DS684
Cloud Computing
Week 07

Regarding Labs and Assignments

- Class participation means more than Zoom attendance. You must actively participate in the discussion and labs, and answer questions.

- Must hit Submit button, otherwise no grade
- If you need extension in time, must send written request (***email***). Otherwise no grade and no makeup. Requests sent over Zoom chat do not count.
- For any technical difficulty (installation, Azure access, etc), you must send written explanation (***email***) before the deadline. Otherwise no grade and no makeup.

Teaching Schedule

Week 7: Azure Synapse Analytics Part I: Data Warehouse

Week 8: Azure Synapse Analytics Part II: Data Engineering

Week 9: Visualization using Power BI

Week 10: Azure Machine Learning

Week 11: Final project presentation

Agenda

- Data Warehouse
  - Business Intelligence
  - Major Data Warehouses
- Data Lake
- Azure Synapse Analytics
  - Demo and Lab

Data Warehouse

A data warehouse is a centralized repository of integrated data from one or more disparate sources. Data warehouses store current and historical data and are used for reporting and analysis of the data.

- Stores data in a format that is optimized for reading and analyzing
- Stores historical data from multiple systems
- Stores the lowest level of data, with aggregated views provided in the warehouse for reporting
- Allows the transactional system to focus on handling writes, while the data warehouse satisfies the majority of read requests.

Data Warehouse

Just because a database has all the data of this company, does not mean it is a data warehouse.

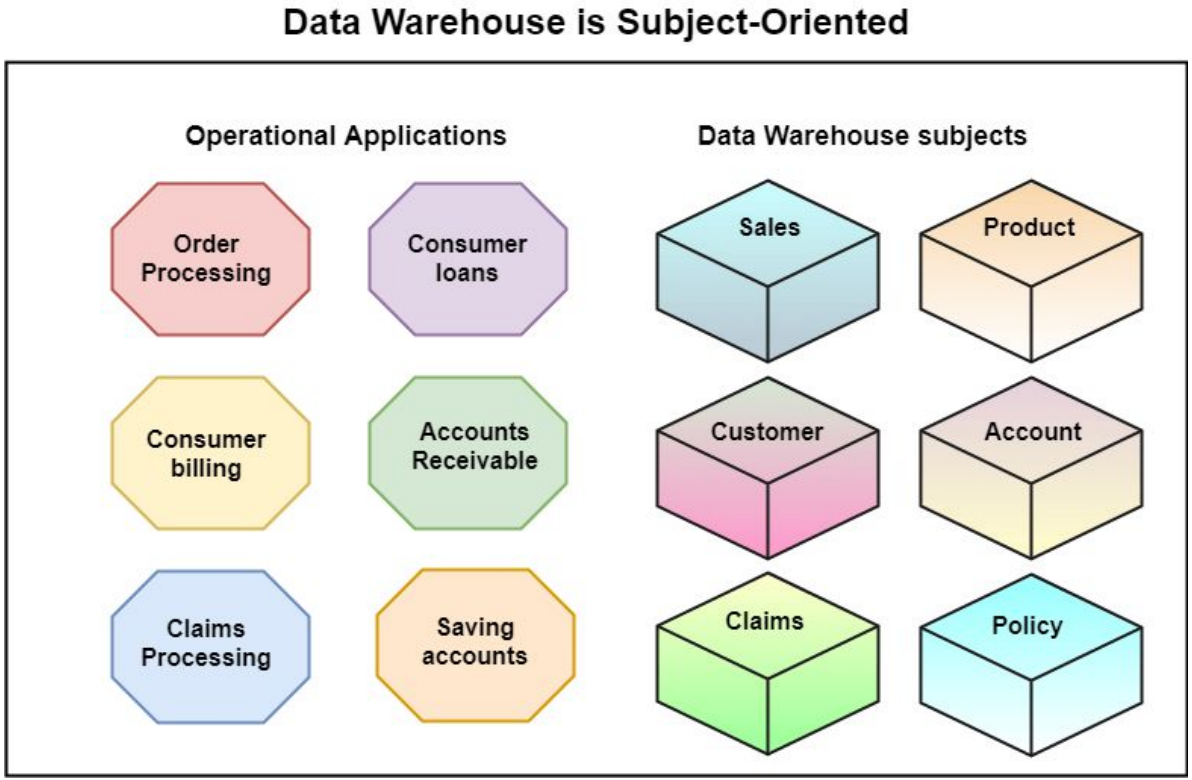
Data in a data warehouse must be properly organized based on analytical needs.

Operational Data Store (ODS)

Central database that provides a snapshot of the latest data from multiple transactional systems.

- Most current business applications run on top of relational databases, usually some common ones like Oracle, SQL Server, PostgreSQL, etc.
- Modern technology allows the users to maintain a copy of operational data in a centralized place
- If the application does not require high performance or integrity or security, it can even directly run on top of ODS
- Used for redundancy, warm backup, operational reporting, etc.

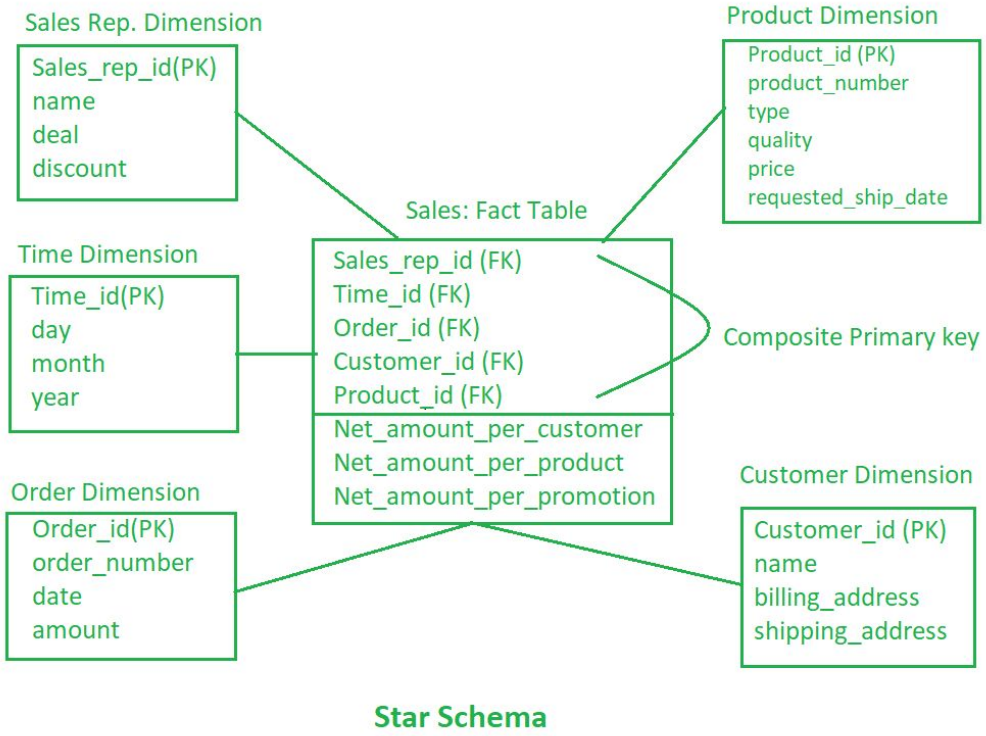
DWH vs ODS


Data Warehouse is Subject-Oriented
Operational Applications
Order Processing
Consumer loans
Consumer billing
Accounts Receivable
Claims Processing
Saving accounts
Data Warehouse subjects
Sales
Product
Customer
Account
Claims
Policy

Datamart

Data stored according to subject area analytical needs

Usually in star schema



Data Warehouse vs Datamart

There are many different opinions on the definitions and approaches of data warehouse and datamart. Datamart will always be star schema because it is created for analytical purpose. But there are many options for data warehouses.

