Data Curation Techniques

Siddharth R

Data Encoding

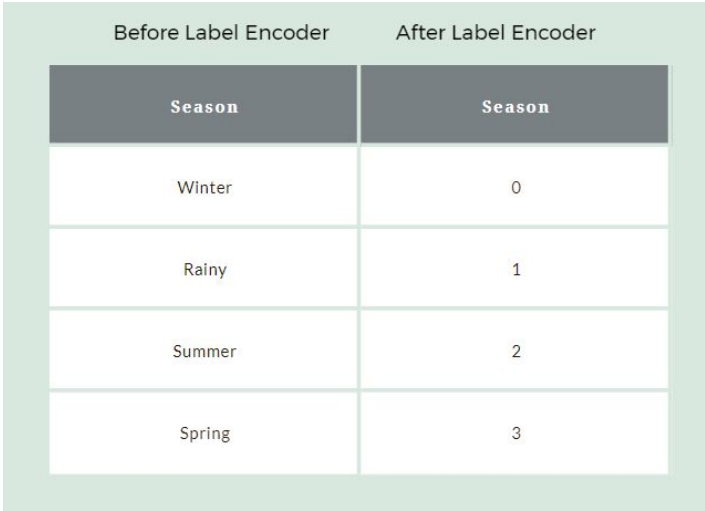
- Transform categorical variables into numerical representations
- Why we need a data encoding?
- Bridges the gap between raw data and ML algorithms

Common Types:

1. Label Encoding
2. One hot encoding
3. Ordinal encoding
4. Binary encoding
5. Frequency encoding

Label Encoding

- When dealing with a classification problem where the target variable (the variable you're trying to predict) has categorical values, label encoding can be used.
- Converts categories into integer values starting from 0 or 1.
- Each category is assigned a unique number.
- Creates ordinal relationships for nominal data (which can be misleading).



Before Label Encoder

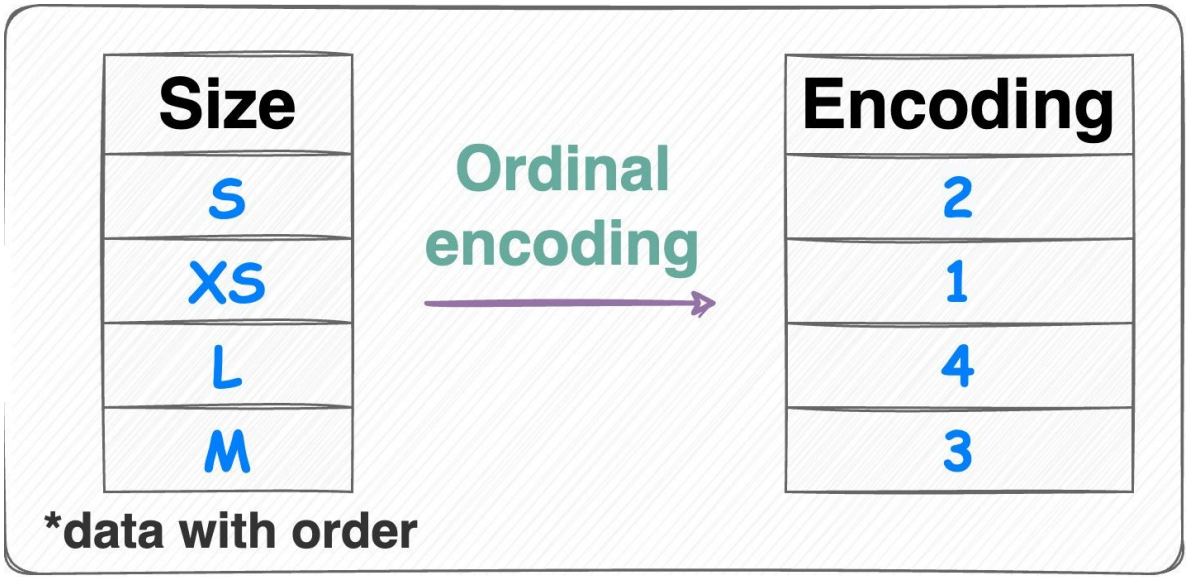
Season
Winter
Rainy
Summer
Spring

After Label Encoder

Season
0
1
2
3

Ordinal Encoding

- Similar to label encoding but used when categories have a meaningful order
- Misleading results if used with nominal data.



