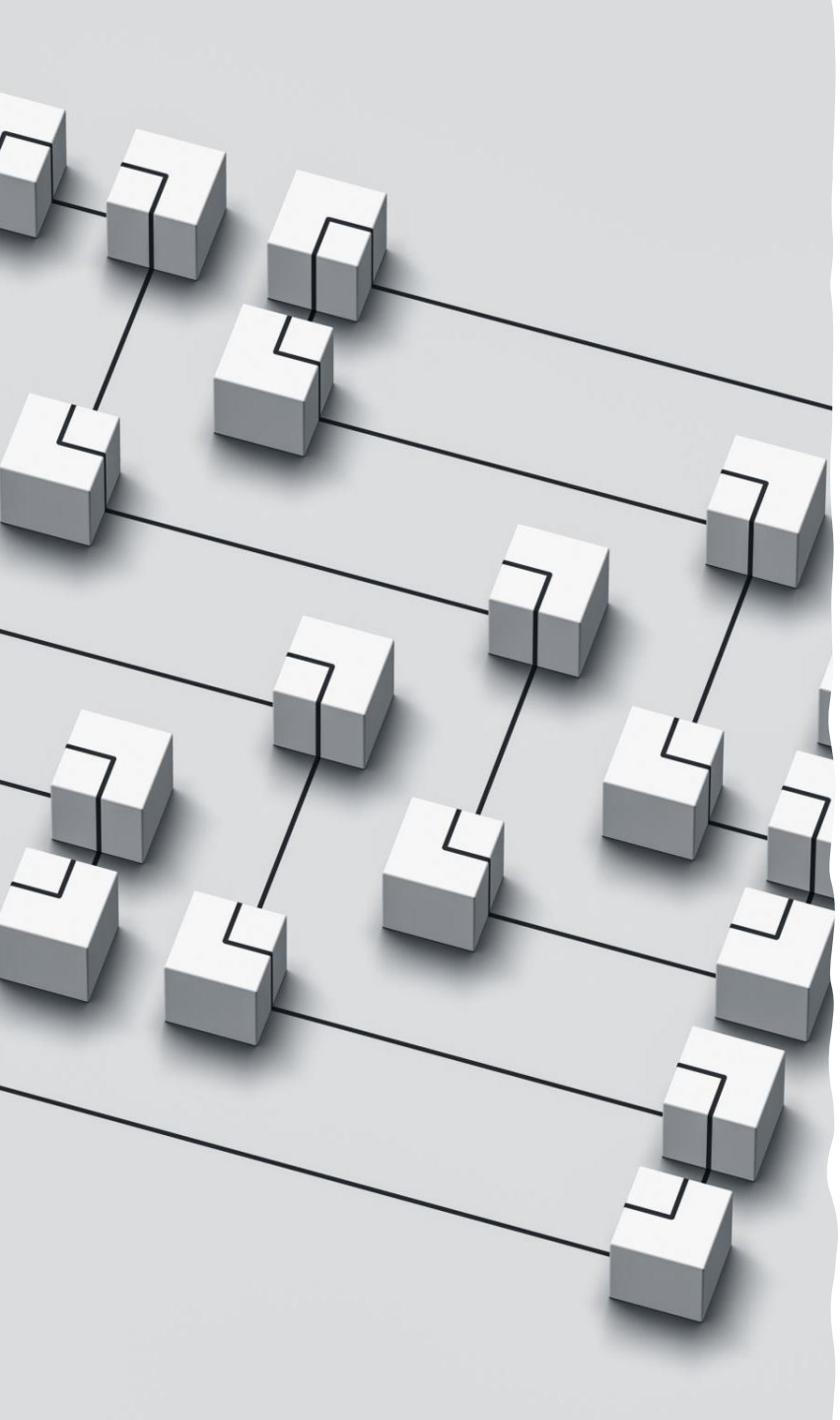
Business Intelligence and Data Warehousing (ANL408)

- By Sabarish Nair

Recap from last week....

