What is a data warehouse?

DATA WAREHOUSING CONCEPTS

Aaren Stubberfield
Data Scientist

datacamp

What you will learn

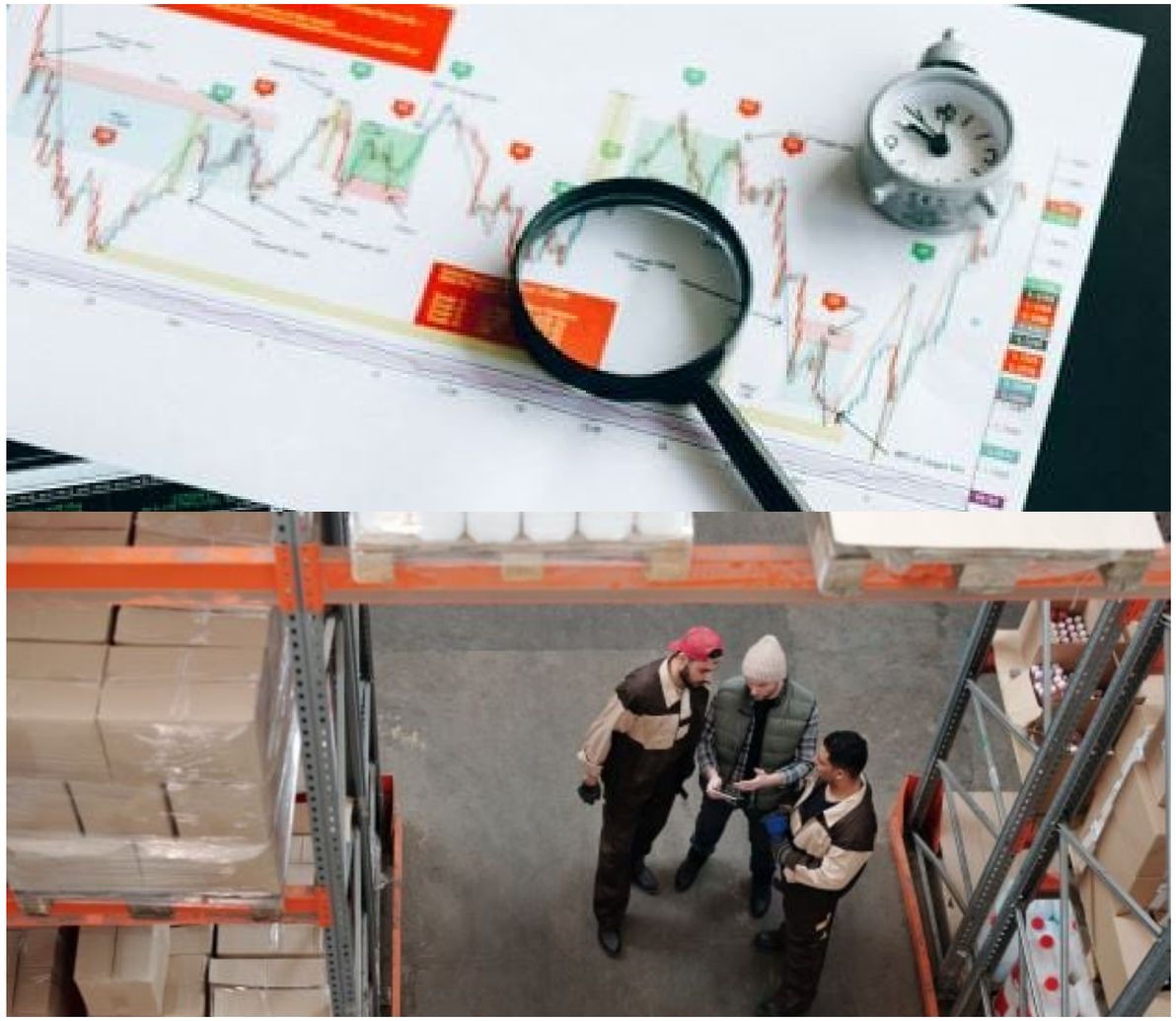
- What is a data warehouse
- Warehouse architectures and properties
- Data warehouse data modeling
- Data prep and cleaning

What is a data warehouse?