One hot Encoding

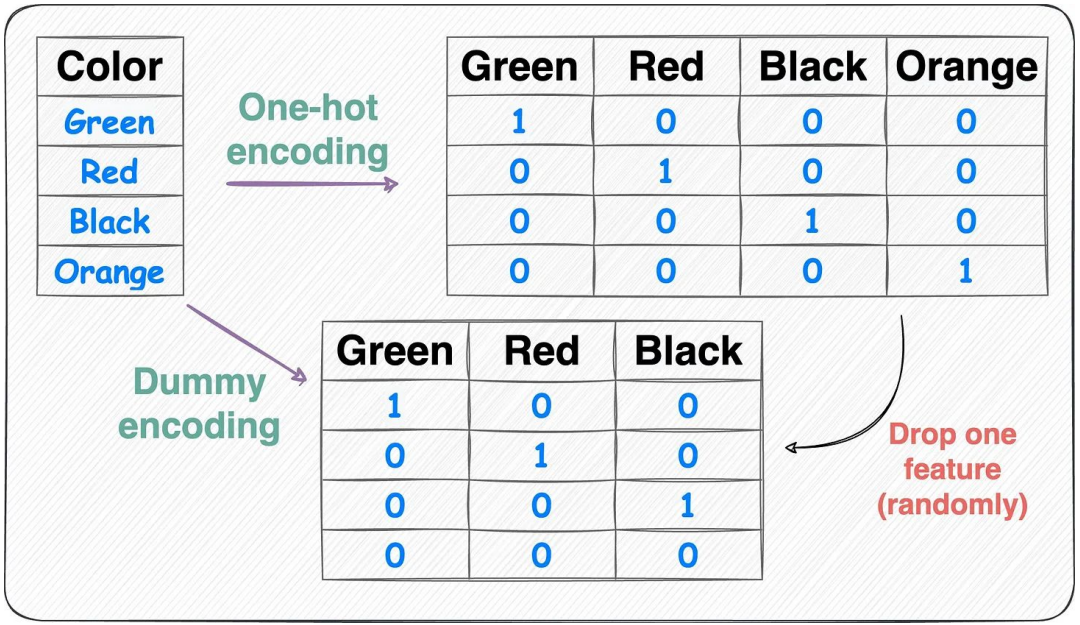
- Converts categories into binary vectors.
- Creates a separate column for each category and assigns 1 or 0.
- Increases dimensionality for high-cardinality features.

id	color
1	red
2	blue
3	green
4	blue

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

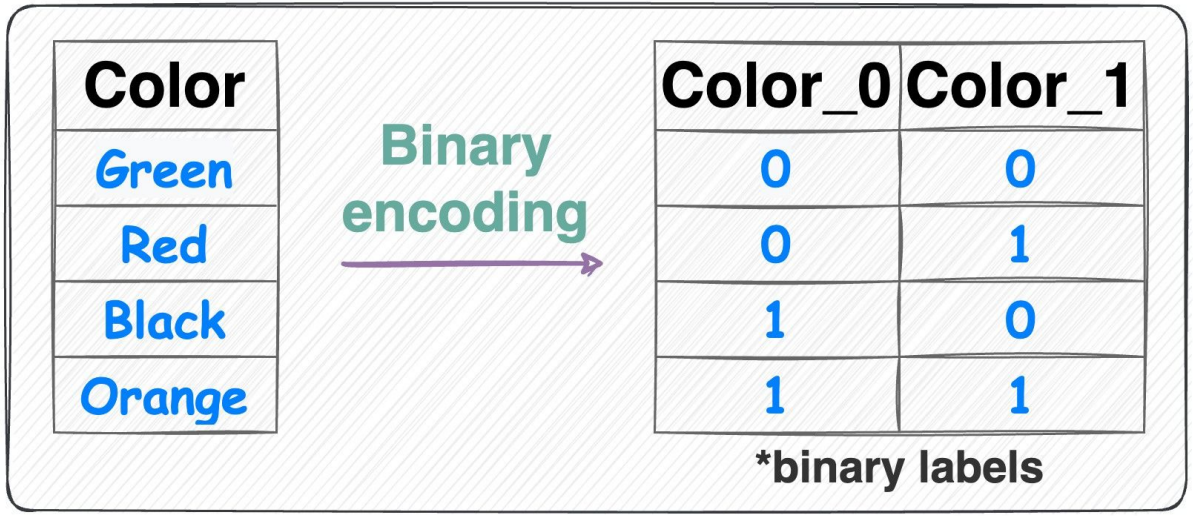
Dummy Encoding

- Dummy encoding uses N − 1 binary variables to represent N categories
- Avoid multicollinearity issues / dummy variable trap



