

B0M33BDT - Azure

Petr Filas

4.12.2024

Agenda



- Fundamentals
- Data Services
- Databricks
- AI/ML Services



Azure (fundamentals)

> Four levels of management

1. Resources

- Instance of a cloud service (e.g. different types of storages, databases, compute services, virtual machines...).

2. Resource groups

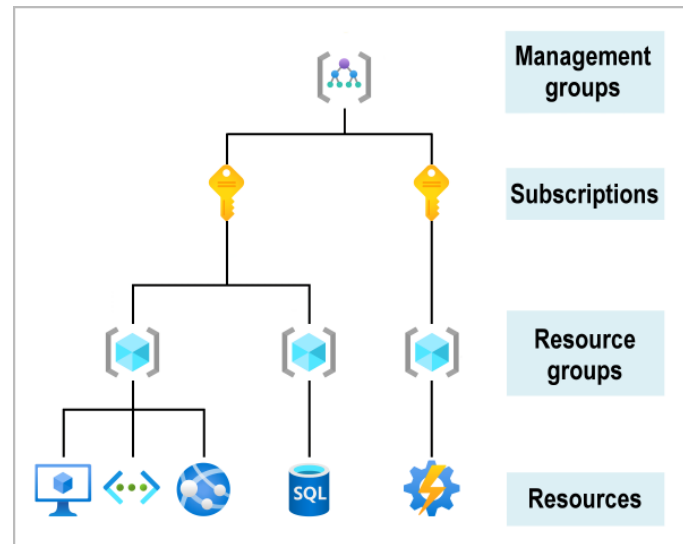
- Containers that hold related resources for an Azure solution.
 - Enables to configure multiple resources at once, delete all resources within it etc.
- Can't be nested.
- Associated with one region and subscription.
 - Can contain resources from different regions!

