DATA WAREHOUSE

Chapter 3 Dimensional Modeling

WHAT IS DIMENSIONAL MODLEING / A DIMENSIONAL MODEL

- Organizing data in a specific way, typically used in data warehouses
- Aimed at improving usability and performance for reporting and OLAP use cases
- Key components:
  - Facts
    - Measured values, such as profit
    - Can be aggregated
  - Dimensions
    - Provide context to facts
    - Examples: time periods (months, years), product categories
- Purpose
  - Facilitates turning measurements into meaningful insights
  - Allows analysis of facts by various dimensions (e.g., profit by year or category)
- Visual Representation
  - Fact tables at the center with multiple dimensions clustered around
  - Resembles a star, hence called a Star Schema

WHAT IS DIMENSIONAL MODEL

- Visual Representation
  - Fact tables at the center with multiple dimensions clustered around
  - Resembles a star, hence called a Star Schema
- Optimizations
  - Fast data retrieval
  - High performance and usability
- Why Use Dimensional Models?
  - Optimized for data warehouses
  - Essential for reporting and OLAP applications

WHAT IS DATA MODELING?
OR „SIMPLIFICATION OF TABLES“

Product ID	Product Name	Category	Price	Stock Level	Supplier
1	Laptop	Electronics	$899.99	50	TechSupplier GmbH
2	Smartphone	Electronics	$499.99	150	MobileCorp
3	Refrigerator	Home Appliances	$399.99	30	HomeAppliance AG
4	Coffee Machine	Home Appliances	$79.99	200	BrewMaster GmbH
5	T-Shirt	Clothing	$19.99	300	FashionWear Inc.
6	Jeans	Clothing	$39.99	120	DenimDesigns
7	Soccer Ball	Sporting Goods	$24.99	75	SportsGear Ltd.
8	Bicycle Helmet	Sporting Goods	$49.99	40	SafeRide Supplies
9	Novel "The Wanderer"	Books	$14.99	500	BookWorld
10	Cookbook	Books	$29.99	200	CulinaryDelights

WHAT IS DATA MODELING?
OR „SIMPLIFICATION OF TABLES“

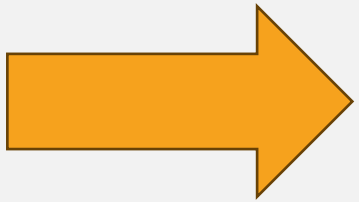
Product ID	Product Name	Category	Price	Stock Level	Supplier
1	Laptop	Electronics	$899.99	50	TechSupplier GmbH
2	Smartphone	Electronics	$499.99	150	MobileCorp
3	Refrigerator	Home Appliances	$399.99	30	HomeAppliance A
4	Coffee Machine	Home Appliances	$79.99	200	BrewMaster Gmb
5	T-Shirt	Clothing	$19.99	300	FashionWear Inc.
6	Jeans	Clothing	$39.99	120	DenimDesigns
7	Soccer Ball	Sporting Goods	$24.99	75	SportsGear Ltd.
8	Bicycle Helmet	Sporting Goods	$49.99	40	SafeRide Supplies
9	Novel "The Wanderer"	Books	$14.99	500	BookWorld
10	Cookbook	Books	$29.99	200	CulinaryDelights

Supplier ID	Supplier Name	Location	Contact Person
101	TechSupplier GmbH	Berlin, Germany	John Doe
102	MobileCorp	New York, USA	Jane Smith
103	HomeAppliance AG	Munich, Germany	Michael Johnson
104	BrewMaster GmbH	Vienna, Austria	Sarah Brown
105	FashionWear Inc.	Paris, France	Emily Wilson
106	DenimDesigns	London, UK	Adam Taylor
107	SportsGear Ltd.	Sydney, Australia	Laura Lee
108	SafeRide Supplies	Toronto, Canada	Peter Jackson
109	BookWorld	Tokyo, Japan	Maria Suzuki
110	CulinaryDelights	Rome, Italy	Marco Rossi

Product ID	Product Name	Category	Price	Stock Level	Supplier ID
001	Laptop	Electronics	$899.99	50	101
002	Smartphone	Electronics	$499.99	150	102
003	Refrigerator	Home Appliances	$399.99	30	103
004	Coffee Machine	Home Appliances	$79.99	200	104
005	T-Shirt	Clothing	$19.99	300	105
006	Jeans	Clothing	$39.99	120	106
007	Soccer Ball	Sporting Goods	$24.99	75	107
008	Bicycle Helmet	Sporting Goods	$49.99	40	108
009	Novel "The Wanderer"	Books	$14.99	500	109
010	Cookbook	Books	$29.99	200	110