> **A computer system designed to store and analyze large amounts of data for an organization.**

What does a data warehouse do?

- Gathers data from different areas of an organization
- Integrates and stores the data
- Make it available for analysis



[1] Photos from Pexel by Nataliya Vaitkevich and Tiger Lily

Why is a data warehouse valuable?

Organizations implement data warehouses in order to:*

- Support business intelligence activity
- Enable effective organizational analysis and decision-making
- Find ways to innovate based on insights from their data

[1] Data Management Book Of Knowledge 2nd Edition

Meet Bravo!

- Hypothetical publicly traded company
  - Sells home office furniture

[1] Photo from Pexel by Pixabay

Common scenarios

- Product sales forecasting
- Governance and regulation adherence
- Insight and growth

Summary

What is a data warehouse?

- A computer system designed to store and analyze large amounts of data for an organization.

What does a data warehouse do?

- Gathers data from different areas
- Integrates and stores the data
- Make it available for analysis

Why is a data warehouse valuable?

- Support business intelligence activity
- Enable effective analysis and decision-making
- Foster data-driven innovation

Let's practice!

DATA WAREHOUSING CONCEPTS

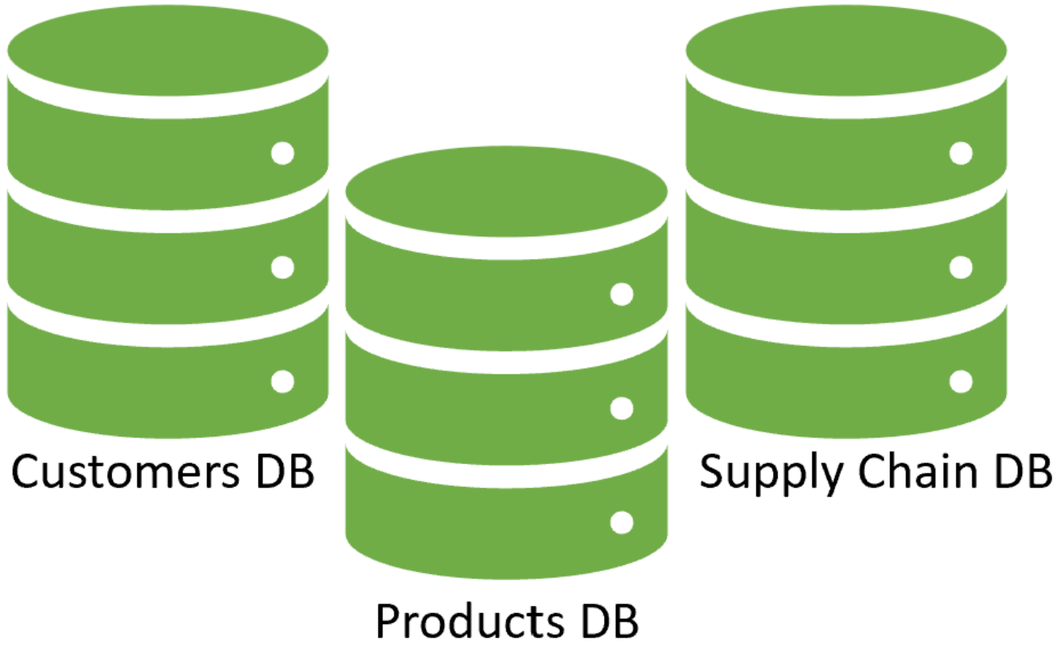
What's the difference between data warehouses and data lakes?