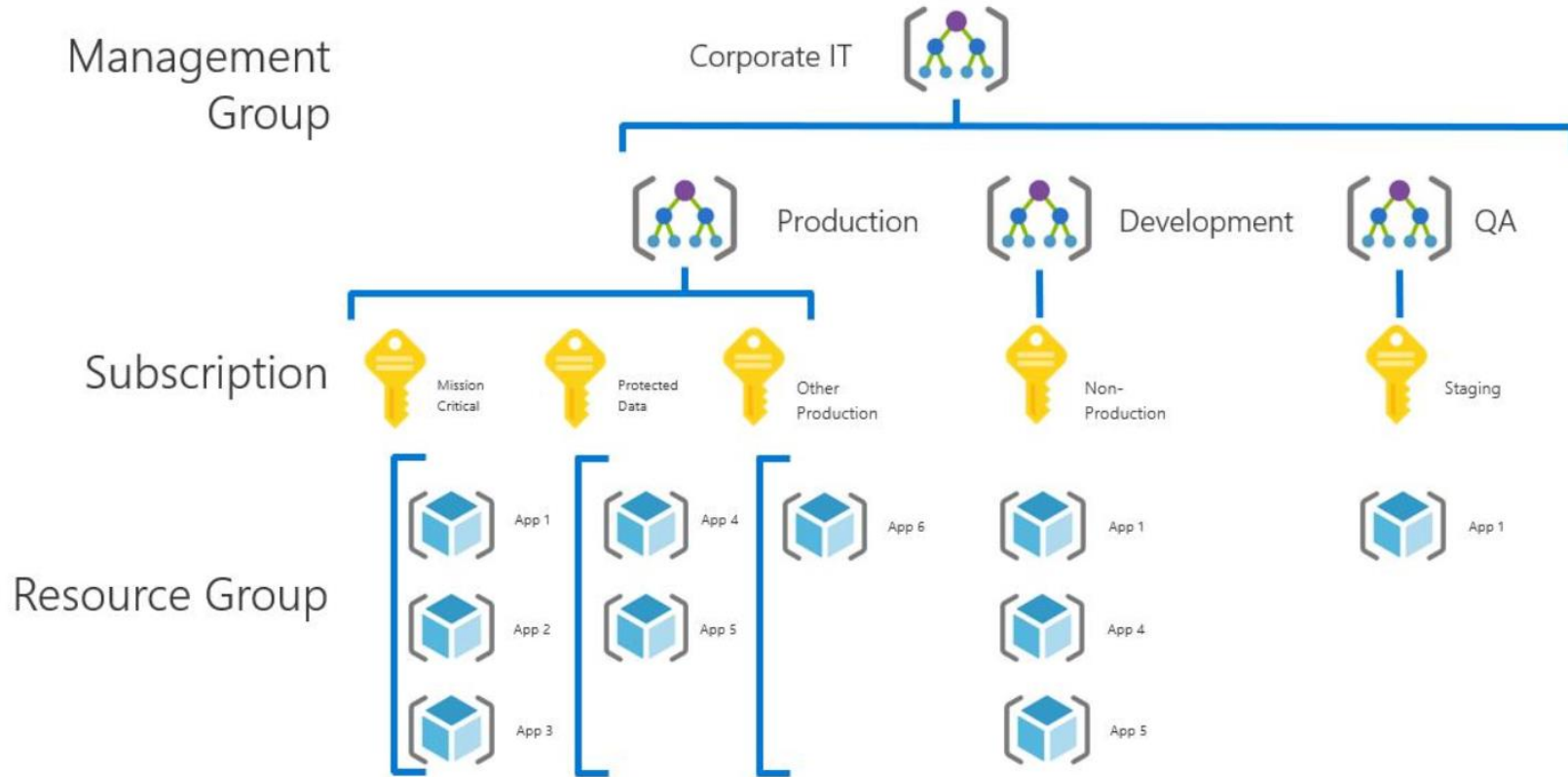
3. Subscriptions

- Groups resource groups.
- Used for billing purposes, setting limits on spending etc.
- Can't be nested.

4. Management groups

- Manage access, policies, and compliance for multiple subscriptions or other management groups (can be nested).
- Max. 7 levels of depth, one root management group.





How to create a resource

- *Create Management Group (not necessary)*
- Create **Subscription**
- Create **Resource Group**
- Create **Resource**
 - **ARM** template (Azure Resource Manager) – resource definition in JSON
 - **Bicep** - Domain Specific Language (DSL)
 - **CLI** (Command Line Interface) – **az** utility
 - Manually in **Azure Portal**

