**Project Documentation**

**Star Schema design : Hotel\_data using Kimball theory**

**Glossary**

1.) Introduction to Datawarehousing concepts and Architecture

2.) Introduction to ETL

3.) Design Approaches in DWH and its Architecture

4.) Types of Schema in DWH

5.) Project Overview

**Introduction to Datawarehousing concepts and its architecture**

What is Data Warehouse?

"A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process."The process of constructing and using data warehouses is called data warehousing.

A common way of introducing data warehousing is to refer to the characteristics of a data warehouse as :

* [Subject Oriented](https://docs.oracle.com/cd/B10501_01/server.920/a96520/concept.htm" \l "49840)

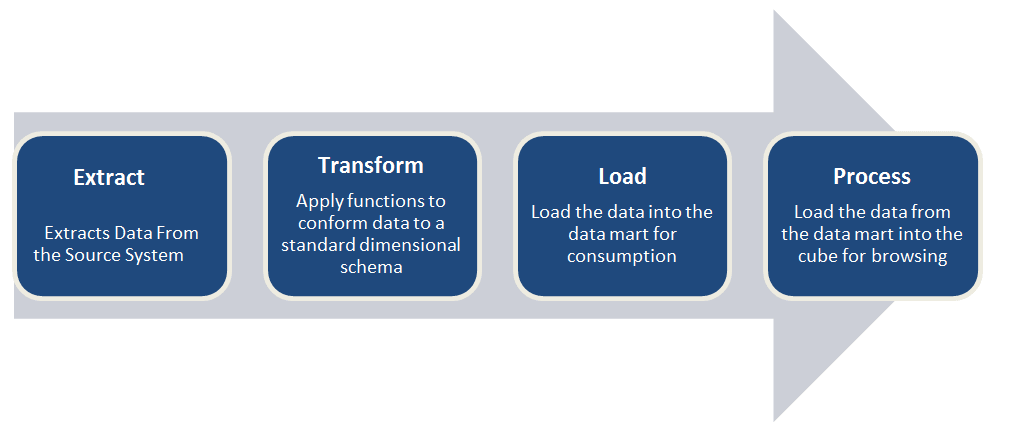
* [Integrated](https://docs.oracle.com/cd/B10501_01/server.920/a96520/concept.htm" \l "49842)

* [Nonvolatile](https://docs.oracle.com/cd/B10501_01/server.920/a96520/concept.htm" \l "49871)

* [Time Variant](https://docs.oracle.com/cd/B10501_01/server.920/a96520/concept.htm" \l "49853)

Components of a DataWarehouse

DataWarehouse Architecture



Introduction to ETL

ETL is defined as a process that extracts the data from different RDBMS source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. ETL full-form is Extract, Transform and Load.

## 1) Extraction

Data is extracted from the source system into the staging area. Transformations if any are done in staging area so that performance of source system in not degraded. Also, if corrupted data is copied directly from the source into Data warehouse database, rollback will be a challenge. Staging area gives an opportunity to validate extracted data before it moves into the Data warehouse.

## 2) Transformation

Data extracted from source server is raw and not usable in its original form. Therefore it needs to be cleansed, mapped and transformed. In fact, this is the key step where ETL process adds value and changes data such that insightful BI reports can be generated.

## 3) Loading

Loading data into the target datawarehouse database is the last step of the ETL process. In a typical Data warehouse, huge volume of data needs to be loaded in a relatively short period.Hence, load process should be optimized for performance.

Design Approaches in DWH and its Architecture

#### Bill Inmon - Top-down Data Warehouse Design Approach

Bill Inmon Methodology is based on a top-down approach and defines data warehouse , This top-down design provides a highly consistent dimensional view of data across data marts as all data marts are loaded from the centralized repository (Data Warehouse). The top-down design has also proven to be flexible to support business changes as it looks at the organization as whole, not at each function or business process of the organization. Generating a new dimensional data marts against the data stored in the data warehouse is a relatively simple task.

Ralph Kimball - Bottom-up Data Warehouse Design Approach

is design methodology is called dimensional modeling or the Kimball methodology. This methodology focuses on a bottom-up approach, emphasizing the value of the data warehouse to the users as quickly as possible. In his vision, a data warehouse is the copy of the transactional data specifically structured for analytical querying and reporting in order to support the decision support system. As per his methodology, data marts are first created to provide reporting and analytical capabilities for specific business\functional processes and later on these data marts can eventually be unioned together to create a comprehensive enterprise data warehouse. The bottom-up approach focuses on each business process at one point of time so the return on investment could be as quick as first data mart gets created.