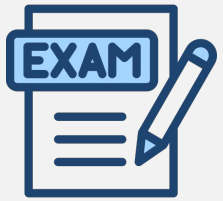
WHAT IS DATA MODELING?
OR „SIMPLIFICATION OF TABLES“

Product ID	Product Name	Category	Price	Stock Level	Supplier
1	Laptop	Electronics	$899.99	50	TechSupplier GmbH
2	Smartphone	Electronics	$499.99	150	MobileCorp
3	Refrigerator	Home Appliances	$399.99	30	HomeAppliance AG
4	Coffee Machine	Home Appliances	$79.99	200	BrewMaster GmbH
5	T-Shirt	Clothing	$19.99	300	FashionWear Inc.
6	Jeans	Clothing	$39.99	120	DenimDesigns
7	Soccer Ball	Sporting Goods	$24.99	75	SportsGear Ltd.
8	Bicycle Helmet	Sporting Goods	$49.99	40	SafeRide Supplies
9	Novel "The Wanderer"	Books	$14.99	500	BookWorld
10	Cookbook	Books	$29.99	200	CulinaryDelights



Product ID	Product Name	Category	Price
001	Laptop	Electronics	$899.99
002	Smartphone	Electronics	$499.99
003	Refrigerator	Home Appliances	$399.99
004	Coffee Machine	Home Appliances	$79.99
005	T-Shirt	Clothing	$19.99
006	Jeans	Clothing	$39.99
007	Soccer Ball	Sporting Goods	$24.99
008	Bicycle Helmet	Sporting Goods	$49.99
009	Novel "The Wanderer"	Books	$14.99
010	Cookbook	Books	$29.99

Product ID	Stock Level	Supplier ID
001	50	101
002	150	102
003	30	103
004	200	104
005	300	105
006	120	106
007	75	107
008	40	108
009	500	109
010	200	110

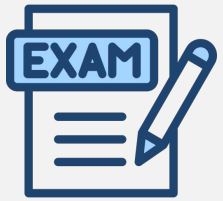


WHAT IS DATA MODELING? OR „SIMPLIFICATION OF TABLES“

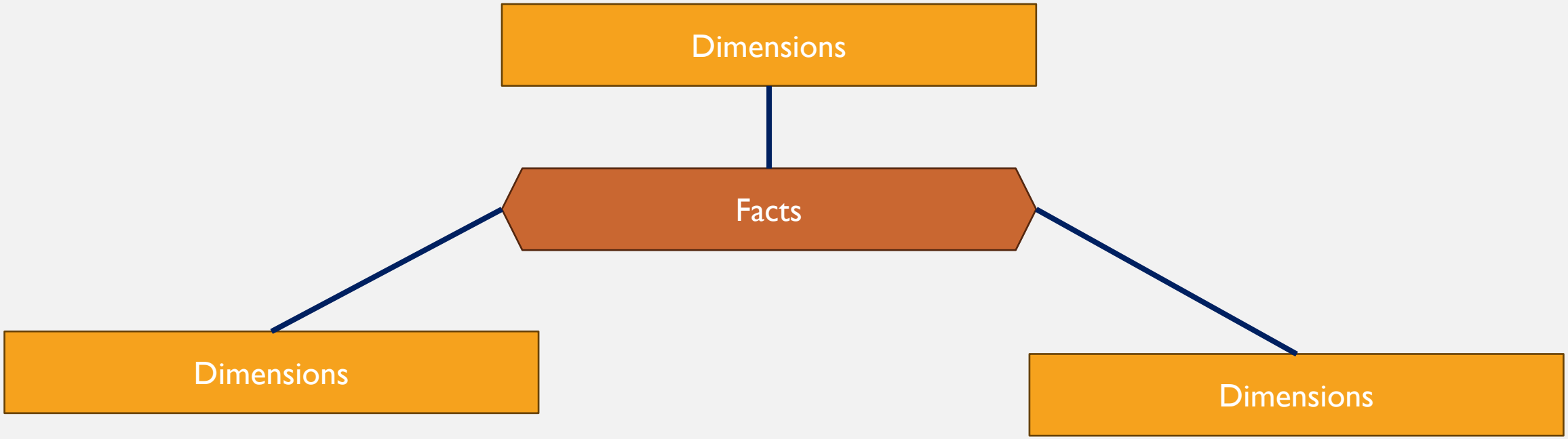
- Benefits of Dimensional Modeling
  - Allows slicing and dicing of profit data by various dimensions (e.g., month, weekday)
  - High performance and usability are key advantages
- Goal of Dimensional Modeling
  - Aimed at achieving fast data retrieval
  - Performance and usability result from the structured organization of data into dimensions and fact tables
- Preferred Technique for Data Warehousing
  - Dimensional modeling is favored for data warehousing due to its suitability for OLAP use cases and reporting
  - High performance and usability are crucial in data warehousing environments

