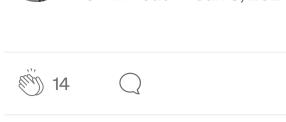
# **Data Warehousing Concepts for Beginners** — Data Engineers Rahul Sounder · Follow 5 min read · Jan 8, 2024



Users / Analysts **Data Marts Data Sources Data Staging** Data Warehouse Operational Sales System Operational Data Meta Data Raw Data Marketing System **Processing** Summary Data Flat Files Engineering **DATA WAREHOUSE ARCHITECTURE** Data warehousing is a process of collecting, storing, and managing data from different sources to support business decision-making. It involves the integration of data from various operational systems into a central repository, known as a data warehouse.

• A centralized repository that stores integrated, historical data from various sources. • It is optimized for reporting and analysis rather than transaction processing.

- The process of extracting data from source systems, transforming it to meet the data warehouse's requirements, and loading it into the data warehouse.
- being stored in the data warehouse.

- **Dimensional Modeling** • A modeling technique used in data warehousing to organize and
- (numeric measures) to create a star or snowflake schema.

- **DimProduct** ProductKey ProductAlternateKey
- **DimDate**

### DimEmployee § EmployeeKey

business requirements. New dimensions can be added, and existing ones can be modified without significant impact on the overall structure. • Scalability: The star schema is scalable, making it suitable for large data warehouses. As the volume of data grows, the star schema can handle it effectively, provided proper indexing and optimization are implemented.

Its simplicity makes it user-friendly for both database administrators and

normalized approach to the structure. The term "snowflake" refers to the shape the schema takes on when visualized: a central fact table surrounded by dimension tables that are further normalized into a branching, snowflake-like pattern. Dimension Table Dimension Table **Dimension Table** 

Fact Table

Revenue

Dealer\_ID

**Date Dim** 

Date\_ID

Year

Month

Quarter

Date

Dimension Table

Variant

Variant\_ID

Variant\_Name Product\_Name Name Fuel type Model\_ID Address Variant\_ID Country Fact Table: At the center of the star schema is the fact table. This table contains quantitative data, often numeric measures or metrics, that

Relationships: The fact table and dimension tables are connected through relationships established by keys. A primary key in a dimension table is linked to a foreign key in the fact table. These relationships allow for the integration of data across different dimensions. Attributes: Each dimension table contains attributes that provide details

However, there are some trade-offs with the snowflake schema: 1. Query Performance: Due to the normalization and increased number of joins, query performance in a snowflake schema may be slightly slower compared to a star schema.

the data warehouse. • Data mining techniques help uncover hidden insights and support predictive analysis.

## ZURE **ATABRICKS** terview

estions and Answers

Rahul Sounder Rahul Sounder **Top 50 Data Modelling Interview Top 25 Databricks Interview Questions and Answers for a Dat... Questions for Data Engineers** Explain the difference between logical and What is Databricks? Answer: Databricks is a physical data models. unified analytics platform that accelerates...

**Python—Data Engineers Coding Interview Questions —Part 1** Q1: Create a Python function to Identify only the unique values inside the list and create a... Dec 20, 2023 \*\*\* 72 See all from Rahul Sounder

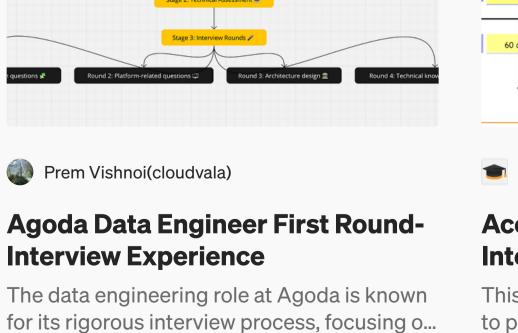
> Captures and Debezium, AWS DMS, ropagates data Synchronizing databases, Google Cloud Data Fusion changes in real-time data replication AWS Lambda, Azure Event Grid, specific events IoT data processing, or conditions automated workflows Google Functions 1uleSoft. Combines data from different Automates ML Kubeflow, model preparation, Predictive analytics, training, and deployment

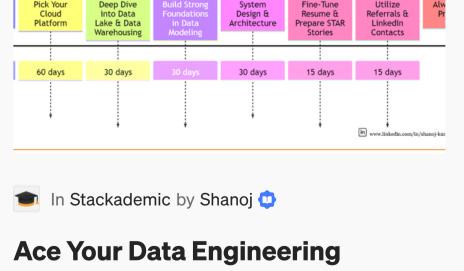
for efficiently processing large-scale data.

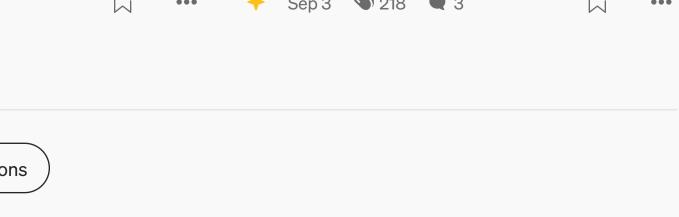
Dec 14, 2023 38

Staff picks 766 stories · 1439 saves Self-Improvement 101 20 stories · 3021 saves

> Sai Parvathaneni **Commonly Asked Snowflake Questions in Interviews #1**







• Data marts can be independent or linked to the main data warehouse. structure data for easy querying and reporting.

A dimensional model where a central fact table is connected to dimension tables, forming a star-like structure.