Types of Schema in DWH

## Star Schema

## The star schema is the simplest type of Data Warehouse schema. It is known as star schema as its structure resembles a star. In the Star schema, the center of the star can have one fact tables and numbers of associated dimension tables. It is also known as Star Join Schema and is optimized for querying large data sets.

## Snowflake Schema

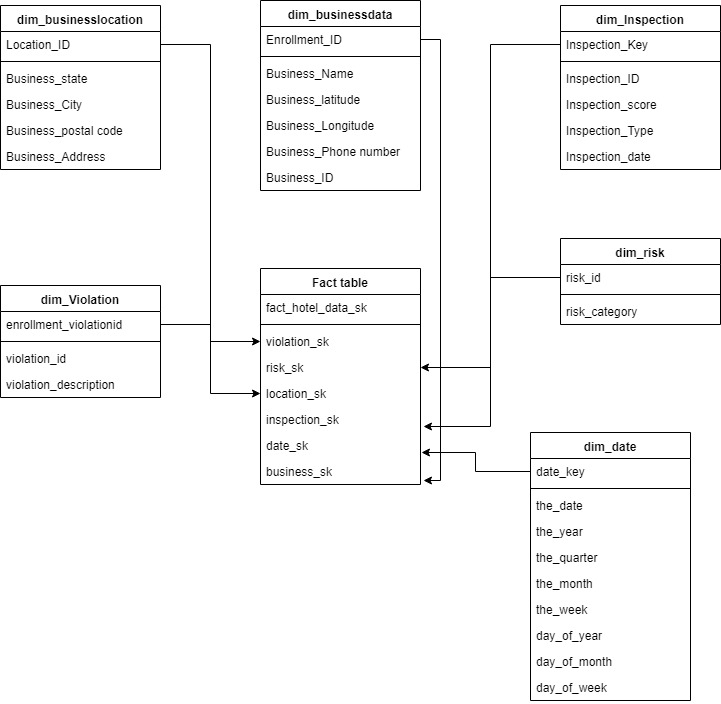
## A Snowflake Schema is an extension of a Star Schema, and it adds additional dimensions. It is called snowflake because its diagram resembles a Snowflake.

## Galaxy schema

## A Galaxy Schema contains two fact table that shares dimension tables. It is also called Fact Constellation Schema. The schema is viewed as a collection of stars hence the name Galaxy Schema.

**Project OverView**

Star Schema : ETL on Hotel\_Data using Kimball Methodology

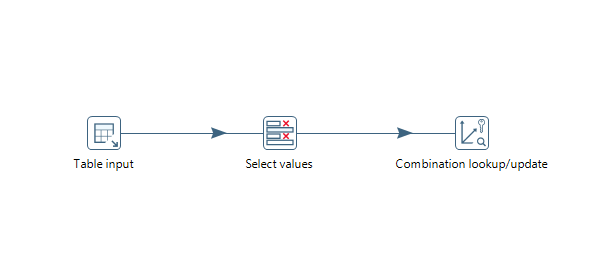


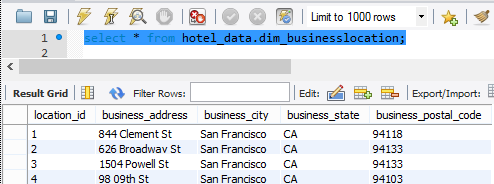
**ETL Execution Instructions**

Setup ETL environment by editing Kettle properties file and add the directory for DIR\_SOURCE\_FILES

