What are the different layers of a data warehouse?

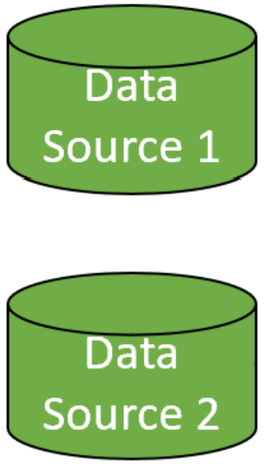
INTRODUCTION TO DATA WAREHOUSING

Aaren Stubberfield
Data Scientist

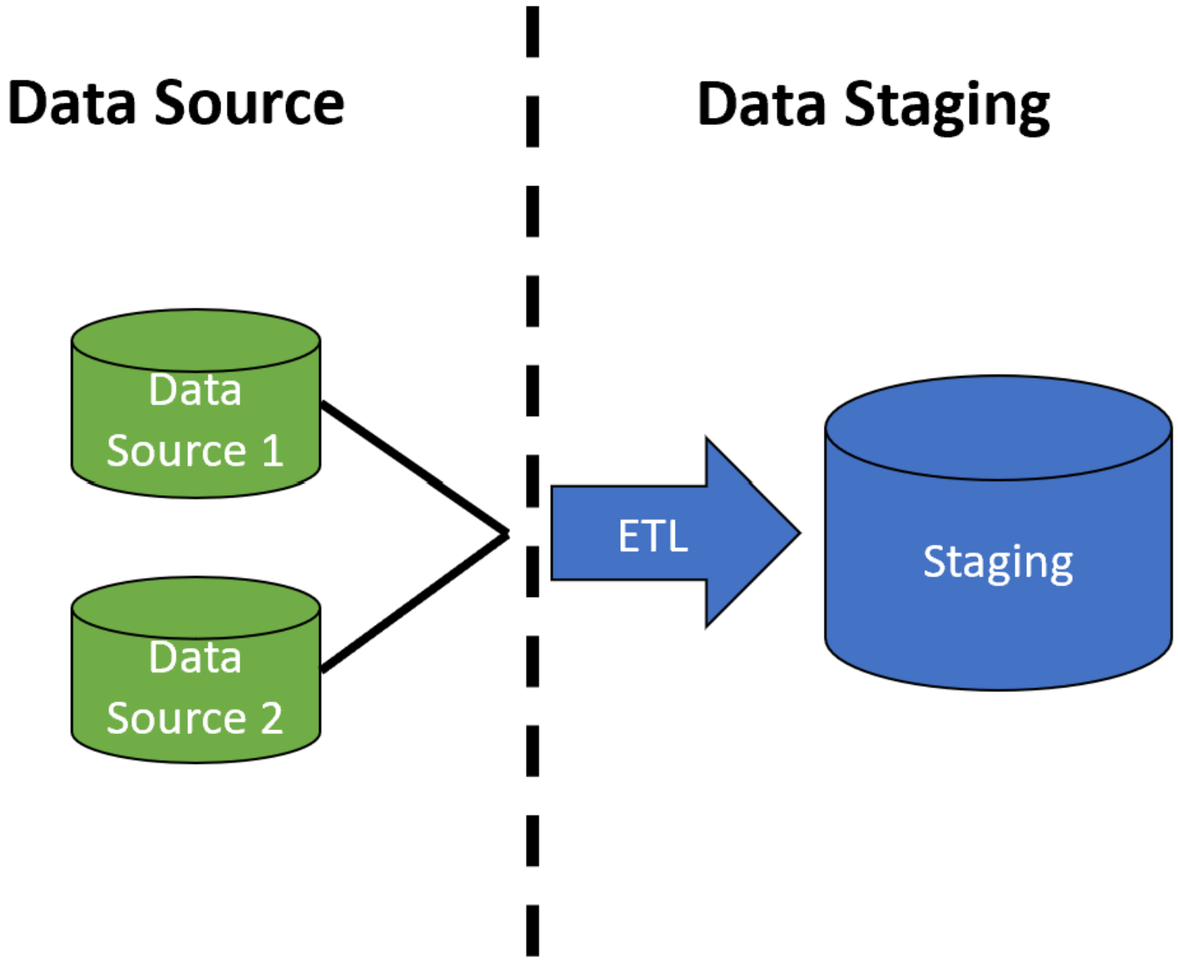
datacamp