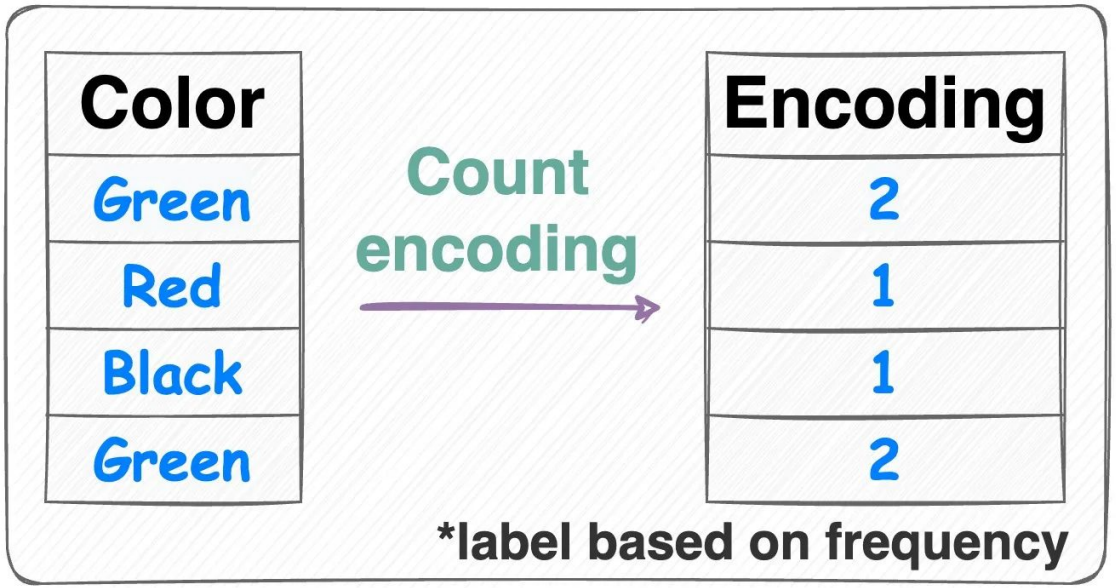
Binary Encoding

- Suitable for high-cardinality categorical data.
- Balances dimensionality and performance.
- Harder to interpret than one-hot encoding.

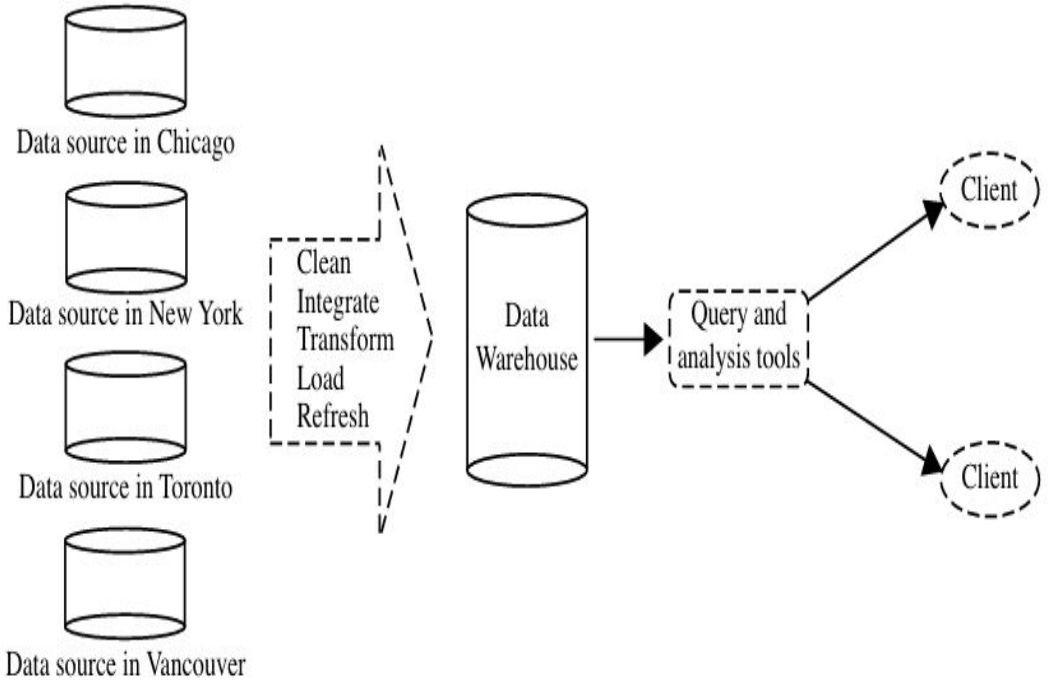


Frequency Encoding

- Encodes categorical features based on the frequency of each category.