- 3NF compliant, relational data warehouse
- Star/Snowflake schema data warehouse (datamart can be conceptual views)
- Logical data warehouse (a layer of views on top of ODS)

Pick whatever is suitable for your situation

Agenda

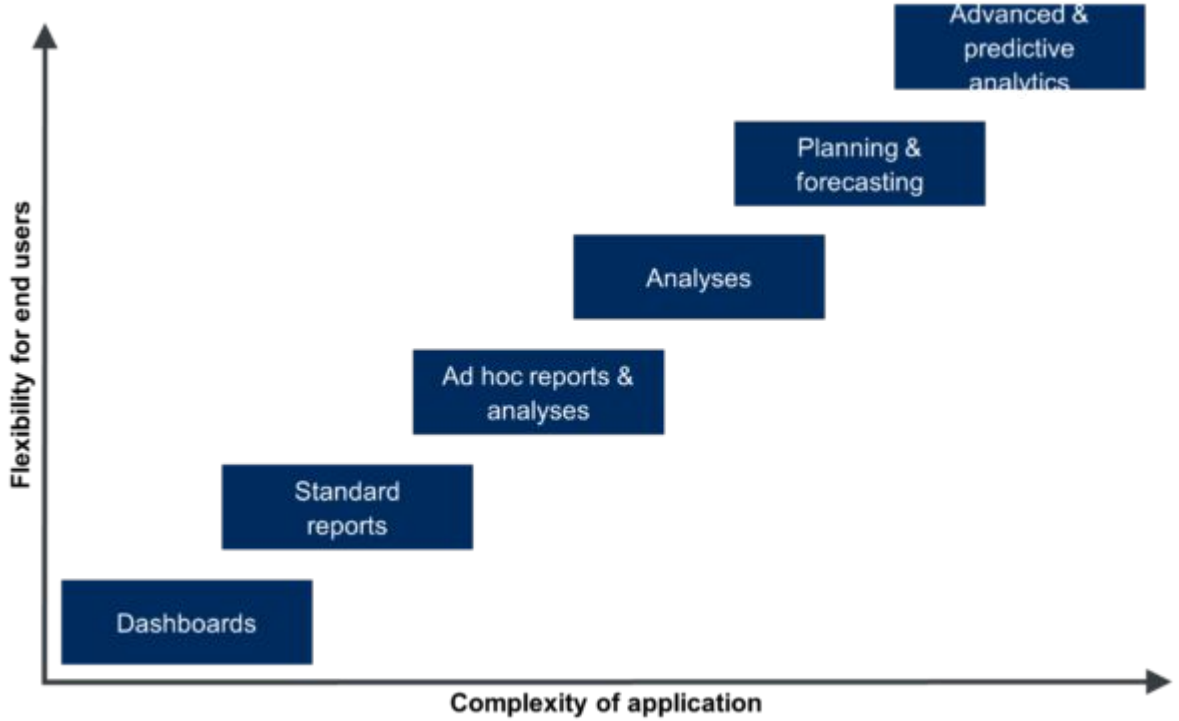
- Data Warehouse
  - Business Intelligence
  - Major Data Warehouses
- Data Lake
- Azure Synapse Analytics
  - Demo and Lab

Business Intelligence

Strategies and technologies used by enterprises for the data analysis and management of business information.

- Many many many different terms in this field, including data analytics, data analysis, data science, etc.
- We will use them interchangeably in this course

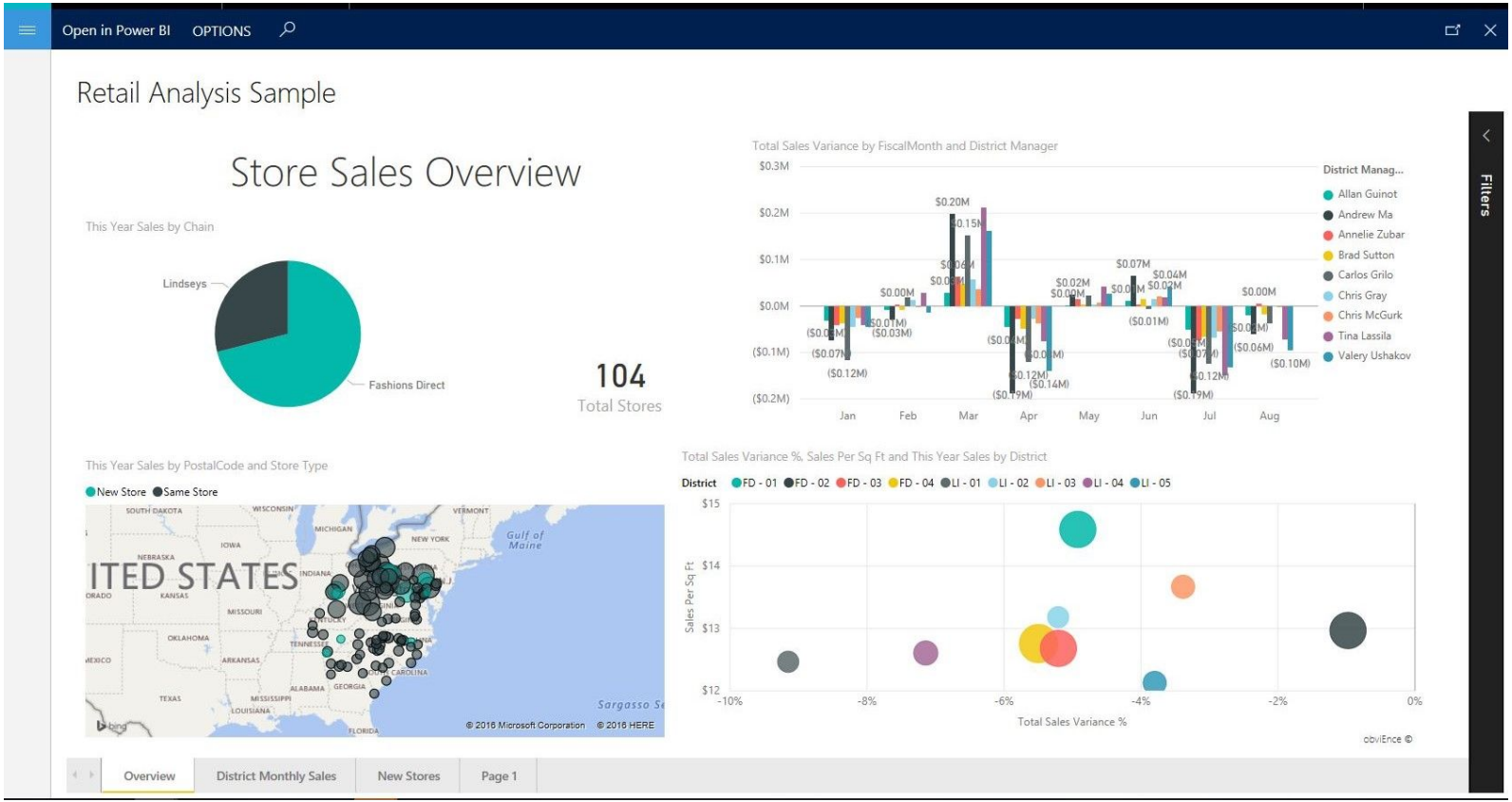
Levels of Analysis



Levels of Analysis

- Dashboard (summarized)
  - Data visualization
- Standard (operational) report
  - Static reports
- Ad hoc reporting and Online Analytical Processing (OLAP)
  - Based on cube or based on databases
- Data analysis
- Planning and Forecasting
  - Data science
- Machine learning/Artificial Intelligence