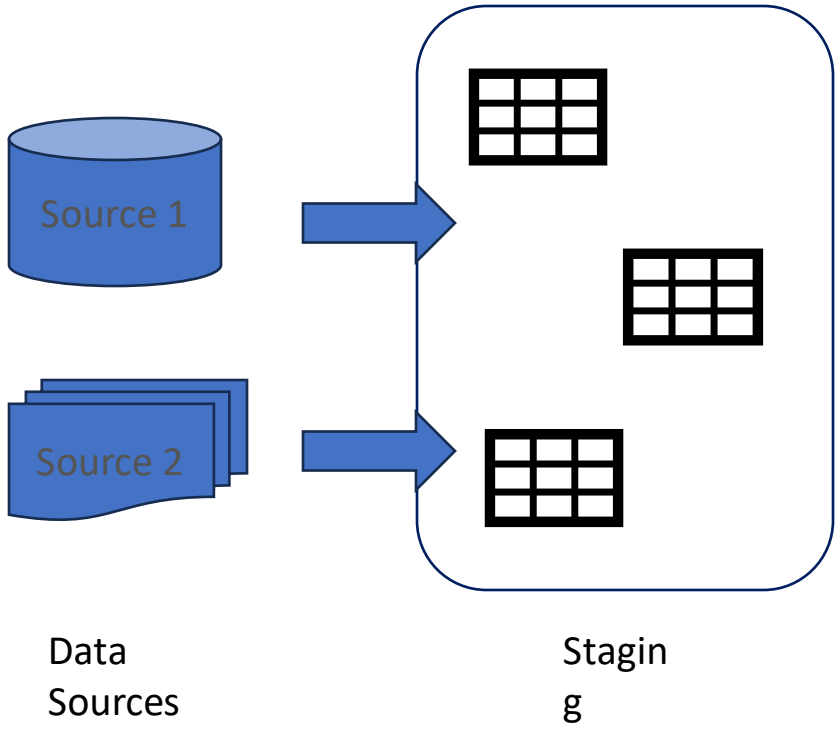
- Slowly Changing Dimensions (SCD)
- Type 0
- Type 1
- Type 2
- Administering Type 2 Dimensions
- Type 1 + Type 2
- Type 3: Additional Attribute
- Practical Example: EDA
- Practical: Check the quality of data
- Practical: Get the data statistics



ETL

- Extract, Transform and Load
- **Extract:** Data is gathered from various sources (E.g. APIs, spreadsheets, applications, etc.)
- **Transform:** Transform the data to fit the schema of data warehouse (E.g. Cleaning the data, converting the data type, etc.)
- **Load:** Load the data into data warehouse
- Automated Process

Extract



- Extract Data from different sources
- Understand the data
- Data is now a part of DWH
- Most commonly transient
- Mostly data is copied and deleted

Extracting Types

INITIAL LOAD: FIRST RUN, LOAD ALL DATA

DELTA LOAD: SUBSEQUENT RUNS, ONLY ADDITIONAL DATA

Initial Load

- Extract all the data from source systems
- Also known as full load or full refresh
- Usually done when setup is done for the first time
- Can be time consuming and resource intensive
- Need to build a strategy for initial load:
  - What data is needed?
  - What is the volume of data?
  - Good time to load the data (Nights, Weekends)?

Delta Load

- Known as incremental load
- Involves only changes/addition (delta)
- Ongoing update to data warehouse
- Deal with only a subset of data
- More efficient than initial load
- Faster load time and reduced resource consumption
- Frequency: Decided by business (Daily, Weekly, etc.)
- How do we identify delta (Date_Created, date_modified, MAX(primary key)??

Transform

Create a consolidated view of the data

Reshape (for analytical or reporting needs)