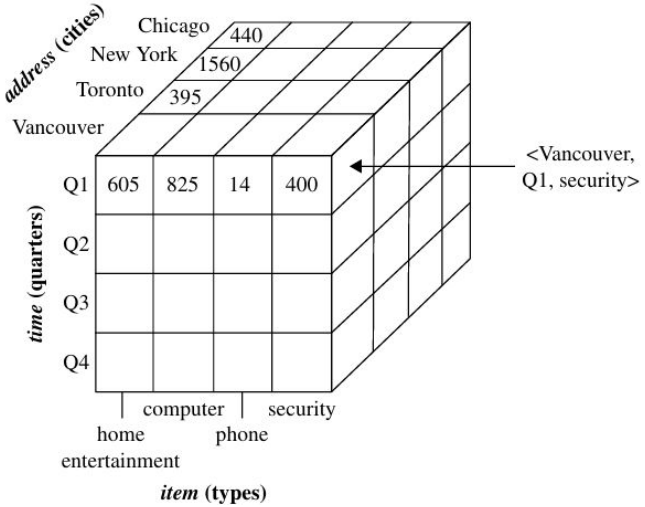


Data Warehouse



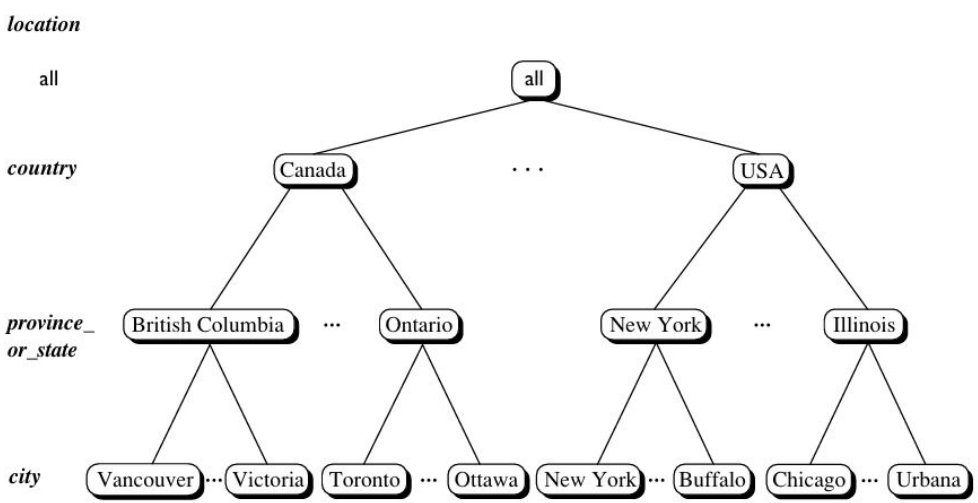
Data Cube

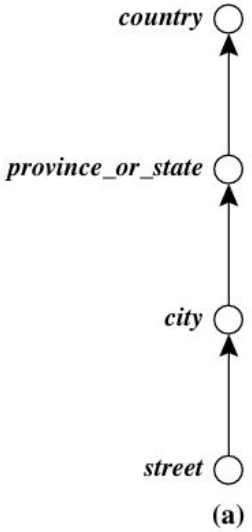
- Multi-dimensional array of data used to represent and analyze information in a data warehouse.
- Analyze data across multiple dimensions
- Facilitates Online Analytical Processing (OLAP), enabling fast querying and data aggregation
- A data cube typically consists of dimensions and facts
  - Dimension represent perspectives or attributes by which t data can be analyzed.
  - Facts represents the measures
- Each cell within the cube holds an aggregated va

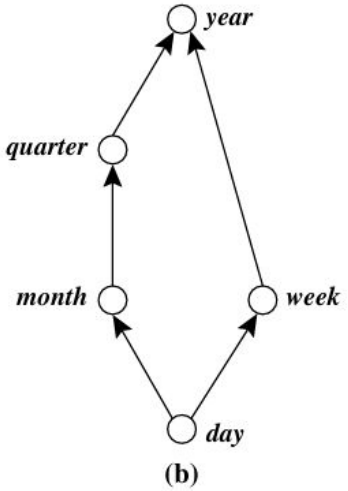


What is concept hierarchy

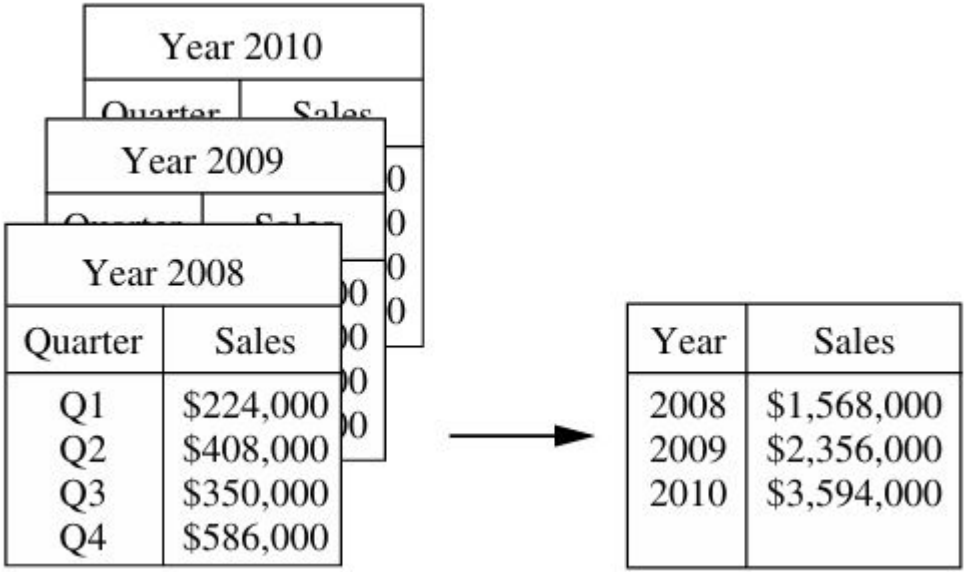
Mappings from a set of low-level concepts to higher-level





