Data warehouse, data lakes and data hubs....critical view

Mohamed Sharaf
Cloud Solution Architect
Microsoft

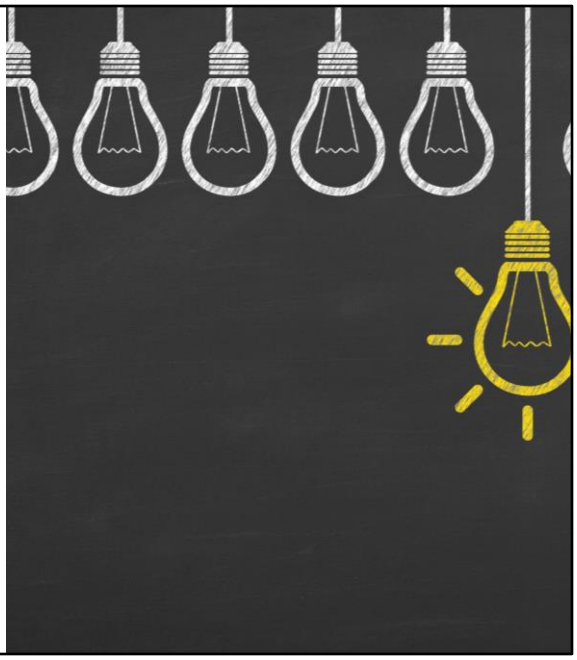
@MohamSharaf

https://www.linkedin.com/in/mosharafms/

Primary Audience

Solution Architects

Data Architects



Agenda

Data Analytics

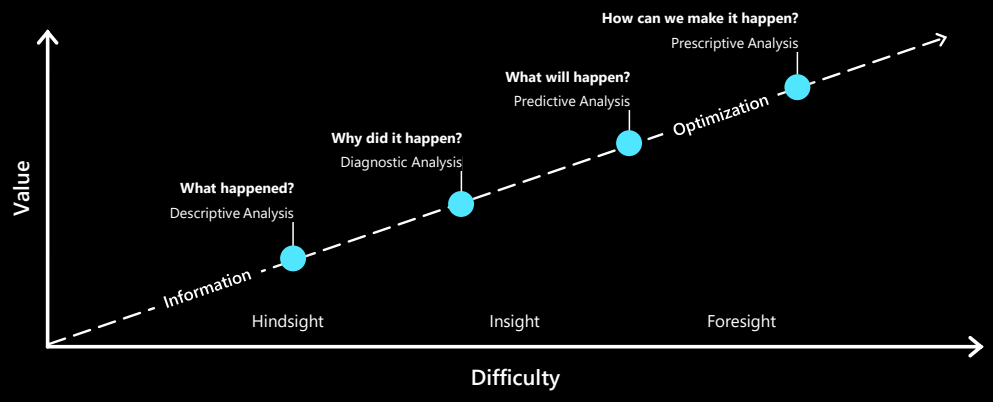
Defining core components

Comparisons

The most important criteria

Practical Wisdom

Analytics


How can we make it happen?
Prescriptive Analysis
What will happen?
Predictive Analysis
Why did it happen?
Diagnostic Analysis
What happened?
Descriptive Analysis
Optimization
Information
Value
Hindsight
Insight
Foresight
Difficulty

Definitions

Operational Stores

ETL/EL(T)

Data Hub

Data Warehouse

Data Lake

Operational Databases/Stores

- Source of your business data
- Typically implemented as relational databases

- Don’t forget about the *others*
- What about dark data
- External (contextual) data

ETL – ELT / EL(T)

- Extract – Transform – Load
- Now we have plenty of storage and we don’t know all questions?
- Extract-Load-Transform
- I want to be free to use multiple transformation tools?
- Extract-Load => Transform

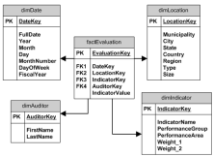
Data Warehouse

A logical architecture where collection of data from different sources integrated together