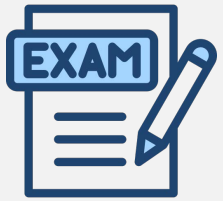
FACT & FACT TABLE

- Understanding Facts and Fact Tables
  - Differentiating between facts and dimensions in the data warehouse
  - Clarifying the distinction between facts and dimensions for better understanding
- Structure in a Star Schema
  - Fact table positioned in the center with dimensions clustered around it
  - Example: Sales table as the fact table with dimension tables revolving around it
- Role of Fact Tables
  - Central role in the data warehouse, containing key measurements (e.g., sales, profit)
  - Foundation for data analysis and aggregation
- Characteristics of Facts
  - Typically additive, allowing for meaningful aggregation
  - Numerical values that can be aggregated (e.g., total units sold)
  - Often event or transaction-based, representing occurrences in the business
- Types of Facts
  - Event-based or transactional, such as sales transactions
  - May include date or time columns to track events over time

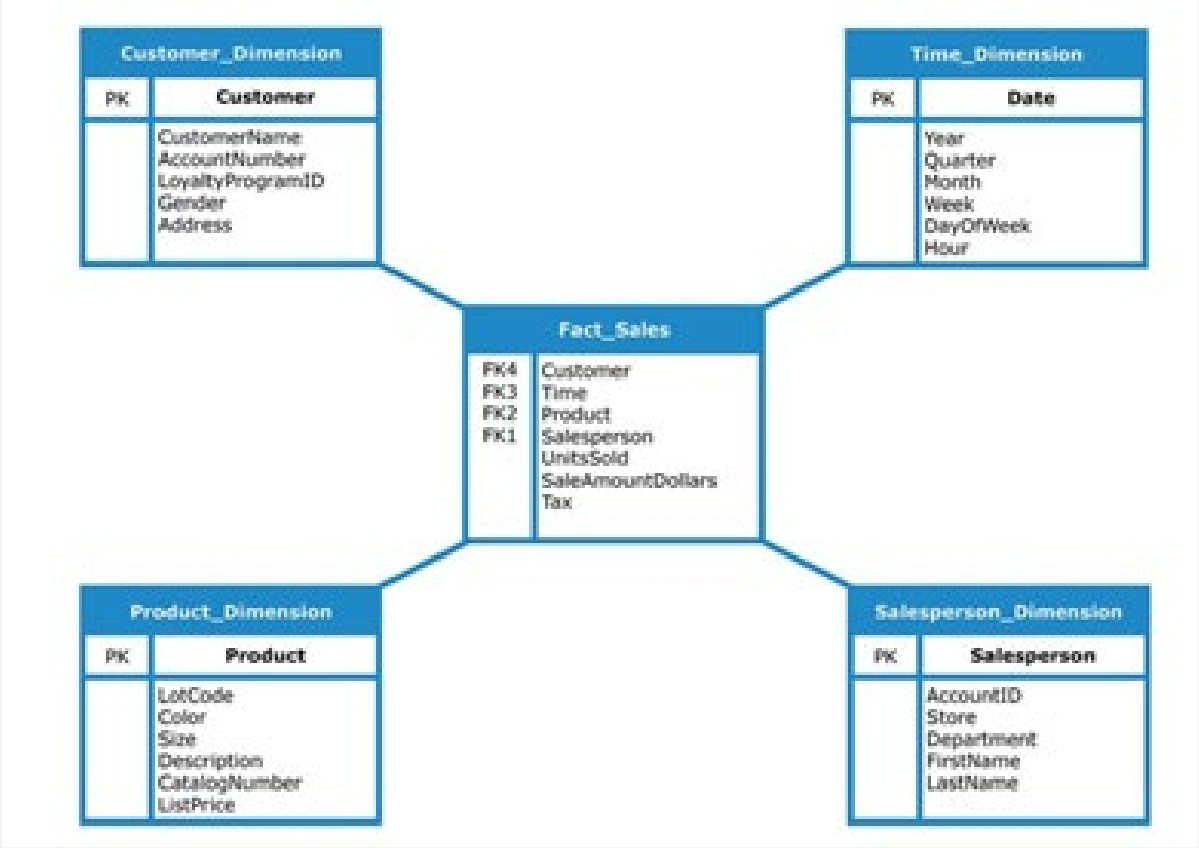


FACT & FACT TABLE
STAR SCHEMA





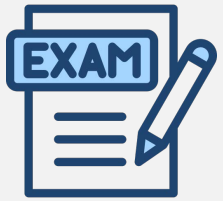
DATA AND DATA TABLE FACTS



- Aggregatable
- Measuarable vs. Descriptive
- Event- or transactional data
- Date/time in a fact table