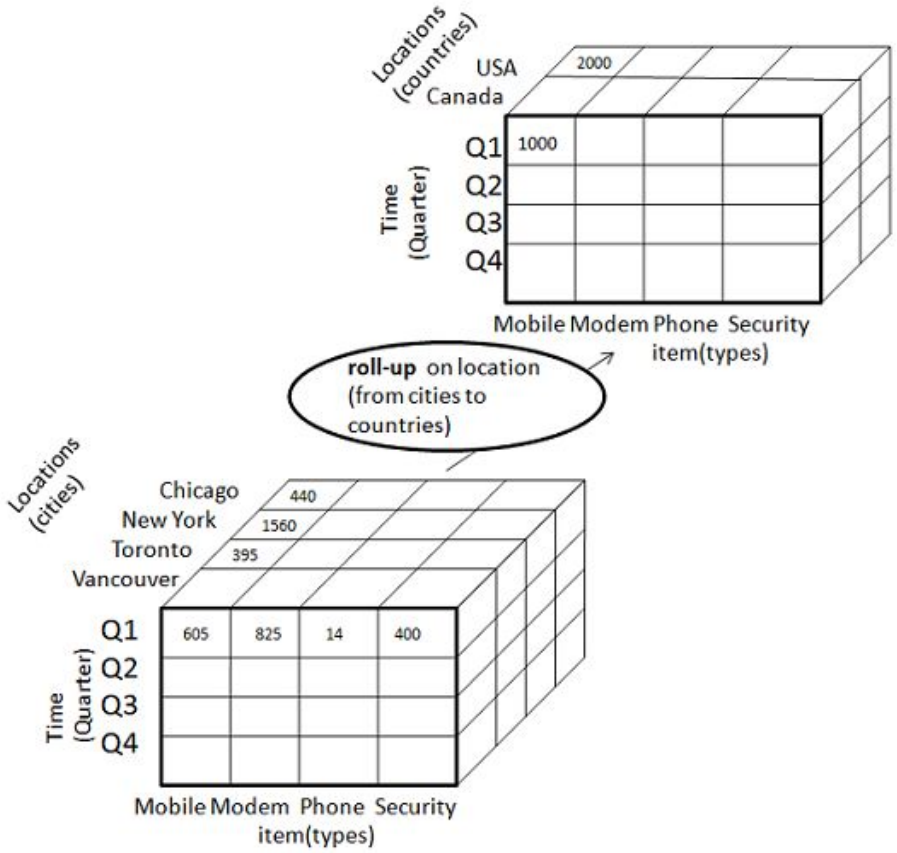


Operations - Aggregation



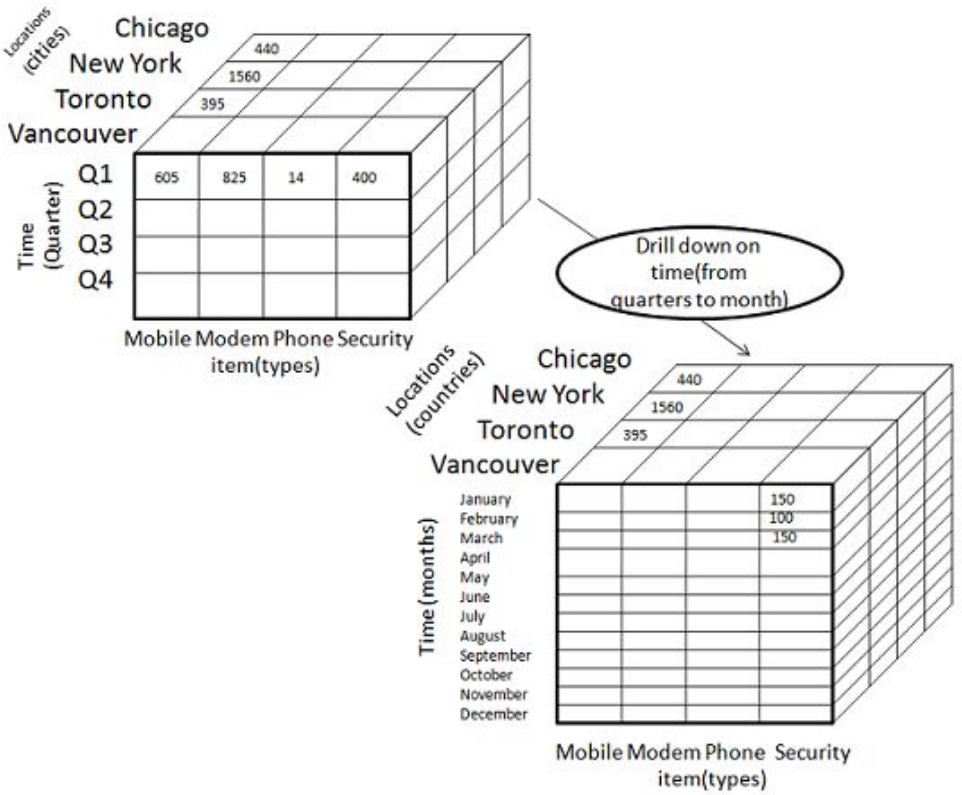
Operations - Roll-Up

- By climbing up a concept hierarchy for a dimension
- By dimension reduction



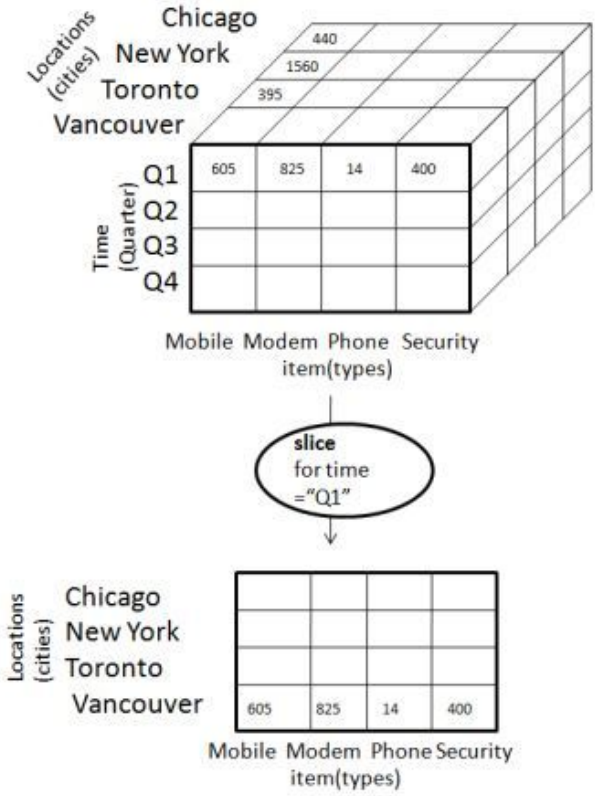
Operations - Drill down

- By stepping down a concept hierarchy for a dimension