Relational representation of the data in a star schema

Typically, represented by one product

Implement MPP



Predictability

Answers specific questions

Azure Synapse, Snowflakes, BigQuery, RedShift

Data Warehouse

- DW is a logical design. Can be implemented in many ways
- The same relational engine used for the operational database can be used for data warehouse
- Some analysts use DW as a lake for raw data of all types

Data Lake

- A data lake is a concept consisting of a collection of storage instances of various data assets. These assets are stored in a near-exact, or even exact, copy of the source format and are in addition to the originating data stores*

- DO NOT Confuse the concept with product
- Core component is a storage that's virtually unlimited
- Structured, semi-structured & unstructured
- Storage alone doesn't provide much
- No indexing, No catalog, No ingestion

*Gartner

Data Lake

- Because it's a concept, typically achieved by multiple products for each vendor
- (Data Factory, Azure databricks, Purview, Azure Storage & Azure Synapse Analytics)
- (AWS Glue, Kinesis, Lake Formation, S3, Athena, Redshift)

Data Hub

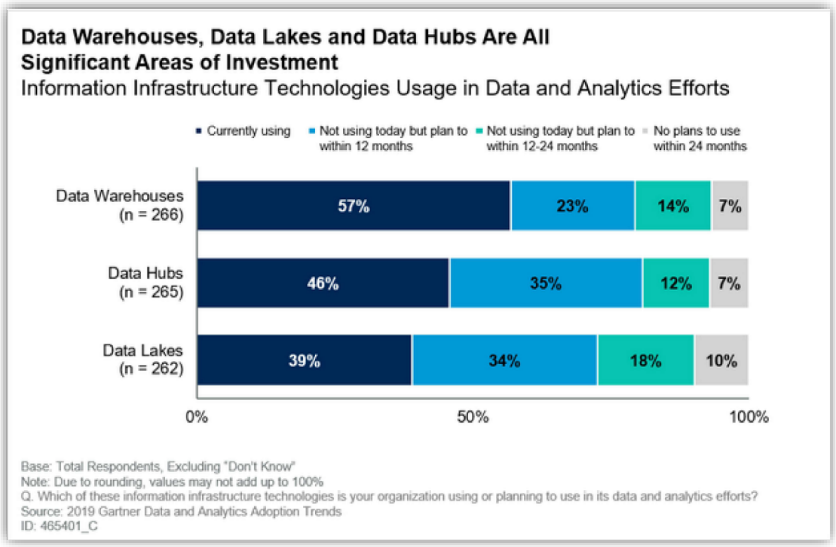
- Logical architecture which enables data sharing by connecting producers of data with consumers of data. Endpoints interact with the data hub, provisioning data into it or receiving data from it. The bug provides a point of mediation and governance and visibility*.

- Subtle component and shows up when sharing is a requirement
  - Especially outside the organization
- Not a data store

*Gartner.Com

Will a Data Hub Solve Your Data Dilemma? (gartner.com)
https://www.gartner.com/en/webinars/26171/will-a-data-hub-solve-your-data-dilemma-

Data Warehouse is dead…Long live Data Lake…or is it?



*Gartner.Com

Gartner's report "Data Hubs, Data Lakes, and Data Warehouses: How They Are Different and Why They Are Better Together".
Published in February 2020

Data Lake vs Data Warehouse

	Data Lake	Data Warehouse
Store structured, non structured and semi structured data	Yes	Yes (not the fastest option)
Multi-thread / Multi-client retrieval	Yes	Yes
Index data	Not in the storage layer (z-ordering and similar techniques help)	Yes
Partition data	Yes	Yes
Switch compute engines	Yes	No. Data warehouse is storage and compute engine
Writing using any schema	Yes	Yes (using XML, JSON,...)
Price	Cheaper + Compute cluster cost	More expensive
Best for	Exploring	Serving gold data

Optimize performance with file management — Databricks Documentation
https://docs.databricks.com/delta/optimizations/file-mgmt.html