Layer overview - data source

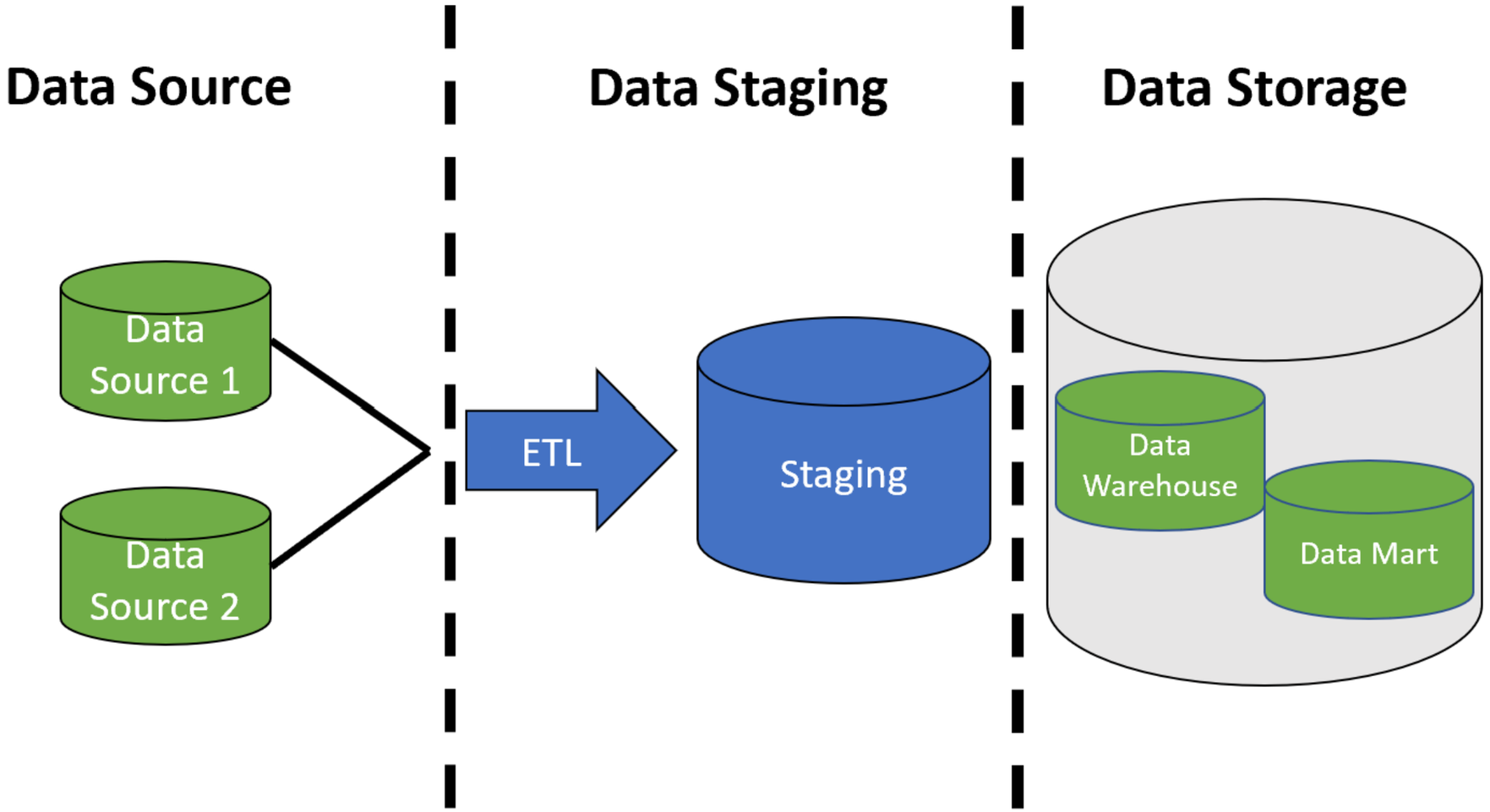
Data Source



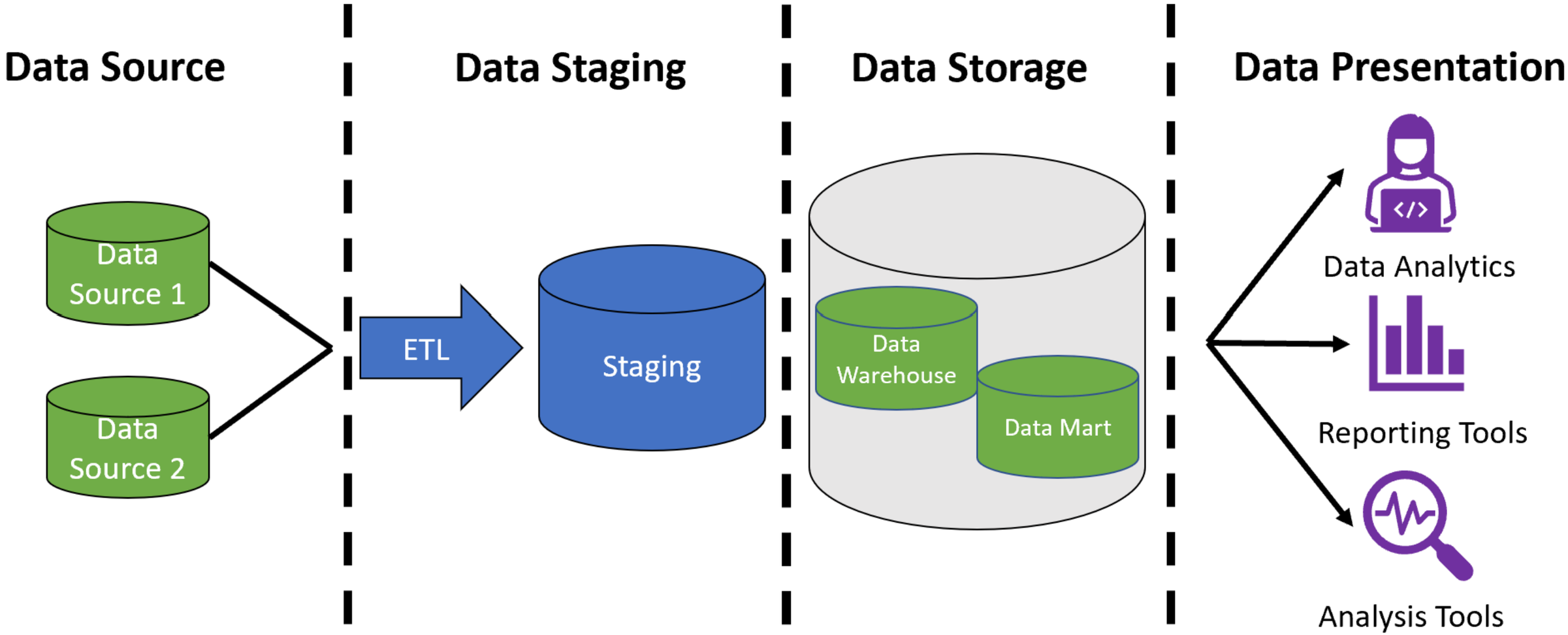
Layer overview - data staging



Layer overview - data storage

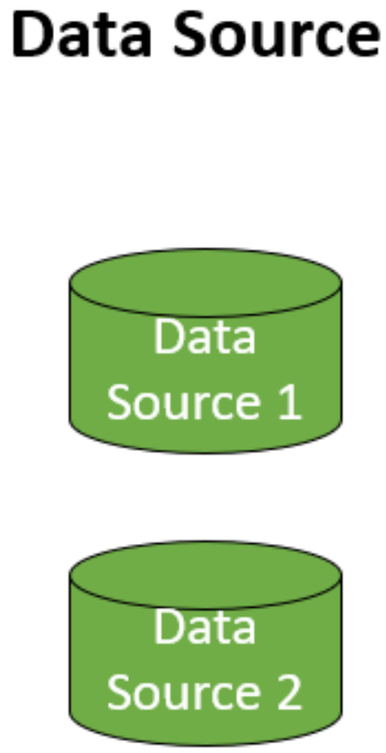