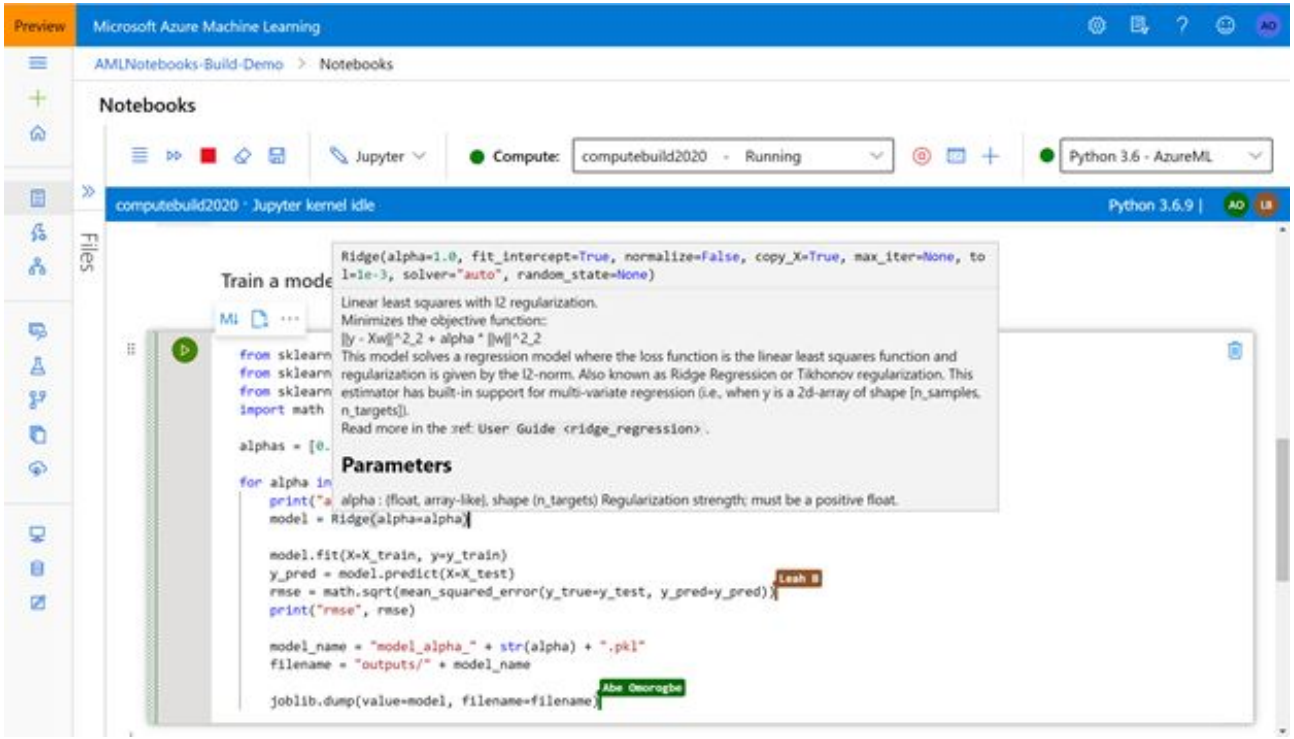
Dashboard


Open in Power BI
OPTIONS
Retail Analysis Sample
Store Sales Overview
This Year Sales by Chain
Lindseys
Fashions Direct
104
Total Stores
Total Sales Variance by FiscalMonth and District Manager
District Manag...
Allan Guinot
Andrew Ma
Annelie Zubar
Brad Sutton
Carlos Grilo
Chris Gray
Chris McGurk
Tina Lassila
Valery Ushakov
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
This Year Sales by PostalCode and Store Type
New Store
Same Store
Total Sales Variance %, Sales Per Sq Ft and This Year Sales by District
District
FD - 01
FD - 02
FD - 03
FD - 04
LI - 01
LI - 02
LI - 03
LI - 04
LI - 05
Sales Per Sq Ft
Total Sales Variance %
Filters
Overview
District Monthly Sales
New Stores
Page 1

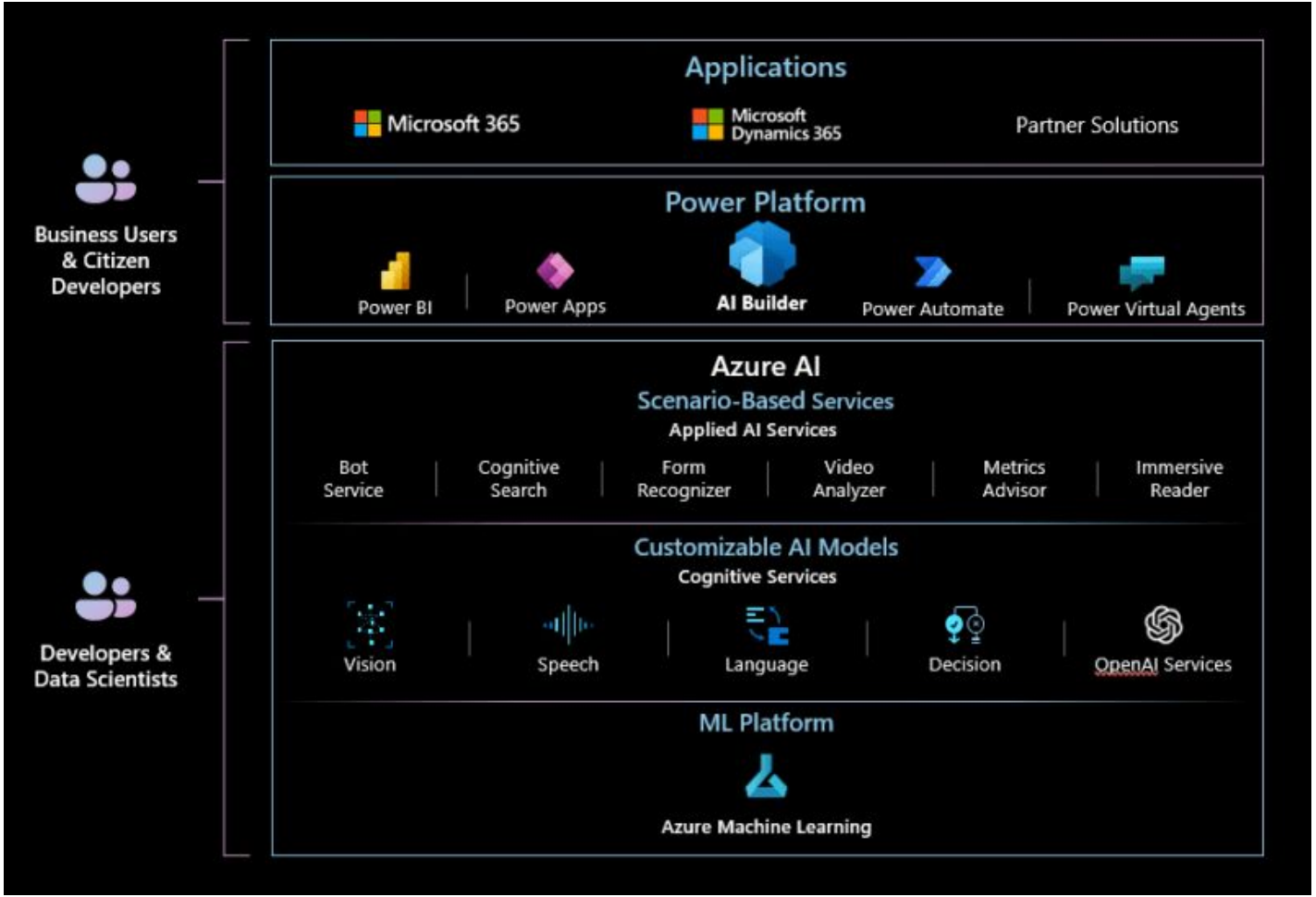
Standard Report

Category	Status	Avg Price	Last Year	This Year	Goal
100-Groceries		$1.36	$810,176	$829,776	$810,176
090-Home		$3.28	$2,913,647	$3,053,326	$2,913,647
080-Accessories		$4.22	$1,273,096	$1,379,259	$1,273,096
070-Hosiery		$3.57	$573,604	$486,106	$573,604
060-Intimate		$4.02	$955,370	$852,329	$955,370
050-Shoes		$13.73	$3,640,471	$3,574,900	$3,640,471
040-Juniors		$7.06	$3,105,550	$2,930,385	$3,105,550
030-Kids		$5.20	$2,726,892	$2,705,490	$2,726,892
020-Mens		$6.89	$4,453,133	$4,452,421	$4,453,133
010-Womens		$6.70	$2,680,662	$1,787,958	$2,680,662
Total		**$5.19**	**$23,132,601**	**$22,051,952**	**$23,132,601**

Azure Data Science Example



Azure Machine Learning Services



Agenda

- Data Warehouse
  - Business Intelligence
  - Major Data Warehouses
- Data Lake
- Azure Synapse Analytics
  - Demo and Lab

Introduction to Major Data Warehouses

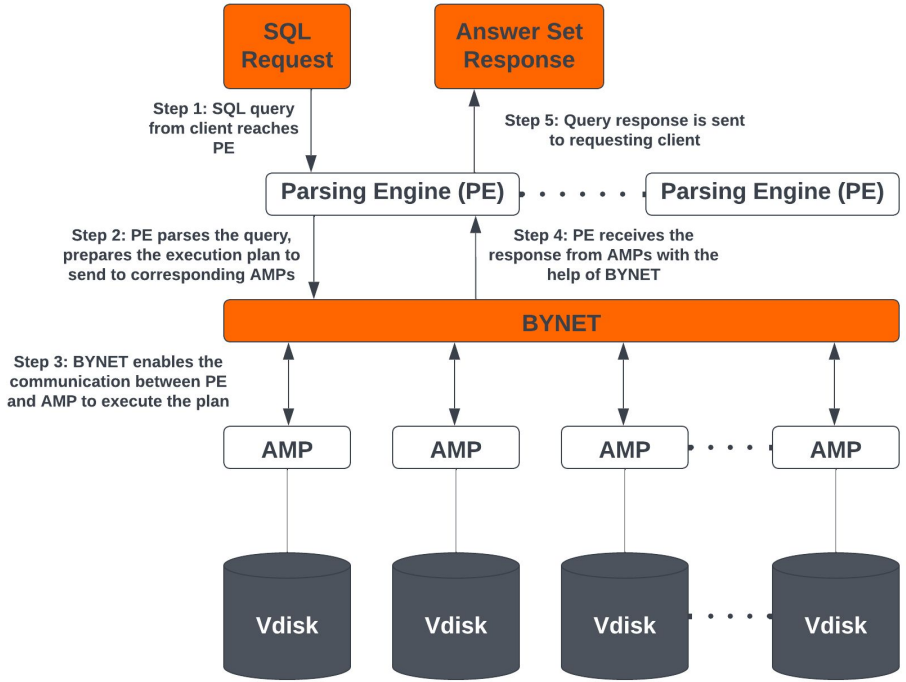
Traditional Data Warehouses

- General purpose RDBMS with DWH support
  - Oracle
  - IBM DB2
- Purposely built Data Warehouse
  - Teradata
  - IBM Netezza

Traditional Data Warehouse Layout

One server with pre-determined CPU, memory, and hard disk

Example: Teradata Vantage



Traditional Data Warehouse Layout

Compute and Storage are tied together.

