

ETL and ELT

DATA WAREHOUSING CONCEPTS



Aaren Stubberfield
Data Scientist

Understanding ETL and ELT names

ETL

1. Extract
2. Transform
3. Load

ELT

1. Extract
2. Load
3. Transform

Understanding ETL - Pros and Cons

- Data transformed during the move
- Uses separate system to process data

Pros:

- Lower data storage costs
- PII security compliance

Cons:

- Transformation errors/changes require new data pulls
- Costs of separate system to process data

Understanding ELT - Pros and Cons

- Data is loaded, then transformed
- Uses the warehouse to transform the data

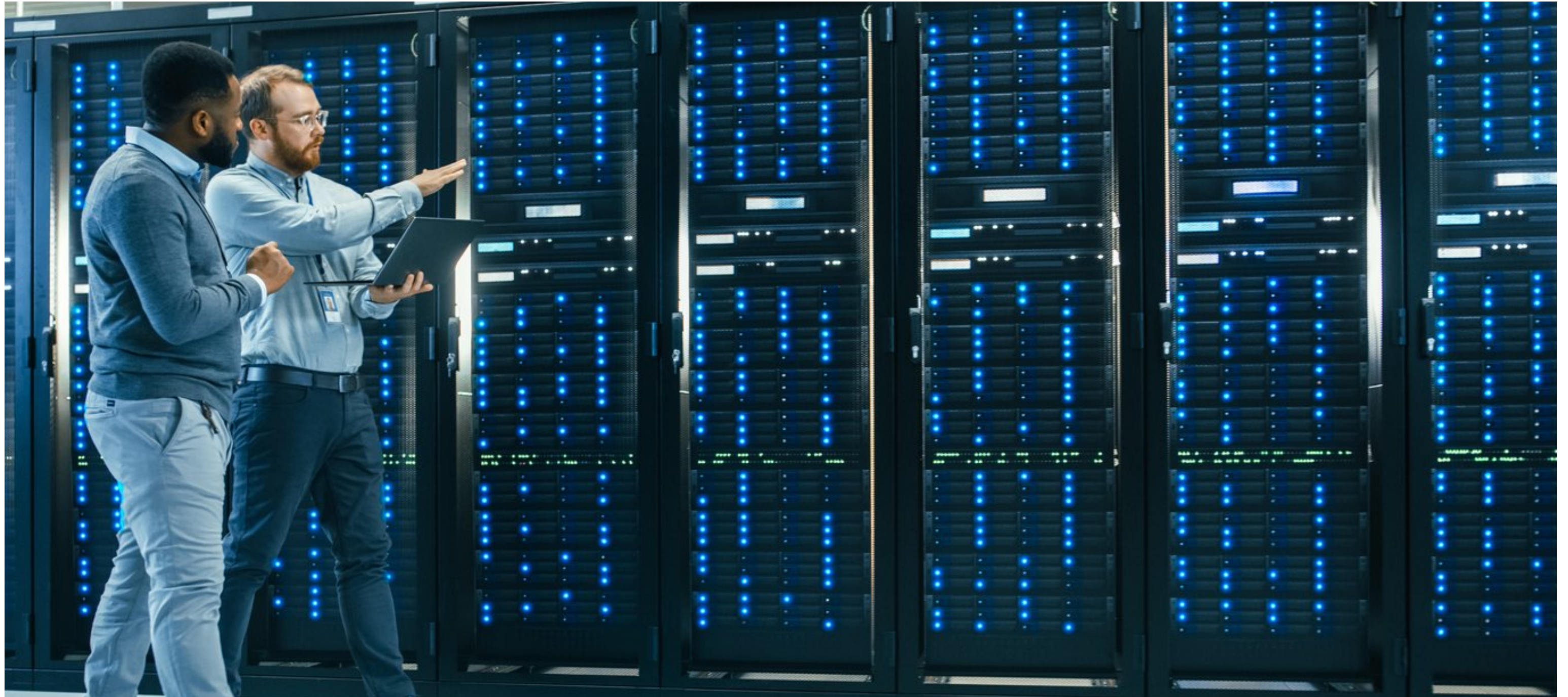
Pros:

- No separate system to process data
- Transformations can be rerun without impacting source systems
- Works well for near real-time requirements

Cons:

- Increased storage needs from raw data
- Compliance with PII security standards

The cloud and ELT



Let's practice!
DATA WAREHOUSING CONCEPTS

Data cleaning

DATA WAREHOUSING CONCEPTS



Aaren Stubberfield
Data Scientist

Video agenda

- Data format revision
- Address parsing
- Data validation
- De-duplication

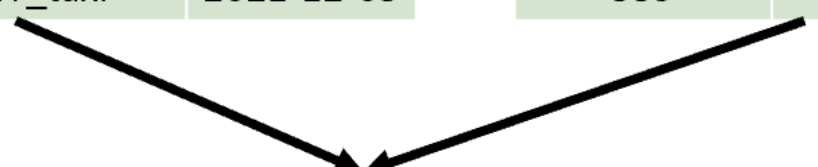
Data format cleaning

- Update values to an expected format
 - Dates
 - Names of options
 - Capitalization
- Ensures output is in a consistent format

Taxi data example

CustomerID	User_Name	Join_Date	CustomerID	Last_Ride_Date
440	CABBY13	2022-06-20	440	7/01/2022
230	taxi#1	2020-08-03	230	8/3/2020
559	NY_taxi	2021-12-05	559	1/31/2021

CustomerID	User_Name	Join_Date	Last_Ride_Date
440	cabby13	2022-06-20	2022-07-01
230	taxi#1	2020-08-03	2020-08-03
559	ny_taxi	2021-12-05	2021-01-31



Address parsing

- Dividing a street address into its components
- Can use tools to validate addresses

Address
1234 S Normal St, Cleveland, OH 44102

Address	City	State	Zip
1234 S Normal St	Cleveland	OH	44102

Data validation

- Range check
 - Is the value within the expected range?
 - Example: A person's age
- Type check
 - Is the value the proper data type?
 - Example: Storing age as string vs number

Age	
300	Not Valid
67	Valid
43	Valid

Age	dtype	
30	String	Not Valid
67	String	Not Valid
43	string	Not Valid

Duplicate row elimination

- This process gets rid of duplicate entries

DoctorID	DoctorName
275	Miach
300	Debbie
310	Berry



DoctorID	DoctorName
274	Hull
275	Miach
276	Clemency
277	Lydon
278	Chapin
279	Noel



DoctorID	DoctorName
274	Hull
275	Miach
276	Clemency
277	Lydon
278	Chapin
279	Noel
300	Debbie
310	Berry