Layer overview - data presentation



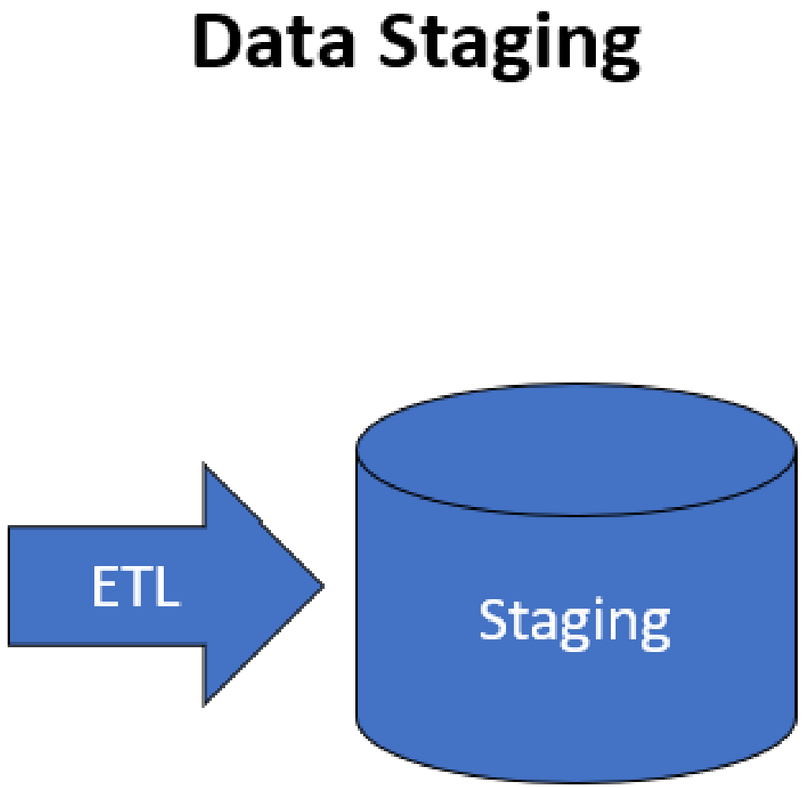
Data source layer

- All data sources for data warehouse
- Examples of data sources:
  - Transactional database
  - Log files
  - Spreadsheets