DATA WAREHOUSING CONCEPTS

Aaren Stubberfield
Data Scientist

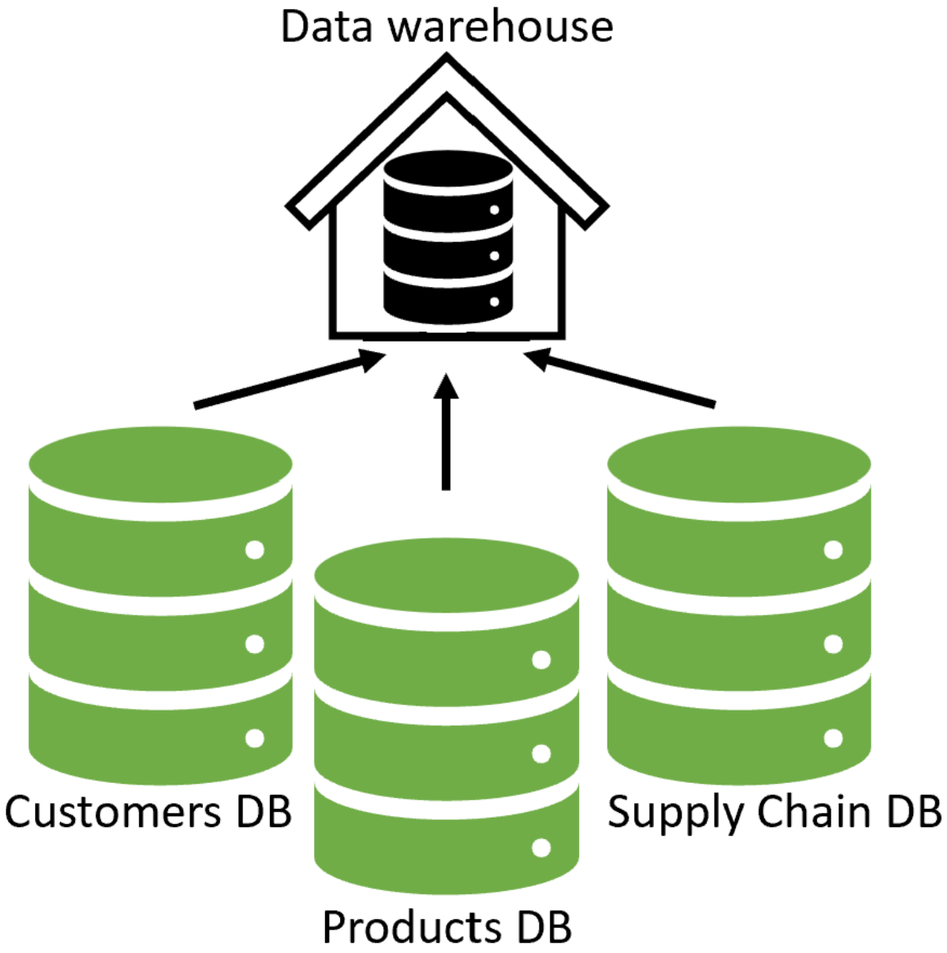
Database

- Structured data in rows and columns
- Transactional databases store transactions



Data warehouse

- Gather data, integrate, and make available for analysis
- Many input data sources
- Stores structured data
- Complex to change
  - Upstream and downstream effects must be considered
- Typically >100 GB in size

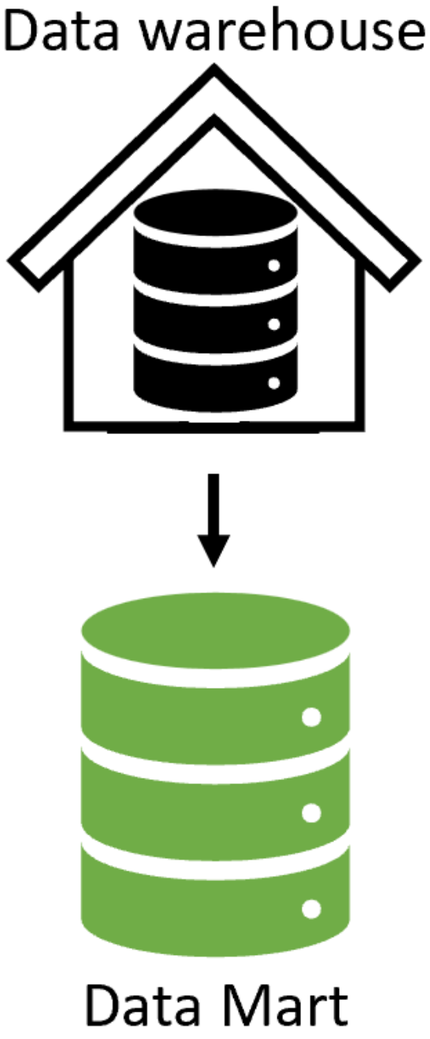


Why the data warehouse?