- Not scalable/elastic
- Don’t grow together
- Not cost efficient.

Solution: Separation of Compute and Storage (sounds familiar?)

Cloud Data Warehouses

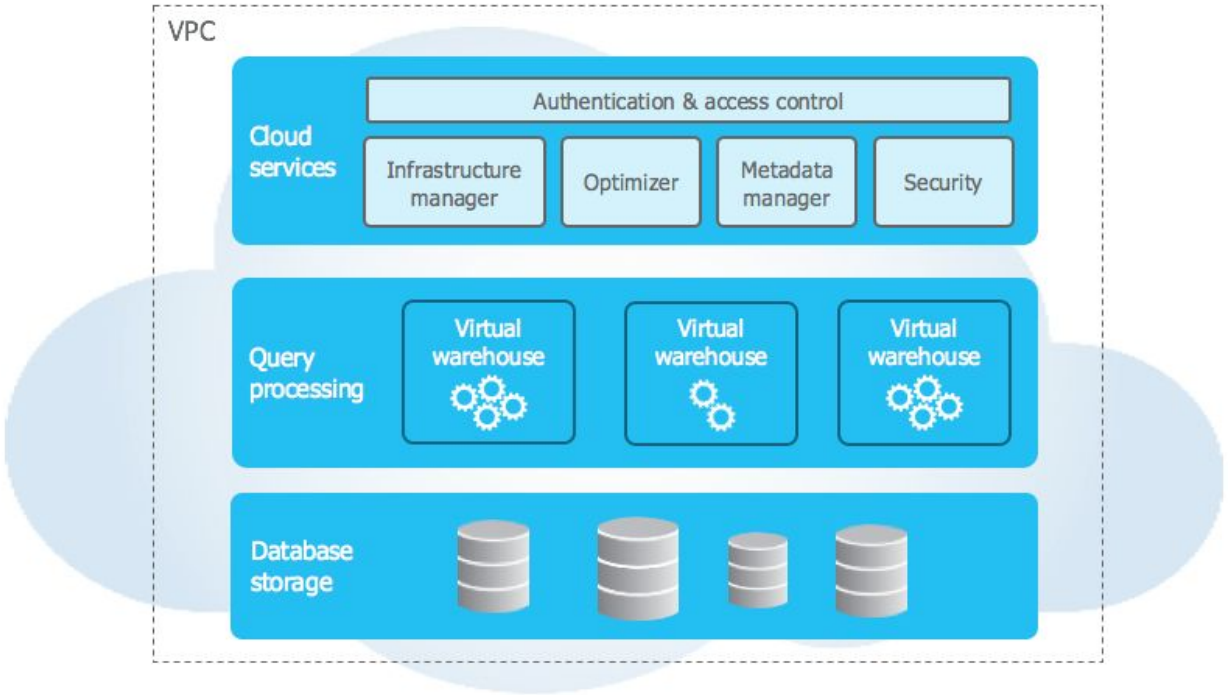
Moving towards separation of compute and storage

- Snowflake
- Google BigQuery
- Amazon Redshift

Snowflake

Most popular cloud data warehouse.

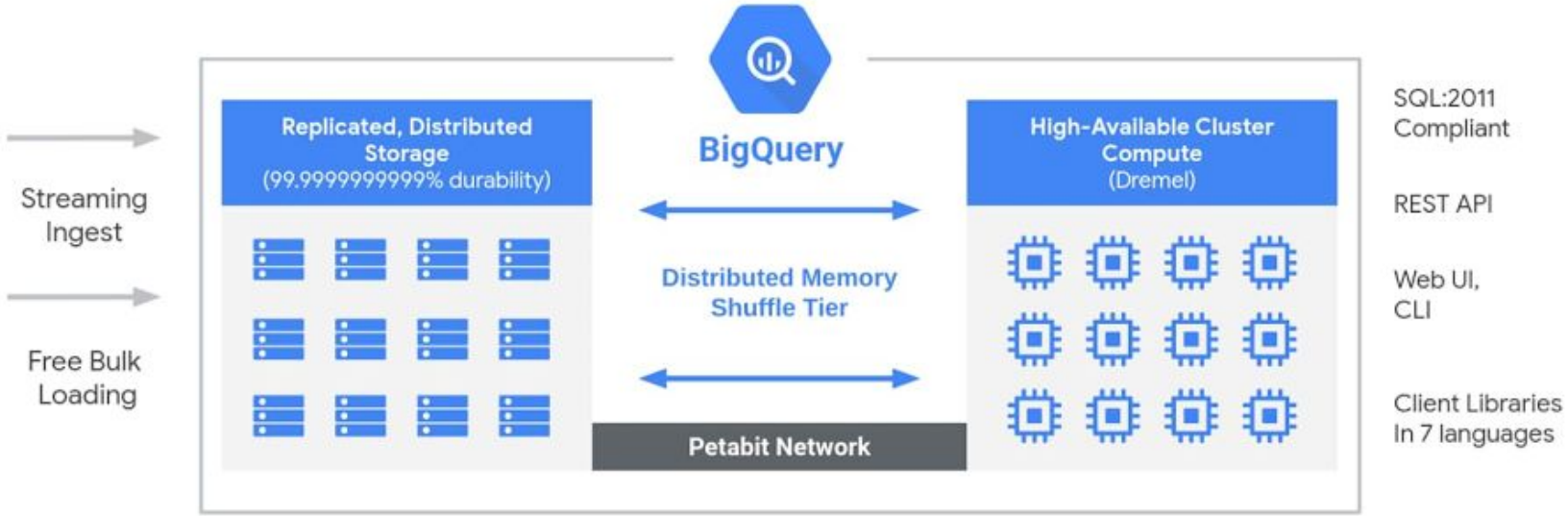
Expensive but well architected.



[Understanding Parallelism in Snowflake]

Google BigQuery

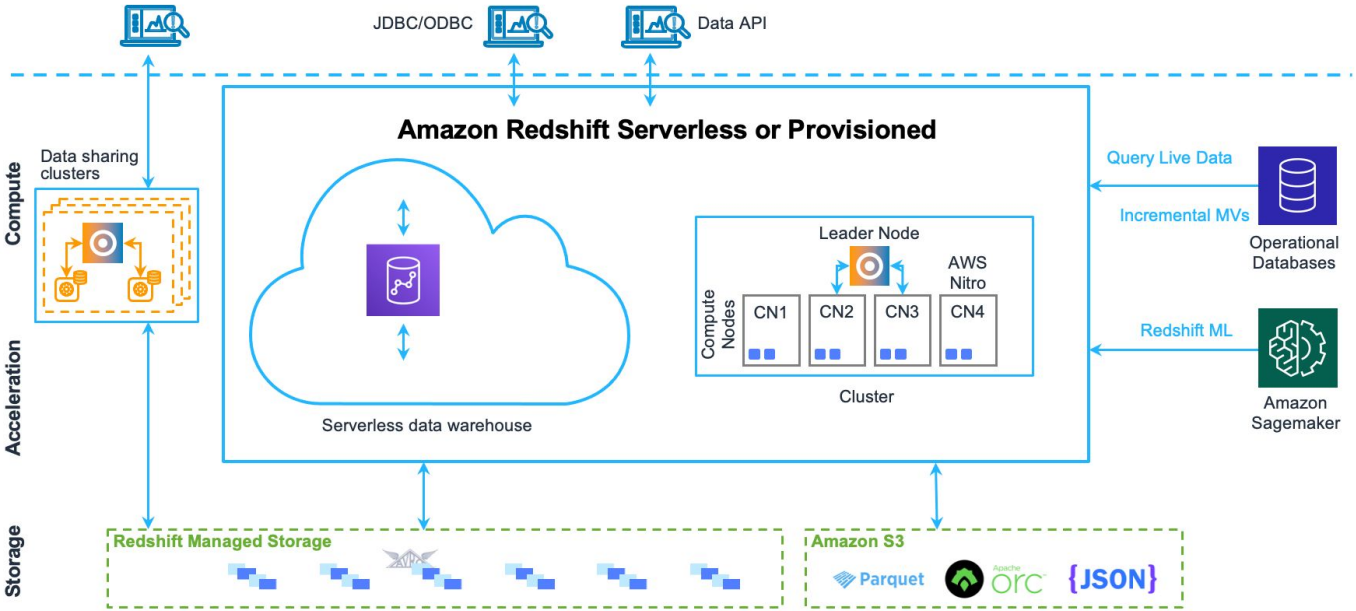
Google technology (both a blessing and a curse). Most powerful and flexible cloud data warehouse.