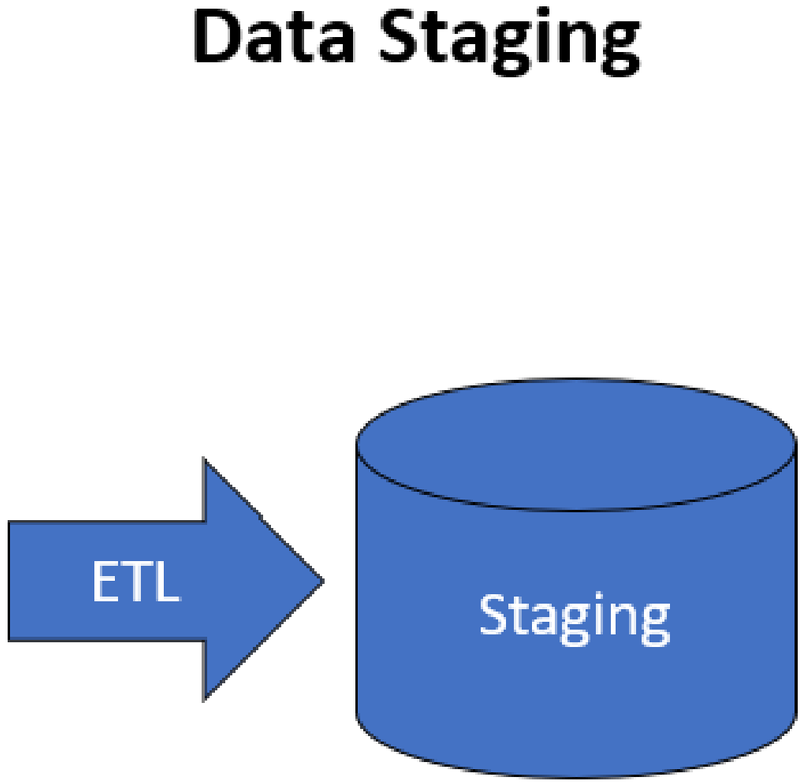
Data staging layer

- Layer extracts, transform, and clean data through ETL process
- Contains ETL process and storage tables



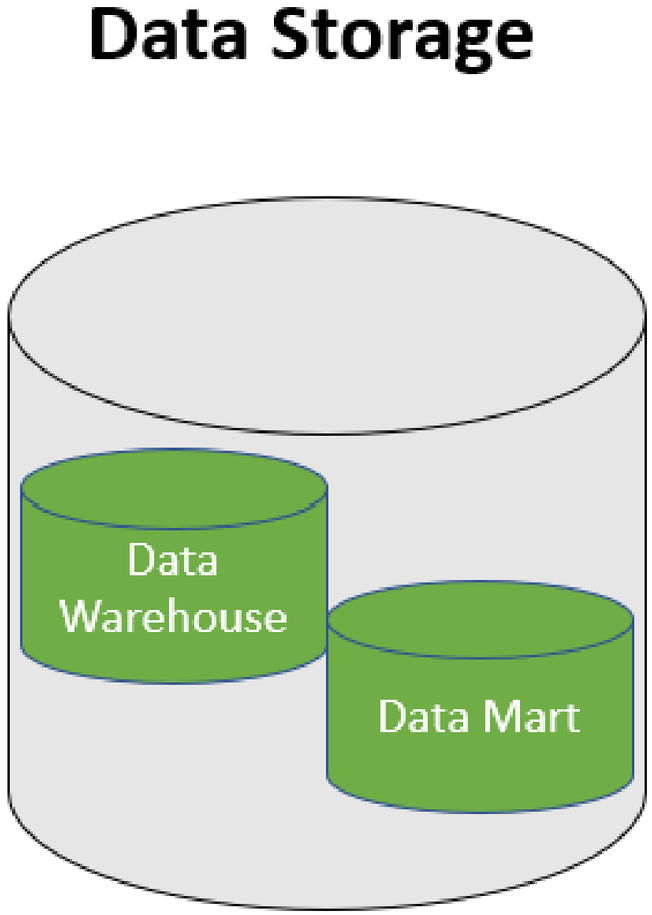
ETL process within data staging layer

- Extracted
- Business rules applied and cleaned
- Staging database often used
- Must be able to extract valid data
- Batch / full loading



Data storage layer

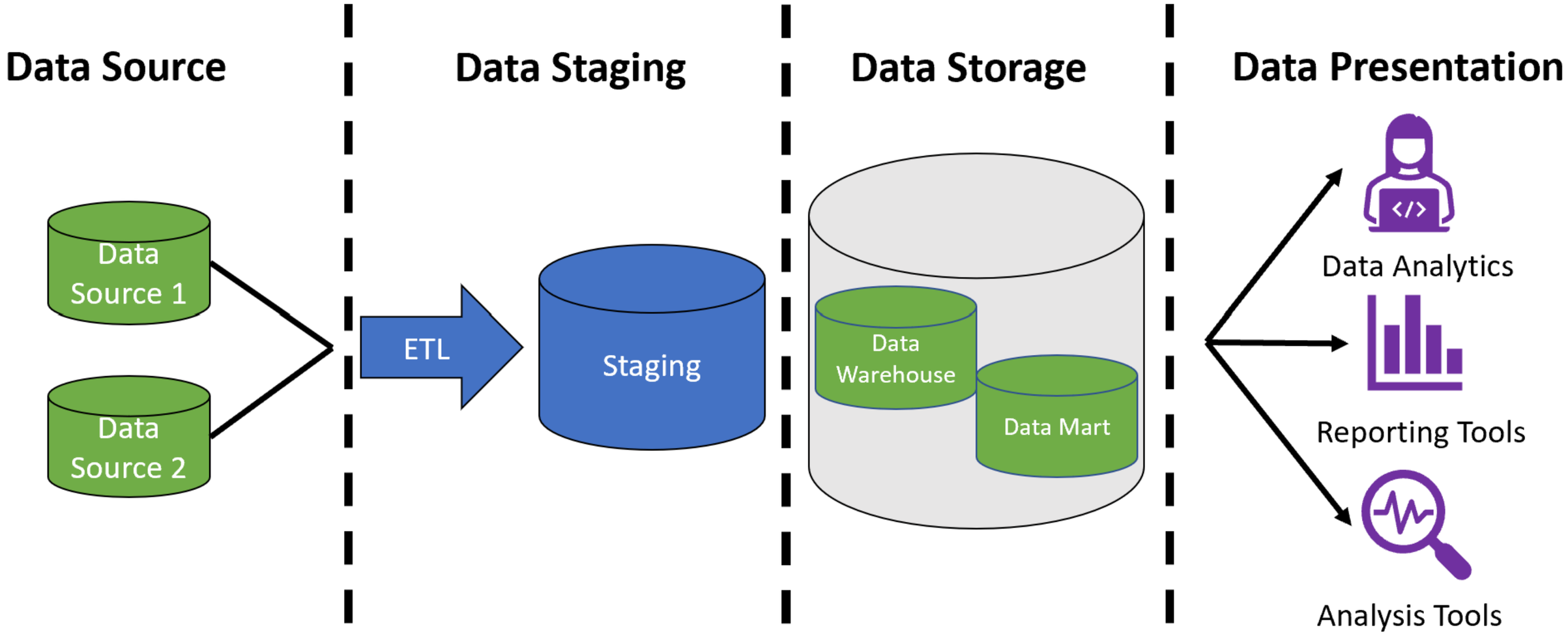
- Data is stored in warehouse and data marts
  - *Data warehouse -> Data mart*
  - *Data mart -> Data warehouse*