Azure stack



Security & Management



Platform Services

Media & CDN



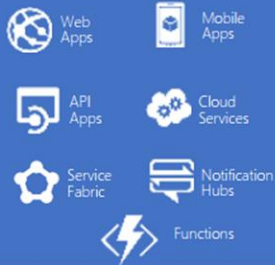
Integration



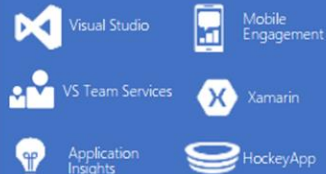
Compute Services



Application Platform



Developer Services



Data



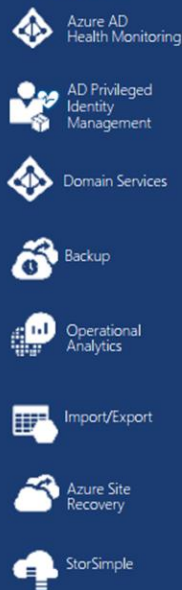
Intelligence



Analytics & IoT



Hybrid Cloud



Infrastructure Services

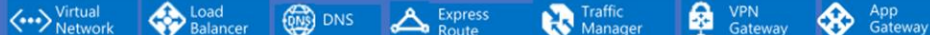
Compute



Storage

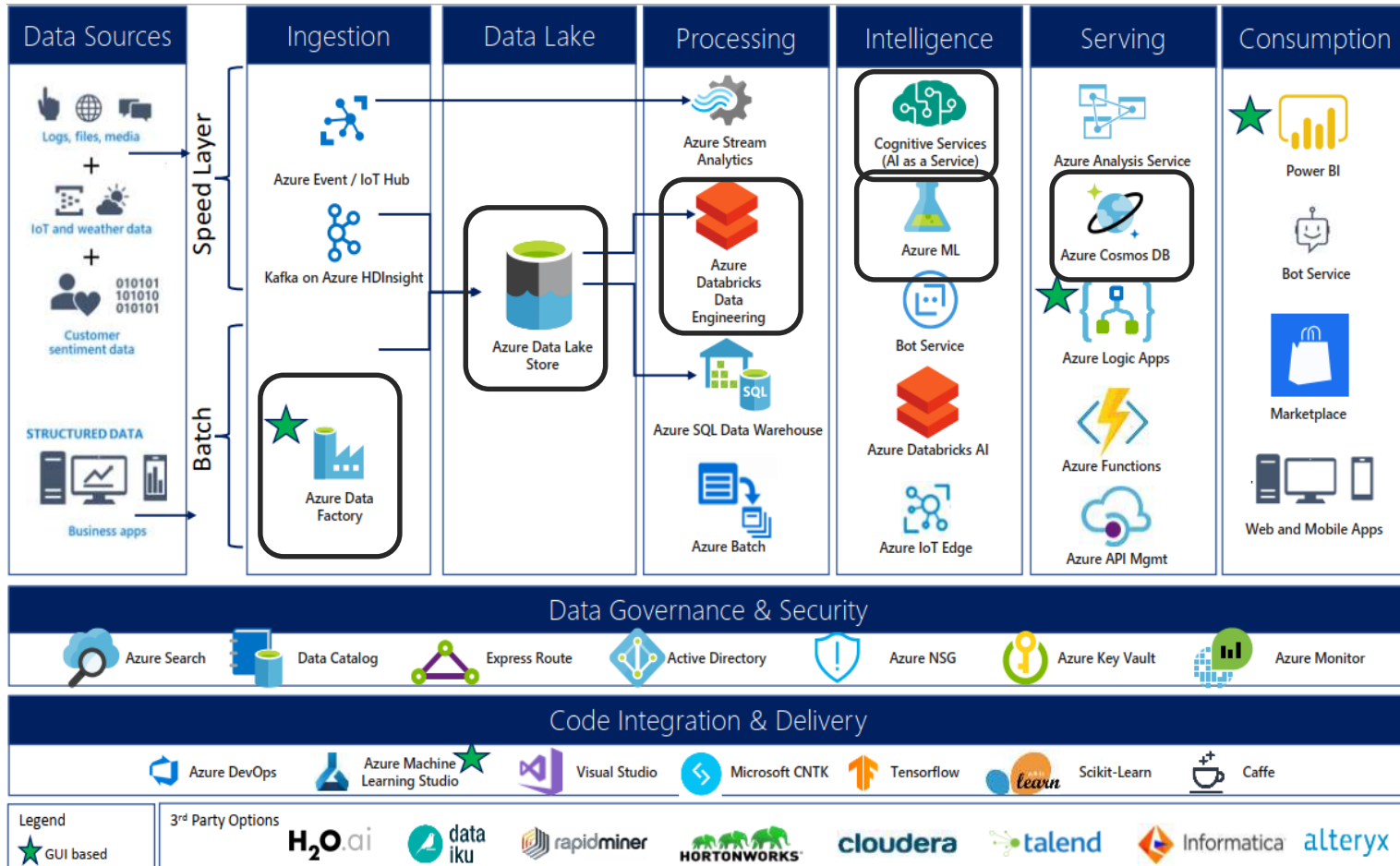


Networking



BDDS Stack

< PROFINIT >



Relevant BD resources in Azure

- Azure Data Factory
- Azure Storage Account / Storage Account Services
- Azure Databases
- Azure Fabric
- **Azure Databricks**
- Azure ML & Azure Cognitive services
- Azure HDInsight



Azure (Big) Data Services

Azure Data Factory (ADF)

- Data integration service.
- GUI for creating **data-driven workflows** for orchestrating data movement/transformation.
- „**Pay-as-you-go**“ billing model.



Code-Free ETL as a service

Ingest



- Multi-cloud and on-premise hybrid copy data
- 100+ native connectors
- Serverless and auto-scale
- Use wizard for quick copy jobs

Control Flow



- Design code-free data pipelines
- Generate pipelines via SDK
- Utilize workflow constructs: loops, branches, conditional execution, variables, parameters, ...

Data Flow



- Code-free data transformations that execute in Spark
- Scale-out with Azure Integration Runtime
- Generate data flows via SDK
- Designers for data engineers and data analysts

Schedule



- Build and maintain operational schedules for your data pipelines
- Wall clock, event-based, tumbling windows, chained

Monitor



- View active executions and pipeline history
- Detail activity and data flow executions
- Establish alerts and notifications

1. Datasets

- Named view of data that simply points or references the data.
 - “pointer to a data structure”
- Used for both input/output.
- E.g...: csv file in Blob Storage, database table, file on the internet (only as input) ...

2. Mapping Data Flows

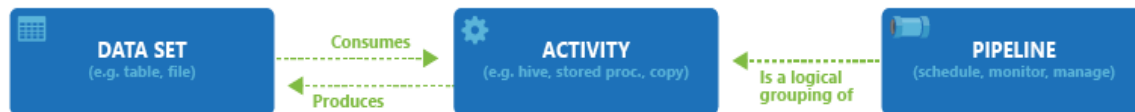
- Graphs of data transformation logic – for transforming data of any size.
- Executed on a Spark cluster that spins-up and spins-down when you need it.
 - No need to ever manage/maintain a cluster.
- Used as an activity.

3. Activities (Tasks)

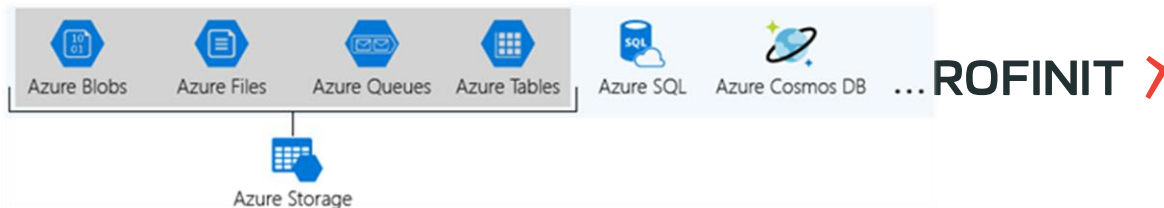
- Processing steps, actions, in a pipeline.
- Almost 100 different activities.
- E.g.: copy data activity (from one dataset to another), Hive activity that runs query on HDInsight cluster, Databricks notebook launch activity...

4. Pipelines

- Logical grouping of activities that performs a unit of work.
- Can be chained together to operate sequentially, or they can operate independently in parallel.
- Can be parametrized, uses variables, control flows (branching, for-each iterators...).