Amazon Redshift Serverless

Amazon data ecosystem.

Cheap, but with enough data warehouse functionalities.



Where is Microsoft (Azure)?

- Microsoft was not very strong in the traditional data warehouse field
- It provides a general purpose RDBMS (SQL Server)
  - Great for small to medium sized data warehouses (less than 1TB)
  - Not so well for large data warehouses, or data warehouses with special requirements
- When it comes to Azure
  - Had a product called SQL Data Warehouse, which is a product somewhat based on SQL Server, but didn't do well
  - Taking two different approaches
    - Working with third party
      - Databricks
      - The new Microsoft-Oracle alliance
    - Data lake

Agenda

- Data Warehouse
  - Business Intelligence
  - Major Data Warehouses
- Data Lake
- Azure Synapse Analytics
  - Demo and Lab

Data Lake

- Data warehouse is very expensive
- Modern data sources are generating well formatted raw data files, either relational, or JSON with reasonably stable schema, so well formatted that we can treat them as relational data, and analyze them using SQL queries.
- So why not keep the data in its raw format and query them directly?
    - This is the origin of data lake
    - It can be a repository of files (using Azure Storage Account)

Hadoop Data Lake

The term “data lake” gained popularity after it was used by Apache Hadoop ecosystem

A Hadoop data lake is one which has been built on a platform made up of Hadoop clusters. Hadoop is particularly popular in data lake architecture as it is open source (as part of the Apache Software Foundation project).

Azure uses Hadoop data lake technology to build its own data lake.

Hadoop Modules Review

1. Hadoop Common: contains libraries and utilities needed by other Hadoop modules
2. Hadoop Distributed File System (HDFS): a distributed file-system that stores data on the commodity machines, providing very high aggregate bandwidth across the cluster
3. Hadoop MapReduce: a programming framework for large scale data processing

HDFS: Data Storage. MapReduce: Data Analysis

Hadoop Distributed File System (HDFS)

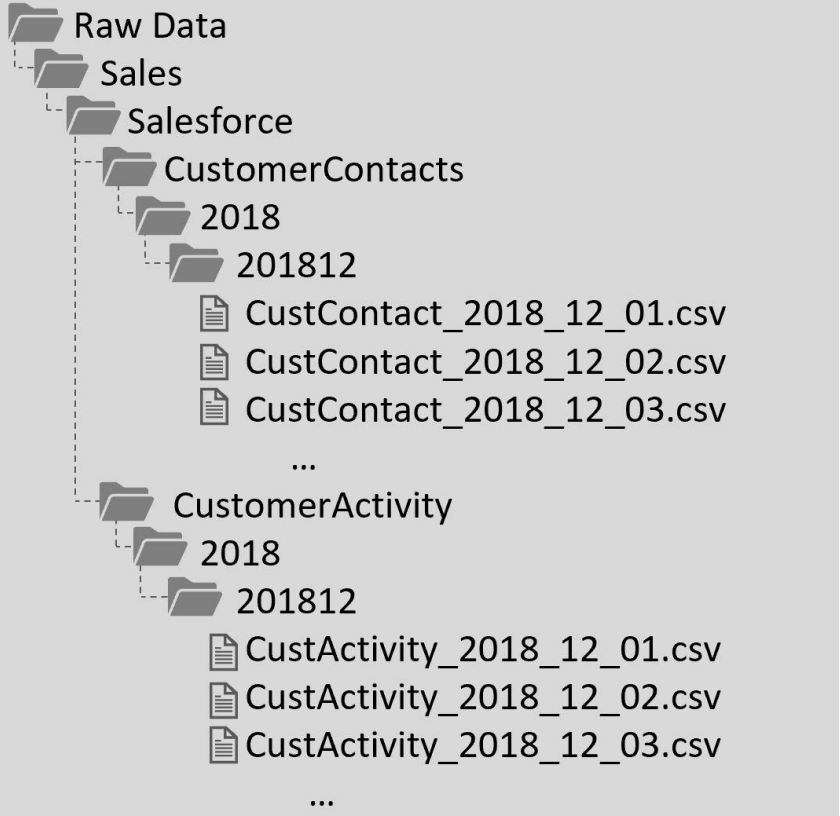
HDFS stores large files (typically in the range of gigabytes to terabytes) across multiple machines (nodes).

- Split data into blocks
- Replicate the data blocks across multiple (default 3) hosts
  - Two on the same rack
  - One on a different rack

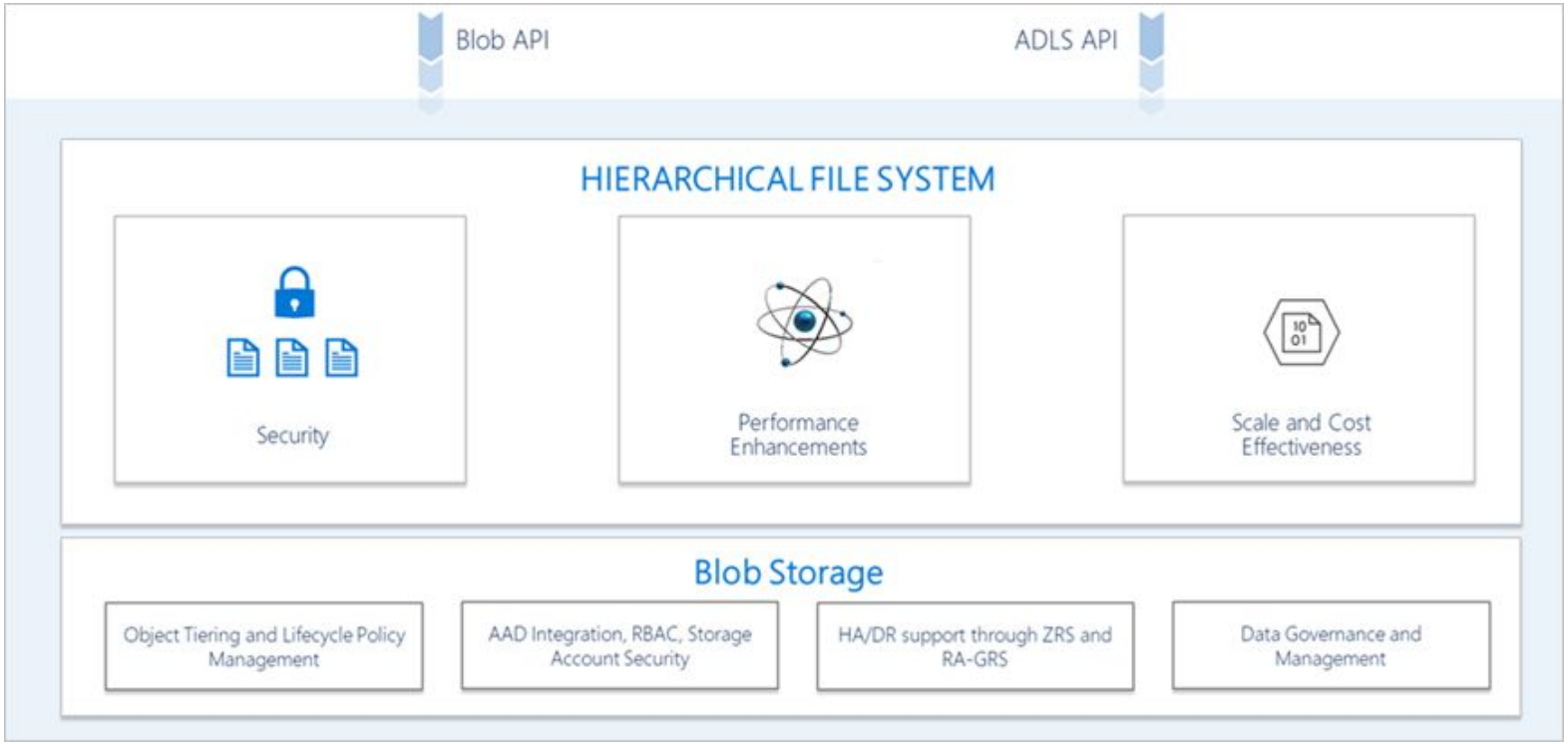
Data nodes can talk to each other to rebalance data, to move copies around, and to keep the replication of data high.