Data governance



Let's practice!
DATA WAREHOUSING CONCEPTS

On premise and cloud data warehouses

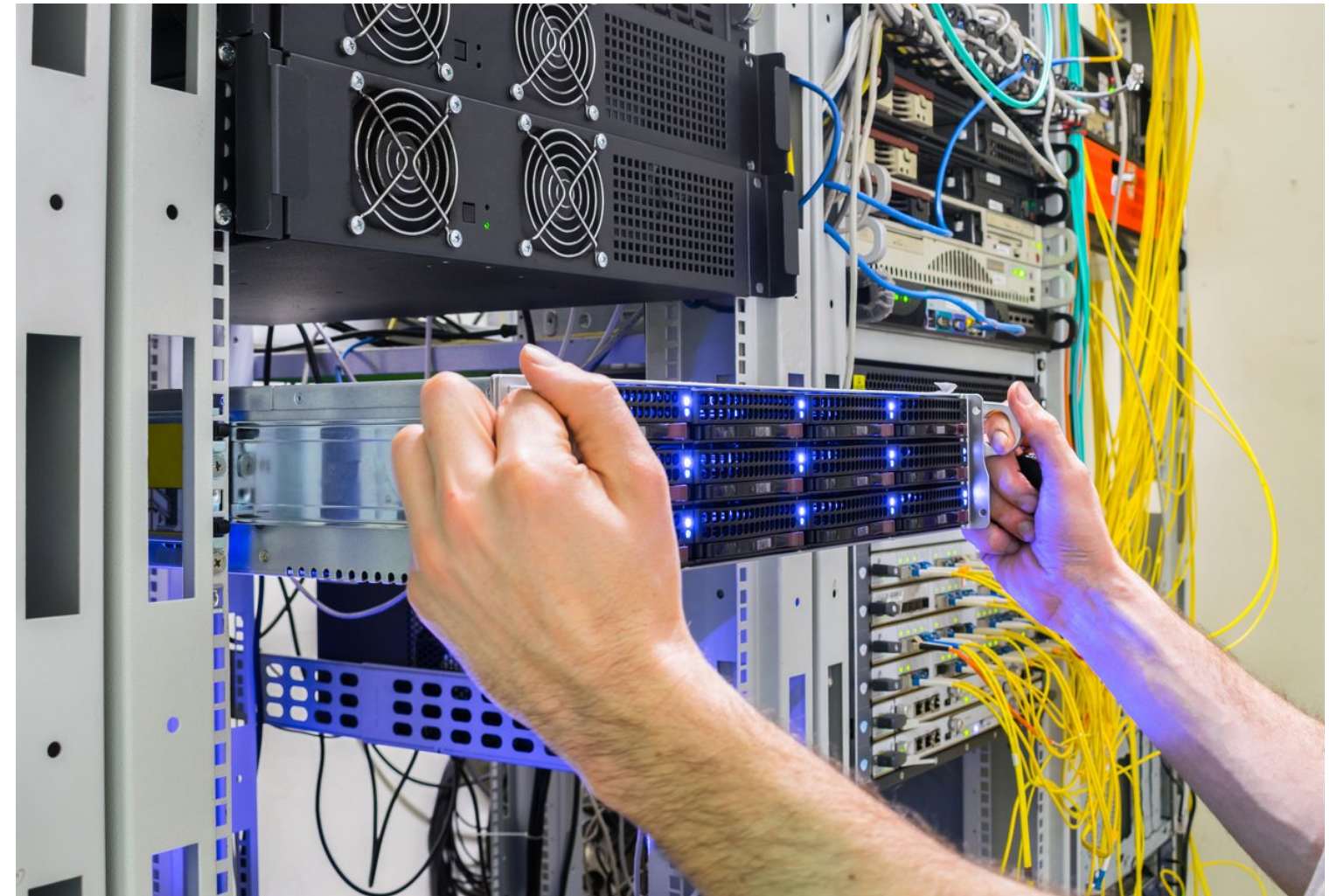
DATA WAREHOUSING CONCEPTS



Aaren Stubberfield
Data Scientist

On premise

- Purchase and install software and hardware
- On the grounds of the organization



On premise - pros and cons

Pros:

- Complete control
- Implement custom data governance
- Local network speeds
- Can optimize for workloads

Cons:

- Upfront hardware and software costs
- Personnel/staff must maintain system
- Must keep up with patches and security

In the cloud

- Rapid growth
- Forecasted continued growth



¹ Gartner Says Four Trends Are Shaping the Future of Public Cloud. Press release: Aug. 2021

In the cloud - pros and cons

Pros:

- No maintaining equipment and infrastructure
- Frees up personnel
- Can scale storage and compute resources
- No upfront investment in equipment/software

Cons:

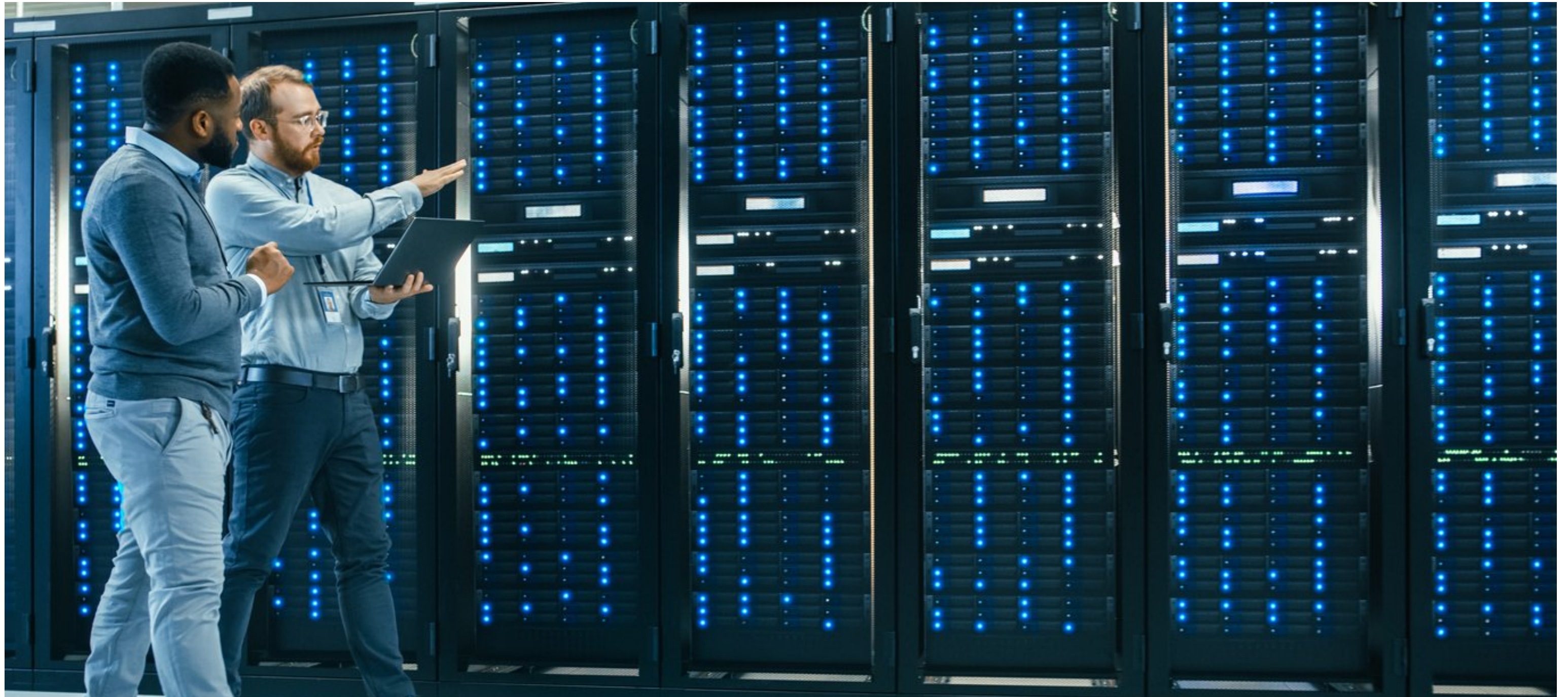
- Less control
- Cannot optimize warehouse workloads
- Possible unanticipated costs

Hybrid approach

- On premise and in the cloud data warehouse
- Reasons for hybrid approach:
 - Backup
 - Disaster recovery



Summary



Let's practice!
DATA WAREHOUSING CONCEPTS

Data warehouse design example

DATA WAREHOUSING CONCEPTS



Aaren Stubberfield
Data Scientist

Let's set the stage

- A new startup company
- Photo sharing app



¹ Photo by Alex Alvarado from unsplash.com

Top-down, or bottoms up approach?

Considerations:

- Vital to show business impact quickly
- Top-down approach has a longer startup process

Decision:

- Bottom-up approach
- Sales data mart must be the priority

Kimball - select the organizational process (step 1)

Considerations:

- What type of customers purchase large volumes of photos?

Decision:

- Develop customer purchases

Kimball - Declare the grain (Step 2)

Considerations:

- Data should be flexible to answer many questions
- Selecting the lowest grain possible

Decision:

- Tracking customer/photo purchases

Kimball - Identify the dimensions (Step 3)

Considerations:

- How do users describe the data that results from the business process?
- Customer prioritization

Decision:

- Customer location (country & state)
- Date customer joined
- Default payment method

Kimball - Identify the facts (Step 4)