- By introducing a new dimension.



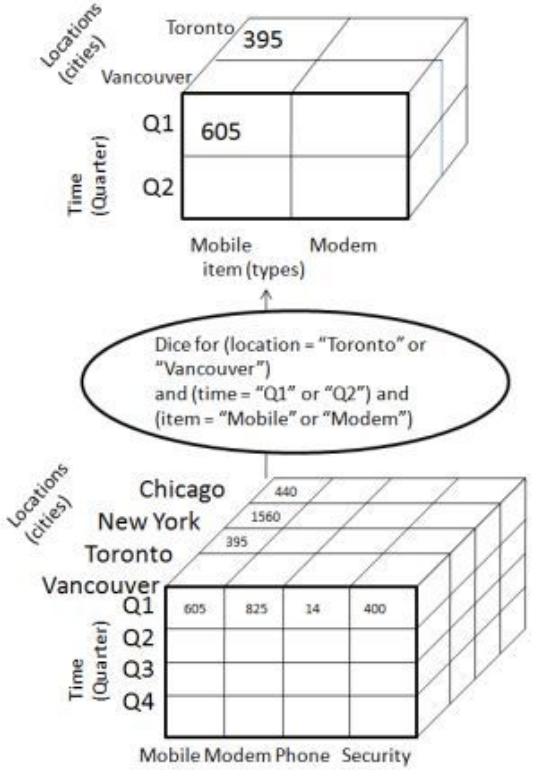
Operations - Slice

Slice selects one particular dimension from a given cube and provides a new sub-cube.



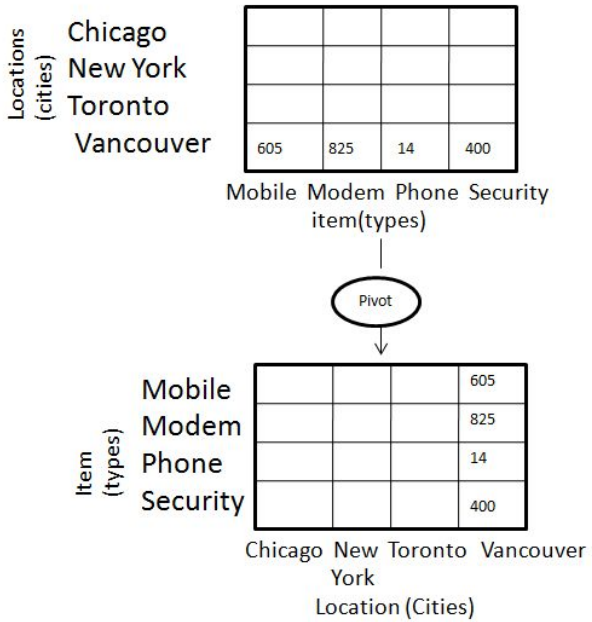
Operations - Dice

Dice selects two or more dimensions from a given cube and provides a new sub-cube.