DATA AND DATA TABLE FACTS IN A STAR SCHEMA

- Role in Star Schema
  - Clustered around the facts
  - Categorize facts to provide meaningful context
- Purpose
  - Supportive and descriptive, not measurable
  - Examples: product name, product category
  - Help in analyzing, filtering, grouping, and labeling data
  - Enable slicing and dicing of data
- Usage in Reporting
  - Used for filters, grouping in charts (e.g., bar charts)



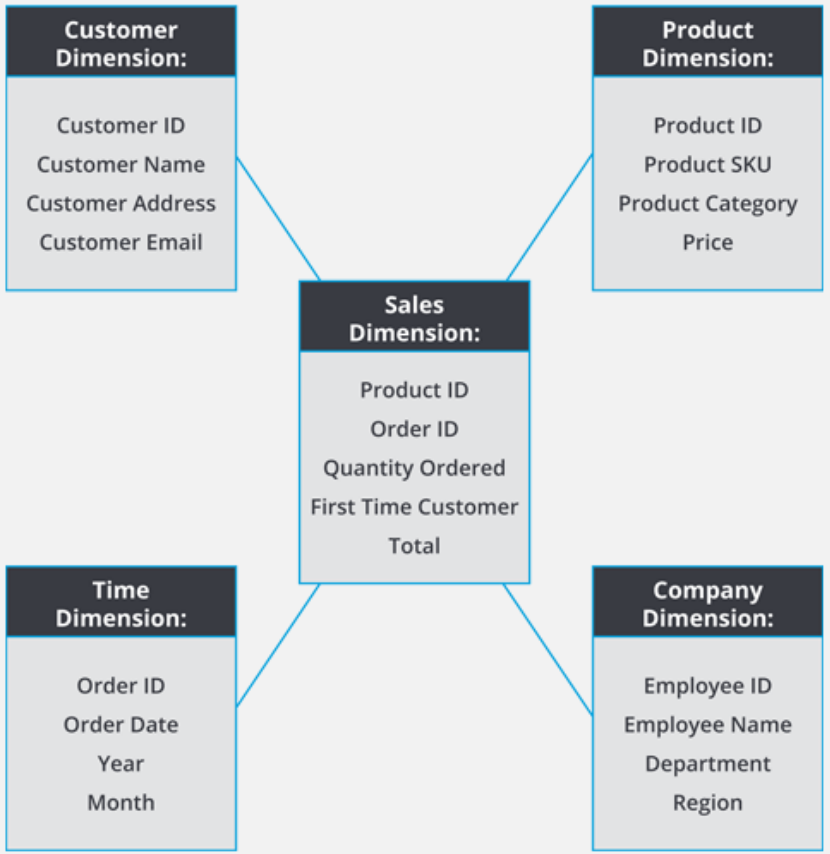
DATA AND DATA TABLE CHARACTERISTICS OF DIMENSIONS

- Non-aggregatable
  - Cannot be summed up meaningfully
  - Example: Years (2019 + 2020 doesn't provide additional insight)
- Descriptive Nature
  - Provide context rather than measurements
- Static Data
  - More static than facts, with fewer changes
  - Examples: product name, product category
  - Changes in dimensions are less frequent and data is generally stable
- Contrast with Facts
  - Facts are aggregatable and measurable
  - Facts represent events or transactions (e.g., sales)
  - Facts contain dynamic data with frequent changes

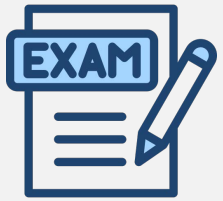
DATA AND DATA TABLE CHARACTERISTICS OF DIMENSIONS

- Structure
  - Contain dimensions in dimension tables
  - Have a primary key to identify each row
  - May include a foreign key for certain scenarios (e.g., snowflake dimensions)
- Use Cases
  - People
    - Employees
    - Customers
    - Managers
  - Products
    - Product categories
  - Places
    - Regions
    - Cities
    - Addresses
  - Time/Date Related
- Example: Customer Dimension
  - Contains customer ID as the primary key
  - Includes various attributes for customers
- Changing Dimensions
  - Sometimes dimensions change, leading to slowly changing dimensions

