Class Note

Data warehouse design — important considerations:

1. grain/granularity of fact tables
2. identifying the dimension tables
3. slowly changing dimensions

Star schema

- One large central fact table
- Various smaller dimension tables are connected to the fact table
- Fact Table
  - Has multiple foreign keys referring to each of the dimension tables (1:N relationship)
  - A composite primary key made of all these FKs
  - Contains measurement data (fact)
- Dimension Table
  - Contains information about each of the facts in the Fact Table
  - Contains criteria for aggregating the measurement data(fact)
- Surrogate Keys, meaningless integers used to connect the fact to the dimension tables.

Procedure of Building A Data Warehouse:

- Designing the data warehouse schema
- Building the ETL pipelines — consume 80% of time
- Creating the BI reports

Granularity of the Fact Table

- Granularity of data = how detail the data is
- Higher granularity implies more rows, while lower granularity implies fewer rows.
- Tradeoff — level of detailed analysis VS. storage/query performance
- Data with higher granularity —> Data with lower granularity
  - e.g Going from days(high) to months(low)

- Data with higher granularity $<$—X— Data with lower granularity

Slowly changing dimensions

Dimensions that changes slowly and irregularly over a period of time.