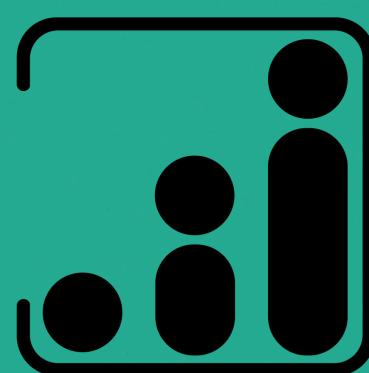


# Fact Data Modeling

Facts	Dimensions
<ul style="list-style-type: none"><li>- Represent something that happened (e.g., a transaction or event).</li><li>- Often 10-100x the volume of dimensional tables due to detailed records.</li><li>- Do not change; facts are immutable once recorded (historical data).</li><li>- Often appear in layers and are then used for aggregations and calculations (e.g., total steps, average steps per day).</li></ul>	<ul style="list-style-type: none"><li>- Represent descriptive or categorical data providing context to facts.</li><li>- Smaller in size as they store unique values or categories.</li><li>- Can change (e.g., player name, team name, or categories).</li><li>- Used for filtering, grouping, or adding descriptive context to fact data.</li></ul>



# Fact Data Modeling

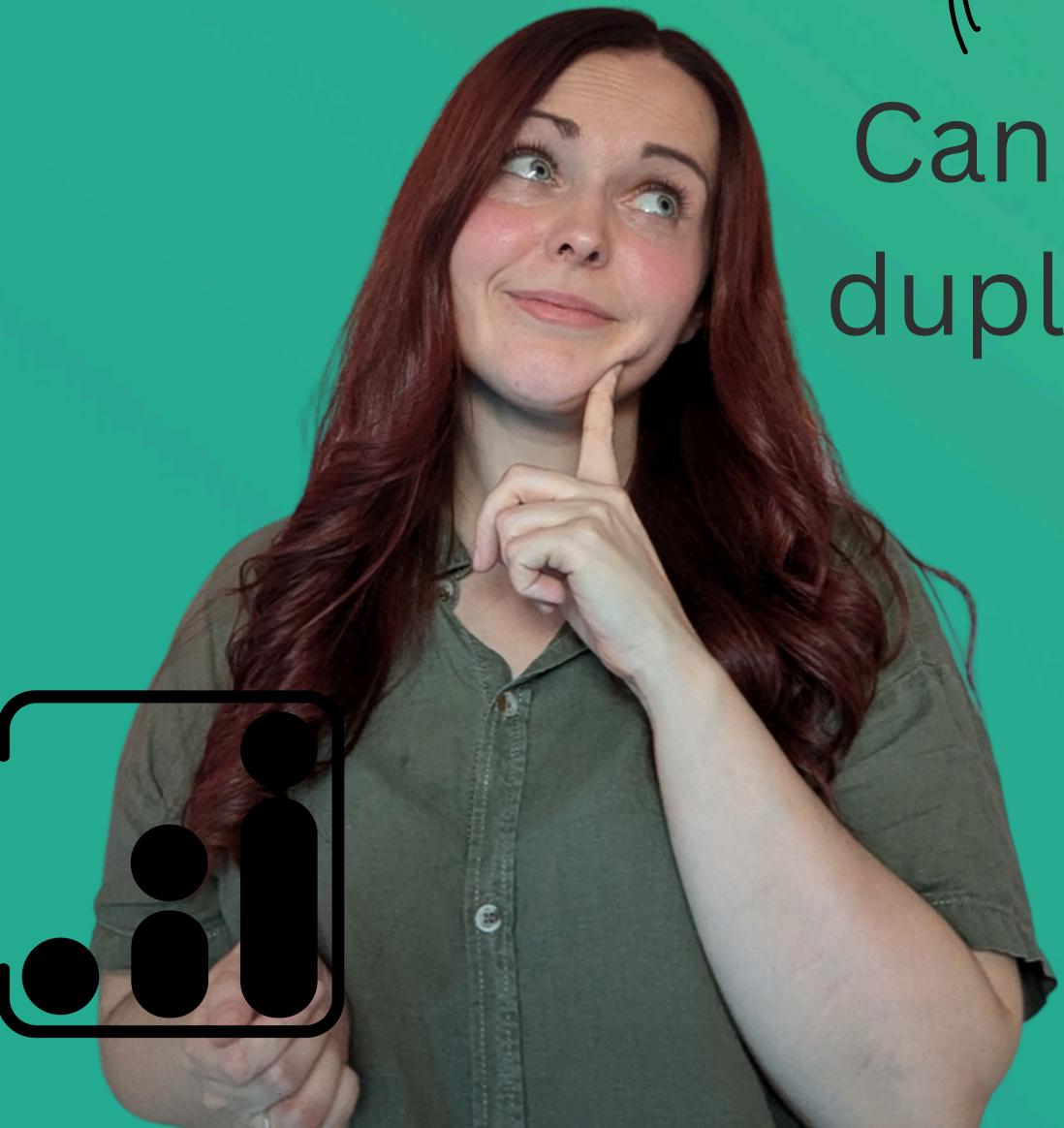
“

## Normalization

No dimensional attributes just IDs  
to get the information

Can be slower!

”



## Denormalization

Bring in attributes into the table  
rather than using dimension tables

”

Can cause  
duplicates!

”

# Key Aspects

Who?

entities or individuals associated with the data

What?

the action, event, or object the data pertains to

Where?

the location of the event, transaction, or data origin

When?

the time the event occurred

How?

the process behind data creation/transformation



# Common Techniques

---

## Sampling

Works well for metric-driven use-cases, but not where governance is critical

## Broadcast Join

Only works when one side of the join is small, otherwise performance issues occur

## Bucketing

Avoids need for expensive reshuffling by pre-partitioning tables