Consolidate Data

Merge data from different systems

Transaction_PK	Amount	Date
1	100	10/03/2024
2	200	10/04/2024

System 1

Transaction_PK	Amount	Date
3	300	10/03/2024
4	400	10/04/2024

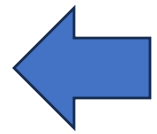
System 2

Reshape Data

Reshape according to business requirements

Transaction_PK	Amount	Date
1	100	10/03/2024
2	200	10/04/2024

Transaction_PK	Amount	Date_FK
3	300	20241003
4	400	20241004



Make it an FK to link it with dimension

Different Transformations

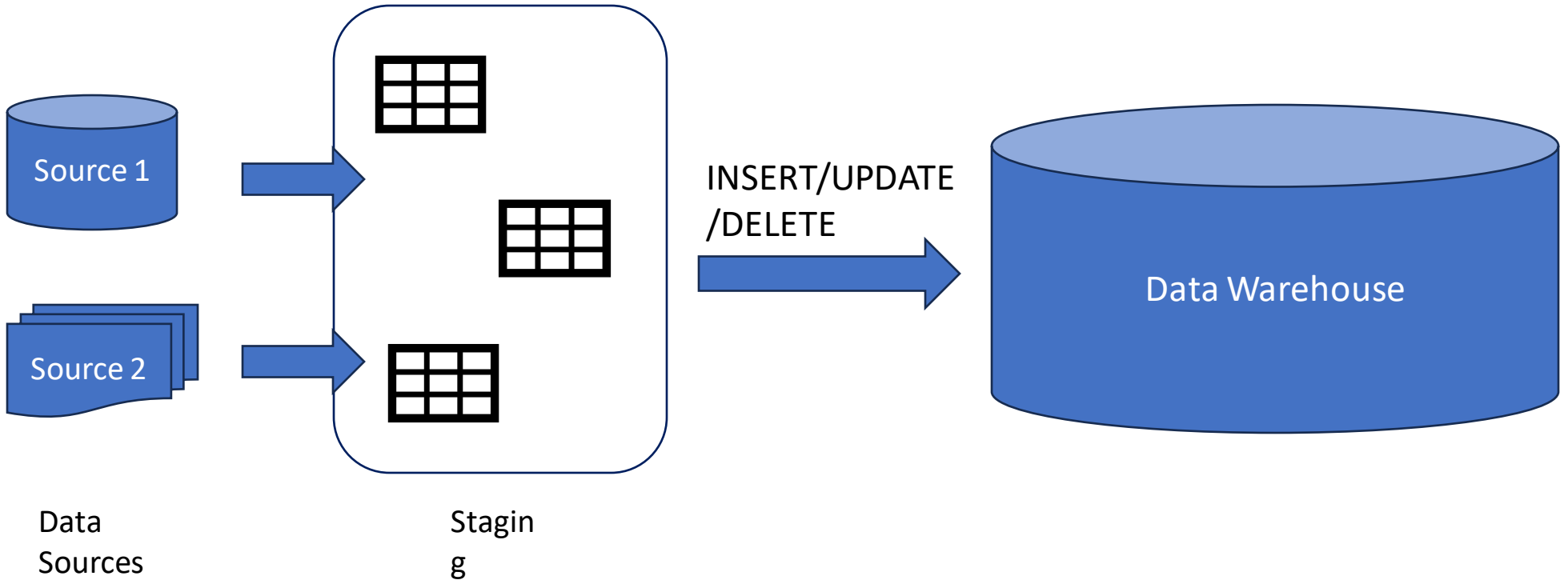
Deduplication
Filtering (rows and columns)
Cleaning
Standardization
Surrogate key generation

Basic Transformations

Joining
Splitting
Aggregating
Creating Derived Fields

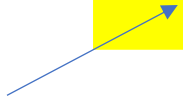
Advanced Transformations

Load


Source 1
Source 2
INSERT/UPDATE
/DELETE
Data Warehouse
Data
Sources
Stagin
g

Load (INSERT/APPEND)

Product_ID	Name	Category
1	Bat	Toy
2	Milk	Beverage
3	Ball	Toy
4	Almond Milk	Liquid



New row added

Load (UPDATE)

Product_ID	Name	Category
1	Bat	Toy
2	**Almond Milk**	Beverage
3	Ball	Toy

Existing Row Updated

Load (DELETE)

Delete is not usually performed

Product_ID	Name	Category
1	Bat	Toy
~~2~~	~~Milk~~	~~Beverage~~
3	Ball	Toy

Deleted Row

ETL Tools

- **Enterprise** (Commercial, most mature, offer good support)
- **Open Source** (Source Code, Often Free, usually no support)
- **Cloud-native** (Cloud technology, data already in cloud, efficiency, flexibility)
- **Custom** (Own development, customized, internal resources and training, not an ideal solution)

ETL Tools

- **Enterprise** (Alteryx, Informatica, Oracle Data Integrator, Microsoft SSIS)
- **Open Source** (Hadoop, Pentaho Data Integration)
- **Cloud-native** (Azure Data Factory, AWS Glue, Google Cloud Data Flow)

Choosing ETL Tools

- Evaluate current situation/needs
- What do you want to improve?
- Data Sources and other tools?
- Define your requirements!
- Who are the users?