Storage Account

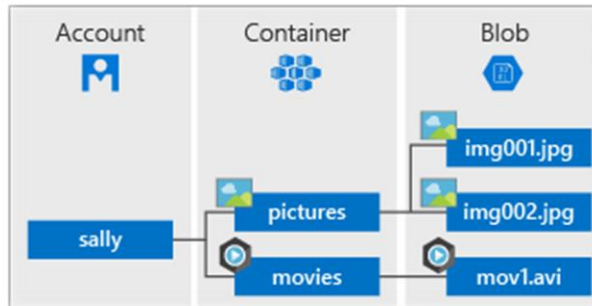


- Azure Storage offers several types of storage services.
- Data can be accessed from anywhere using HTTP/HTTPS.
 - `https://<storage-account>.blob.core.windows.net`
 - Globally unique name!
- Data in SA are **highly available**, **secured** a **massively scalable**.
- Configurable redundancy, disk types.
 - Locally redundant, zone-redundant, geo-redundant...

Storage Account services

1. Blob storage

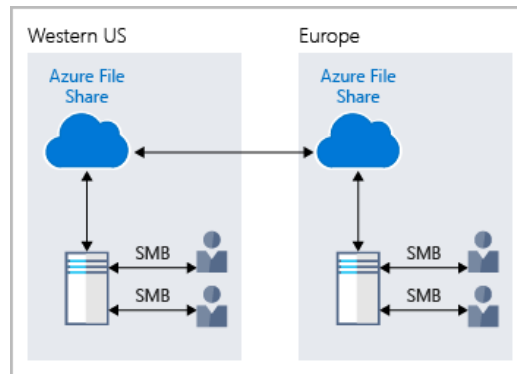
- Good for large unstructured data.
- Cheapest storage option on Azure.
- **Container** – structure for organizing blobs.

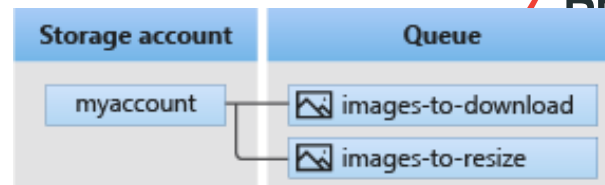


FINIT >

2. Azure files

- Fully managed Azure file share service.
- Can be easily mounted on Windows, Linux, MacOS.
- Server Message Block (SMB) Protocol
 - network file sharing protocol



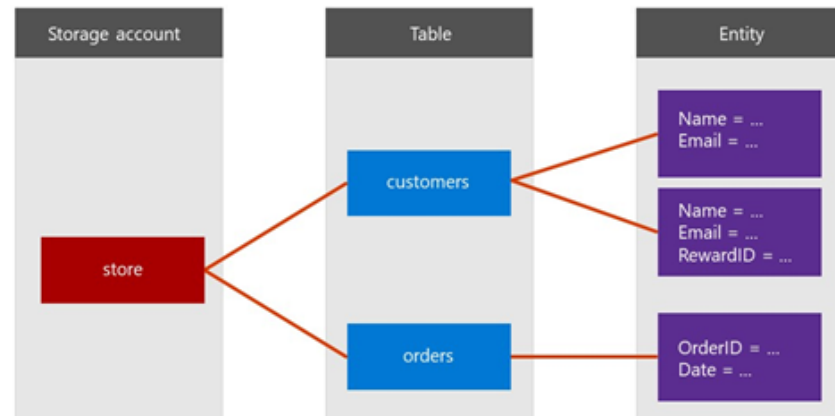


3. Azure Queue

- Storing large numbers of messages (to 64 KB).
- Communication between apps (on-premise + cloud).

4. Azure Tables

- Structured NoSQL data.
- Schema-less design.



5. Azure Data Lake Storage (ADLS)

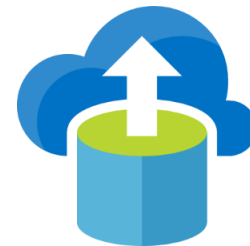
- Comprehensive, scalable, and cost-effective data lake solution for high-performance big data analytics built into Azure
- Gen1 (Apache HDFS), **Gen2**
- **ADLS Gen2 = Azure Blob Storage + ADLS Gen1**



Azure Data Lake Storage Gen2

Azure Databases

> Many SQL/NoSQL databases for various use cases.



> Azure Database Migration Service

- Tool to simplify, guide, and automate database migration to Azure.
- Convenient migration with near-zero downtime.
- MySQL, PostgreSQL, MongoDB...



Homogeneous
sources



Heterogeneous
sources



Orchestration



Scale migration



Near-zero
downtime

- Azure Database for MariaDB, Azure Database for PostgreSQL, different SQL options
 - Depends on how much control over the OS we need.
 - Oracle – Oracle Cloud Infrastructure
- IaaS
- PaaS
 - Auto scaling.
 - Pay only for what you really use.