- How quickly the query will run on a large amount of data
- Avoid slowing down transactional database

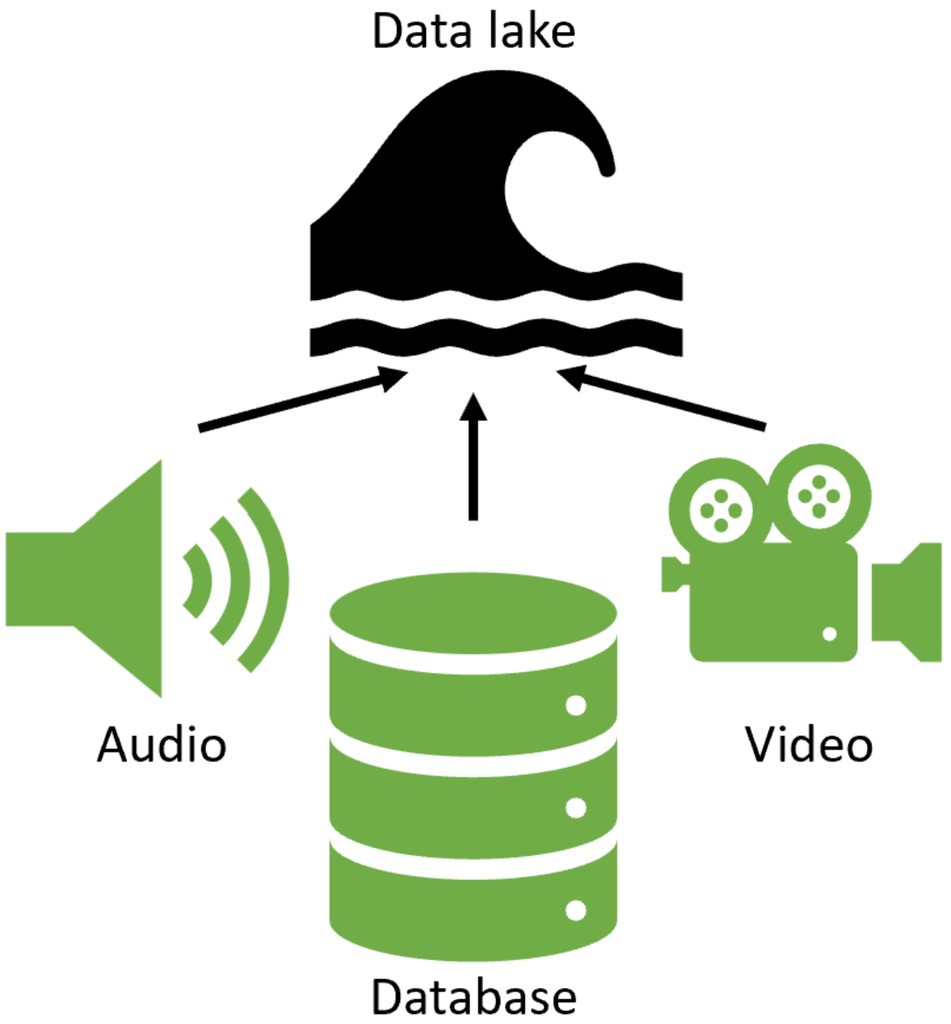
Data marts

- A relational database for analysis
- Data is focused on one subject area
- Few input data sources
- Typically <100 GB in size



