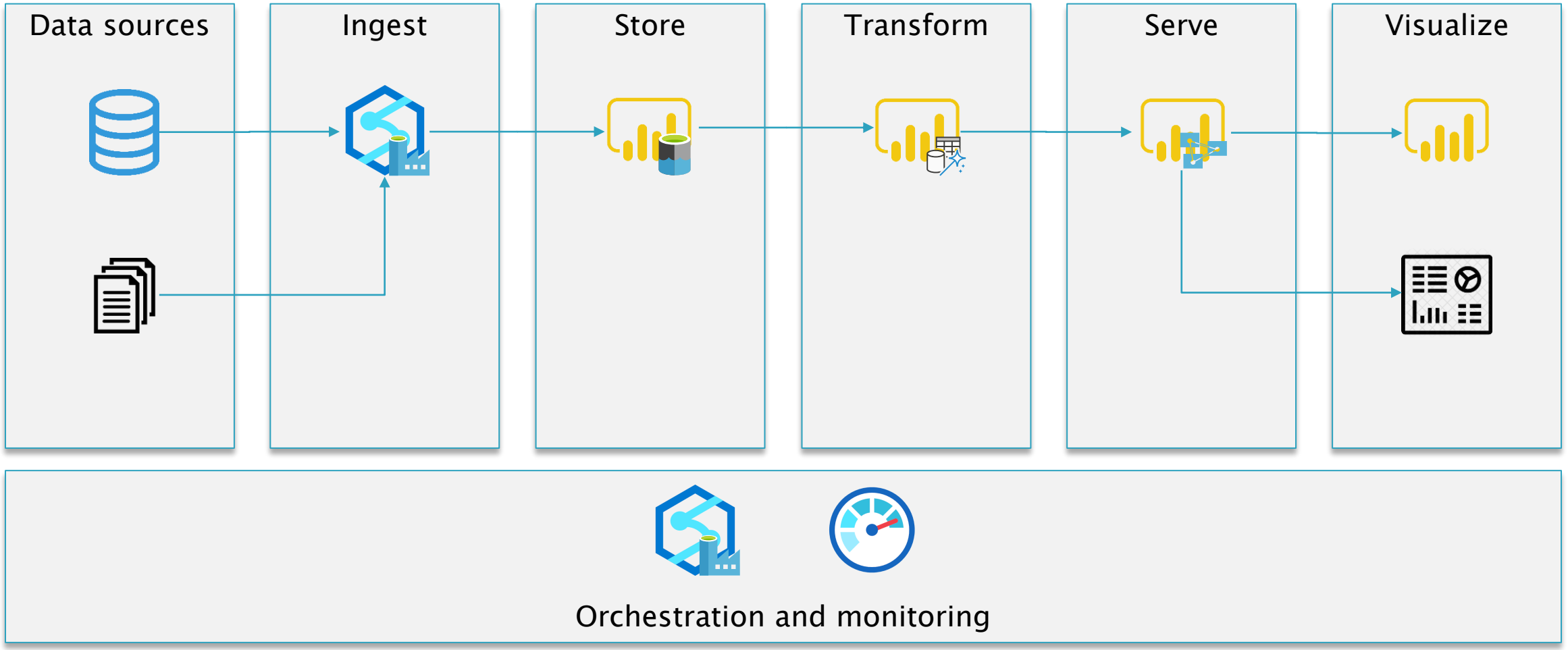
ORANGEMAN

Dataflows in Power BI



ORANGEMAN

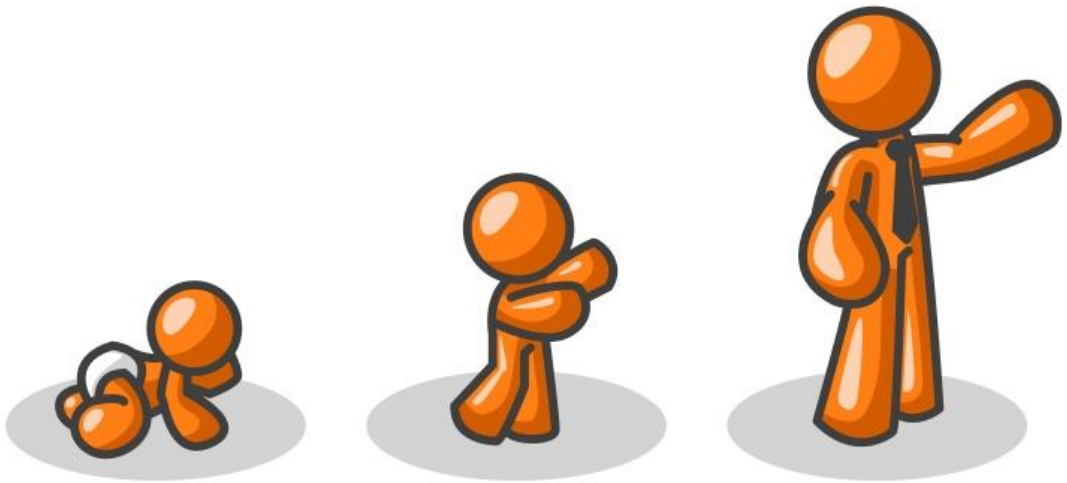
Dataflows in Power BI with Data Factory



ORANGEMAN

Pricing

- Estimate



ORANGEMAN

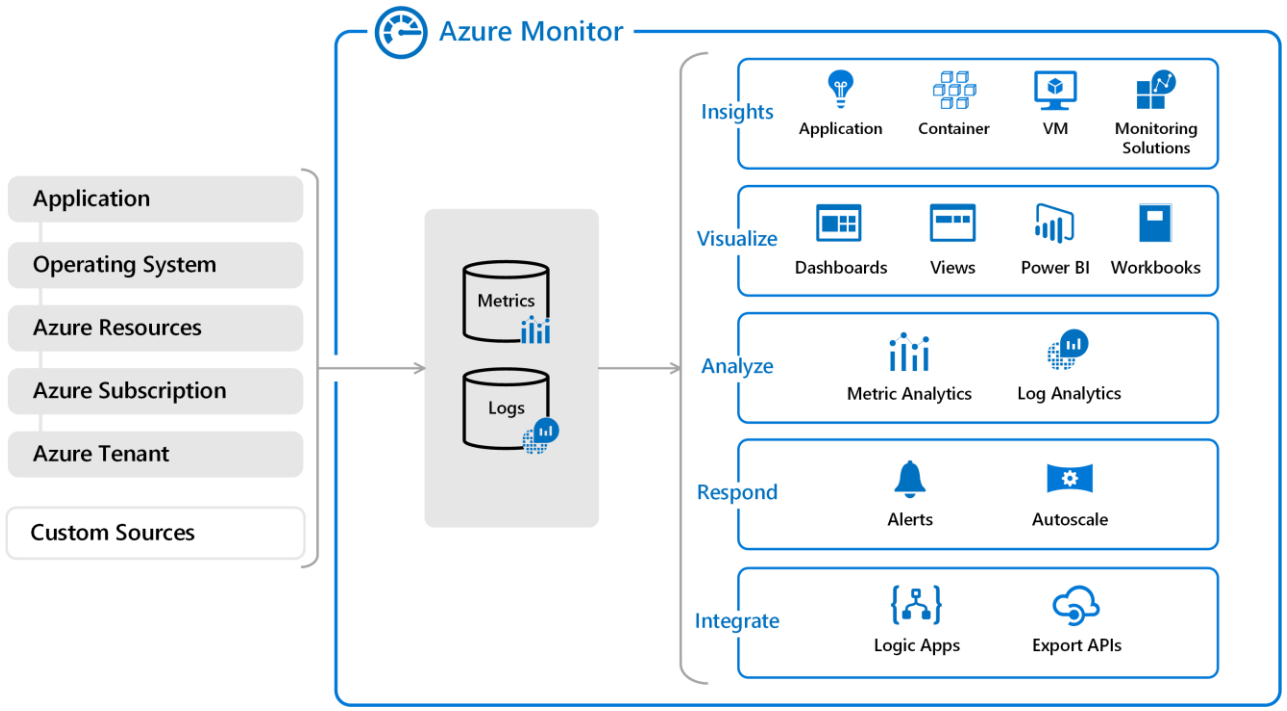
Pricing estimate: Traditional Modern DW

Service type	Description	Estimated Cost
Storage Account	Data Lake Storage Gen2, Standard, LRS Redundancy, Hot Access Tier, Hierarchical Namespace File Structure, 100 GB Capacity	$ 3.71
Azure SQL Database	Single Database, DTU Purchase Model, Standard Tier, S0: 10 DTUs, 250 GB included storage per DB, 1 Database(s) x 640 Hours	$ 12.90
Azure SQL Database	Single Database, DTU Purchase Model, Standard Tier, S6: 400 DTUs, 250 GB included storage per DB, 1 Database(s) x 90 Hours	$ 72.59