Data presentation layer

- Users interact with stored data
- Users:
  - Use BI (Business Intelligence) tools
  - Use data mining tools
  - Create direct queries

Summary


Data Source
Data Staging
Data Storage
Data Presentation
Data Source 1
Data Source 2
ETL
Staging
Data Warehouse
Data Mart
Data Analytics
Reporting Tools
Analysis Tools

Let's practice!

INTRODUCTION TO DATA WAREHOUSING

The presentation layer

INTRODUCTION TO DATA WAREHOUSING

Aaren Stubberfield
Data Scientist

datacamp

Presentation layer tools

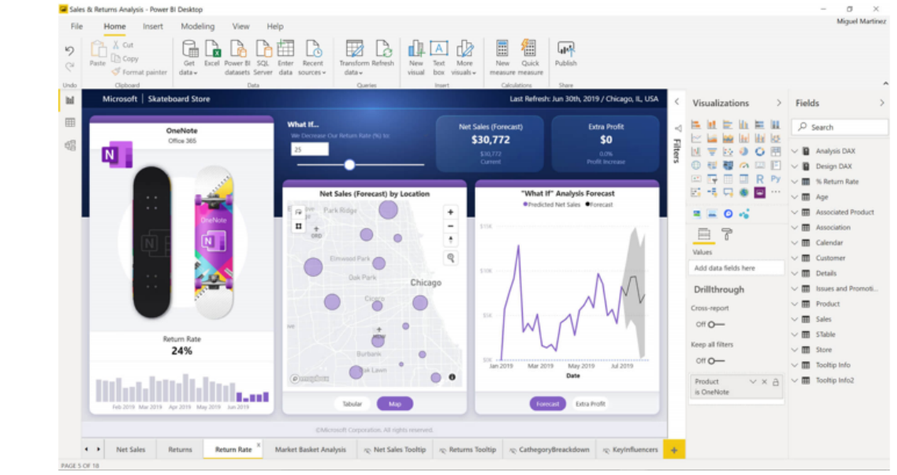
- Users interact with the presentation layer
  - Area of constant development

Presentation Layer Groups:

- Automated reporting/dashboarding tools
- BI/data analytics
- Direct queries

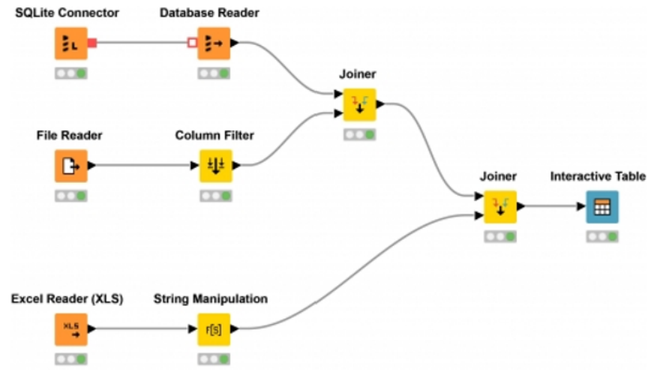
Automated reporting/dashboarding

- Goal:
  - Create reports needed for decision making making
  - Create dashboards using historical data
- Users:
  - Analysts
  - Citizen Data Scientist



