Database Design Fundamentals

Introduction:

As today organization is highly dependent to data and strategic made decisions, the need for more sustainable and soft ways to store the huge amount of datasets urged the creation of more modern softwares for processing them.

Database types and their functionalities:

Relational Databases :

On 1970 Edgar Codd´s paper named: “A Relational Model of Data for Large Shared Data Banks” proposed a new way of storing data using a mathematical concept of relational algebra, considering databases as a set of relationships of tuples each one with consistent attributes. A tuple containing 6 attributes for example would be called a 6-tuple and so forth. Database professionals translate those “attributes” to a set of records (rows) each one with consistently relational attributes.

Records or tuples are uniquely identified by a *primary key* which stores the attributes or their combinations. Keys could consist of name, address, street, house number,

postcode etc... Putting them together to uniquely identify an user or record, those keys are mostly used to set an ID for the table tuples avoiding data repetitions.

- ***Primary keys are unique identifiers of the tables***, meaning that they store the tables ID's. Those are specially important to load all the data stored at the table when the system make a specific request.
- ***Foreign key are the relational keys of the database*** which means they make a relationship with the existing primary keys (identifiers) to link the data.

Databases of this type follow a mathematical approach known as relational algebra, which later on give us the schema for the SQL (Structured Query Language) creation on 1986 by the ANSI (American National Standard Institute).

> "SQL has a direct relationship to relational
> algebra. This describes tables (in relational algebra terms—relations), records (tuples) and the relationship between them."

Employee

Employee ID	Surname	Given Name	Initial	Department ID	DOB	Salary
35411	Felsky	Warren	J	22	25/2/1983	21000.00
35567	Cash	Martin	B	21	1/6/1989	19750.00
35582	Kelly	Edward	T	22	18/7/1961	45450.00

Department

Department ID	Department Name	Department Location
21	Accounts	Head office
22	Payroll	Head office
23	Sales	Latrobe

Fig. 2.5 Relational database consisting of two tables or relations

By using relational databases we have more stability and precision with data and their possible relationships, we usually use modern DBMS (Database Managing Systems) such as MySQL, SQL Server, Oracle and PostgreSQL for developing a system and commands like:

```
SELECT * FROM property_for_rent WHERE type_property = 'House';
```

Where **Select** is the command given to the database, **FROM** is the specification for the proper command execution, the next are the table and the property of it's selection.

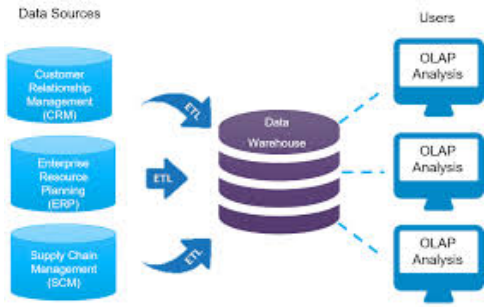
Data Warehouse:

Data warehouses are organizational datasets that are used for analyze data from heterogenous datasets, meaning that it contains historical records which can be analyzed using business intelligence or data analytics.

This type of structure is maintained separate from the main database because the database is continuously operating and updating, the analysis can be difficult to take place. So we can make proper trend reports, I quote:

> "A data warehouse is a central repository of data in an organization storing historical data and being constantly added to by current transactions. The data can be analyzed to produce trend reports. It can also be analyzed using data mining techniques

> for generating strategic information. Unlike data in a relational database, data in a data warehouse is organized in a de-normalized format”



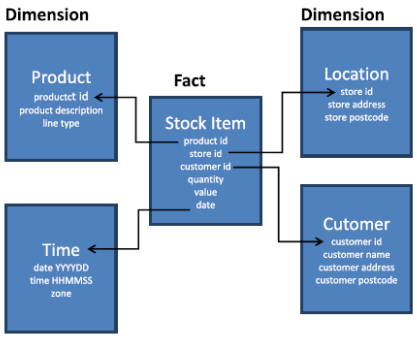
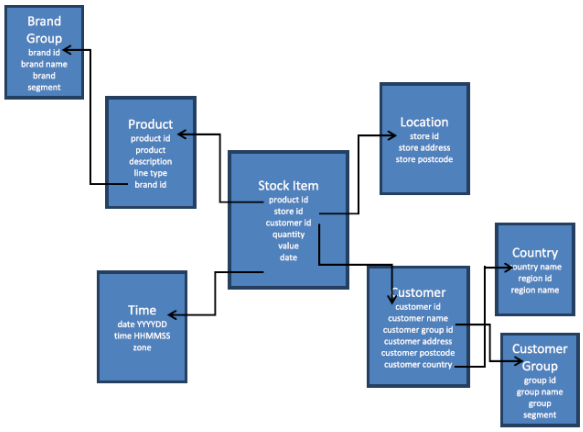


Fig. 2.6 Star schema



We can have three main schema for data warehouses:

- Star schema: Tables are classified as fact tables that record the data about specific events and dimension tables which information relating with the attributes in the previous one.
- Snowflake schema: based on the same principle mentioned above, we have more tables relating to the dimension tables.
- Star-flake schema: which is a snowflake schema where only some of the dimension tables have been de-normalized

Now we will detail the data mining process (more commonly known as data analytics), these techniques are used to remove duplicate and erroneous data from the database before applying the data analysis and management for future business prospects. Thus we have :

- Data cleansing: that’s the process of removing erroneous and duplicate data, such as identical or similar records on the database.
- Initial Analysis: this refers to the analysis of the quality of the data distribution among the system, determining if there exists any bias or other problems.
- Main analysis: where the statistical models are applied to the data sample for asking some question, for example how costumers react to the product marketing based upon the number of clicks of the ads, with the data generated we have to apply statistics to know how to proceed in business.