SQL Server on Azure
Virtual Machines



Best for re-hosting and apps requiring
OS-level access and control
Automated manageability features and
OS-level access

Infrastructure as a Service

Azure SQL
Managed Instance



Best for modernizing existing apps
Offers high compatibility with SQL Server
and native VNET support

Platform as a Service

Azure SQL
Database



Best for building new apps in the cloud
Pre-provisioned or serverless compute
and Hyperscale storage to meet
demanding workload requirements

Azure Cosmos DB

- Globally distributed, multi-model database service for any scale.
- NoSQL, big data.
- Guaranteed single-digit millisecond response times at any scale and 99.999-percent availability (99.999 SLA – highest on Azure).
- Automatic and instant scalability.
- SQL API, MongoDB API, Gremlin API (graphDB), Cassandra API, Table API.
- Good for data consumers (cache)



Azure Synapse Analytics

< PROFINIT >

- Platform that brings together data integration, enterprise data warehousing, and big data analytics.
- SQL pool & Spark pool & Synapse pipelines (Data Factory)
- Challenger for Databricks



Azure Fabric

< PROFINIT >

- All-in-one analytics solution that covers everything from data movement to data science, Real-Time Analytics, and business intelligence
- OneLake + Azure Data Services
- A real Challenger for Databricks?



Azure components availability

- Differs region to region
- See: <https://azure.microsoft.com/en-us/global-infrastructure/services/>

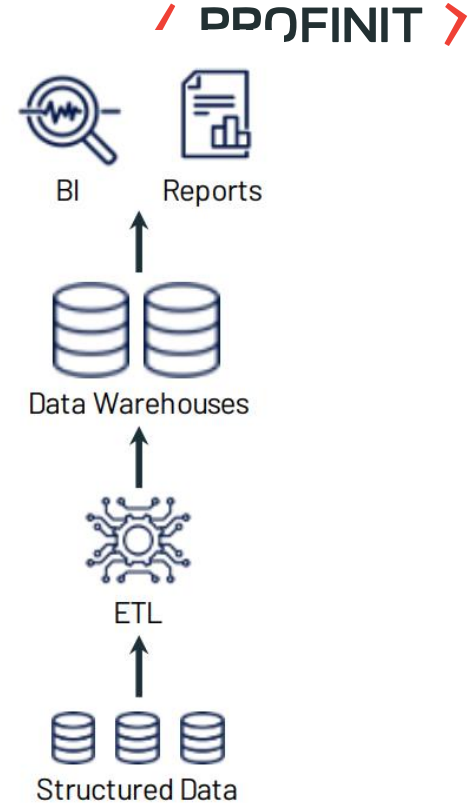


Databricks

Data Warehouse

- Minimal support for video, audio, text
- Long history in decision support and business intelligence applications
- limited support for streaming

→ therefore, most data stored in **Data Lakes**,
only subset of it in DWHs



- - Does not support transactions.
- - Does not enforce data quality.
- Where DWH is good tool, Data Lake not so much:
 - poor BI support,
 - complex to set up and configure,
 - poor performance (when used i.e., for BI).
- Why to use it then? + Can handle all data
 - for ML, data science ...

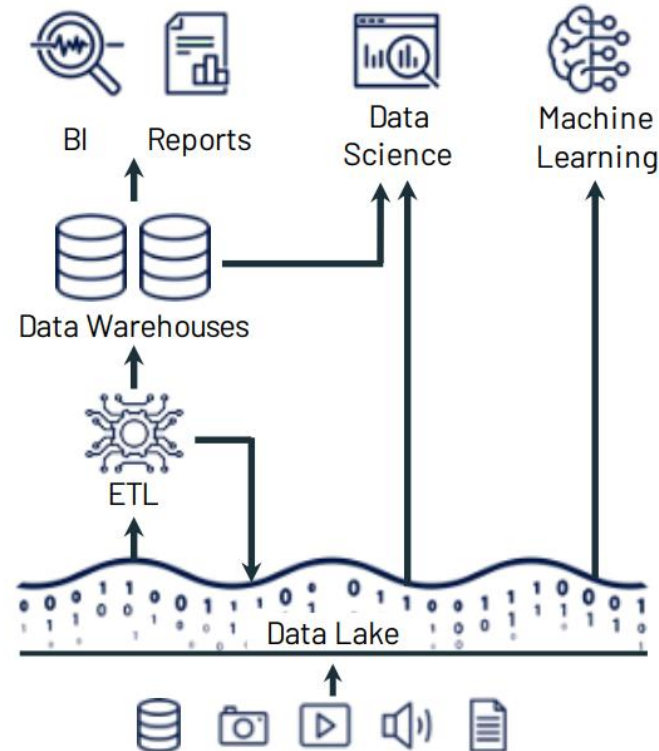
DWH + Data Lake

> Companies have **Data Lakes** and **Data Warehouses** side by side

- Reusing current setup,
- data duplication,
- higher cost, administrative burden..
- NOT ideal setup.