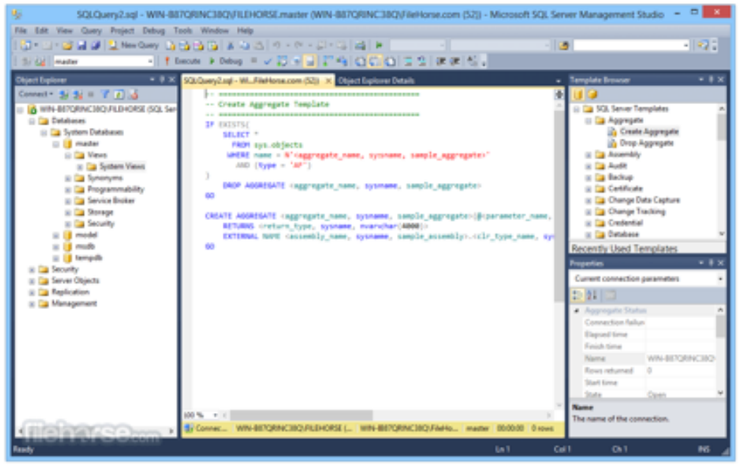
BI/data analytics

- Goal:
  - Tools for exploration
  - Looking for patterns
- Users:
  - Analysts
  - Data Scientist



Direct queries

- Goal:
  - Sophisticated tools for exploration
- Users:
  - Analysts
  - Data Scientist
  - Data Engineer



Let's practice!

INTRODUCTION TO DATA WAREHOUSING

Data warehouse architectures

INTRODUCTION TO DATA WAREHOUSING

Aaren Stubberfield
Data Scientist

datacamp

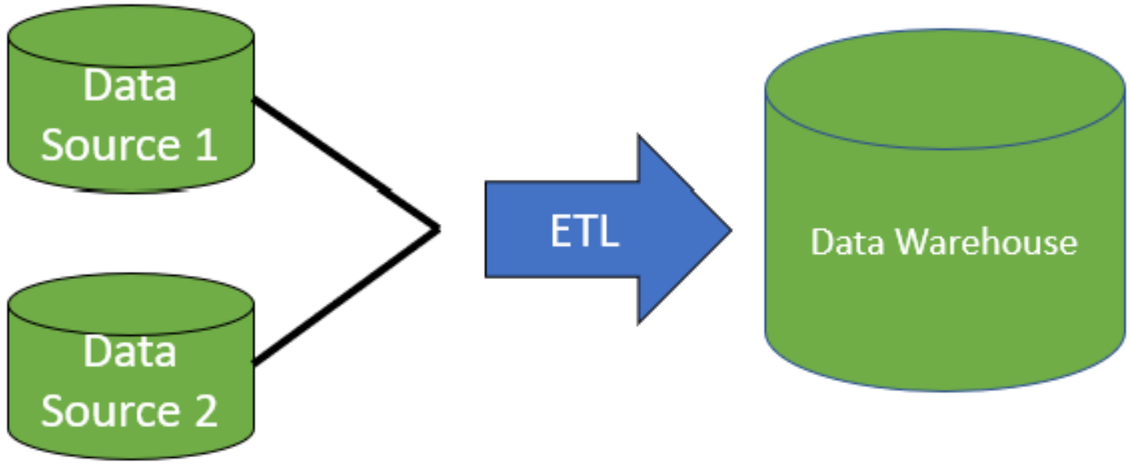
Inmon - top-down

Inmon - top-down

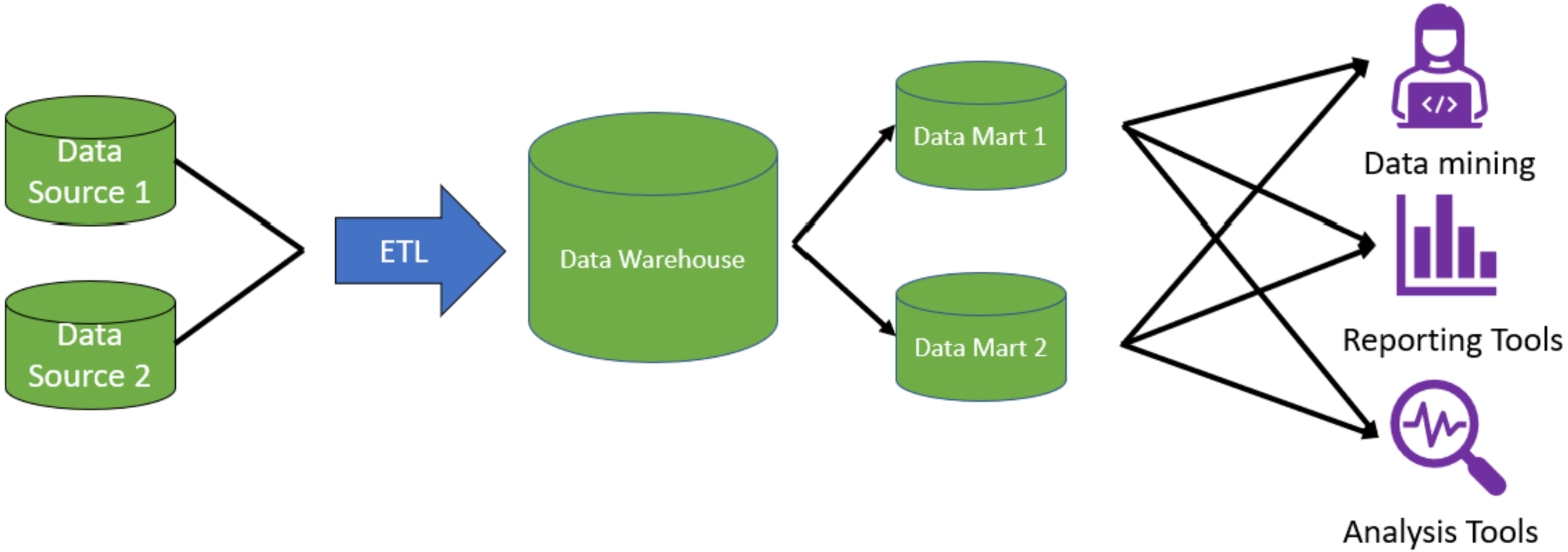
Must decide:

- On all data definitions, cleaning, and business rules
- Before any data enters warehouse

Inmon - top-down



Inmon - top-down



Pros and cons of top-down

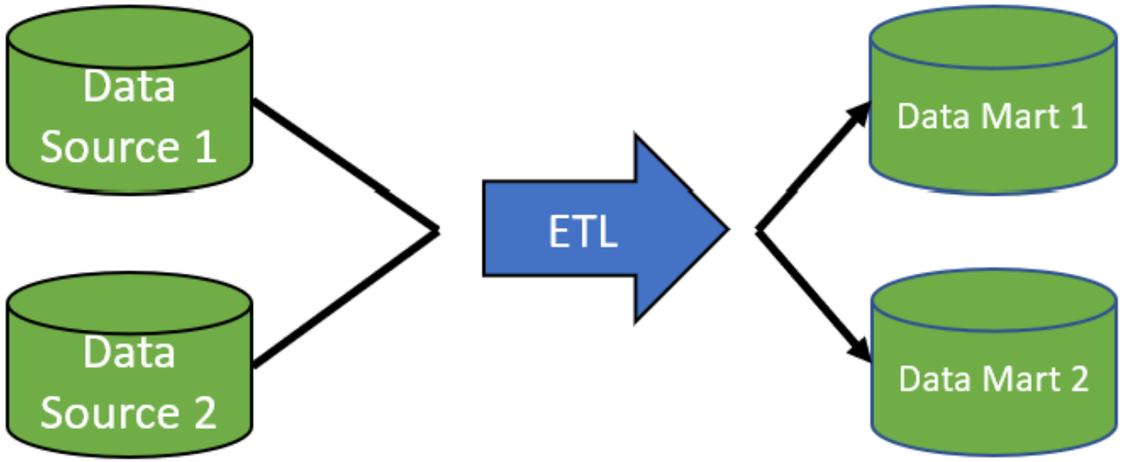
Advantages:

- Single source of truth for organization
- Normalization = less storage
- Easy to change data marts to support reporting changes

Disadvantages:

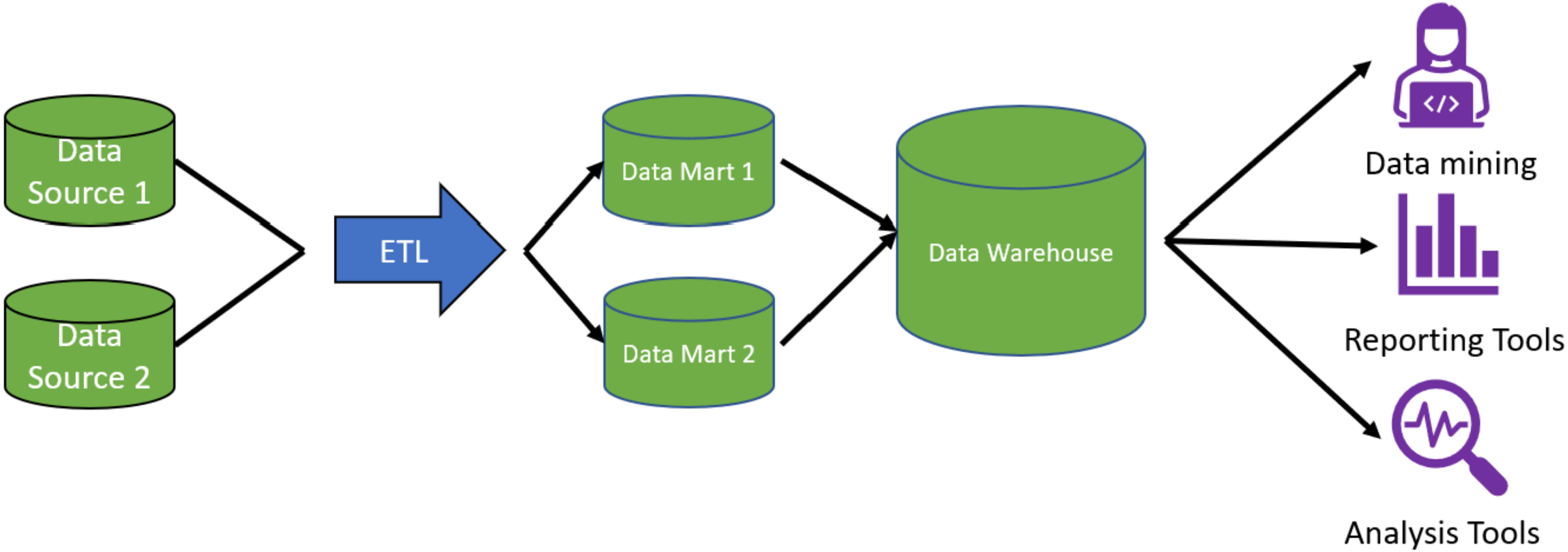
- More joins = slower response time
- Lengthy upfront work
  - Higher startup cost

Kimball - bottom-up



- Denormalizes data
- Focus on departmental data mart
- Data moves directly from ETL to data marts

Kimball - bottom-up



Pros and cons of bottom-down

Advantages:

- Upfront development speed
  - Lower startup cost
- Denormalized = user friendly

Disadvantages:

- Increased ETL processing time
- Greater possibility of duplicate data
- Ongoing development needed

Let's practice!