DATA AND DATA TABLE CHARACTERISTICS OF DIMENSIONS



- Non-Aggregatable
- Measureable vs. Descriptive
- (More) static



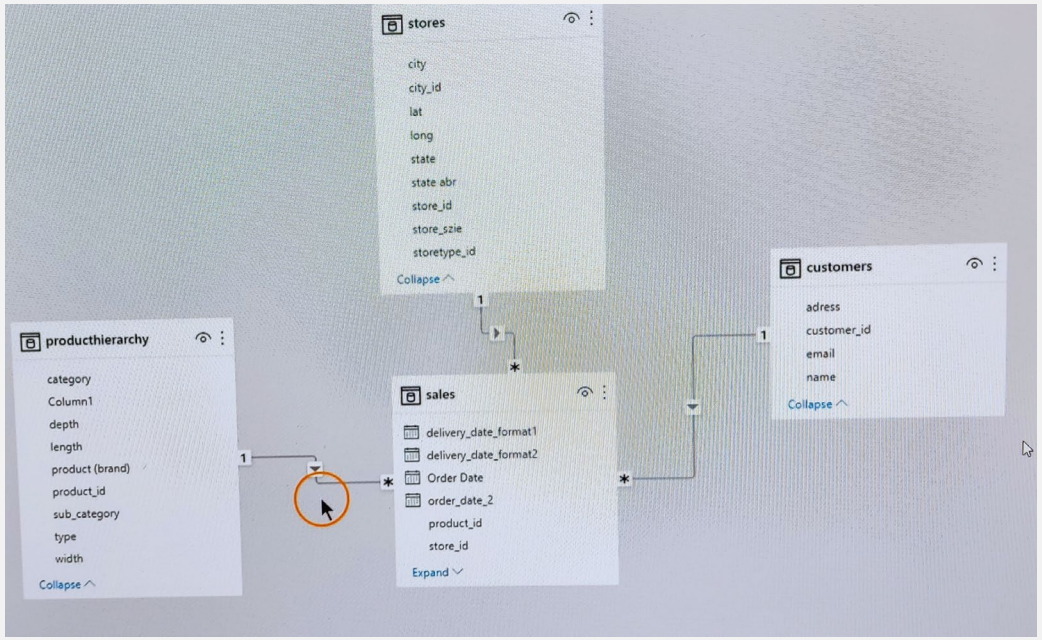
STAR SCHEMA

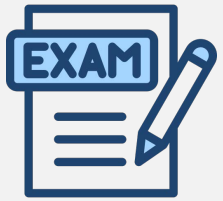
- Introduction to the Star Schema
  - Most important schema in data warehousing, especially in data marts
  - Data arranged and structured into facts and dimensions
- Example Structure
  - Fact Table (e.g., Sales Table)
    - Contains important facts
    - Uses foreign keys to create relationships with dimension tables
  - Dimension Table (e.g., Product Table)
    - Contains attributes related to facts
    - Primary key establishes connections with fact table
- Relationships in Star Schema
  - One-to-many relationship between fact and dimension tables
    - One Side (Dimension Table)
      - Unique values in the primary key column (e.g., Product ID)
    - Many Side (Fact Table)
      - Multiple occurrences of values in the foreign key column
- Hierarchy and Data Redundancy
  - Only one level of hierarchy in star schema
  - Potential for data redundancy
    - Example: Category column repeated in multiple rows

STAR SCHEMA NORMALIZATION VS. DENORMALIZATION

- Normalization
  - Reduces data redundancy
  - Lowers storage costs
  - Benefits write operations, maintenance, and updates
  - Involves more tables and complex queries
  - Less suitable for read operations and visualization use cases
- Denormalization
  - Accepts some data redundancy for better read performance
  - Simplifies queries and enhances usability
- Star Schema Characteristics
  - Single level of hierarchy with direct connections between fact and dimension tables
  - Better read performance and usability compared to fully normalized schemas
- Future Topics
  - Introduction to Snowflake Schema as an alternative
  - Discussing normalization and its impact on performance and usability