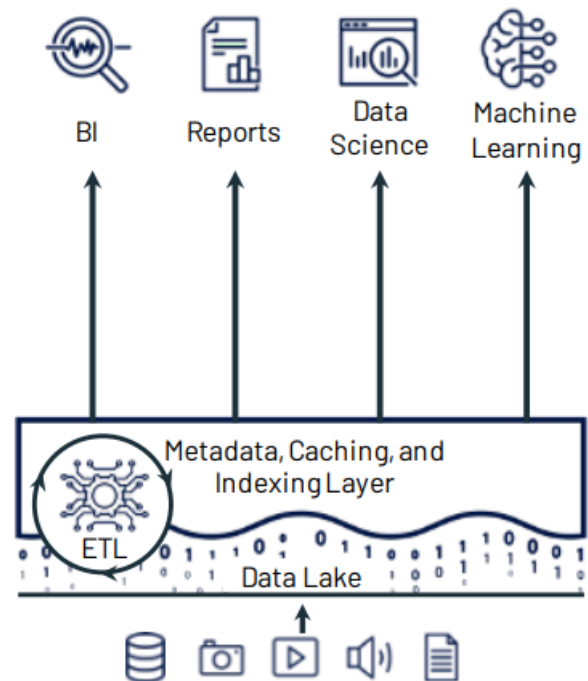
– So, how it can be improved?

- **Lakehouse**



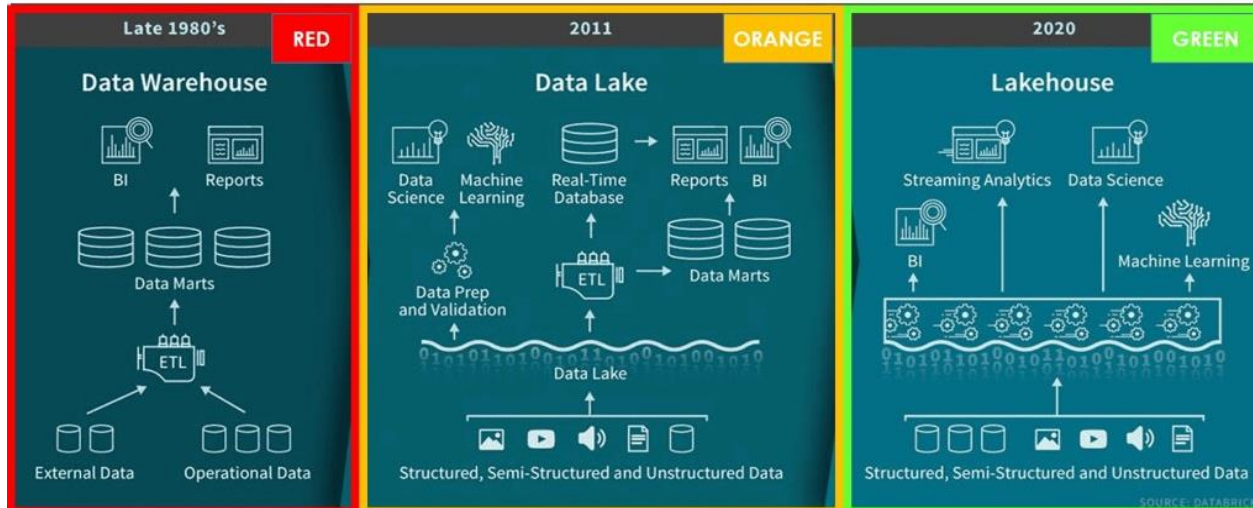
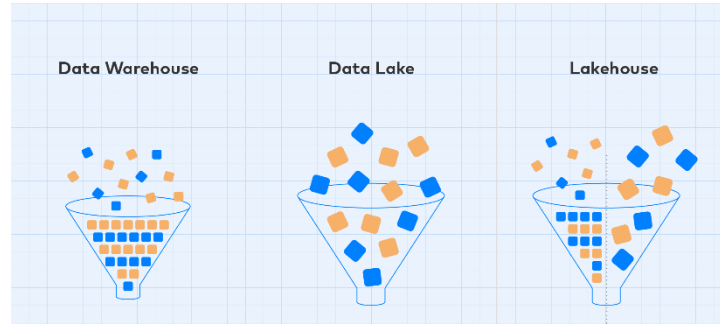
Data Lakehouse

- > Paradigm that brings the best of the two
- > + Has **structured transactional layer**
- > + Reduces data movement and complexity
- > + Reduces cost
- > + Support for diverse workloads
 - data science, machine learning, SQL, analytics ...
- > + Openness
 - open and standardized formats



As time passes...

< PROFINIT >



Difference recap

	Data Warehouse	Data Lake	Data Lakehouse
Storage Data Type	Works well with structured data	Works well with semi-structured and unstructured data	Can handle structured, semi-structured, and unstructured data
Purpose	Optimal for data analytics and business intelligence (BI) use-cases	Suitable for machine learning (ML) and artificial intelligence (AI) workloads	Suitable for both data analytics and machine learning workloads
Cost	Storage is costly and time-consuming	Storage is cost-effective, fast, and flexible	Storage is cost-effective, fast, and flexible
ACID Compliance	Records data in an ACID-compliant manner to ensure the highest levels of integrity	Non-ACID compliance: updates and deletes are complex operations	ACID-compliant to ensure consistency as multiple parties concurrently read or write data

As Azure stack...