ETL Tool Evaluation Matrix

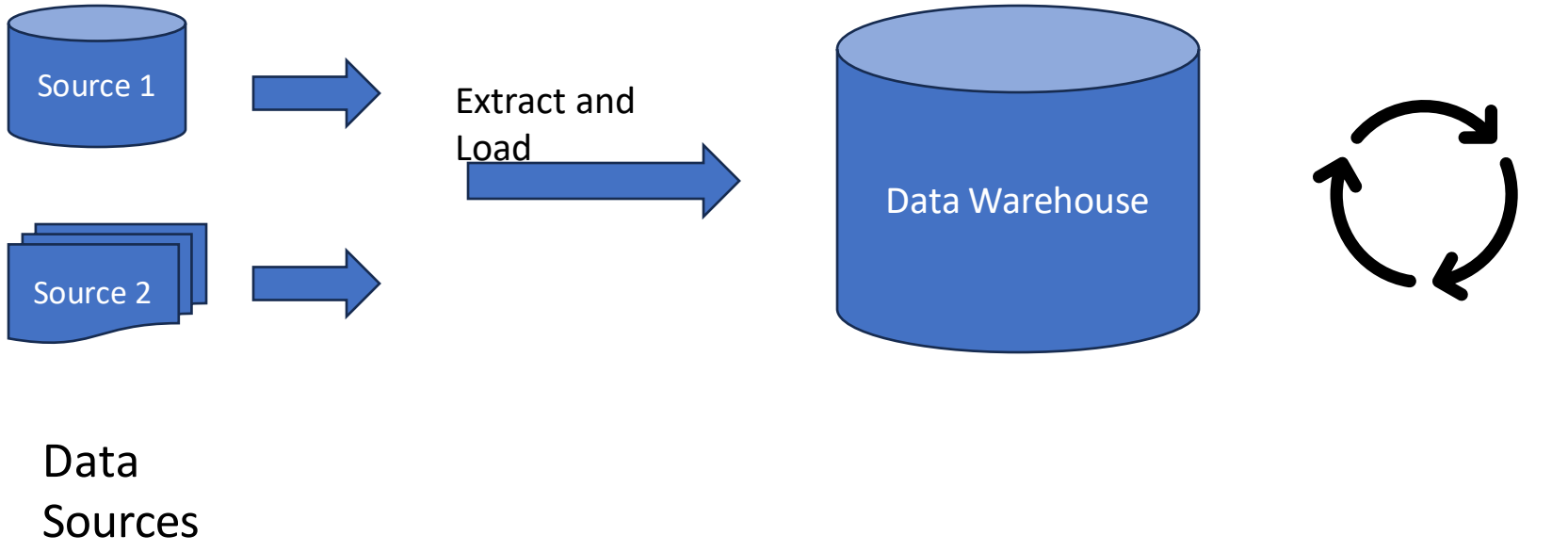
	Text	Must Haves?	Weight/Importance	Rating
Cost			1-5	1-5
Connectors				
Capabilities				
Ease of use/work				
Reviews				
Support/Extras				

Total Weighted Score:

TEST, DEMO, TRIAL!!!

ELT

Extract Load and Transform



Data Sources

ETL vs ELT

Parameter	ETL	ELT
Sequence	Extract -> Transform -> Load	Extract -> Load -> Transform
DW Role	Primarily serves as storage and query engine	Serves as storage, query engine and processing engine
Resource Usage	Requires additional computational resources for ETL	Leverages computational resources of the data warehouse
Complexity	Can be complex due to separate transformation logic	May be simpler in terms of initial setup, but more complex within the warehouse
Nature of Transformations	More Stable Transformations	Transformations change quickly
Security	More Secure	Less Secure since we are loading all the data
Usage	Reporting	Data Science, ML (No specific engineering transformations, real time)


Thank you