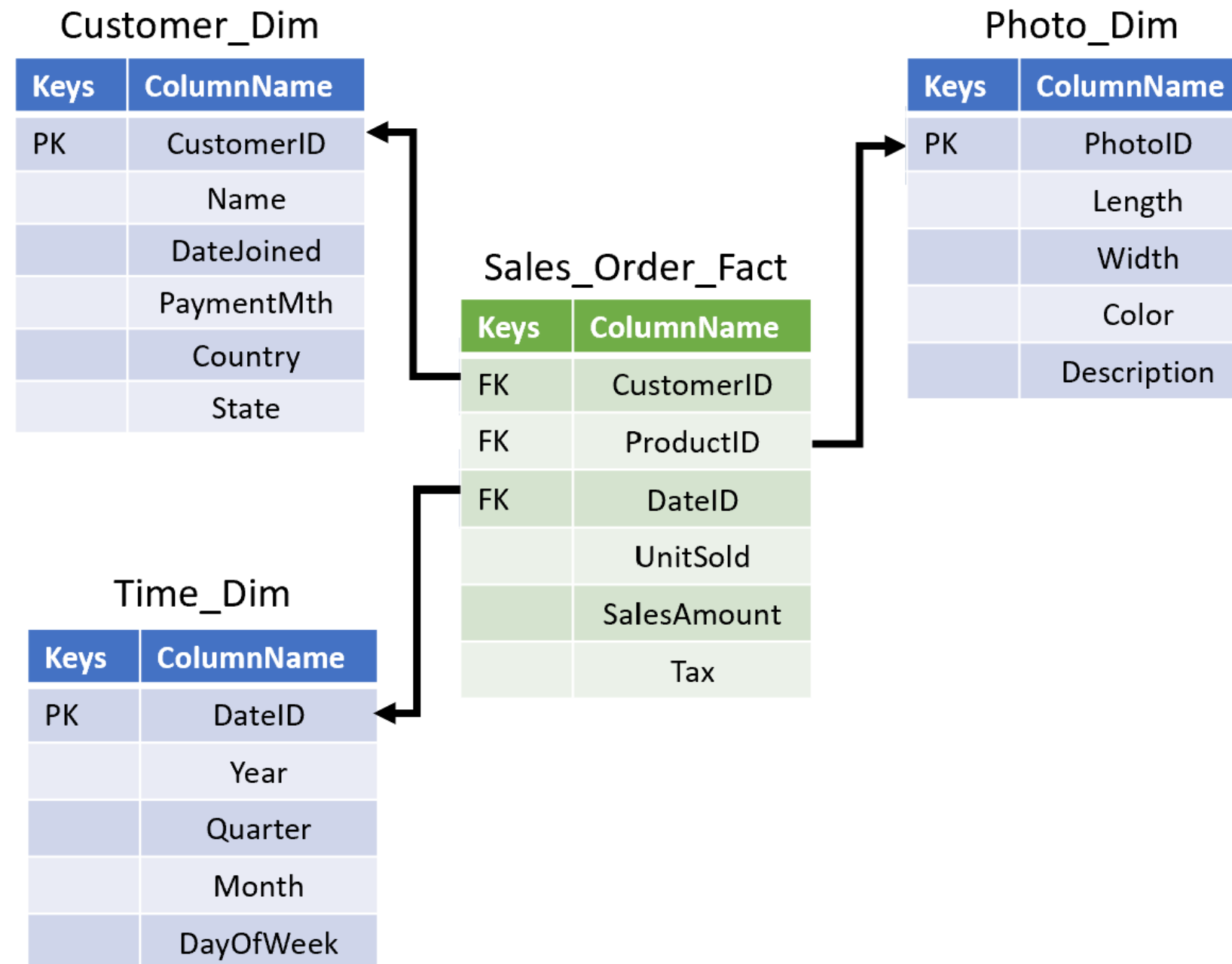
Considerations:

- What are we answering?

Decision:

- Time spent viewing photo before purchase
- Photo cost and tax
- Date of purchase

Fact and dimensions tables



On-premise or cloud implementation

Considerations:

- We do not want upfront costs for hardware / software infrastructure
- Small team - focus on high value activities

Decision:

- Cloud implementation

ETL or ELT implementation

Considerations:

- Keep all data
- Cloud implementation allows us to scale compute as needed

Decision:

- ELT implementation

Summary

- Planning is critical
- Tailor your approach based on the situation

Let's practice!
DATA WAREHOUSING CONCEPTS

Wrap-up

DATA WAREHOUSING CONCEPTS



Aaren Stubberfield
Data Scientist

What we covered - Chapter 1

Data Warehouse Basics

- What is a data warehouse?
- Data warehouse life-cycle
- Comparing to data lakes and marts