Modern Data Warehouse

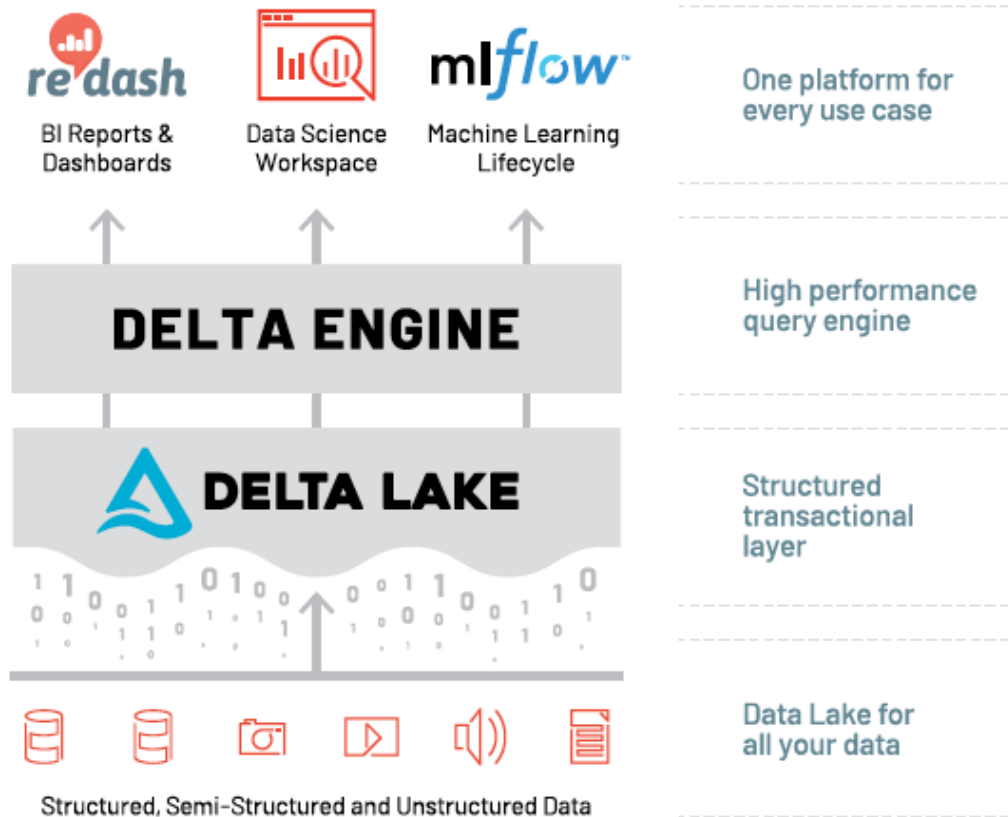


Data Lakehouse



(Data) Lakehouse by Databricks

< PROFINIT >



- Open-source storage format that brings **ACID** transactions to Spark and big data workloads
- **Spark under the hood**, Indexing techniques, schema validation & expectations
- Foundation of a **cost-effective**, highly scalable lakehouse
- Single home for structured, semi-structured and unstructured data
- Unifies batch and streaming data processing
- Open format based on parquet

<https://en.wikipedia.org/wiki/ACID>



- **Collaborative** cloud unified data platform
 - for the entire data team (data scientists, analysts, engineers...)
- Builds on top of open-source software (Spark, MLFlow, Delta Lake...)
- Multilanguage notebooks (similar to Jupyter or Zeppelin)
 - Python, R, Scala, SQL
- Databricks SQL
 - Easily explore Delta Lake table schemas and optimize it
- Dashboards



What are all the Delta things in Databricks?

- <https://docs.databricks.com/en/introduction/delta-comparison.html>

> Databricks Demo Hub

- Product demos various use cases.
- Simple notebooks, easily importable to your workspace using an URL.

> Certifications - Databricks

- Be Databricks Certified professional!

Azure Databricks vs Azure Synapse Analytics



- > Proprietary data processing engine (**Databricks Runtime**) built on a highly optimized version of Apache Spark offering up to 10x performance
- > Databricks notebooks
 - Changes in real time
- > Data lake must be mounted in order to use it
- > Open-source Apache Spark (vanilla version)
- > Nteract notebooks
 - One user has to save before changes are displayed
- > Data lake can be added as a *Linked Service*
 - Direct access