INTRODUCTION TO DATA WAREHOUSING

OLAP and OLTP systems

INTRODUCTION TO DATA WAREHOUSING

Aaren Stubberfield
Data Scientist

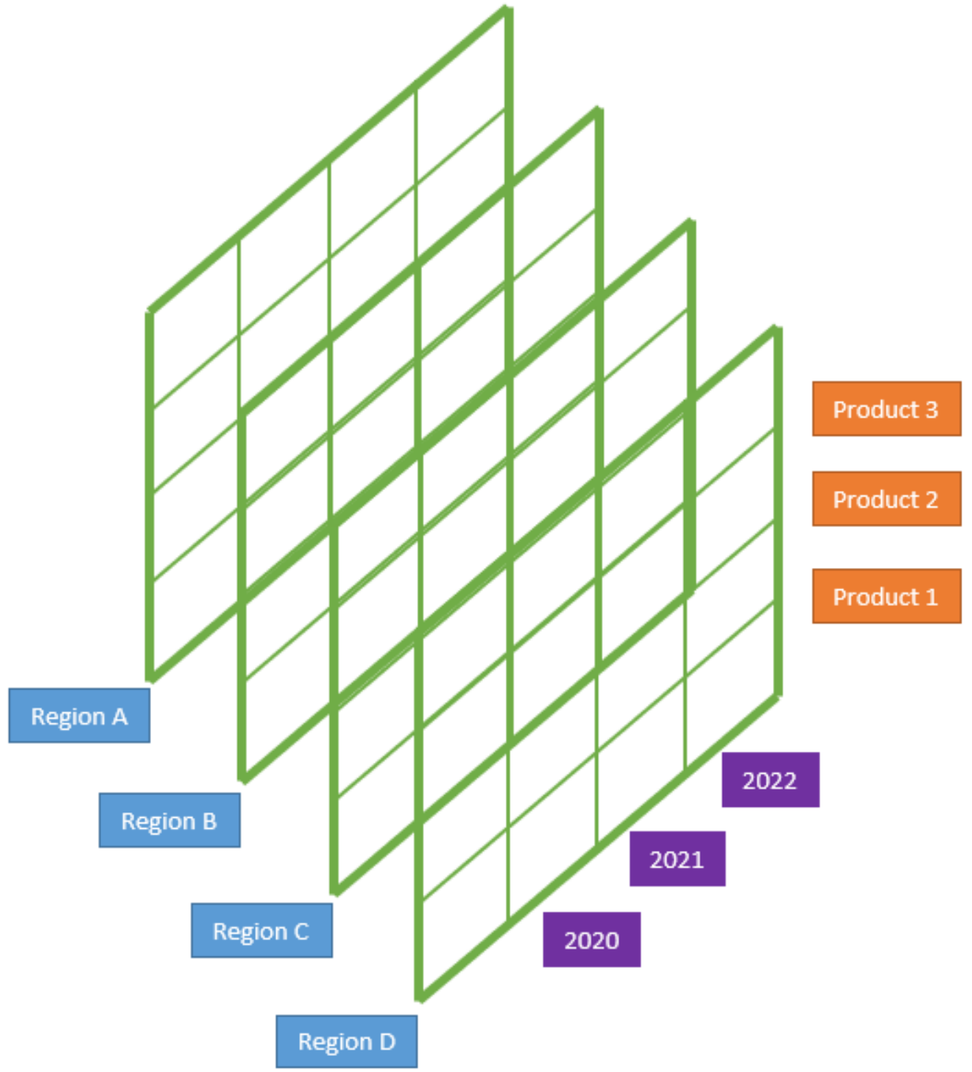
datacamp

OLAP systems

- OLAP (*online analytical processing*)
- Designed to support analysis of large amounts of data
- Example dimensional organization:
  - country, state, city
  - years, months, days
- OLAP reorganizes data into multidimensional format

OLAP cube

- OLAP cube key to OLAP system
- Faster processing vs. traditional relational databases
- Hypercubes have more than three dimensions



OLTP

- OLTP (*online transaction processing*)
- Designed to processing simple database queries
- Used in source systems to data warehouse

[1] Photo by Rodnae-Productions on Pexels

Example for a credit card company

OLTP:

- System tracks customers purchase
- Processes large amount of simply database updates to account balances

OLAP:

- Designed for analyzing purchase data
- Data organized by multiple dimensions

Summary

Differences	OLAP	OLTP
Optimization:	Complex read-only queries for analysis	Simple queries
Data Representation:	Multidimensional	Rows and columns

Let's practice!

INTRODUCTION TO DATA WAREHOUSING