Data Lake

Example of date partitioning at year/month level:

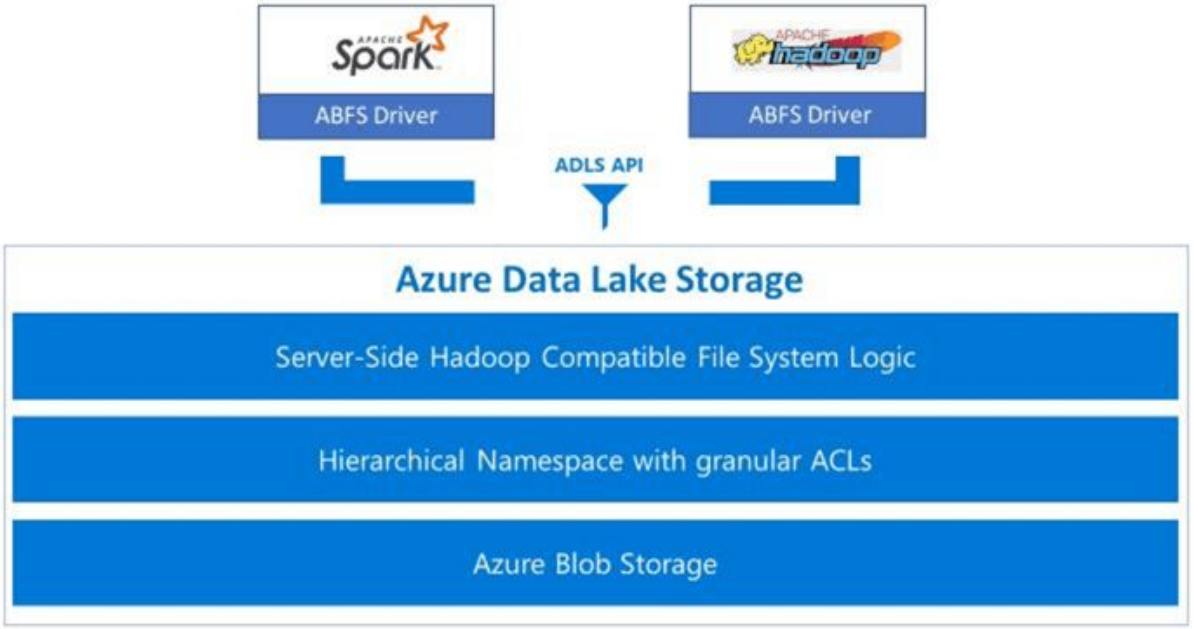


Data Lake


Blob API
ADLS API
HIERARCHICAL FILE SYSTEM
Security
Performance Enhancements
Scale and Cost Effectiveness
Blob Storage
Object Tiering and Lifecycle Policy Management
AAD Integration, RBAC, Storage Account Security
HA/DR support through ZRS and RA-GRS
Data Governance and Management

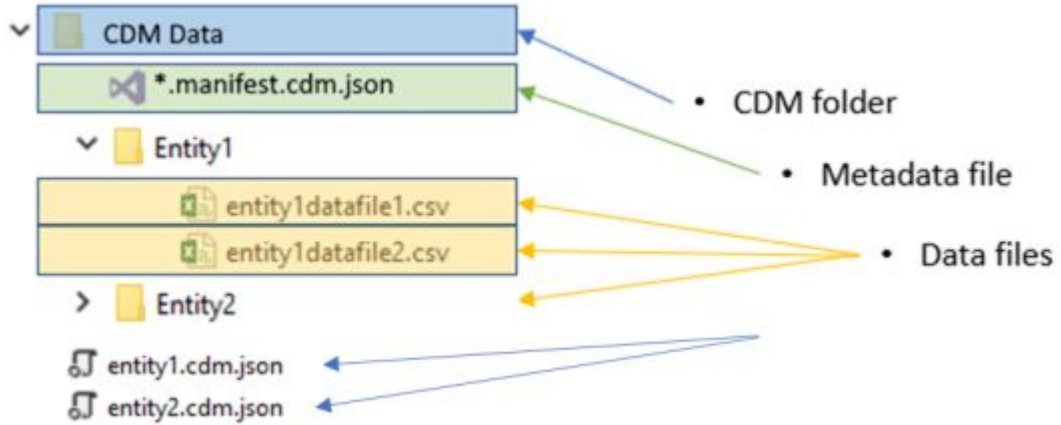
Data Lake

You can run SQL or Spark jobs against the data lake



Data Lake

Azure also adds more technologies to the original Hadoop system



File Format

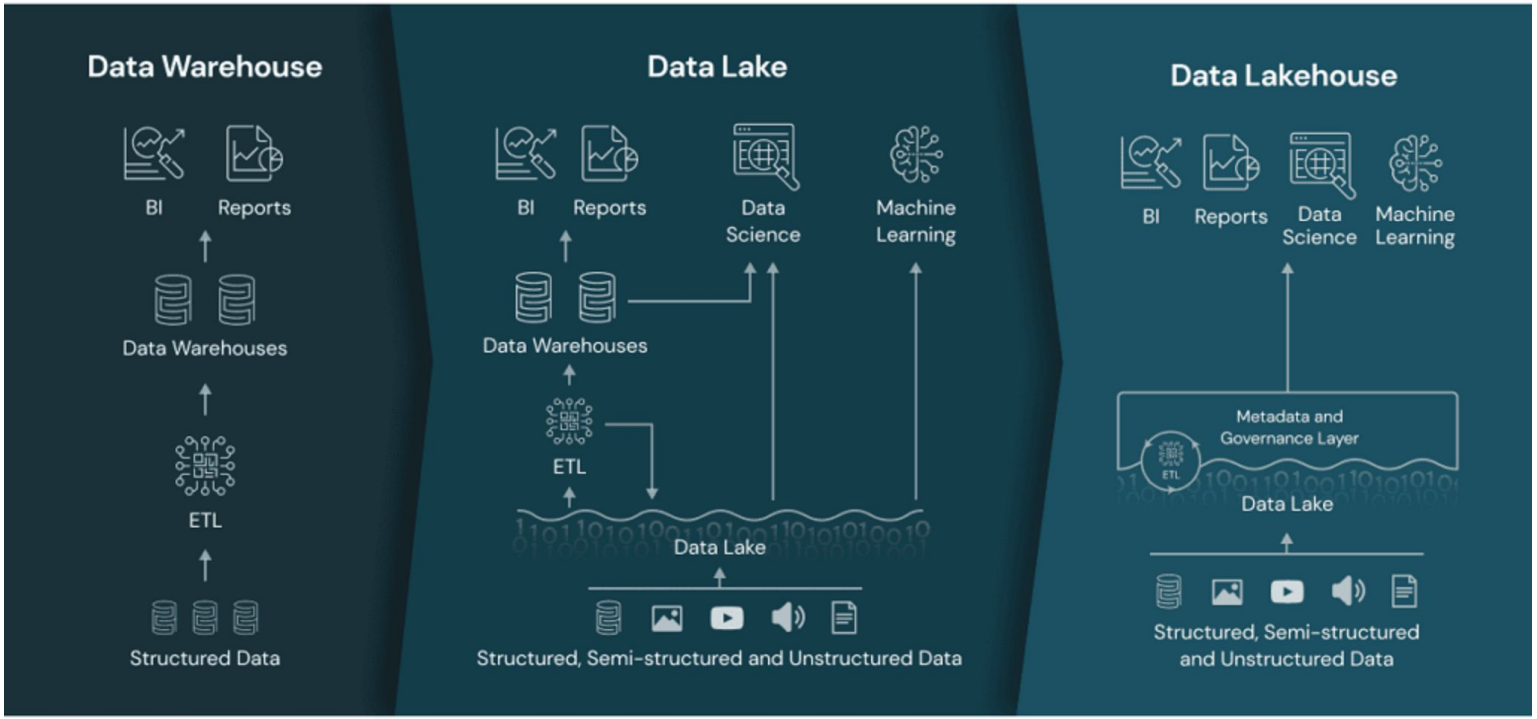
- CSV: row-based data format that is simple to read and write
- Parquet: columnar data format that uses compression and encoding scheme for fast data storing and retrieval. Better for Spark jobs
- Avro: row-based serialization with the data’s schema in the same file. Ideal for write-heavy operations
- ORC: columnar storage with compression. Better for HIVE jobs.