ResellerKey **EmployeeNationalIDAlternateKey** ResellerAlternateKey Fact Table: At the center of the star schema is the fact table. This table contains quantitative data, often numeric measures or metrics, that represent the business processes being analyzed. Examples include sales revenue, quantity sold, or profit. Dimension Tables: Surrounding the fact table are dimension tables. Each dimension table represents a specific aspect or attribute related to the business process. Dimensions are descriptive and provide context to the data in the fact table. Examples of dimensions could be time, geography, product, or customer. Relationships: The fact table and dimension tables are connected through relationships established by keys. A primary key in a dimension table is linked to a foreign key in the fact table. These relationships allow for the integration of data across different dimensions. Attributes: Each dimension table contains attributes that provide details

# **Snowflake Schema** The snowflake schema is another type of dimensional data model used in data warehousing, similar to the star schema. Like the star schema, it organizes data

represent the business processes being analyzed. Examples include sales revenue, quantity sold, or profit. Dimension Tables: Surrounding the fact table are dimension tables. Each dimension table represents a specific aspect or attribute related to the business process. Dimensions are descriptive and provide context to the data in the fact table. Examples of dimensions could be time, geography, product,

Sub-Dimensions: Each dimension table in the snowflake schema can have

sub-dimensions or related tables that store additional attributes. These sub-

about the dimension. For example, a time dimension might have attributes such as year, quarter, month, and day. Attributes in dimension tables are used for filtering, grouping, and labeling data in the fact table. **Advantages of the Snowflake Schema:** 

2. Complexity: The snowflake schema can be more complex to understand

and work with, especially for users who are not familiar with the

warehouse. **Data Mining** 

# Pipelines Integration Learning Detects and Archana Goyal **Data Engineering Series 4—Data**

Welcome to Part 4 of our 10-part series on

Stories to Help You Level-Up

19 stories · 865 saves

20 stories · 2563 saves

**Productivity 101** 

Data Engineering concepts. In this...

at Work

**Pipelines** 

+ Jun 30 **1**04

snowflak



→ Sep 3 \*\*\* 218 \*\*\* 3

Data Warehouse

ETL (Extract, Transform, Load)

• ETL ensures that data is cleansed, standardized, and integrated before **Data Mart** • A subset of a data warehouse that is designed for a specific business line, department, or function.

• It involves defining dimensions (descriptive attributes) and facts **Star Schema** 

DimSalesTerritory SalesTerritoryKey SalesTerritoryAlternateKey

about the dimension. For example, a time dimension might have attributes such as year, quarter, month, and day. Attributes in dimension tables are used for filtering, grouping, and labeling data in the fact table. **Advantages of the Star Schema:** • Simplicity: The star schema is straightforward and easy to understand.

end-users.

• Query Performance: The star schema is designed for optimal query performance. Queries can be executed quickly because the structure allows for efficient joins between the fact table and dimension tables. • Flexibility: The star schema is flexible and adaptable to changes in

for efficient querying and reporting, but the snowflake schema takes a more Location Location\_ID Region **Dimension Table** 

Country

Country\_ID

or customer.

dimensions help to reduce redundancy by separating data into different tables. Normalization: The snowflake schema employs normalization techniques by breaking down dimension tables into smaller, related tables. This reduces data redundancy and improves data integrity but can result in more complex queries due to additional joins.

by normalizing dimension tables. This can save storage space and improve data integrity. 2. Easier Maintenance: Because of the normalization, making changes to the schema, such as updating attributes or adding new ones, can be more

3. Improved Data Integrity: Normalization can enhance data integrity by reducing the risk of update anomalies that can occur when redundant data is stored in multiple places.

structure.

The choice between a star schema and a snowflake schema depends on factors such as the nature of the data, the specific business requirements, and the balance between simplicity and normalization needs in the data warehouse design. Metadata • Information about the data in the data warehouse, including its source, transformation rules, and usage.

14 Q

Jul 10 👋 21

**Recommended from Medium** 

**Practices in Data Engineering** Performance optimization is crucial in data engineering to ensure efficient data...

> Snowflake is a popular cloud-based database platform that provides robust data storage,...

> > Sep 24 👋 5

Oct 1 💜 164 🗨 1 See more recommendations

FactResellerSales DateKey SalesOrderNumber FullDateAlternateKey OrderDateKey DueDateKey ShipDateKey ResellerKey EmployeeKey SalesTerritoryKey OrderQuantity TotalProductCost SalesAmount DimReseller

Model\_ID Country\_Name **Dimension Table** Branch\_ID **Dimension Table** Date\_ID Units\_Sold Product **Branch Dim** Revenue Product\_ID Branch\_ID

Dealer Dealer\_ID

Location\_ID Country\_ID

Dealer\_NM

Dealer\_CNTCT

1. Reduced Redundancy: The snowflake schema reduces data redundancy

4. Suitability for Hierarchical Data: The snowflake schema is well-suited for representing hierarchical relationships within dimensions.

straightforward and less prone to errors.

• Metadata helps users understand and interpret the data in the • The process of discovering patterns and relationships in data stored in

Written by Rahul Sounder 208 Followers · 6 Following Senior Engineering Manager - Data at Xiaomi Technology | Ex-Amazon, Merck | SAFe® 5 Agilist | Certified AWS Solutions Architect More from Rahul Sounder

**Advanced Data** 

**Modeling Interview** 

**Questions** 

Follow

Rahul Sounder Rahul Sounder **PySpark Optimization Techniques for Data Engineers** Optimizing PySpark performance is essential

Lists

→ Nov 6

In Data Engineer Things by Jagadesh Jamjala

**8 Performance Optimization Best** 

Mayurkumar Surani **Ace Your Data Engineering** Interview: 20 Questions and... So, you're gearing up for a data engineering interview? Congratulations! It's an exciting...

+ Sep 28 **3**0

Help Status About Careers Press Blog Privacy Terms Text to speech Teams