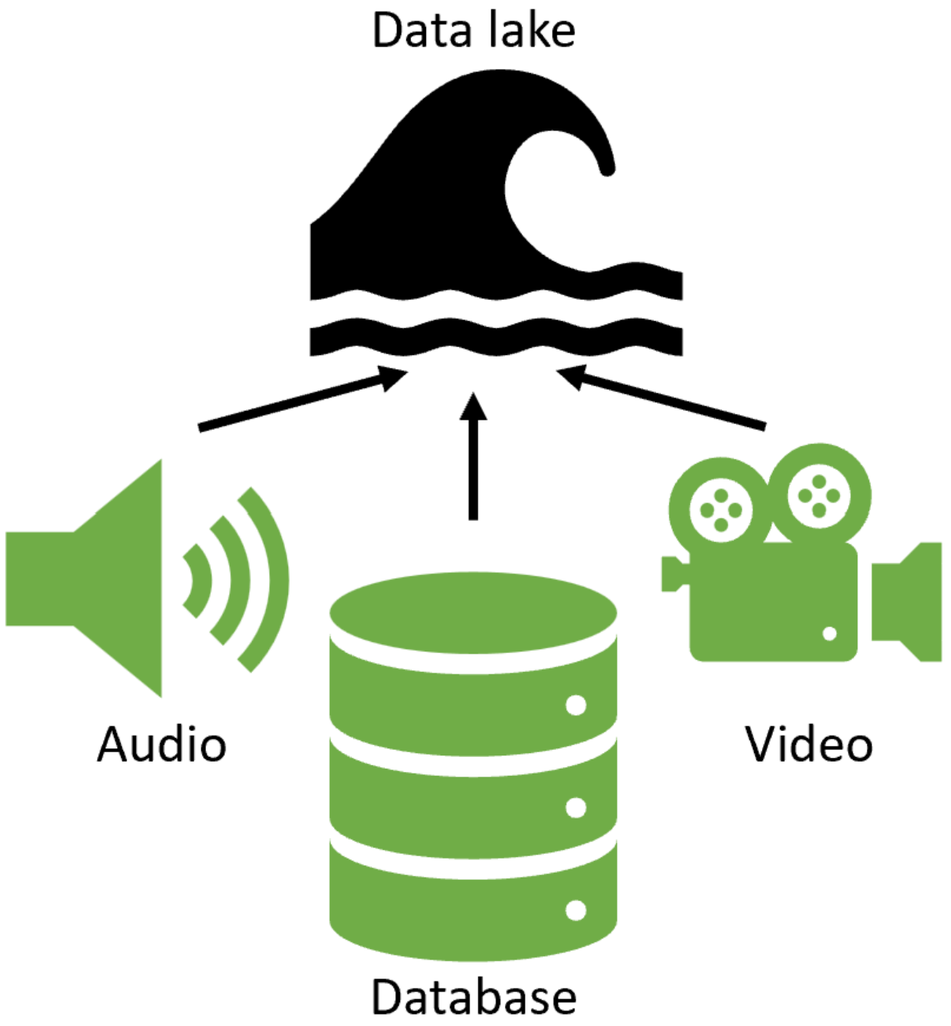
Data lake

- Entire organization store of data
  - Contains data from many departments
  - Many data input sources
  - Typically >100 GB in size
- Stores structured and unstructured data
  - Examples: video, audio, and documents



Data lake

- Less complex to make changes
  - Fewer upstream and downstream effects to consider
- Purpose to store data may not be known
  - Less organized



Summary

Feature	Data Warehouse	Data Mart	Data Lake
Data structure	Structured	Structured	Structured & Unstructured
Complexity to change	Complex	Complex	Less complex
Purpose of data	Known	Known	May not be known
Coverage of departments	Covers many	Covers only one	Covers many
Data sources	Many source systems	Few sources	Many source systems
Typical size	>100 GB	<100 GB	>100 GB

Let's practice!