Operations - Pivot

Rotates the data cube to visualize data from different perspectives.

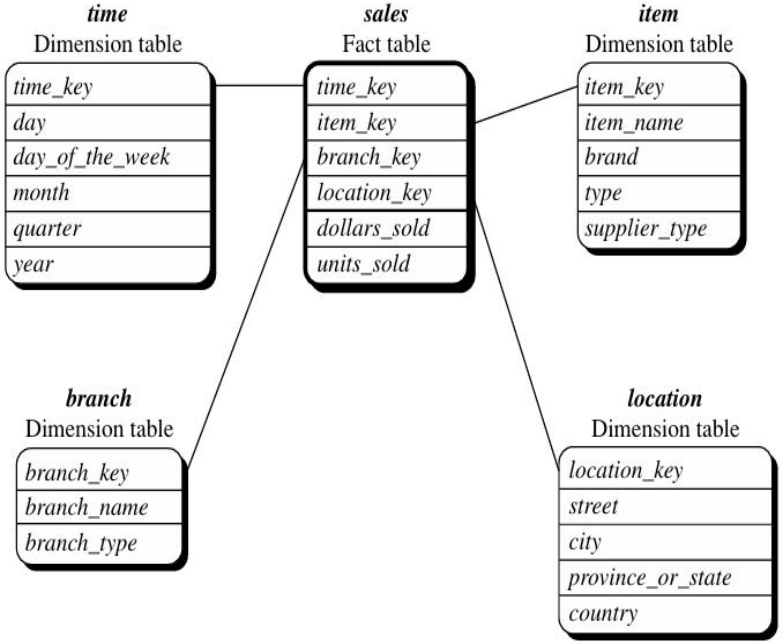


Schemas in Data Warehousing

- A schema in data warehousing defines the logical structure of data.
- It organizes data into tables, relationships, and constraints, ensuring efficient storage, querying, and reporting.
- The entity-relationship data model is commonly used in the design of relational databases
- The most popular data model for a data warehouse is a multidimensional model,which can exist in the form
    - Star schema
    - Snowflake schema
    - Fact constellation schema

Star Schema

The data warehouse contains a large central table (fact table) containing the bulk of the data, with no redundancy and a set of smaller attendant tables (dimension tables), one for each dimension.

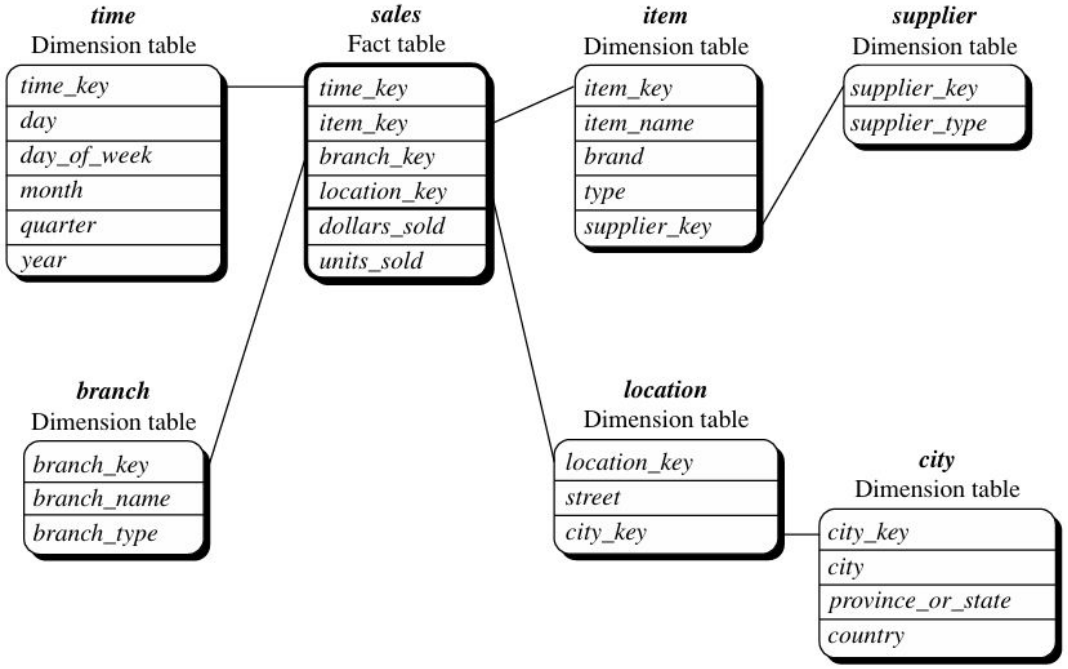


Star schema of *sales* data warehouse.

Snowflake Schema

The snowflake schema is a variant of the star schema model, where some dimension tables are **normalized**, thereby further splitting the data into additional tables. The resulting schema graph forms a shape similar to a snowflake.

Updation, Insertion and Deletion Anomalies in DBs



Snowflake schema of a *sales* data warehouse.