Azure Data Factory	Azure Data Factory V2 Type, Data Pipeline Service Type, Azure Integration Runtime: 1 Activity Run(s), 30 Data movement unit(s), 90 Pipeline activities	$ 9.72
Power BI Embedded	Power BI Embedded, Node type: A1, 3GB Memory, 10 Hours	$ 10.08
Azure Monitor	Log analytics:	$ 1.60
	Monthly Total	$ 110.60

ORANGEMAN

Monitoring with Azure Monitor

- Gather diagnostic from Azure Services
- Analyze with Log Analytics
- Setup alerts
- Export with Power Query



ORANGEMAN

Usefull ressources

- Enterprise Data Warehouse: https://docs.microsoft.com/en-us/azure/architecture/solution-ideas/articles/enterprise-data-
- Traditional Modern Data Warehouse on GitHub : https://github.com/justBlindbaek/TraditionalModernDW
- Pricing estimate: https://azure.com/e/9d56929b959546b384594e1074a1f506

Questions

ORANGEMAN

SQL Day

14 edycja konferencji SQLDay

9-11 maja 2022, WROCŁAW + ONLINE

Data Community

partner złoty

Future Processing

Infinite DATA

KPMG

partner srebrny

Objectivity

summ-it

UBS

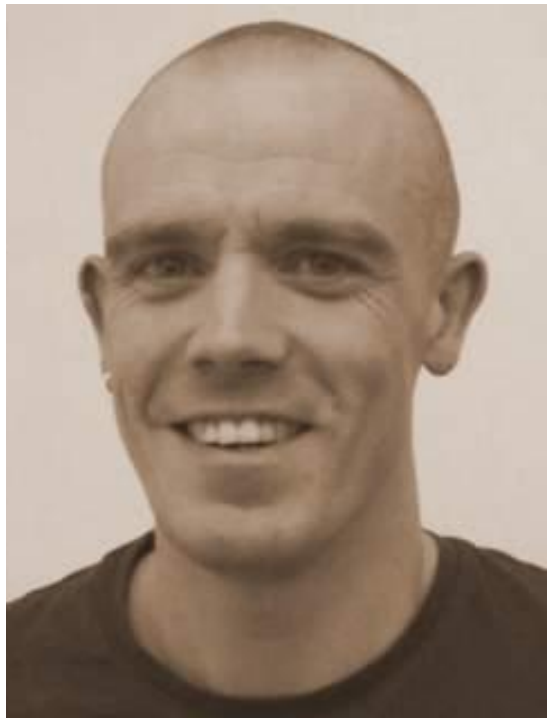
dbWatch DATABASE CONTROL

partner brązowy

elitmind think bright

Just Blindbæk

- Self-employed BI consultant in justB
- Trainer at Orange Man
- Founder
  - Danish Microsoft BI Community (MsBIP.dk)
  - Power BI UG Denmark (PowerBI.dk)
- Strong focus on
  - Azure BI architecture
  - Analysis Services
  - Reporting Services
  - Power BI
- just@justB.dk / blog.justB.dk / @justblindbaek / youtube.com/c/justblindbaek



Microsoft CERTIFIED Trainer

Microsoft CERTIFIED IT Professional

SSAS MAESTRO by Microsoft

MVP Microsoft® Most Valuable Professional

ORANGEMAN