Database vs Data Warehouse

	Database	Data Warehouse
Store, index, retrieve data using multi-threads/multi clients	Yes	Yes
Optimized for multiple compute nodes	No	Yes (MPP-check your size)
Main workload	Short – many – concurrent	Long- few – concurrent (limited)
Concurrency	High	Low
GeoSpatial & Specialized data types	High support	Low support
Specialized storage (NoSQL, SQL, Graph..)	High support by having many engines	Low Support as DW usually share one engine
Schema	Highly normalized (no redundancy)	Denormalize to optimize for fast retrieval
Best for	Transaction processing	Analytics

Memory and concurrency limits - Azure Synapse Analytics | Microsoft Docs
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/memory-concurrency-limits

Overview of Warehouses — Snowflake Documentation
https://docs.snowflake.com/en/user-guide/warehouses-overview.html

Amazon Redshift clusters - Amazon Redshift
https://docs.aws.amazon.com/redshift/latest/mgmt/working-with-clusters.html

-- Geospatial
Using Spatial Data with Amazon Redshift | AWS News Blog
https://aws.amazon.com/blogs/aws/using-spatial-data-with-amazon-redshift/

Introduction to BigQuery GIS | Google Cloud
https://cloud.google.com/bigquery/docs/gis-intro

What is the practical most important factor?

YOUR SKILLSET

- Personal – Group – Enterprise Skillset shape how we use technology
- Thorough understanding of what is the current and potential skillset is crucial to succeed.

Practical Wisdom

"The architect who doesn't code is an architect who believes everything is possible"

-Anders Hejlsberg
Microsoft Technical Fellow & C# Architect

Practical Wisdom

Scenario
- Small data warehouse on-prem based on SQL Server.
- SQL skillset is the dominant
- Current size 4 TB and expected to increase to 30 over 3 years

Solution
- Respect the skillset. Unless there's a realistic reskilling plan, use only relational-based solutions

Practical Wisdom

Scenario
30 TB data warehouse based on SQL
SQL skillset is the dominant
3 reporting solutions, 100s of reports and 120 users
Historical data is rarely queried

Solution

- Most cloud-based data warehouses have concurrency limitations. SQL Database (SMP) can scale up to 100TB
- These requirements make us think if it is a Warehouse?

Practical Wisdom

Scenario
180 TB data warehouse based on SQL
SQL skillset is the dominant
3 reporting solutions, 100s of reports and 120 users
Historical data is frequently queried
Data can be divided into domains

Solution
- Hot data size requires MPP solution
- These requirements make us think of a mix between data warehouse + data marts

Practical Wisdom

Scenario
180 TB data warehouse based on SQL
Social media impressions & website logs data planned to be added
SQL skillset is the dominant, there's a plan to hire data science head
3 reporting solutions, 100s of reports and 120 users
Historical data is frequently queried
Data can be divided into domains

Solution

- Hot data size requires MPP solution
- These requirements make us think of a mix between data warehouse + data marts
- Adding data lake to host unstructured data.
- Need data catalog now more than before

Practical Wisdom

Scenario

Different data types from different sources

Not known what questions to answer. No previous DW implementation

Fresh team to be recruited

Plan is to use different compute engine for different tasks and unify later. Use the storage for governance

Solution

- Uncertainty and exploration are the key for choosing Data Lake. Not the different data types
- Hadoop, Spark, Dask are all available to you to choose which one to choose as compute engine.
- Data cataloging and searching are crucial.

Where is the Data Hub?

- Do you want to share or receive data externally outside the organization?
  - Data hub is mandatory to have automation and governance
- Not a requirement
  - Optional.
  - Think of data hub as building API layer for a web app.
- Can be different implementations
  - Stored Procedures & Views
  - Data Lakes
  - REST API / GraphQL
  - Streaming (Kafka)
  - ……

Thank you!

Mohamed Sharaf
Cloud Solution Architect
Microsoft

@MohamSharaf

https://www.linkedin.com/in/mosharafms/