Data Lakehouse

Data lake and data warehouse are not exclusive to each other. You can build them side by side:

- Data warehouse: storage of aggregated, cleansed data (Golden Records)
- Data lake: storage of detailed original data and data that does not need strong consistency

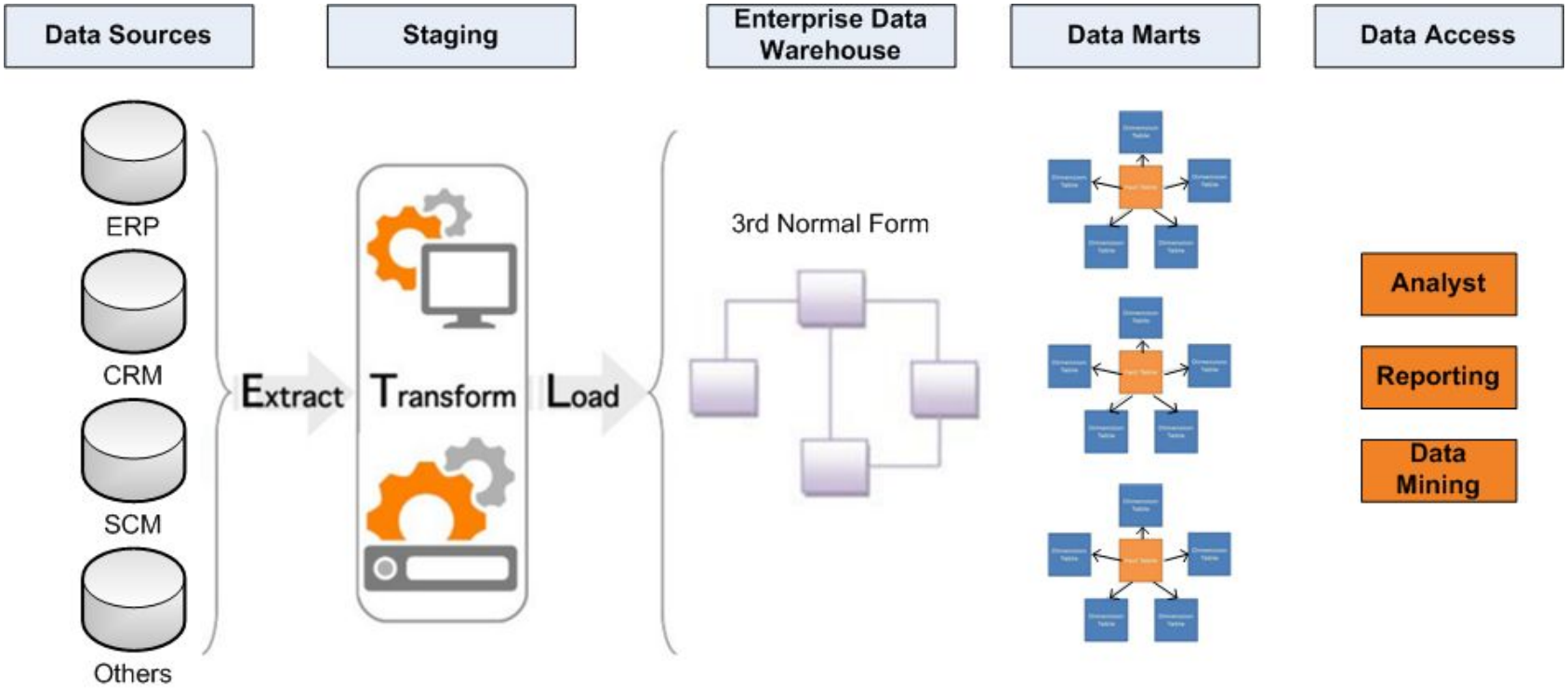
Data Lakehouse



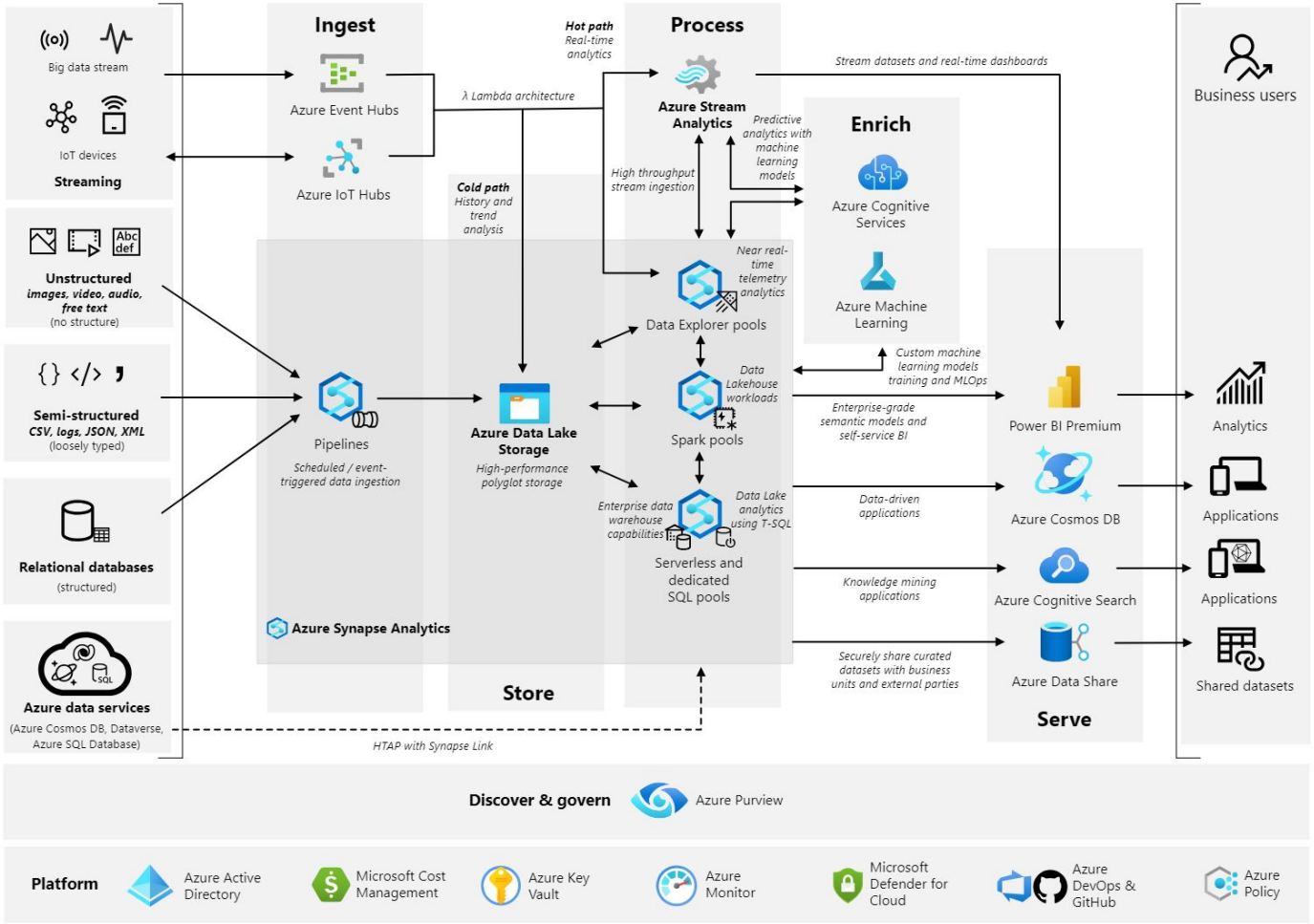
Agenda

- Data Warehouse
  - Business Intelligence
  - Major Data Warehouses
- Data Lake
- Azure Synapse Analytics
  - Demo and Lab