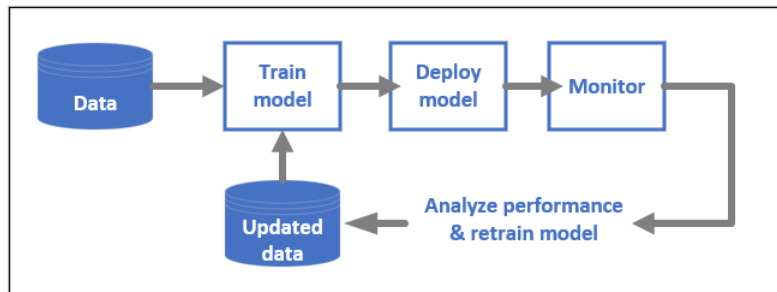


Azure AI/ML Services

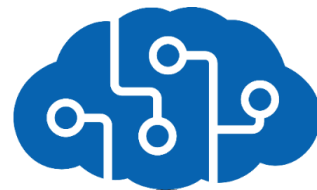
Azure ML



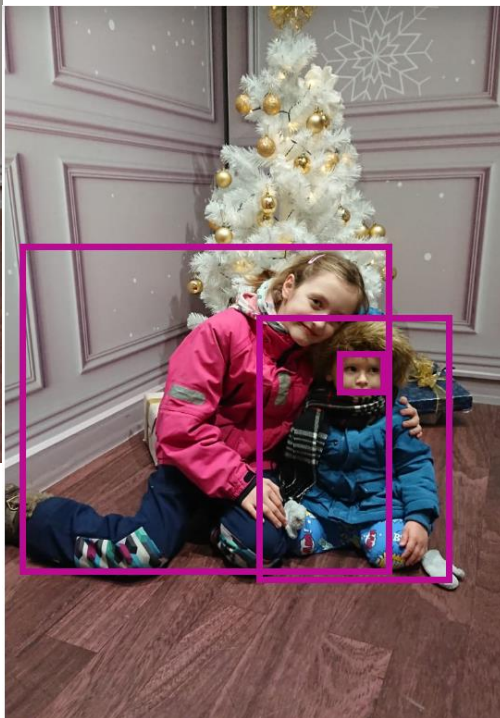
- Service for accelerating and managing ML project lifecycle (end-to-end).
- **Collaboration for machine learning teams**
 - Shared notebooks, compute resources, data, and environments.
 - Tracking and auditability that shows who made changes and when.
 - Versioning.
- Supports PythonSDK, R and other ML frameworks.
- Multinode distributed training, compute clusters with latest GPU options.
- MLOps



- Customizable, pretrained models built with breakthrough AI research (ML expertise not needed).
 - **Speech** (text ↔ audio, translations, speaker verification...)
 - **Computer vision** (detect and identify people, identify emotions...)
 - **Search** (autosuggest, visual search, web search...)
 - **and others**
- Supports a wide range of cultural languages at the service level.
- Examples:
 - <https://azure.microsoft.com/cs-cz/services/cognitive-services/computer-vision/#features>
 - <https://azure.microsoft.com/cs-cz/services/cognitive-services/face/#demo>



Cognitive Services example



```
"smile", "confidence": 0.826899648 }, {  
  "name": "baby", "confidence":  
    0.817665756 }, { "name": "christmas",  
    "confidence": 0.7937247 }, { "name":  
    "child", "confidence": 0.6052515 } ]
```

```
{ "tags": [ "person", "indoor", "child",  
  "baby", "sitting", "table", "little",  
  "stuffed", "holding", "toy", "small",  
  "teddy", "girl", "floor", "young", "boy",  
  "bear", "wooden", "playing", "bed",  
  "laying", "blue", "white" ], "captions": [  
  { "text": "a baby holding a stuffed  
    animal", "confidence": 0.6358383 } ] }
```

- REST API access to OpenAI's powerful language models
 - GPT-4, GPT-3, Embeddings, DALL-E, and Whisper
- Azure OpenAI Services vs OpenAI
 - Azure OpenAI Services = Azure + OpenAI
 - Benefits of integration to other Azure Services
- Key concepts
 - Prompts & completions
 - Tokens (\$\$\$)
 - Resources and Deployments
 - Prompt engineering
 - Models





Obsolete Azure Services

- “Azure Hadoop” 😊
- Managed, full-spectrum, open-source analytics service.
- Open-source frameworks such as:
 - Hadoop, Apache Spark, Apache Hive, LLAP, Apache Kafka, Apache Storm, R, and more, **in Azure environment**.
- Seamless integration with the most popular big data solutions with a one-click deployment.
- Low-cost and scalable.
 - Decoupled compute and storage.
 - Clusters on demand.





Summary

Summary

- Storage is cheap – Compute is expensive
- Unlimited scalability, support of any architecture
- Common architectures for Data Engineering and Data Science
- Some services compete with each other
 - Price vs performance
 - Open-source vs proprietary

Questions?

Profinit EU, s.r.o.
Tychonova 2, 160 00 Praha 6

Tel.: + 420 224 316 016, web: www.profinit.eu

 LINKEDIN
linkedin.com/company/profinit

 TWITTER
[@profinit_EU](https://twitter.com/profinit_EU)

 FACEBOOK
facebook.com/Profinit.EU

 YOUTUBE
Profinit EU, s.r.o.