DATA WAREHOUSING CONCEPTS

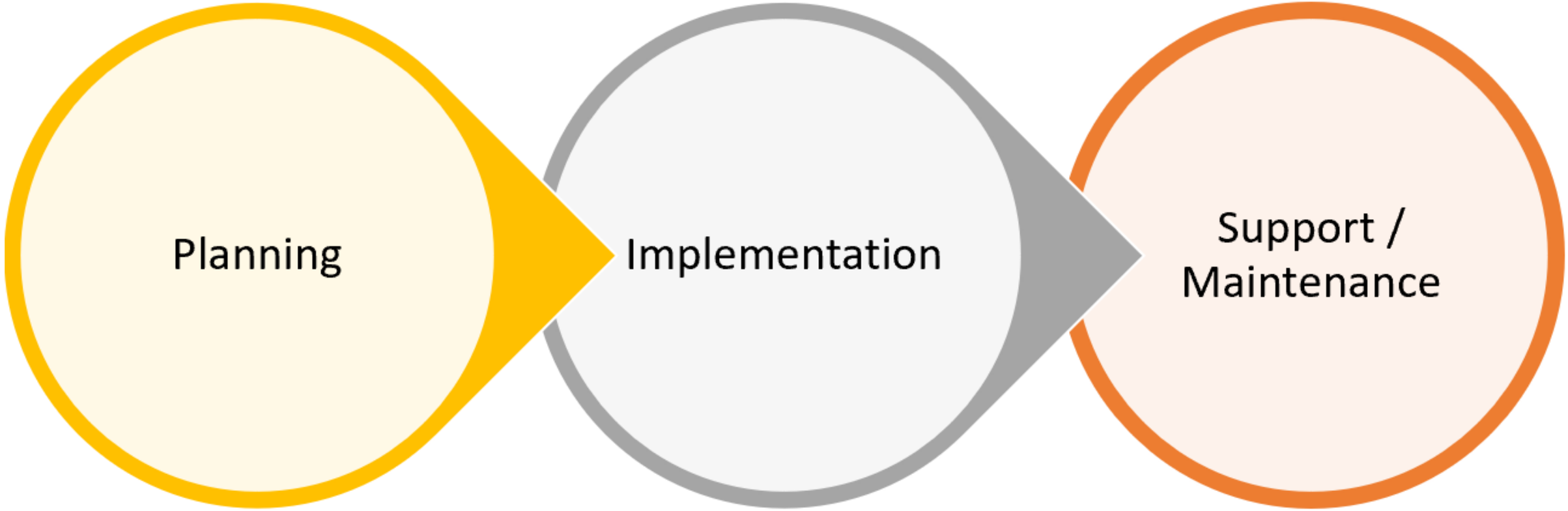
Data warehouses support organizational analysis

DATA WAREHOUSING CONCEPTS

Aaren Stubberfield
Data Scientist

datacamp

High-level life cycle



- Business Requirements
- Data Modeling

- ETL Design & Development
- BI Application Development

- Maintenance
- Test & Deploy

Planning - business requirements

1. Business Requirements:
   - Understanding the organizational needs
   - Personas:
     - Analyst & Data Scientist - collect requirements

Christina: Data Analyst

Alex: Data Scientist

Planning - data modeling

1. Data Modeling:
   - Planning and organizing on integrating data
   - Personas:
     - Data Engineer & Database Admins - design data pipeline
     - Analyst & Data Scientist - data business knowledge

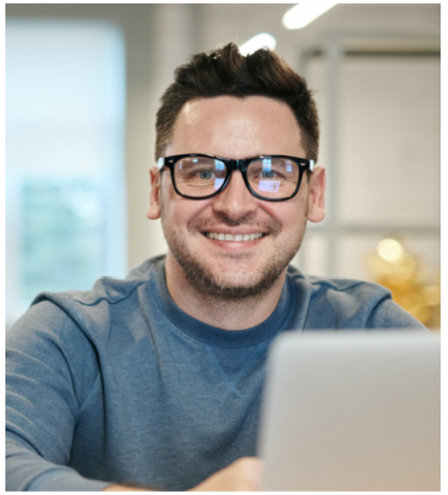
Stacy: Data Engineer Derrick: Database Admin

Christina: Data Analyst Alex: Data Scientist

Implementation - ETL Design & Development

1. ETL Design:
   - Implement data pipelines and ETL process
   - Personas:
     - Data Engineer & Database Admins - implement data pipeline

Stacy: Data Engineer



Derrick: Database Admin

Implementation - BI Application Development

1. BI Application Development:
   - Setup business intelligence (BI) tools
   - Personas:
     - Analyst & Data Scientist - consult on BI tool setup

Christina: Data Analyst

Alex: Data Scientist

Support / Maintenance - Maintenance

1. Maintenance:
   - Make any needed modifications
   - Personas:
     - Data Engineer - modify as needed

Stacy: Data Engineer

Support / Maintenance - Test & Deploy

1. Test & Deploy:
   - Testing
   - Personas:
     - Analyst & Data Scientist - consult on BI tool setup
     - Data Engineers - deploy the data warehouse

Christina: Data Analyst

Alex: Data Scientist

Stacy: Data Engineer

Persona matrix

Life cycle step	Analysts	Data Scientist	Data Engineers	Database Administrators
Business Requirements	X	X		
Data Modeling	X	X	X	X
ETL Design & Development			X	X
BI Application Development	X	X		
Maintenance			X	
Test & Deploy	X	X	X	

Let's practice!

DATA WAREHOUSING CONCEPTS