What we covered - Chapter 2

Warehouse architectures and properties

- Warehouse layers
- Inmon vs Kimball designs
- OLAP and OLTP systems



What we covered - Chapter 3

Data warehouse data modeling

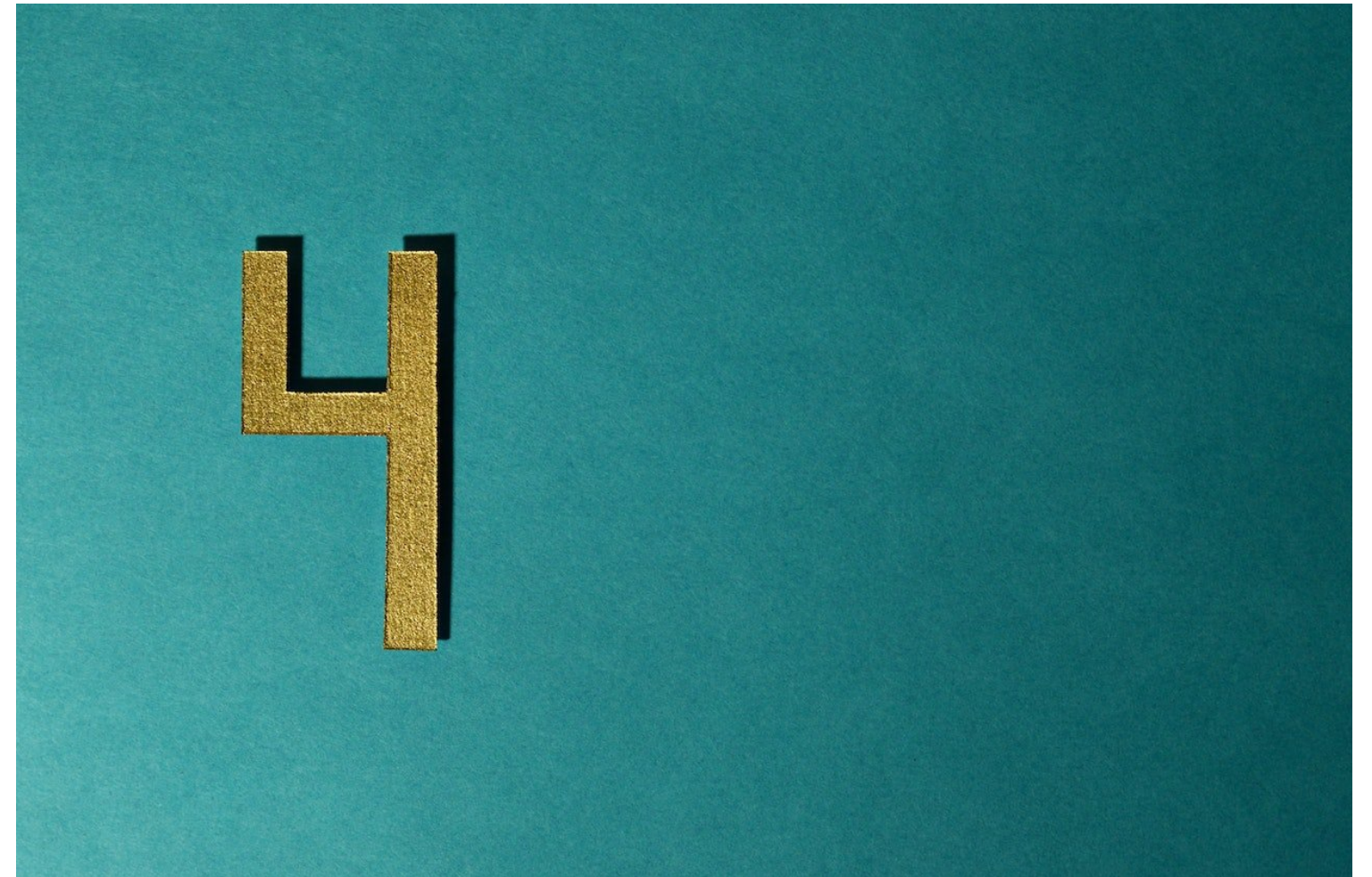
- Fact and dimension tables
- Slowly changing dimensions
- Row / column store
- Kimball's four-step process



What we covered - Chapter 4

Data cleaning and other considerations

- ETL / ELT processes
- Data cleaning
- On-premises / in the cloud implementations



Things we didn't cover

- SQL fundamentals
- Building data pipelines
- Analyzing warehouse data

Learning more

Books:

- **The Data Warehouse Toolkit, Third Edition: The Definitive Guide to Dimensional Modeling** (by Kimball)
- **Building the Data Warehouse** (by Inmon)
- **DAMA - Data Management Body of Knowledge** (by DAMA International)

DataCamp courses:

- Understanding Data Engineering
- Introduction to Power BI
- Intermediate SQL Server

Good bye and thank you!

DATA WAREHOUSING CONCEPTS