DBMS Operations

Security

- Security

  WHO can do WHAT action to WHICH object

  - WHO: Identity, user
  - WHAT: Privilege
  - WHICH: Access
- GRANT/DENY provide/revoke privilege to/from objects for certain user(s)

```
GRANT SELECT ON REP TO Prof_Shan
REVOKE DELETE ON REP TO Student1
```

Security for Table

```
GRANT
    { SELECT | INSERT | UPDATE | DELETE | TRUNCATE
             | REFERENCES | TRIGGER | ALL }
     ON { table_name | ALL TABLES
          IN SCHEMA schema_name }
     TO { [ GROUP ] role_name | PUBLIC }
```

Security for Function

```
GRANT { EXECUTE | ALL [ PRIVILEGES ] }
    ON { FUNCTION function_name | ALL FUNCTIONS
           IN SCHEMA schema_name }
    TO { [ GROUP ] role_name | PUBLIC }
```

- Similar GRANT to stored procedure

System Catalogs

- Central location to store all related information in DBMS
  - What tables do we have?
  - What are their definitions?
  - Who can access them?
  - What is the last time someone read from/write to this table?

System Catalogs

- What tables do we have?

```
SELECT *

FROM pg_catalog.pg_tables
```

schemaname name	tablename name	tableowner name	tablespace name	hasindexes boolean	hasrules boolean	hastriggers boolean	rowsecurity boolean
public	customer	postgres	[null]	true	false	true	false
public	inspecting	postgres	[null]	true	false	true	false
public	inspector	postgres	[null]	true	false	true	false
public	manager	postgres	[null]	true	false	true	false
public	managerphone	postgres	[null]	true	false	true	false
public	product	postgres	[null]	true	false	true	false

Different Types of Data Usage

Different Types of Data Usage

- Online Transaction Processing (OLTP)
  - Transactions, sales, delivery notices, social network interactions, etc.
  - Requires short and fast processing: Small table. Shallow query.
- Online Analytical Processing (OLAP)
  - Complex data pulling of historical data, over a wide range of attributes
  - Requires long and complex processing: Large table. Deep query.

Different Levels of Analytics

- Query & Reporting
  - That's what most data scientists hate to do but have to do.
- Business Intelligence/OLAP: Sorting, slicing/dicing, pivoting, aggregation, etc.
  - That's what most data scientists do.
- Data Mining/Analytics/Statistics
  - That's what most data scientists want to do but never has a chance

Data Mining/Analytics/Statistics

- More data, deeper analysis
  - Data Mining/Analytics/Statistics
  - Machine Learning

BI/OLAP

- Relational OLAP vs Multi-dimensioanl OLAP (ROLAP vs MOLAP):
  - Where does the analysis happen, database or vendor defined space?
  - ROLAP: Slow. Less functionalities.
  - MOLAP: Limited data volume
- Most vendors are Hybrid OLAP (HOLAP) now.
  - MOLAP for most commonly used data
  - ROLAP for data not available in MOLAP

MOLAP Characteristics

- Cube: Multi-dimensional, multi-leveled data. Looks like a cube.
- Multi data source MOLAP
- Server centric: multi tier architecture
    - Desktop vs Server

Visualization

- A technology that is used all across different types of analytics

Dimensional Modeling

Two Types of Analytical Data

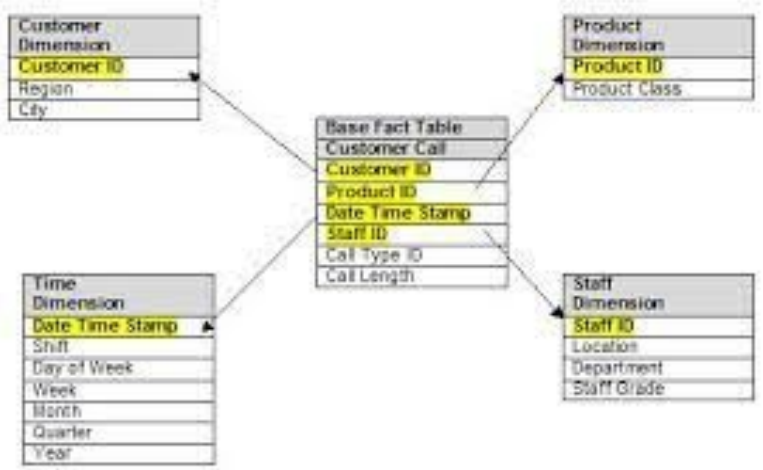
- Fact: Calculable values such as sales $, shopping unit, etc.
- Dimension: Descriptive information such as date/time, geo-location, product label, customer information, etc.
- Facts are calculated along dimensions, e.g., sales of a year/month/week.