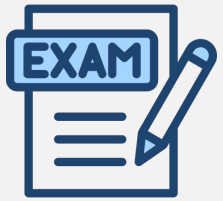
STAR SCHEMA
POWER BI EXAMPLE WITH 1 FACT TABLE



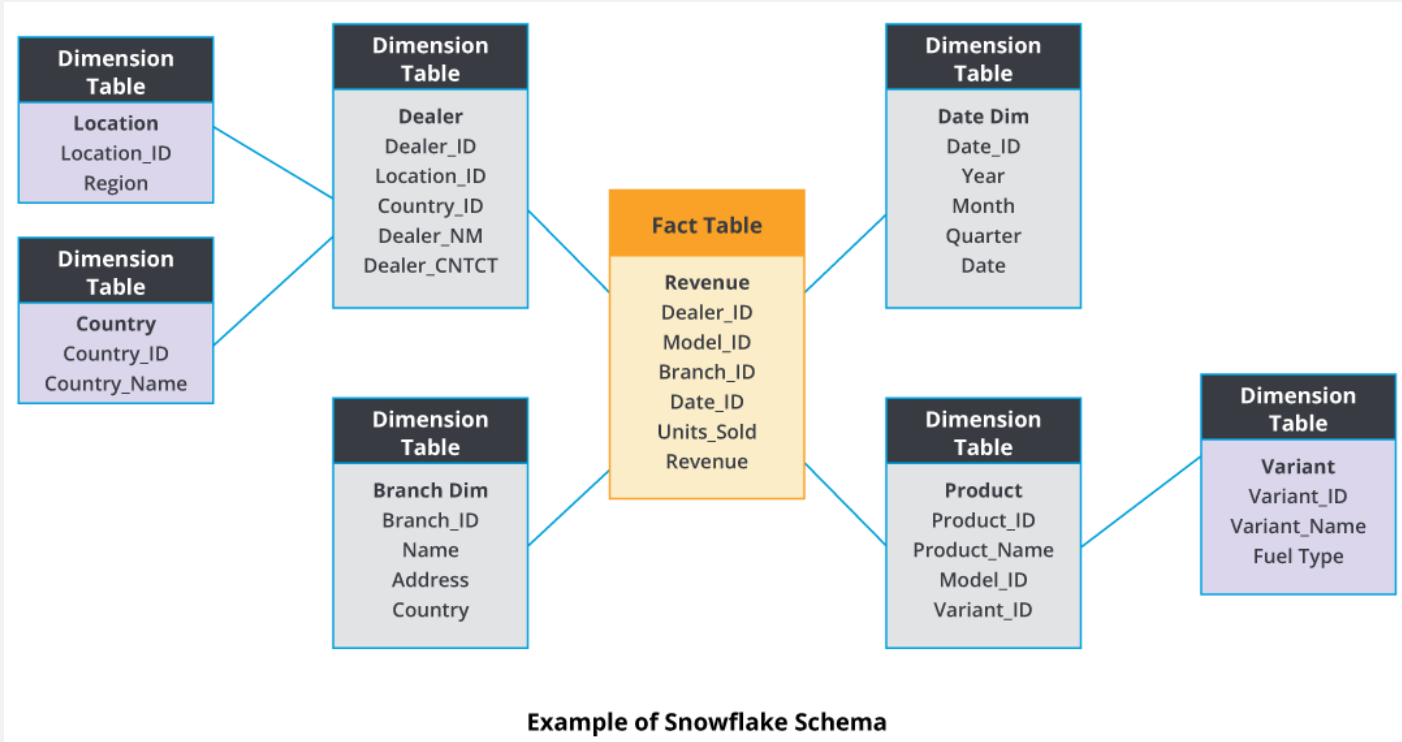


STAR SCHEMA

- Most Common Schema in Data Marts
  - Combines usability (user friendliness) and high query performance
  - Simplest form compared to more complex schemas like the Snowflake Schema
- Optimization and Ideal Use
  - Best suited for specific needs with a known set of queries
  - Common queries: visualizing profit by year, profit by categories
  - Not designed for super complex queries
- Benefits
  - High performance and usability for simple and defined queries
  - Widely used due to these advantages
- Comparison with Snowflake Schema
  - Snowflake Schema is more complex
  - Details on Snowflake Schema will be covered in the next lecture

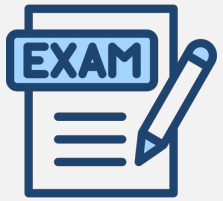


SNOWFLAKE SCHEMA
INTRODUCTION TO SNOWFLAKE SCHEMA



Example of Snowflake Schema

- Relationship to Star Schema
  - Star schema is a special case of the Snowflake schema
  - Snowflake schema allows for multiple levels in the hierarchy
  - Star schema is a Snowflake schema with only one level in the hierarchy
- Prevalence
  - Star schema is much more common in practice

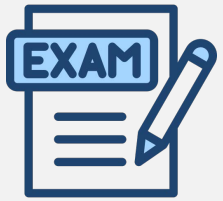


SNOWFLAKE SCHEMA

sales_id	product_id	customer_id	units	price
1	101	201	5	50.00
2	102	202	3	30.00
3	103	203	2	20.00

product_id	name	category_id	sub_category
101	ProductA	100	SubCategory1
102	ProductB	101	SubCategory2
103	ProductC	102	SubCategory3

category_id	category
100	Mobile Phones
101	Laptops
102	Tablets



SNOWFLAKE SCHEMA

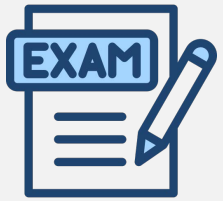
- Reduction of Data Redundancy
  - Redundancies in the Star schema are reduced by keeping IDs instead of written-out text
  - Less disk space is used, reducing storage requirements
- Hierarchical Structure
  - Related information is stored once, creating a second level of hierarchy
  - Resembles the shape of a snowflake, hence the name
- Normalization
  - Snowflake schema is more normalized compared to the Star schema
- Advantages and Use Cases
  - Storage Efficiency
    - More efficient in terms of storage due to reduced redundancy
  - Normalization Benefits
    - Easier to maintain and update data