Typical Enterprise Data Flow



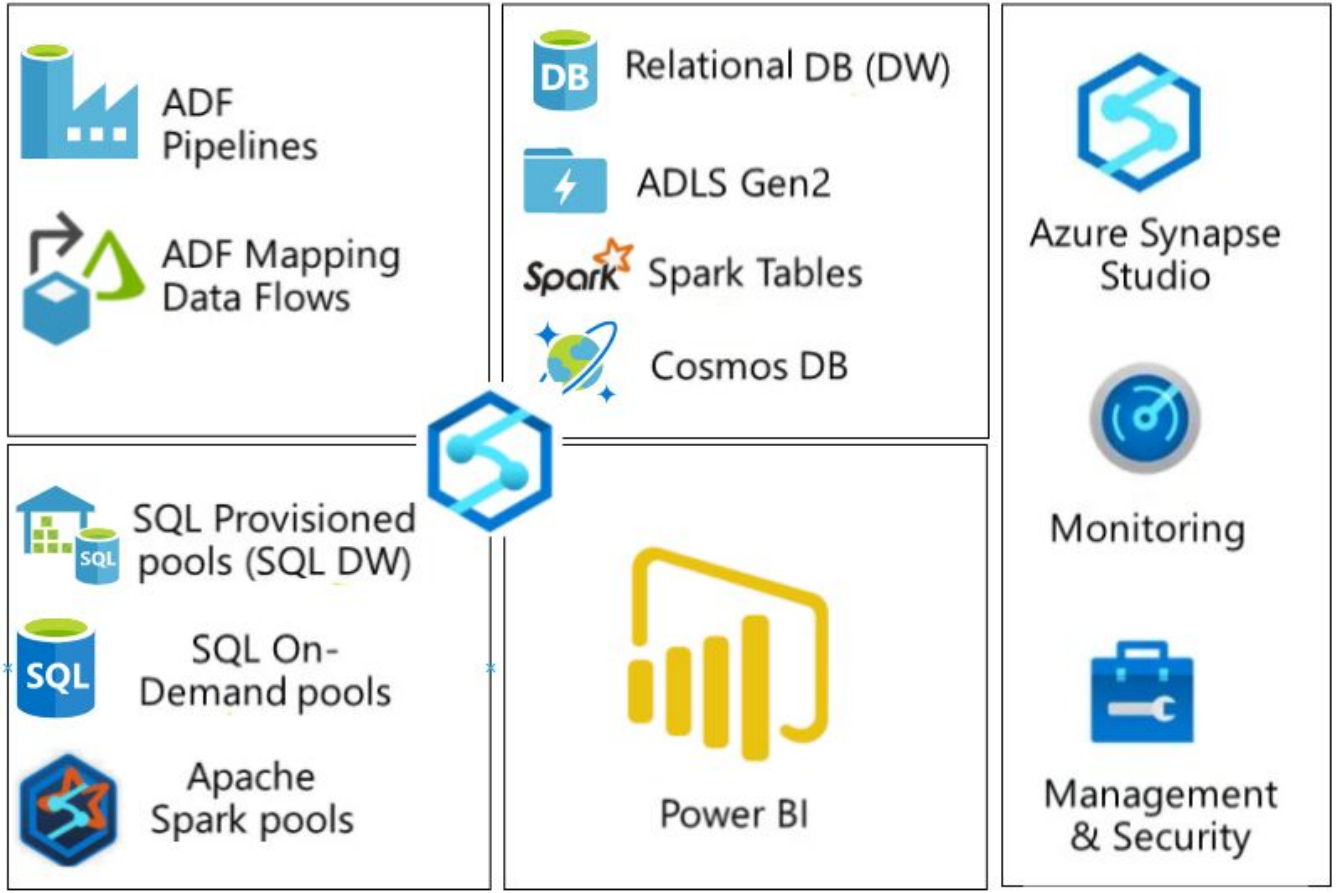
Azure Data Related Services



Microsoft Azure

Cloud Data Warehouse/ Lakehouse

Azure Synapse



Compute in Azure Synapse Analytics

Dedicated SQL Pool: Formally SQL Data Warehouse. A full scale data warehouse product

Serverless SQL Pool: Lakehouse solution

Spark Pool: Running Spark ecosystem

Spark Data Processing

Spark is widely used for ETL and Data Analysis.

- Fast data processing
- Data processing requiring compute power: Graph processing, Machine Learning, Joining

Why not build a data lake processing layer (compute) on top of the data lake files (storage)

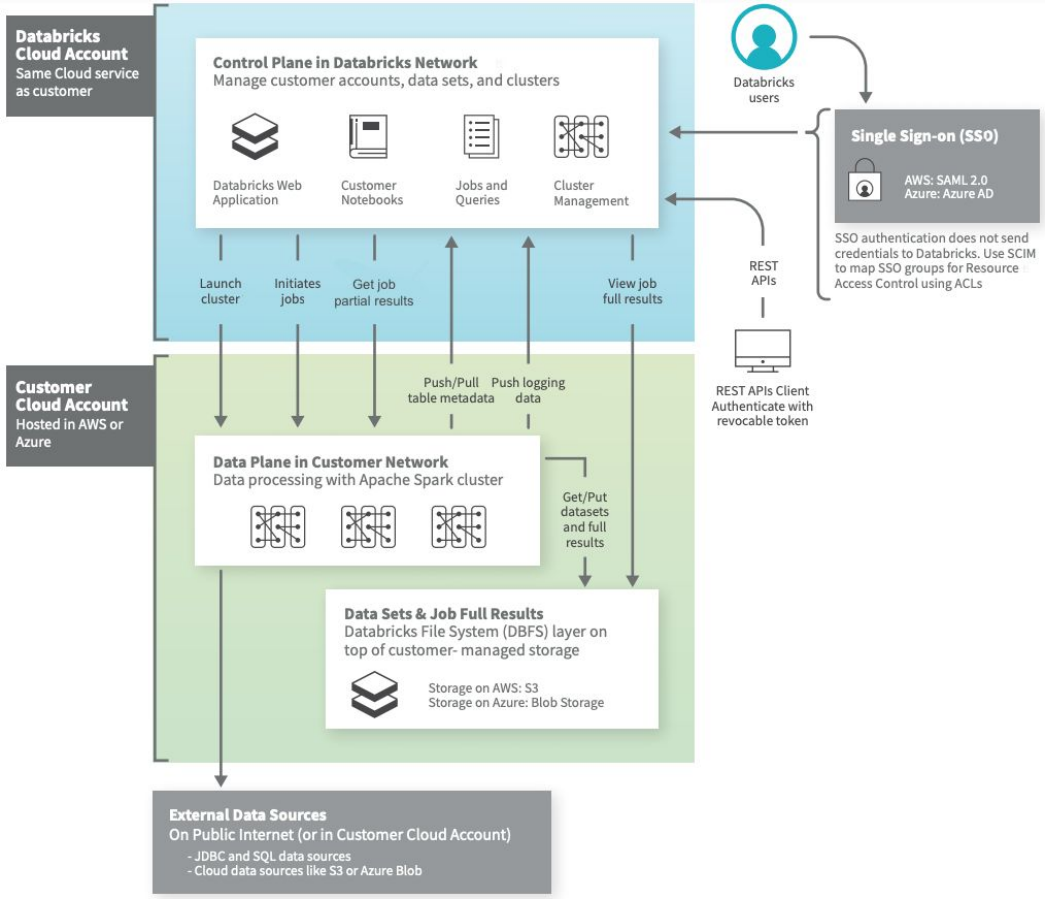
Databricks

American enterprise software company founded by the creators of Apache Spark.

Product: Delta Lake

- Web-based platform based on Spark
- File based data lake storage
- Cloud based compute cluster for Spark job execution
  - SQL platform and IPython-style notebooks.
- Automated cluster management

Databricks


Databricks Cloud Account
Same Cloud service as customer
Control Plane in Databricks Network
Manage customer accounts, data sets, and clusters
Databricks Web Application
Customer Notebooks
Jobs and Queries
Cluster Management
Databricks users
Single Sign-on (SSO)
AWS: SAML 2.0
Azure: Azure AD
SSO authentication does not send credentials to Databricks. Use SCIM to map SSO groups for Resource Access Control using ACLs
REST APIs
REST APIs Client Authenticate with revocable token
Launch cluster
Initiates jobs
Get job partial results
View job full results
Push/Pull table metadata
Push logging data
Customer Cloud Account
Hosted in AWS or Azure
Data Plane in Customer Network
Data processing with Apache Spark cluster
Get/Put datasets and full results
Data Sets & Job Full Results
Databricks File System (DBFS) layer on top of customer- managed storage
Storage on AWS: S3
Storage on Azure: Blob Storage
External Data Sources
On Public Internet (or in Customer Cloud Account)
- JDBC and SQL data sources
- Cloud data sources like S3 or Azure Blob

Databricks

Advantages:

- Cheaper than Snowflake
- Simpler than Synapse

Disadvantages:

- Ecosystem not complete

Agenda

- Data Warehouse
  - Business Intelligence
  - Major Data Warehouses
- Data Lake
- Azure Synapse Analytics
  - Demo and Lab

Synapse Demo and Lab

- Upload files to Storage Account
- Connect to Synapse
- Directly query data lake files using serverless SQL pool
- Create Integration Dataset
- Create Dataflow
  - Create source
  - Create sink
- Create pipeline and execute
- Verify result

Final Project

Review Assignment 06

- Data model

Discuss Assignment 07