Requirements of Analytics

- Focusing on facts
- Frequently run across multiple dimensions
  - Historical
  - Aggregated
  - Filtered
  - Sorted

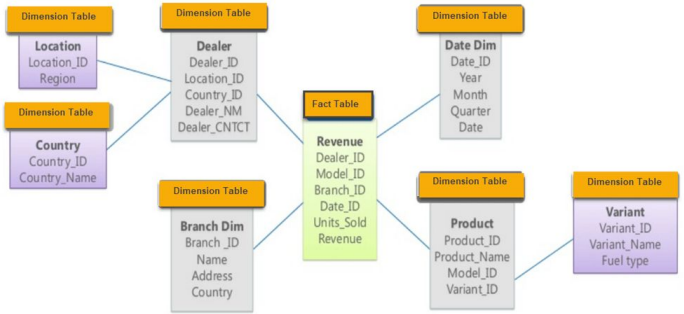
Star Schema

- Since calculations are centered on facts, facts are placed at the center of the universe, and dimensions are placed around the facts as lookup tables.
  - Looks like a star
  - Lookup tables are not 2NF/3NF



Snowflake Schema

- Each dimension can be break down into a series of lookup tables, each contains a certain level of tables
  - Forms a snowflake
  - If facts are aggregated to a higher level, the higher level lookup tables can be used to join the higher level facts



Data Warehouse

Need for Data Warehouse

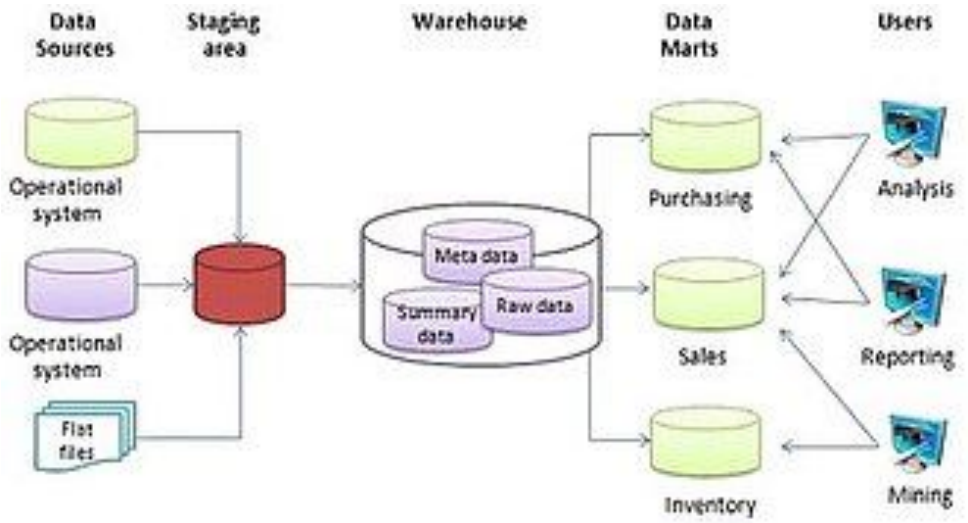
- Transactional databases are not suitable for data analysis.
  - Tables are designed to be relational, not dimensional.
  - Schemas are designed as Entity-Relationship, not star/snowflake.
  - Database is optimized for fast writing of small, current data, not for long querying of historical data.

Data Warehouse

- Data Warehouse: A place for analytical data storage
  - Dimensional tables
  - Star/snowflake schemas
  - Databases designed for complex queries

DWH Architecture

- DWH Architecture
  - Source, ETL, Staging area, Datamart, BI
  - Implementation: Kimball methodology



Datamart

- A subject area of data warehouse:
  - Contains a smaller subset of tables
  - Some tables may be shared among different datamarts
- If implemented poorly (usually because they are implemented independently by different teams), different datamarts may have conflicting data.
  - Very common issue. You will run into it sooner or later.

Other Terminologies

- Operating Data Store (ODS)
  - Databases with transactional data, but somehow can be used as source for analytical operation
  - Somewhere between transactional database and data warehouse
- Logical DWH
  - Use transactional table as fact, with dimensional tables implemented.

Data Lake

- Today's data volume is so large, dumping everything into data warehouse is not practical
- Meanwhile, most of today's data sources are relational.
- Solution: Directly query the text files generated by data sources using SQL, instead of uploading them into data warehouse and query.
    - The collection of text files is called data lake

ETL Overview

- Extract
- Transform
- Load

ETL Options

- Must run in an automated way, usually from a client tool
  - DBMS native tools: psql
    - SQL/Stored Procedure/View/Materialized View
  - Customized tools written in programming languages: python
    - SQL/Stored Procedure/View/Materialized View
    - Programming language processing