SNOWFLAKE SCHEMA PERFORMANCE AND USABILITY CONSIDERATIONS

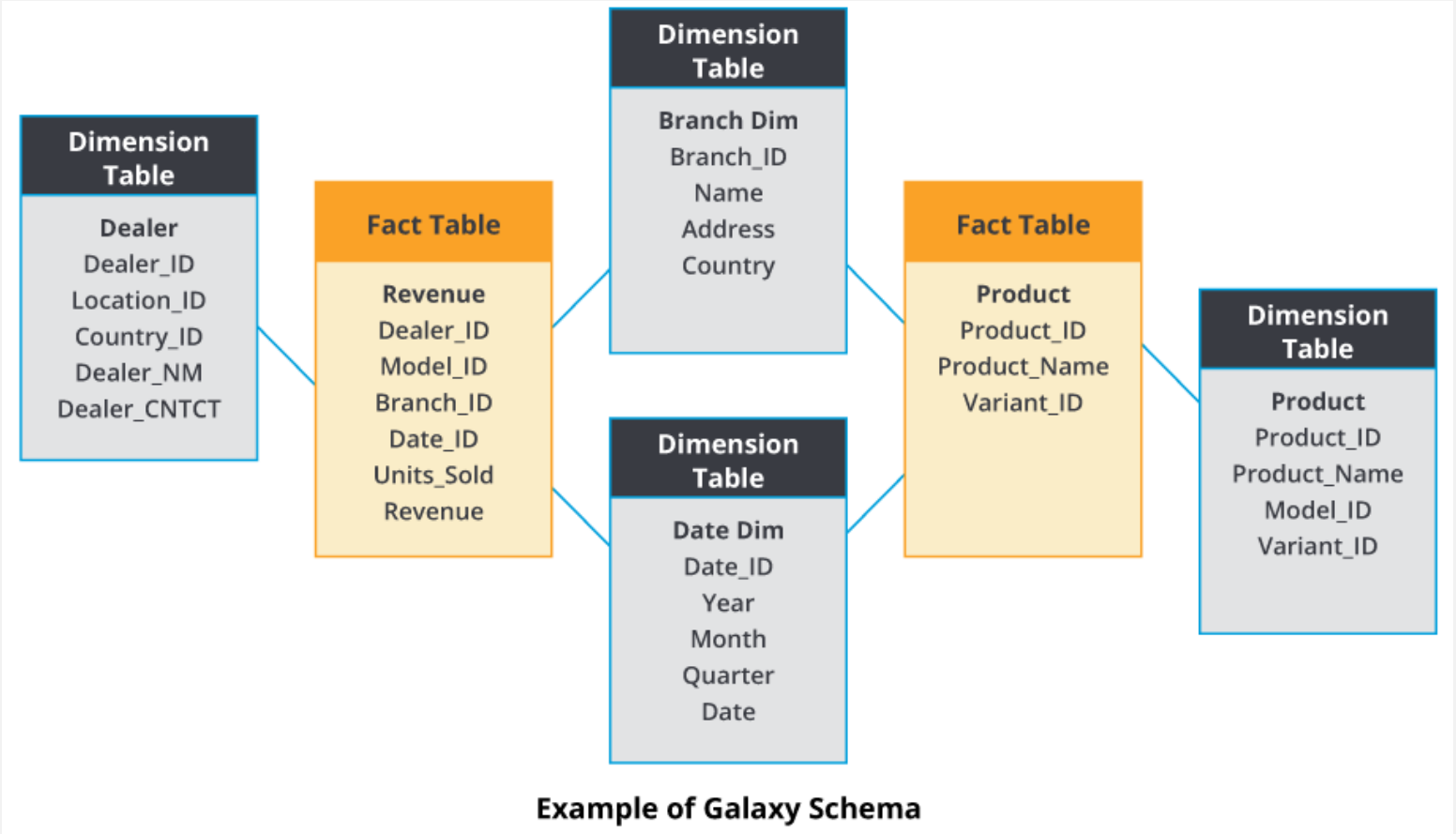
- Impact on Data Mart
  - Lower performance and usability compared to Star schema
  - Snowflake schema is not typically preferred in data marts
- Preference for Star Schema
  - Star schema is usually the better option in data marts
  - Provides better performance and usability
  - Default choice for modeling data
- Exceptional Cases for Snowflake Schema
  - Challenges with Write Operations
    - Snowflake schema may be used if write operations are problematic in Star schema
    - Can address issues with data maintenance
  - Storage Cost Challenges
    - Rare cases where storage costs pose a challenge may warrant consideration of Snowflake schema
    - Requires remodeling data loading in the data mart

SNOWFLAKE SCHEMA RECOMMENDATION

- Default preference should be given to Star schema due to its benefits
- Understanding Snowflake schema is beneficial for awareness and informed decision-making if encountered in practice