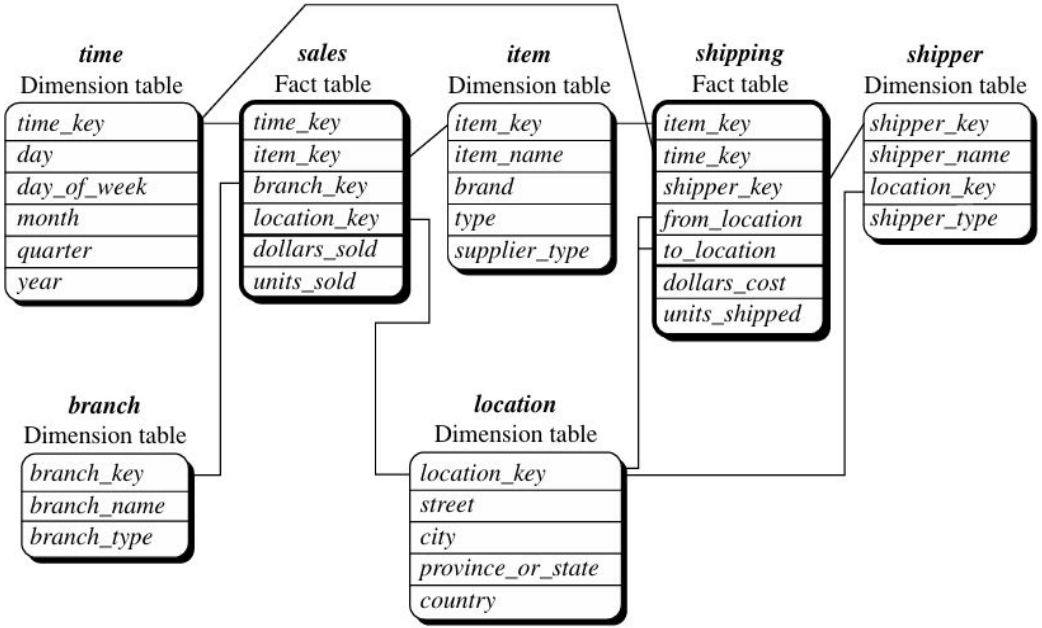
When Snowflake Schema is Better Than Star Schema

- Frequent Data Updates
- Large Volume of Data
- Complex Relationships

But when query speed is more critical and data redundancy is acceptable, you can select star schema because query performance of snowflake schema is affected by multiple joins.

Fact constellation

Sophisticated applications may require multiple fact tables to share dimension tables. This kind of schema can be viewed as a collection of stars, and hence is called a galaxy schema or a fact constellation.



Fact constellation schema of a sales and shipping data warehouse.

ETL vs ELT

- Move data from multiple sources to a data warehouse for analysis and reporting.
- ETL (Extract, Transform, Load): Transform data before loading it into the target system.
- ELT (Extract, Load, Transform): Load raw data into the target system first and transform it after loading.
- ETL is best suited for structured, relational databases and data must be validated before entering the data warehouse.
- ELT is ideal for large, unstructured datasets. faster data loading since transformations happen after loading.

Which one you will choose?

1. A manufacturing company installs IoT sensors on its machines that generate millions of data points every minute. The data is in semi-structured JSON format and needs to be stored in a cloud-based data warehouse for later analysis.
2. A healthcare organization needs to move patient records and medical history data from different hospital databases to a centralized system. Due to regulatory compliance (HIPAA/GDPR), the data needs to be cleaned and validated before entering the target system.
3. Which one is more efficient in space?

Data Governance

- Foundation for data-driven business practices
- Engages **people, processes, and technologies** to maximize data value across an organization while protecting data with appropriate security controls.
- Three core components
  - Discoverability
  - Security
  - Accountability

Discoverability

- Metadata management and master data management
- Metadata - “data about data”
- Metadata - can be either manual or automated
- DMBOK identifies four main categories of metadata that are useful to data engineers
  - Business metadata - to answer non technical questions about who, what, where, and how.
  - Technical metadata - Pipeline metadata, Data lineage and Schema metadata
  - Operational metadata - statistics about processes, job IDs, application runtime logs, data used in a process, and error logs.
  - Reference metadata - used to classify other data. internal codes, geographic codes, units of measurement, and internal calendar standards.

Accountability & Security

- Data ownership & stewardship
- Roles and responsibilities
- Data quality
  - Accuracy
  - Completeness
  - Timeliness
- Compliance and security
  - GDPR (General Data Protection Regulation) – Europe
  - HIPAA (Health Insurance Portability and Accountability Act) – Healthcare industry

Master data Management

- As organizations grow larger and more complex through organic growth and acquisitions, and collaborate with other businesses, maintaining a consistent picture of entities and identities becomes more and more challenging.
- Master data management (MDM) is the practice of building consistent entity definitions known as golden records.
- For example, an MDM team might determine a standard format for addresses, and then work with data engineers to build an API to return consistent addresses

Thank You !!!