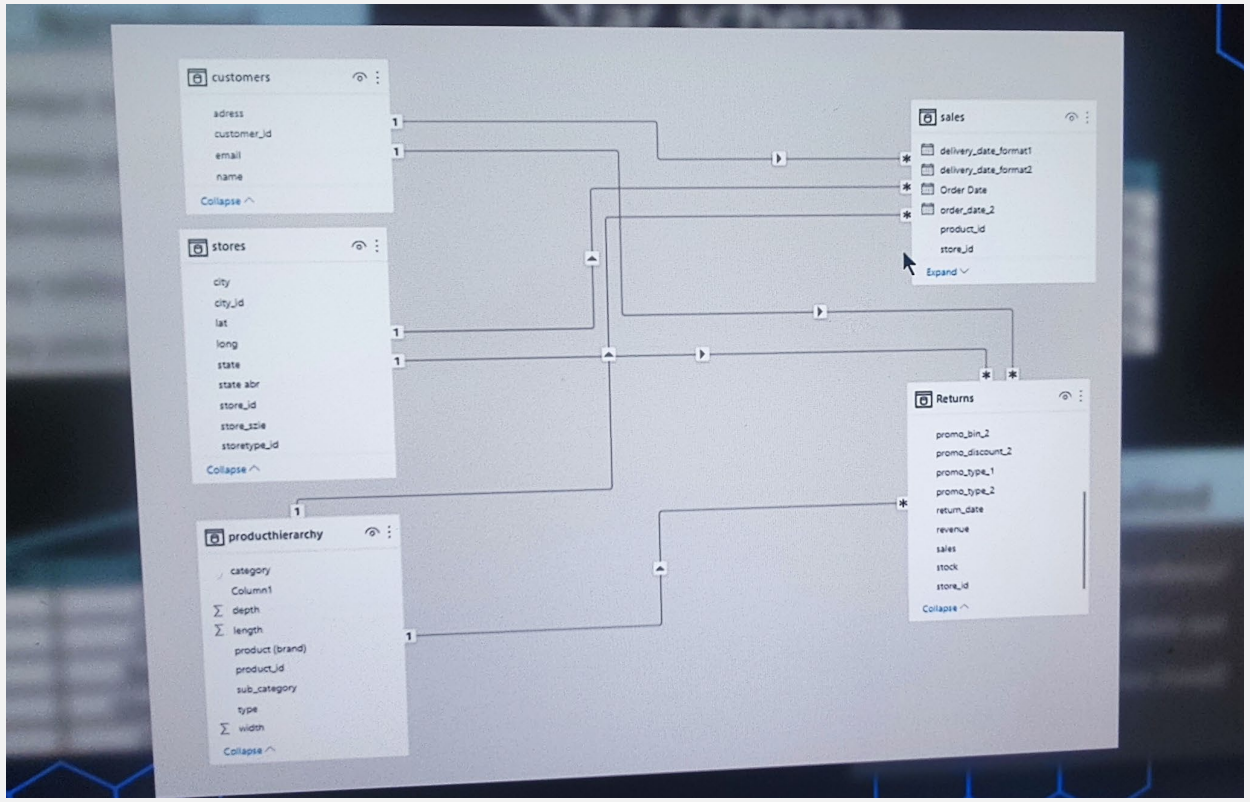


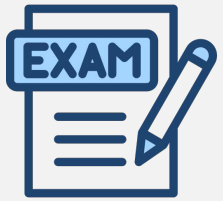
2 FACTS = GALAXY SCHEMA



Example of Galaxy Schema

POWER BI EXAMPLE WITH 2 FACT TABLES





COMPARISON

	Star Schema	Snowflake Schema	Galaxy Schema
Elements	Single Fact Table connected to multiple dimension tables with no sub-dimension tables	Single Fact Table connects to multiple dimension tables that connects to multiple sub-dimension tables	Multiple Fact Tables connects to multiple dimension tables that connects to multiple sub-dimension tables
Normalization	Denormalized	Normalized	Normalized
Number of Dimensions	Multiple dimension tables map to a single Fact Table	Multiple dimension tables map to multiple dimension tables	Multiple dimension tables map to multiple Fact Tables
Data Redundancy	High	Low	Low
Performance	Fewer foreign keys resulting in increased performance	Decreased performance compared to Star Schema from higher number of foreign keys	Decreased performance compared to Star and Snowflake. Used for complex data aggregation.
Complexity	Simple, designed to be easy to understand	More complicated compared to Star Schema – can be more challenging to understand	Most complicated to understand. Reserved for highly complex data structures
Storage Usage	Higher disk space due to data redundancy	Lower disk space due to limited data redundancy	Low disk space usage compared to the level of sophistication due to the limited data redundancy
Design Limitations	One Fact Table only, no sub-dimensions	One Fact Table only, multiple sub-dimensions are permitted	Multiple Fact Tables permitted, only first level dimensions are permitted