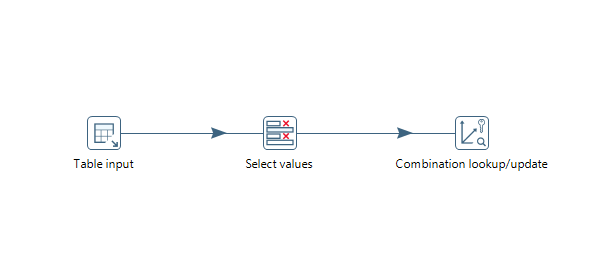
**The Dimension Transformations**

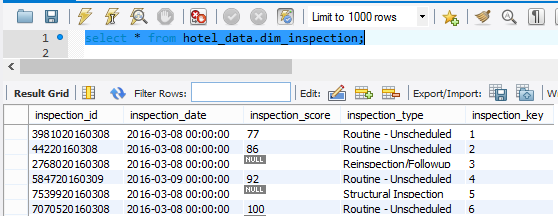
**1.) Dim\_Businesslocation**

****

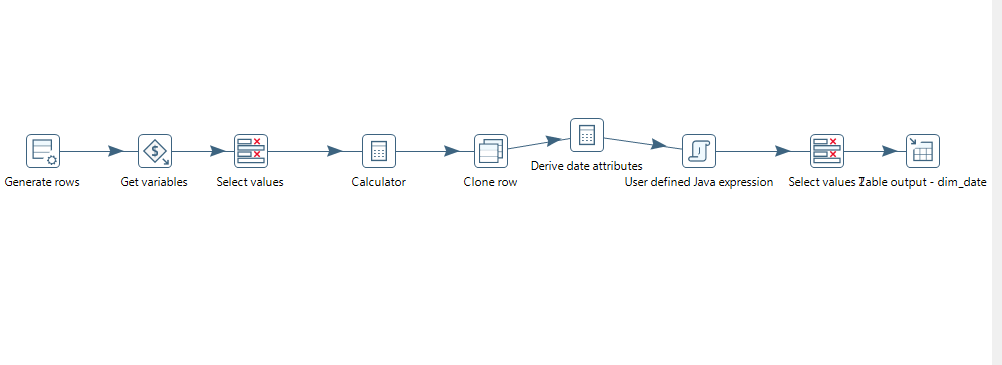
****

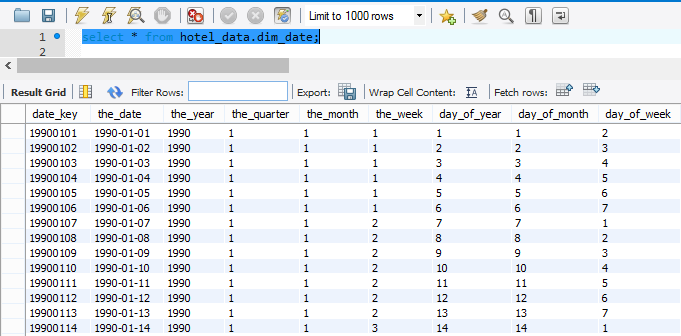
**2.) Dim\_inspection**

****

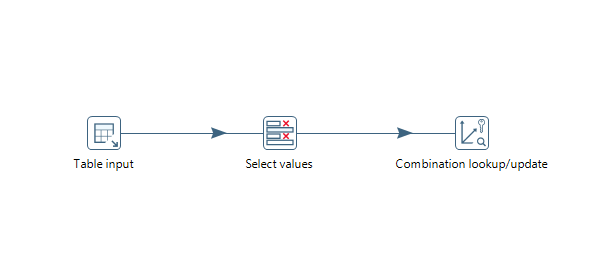
****

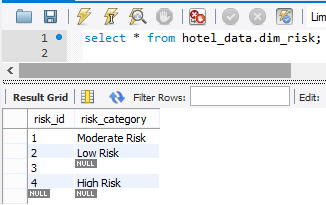
3.) Dim\_date



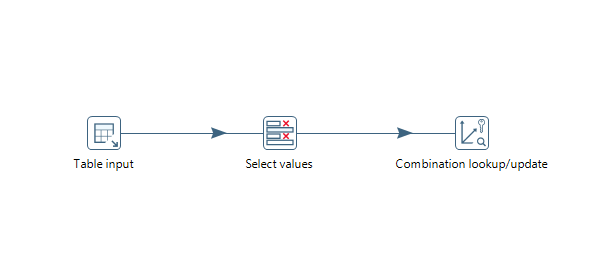


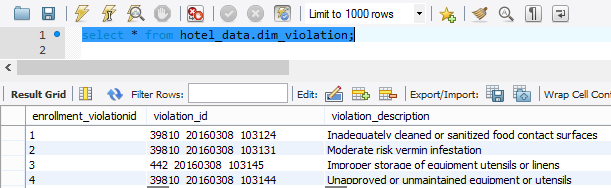
4.)Dim\_Risk



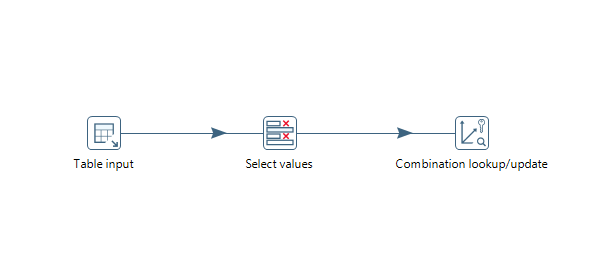


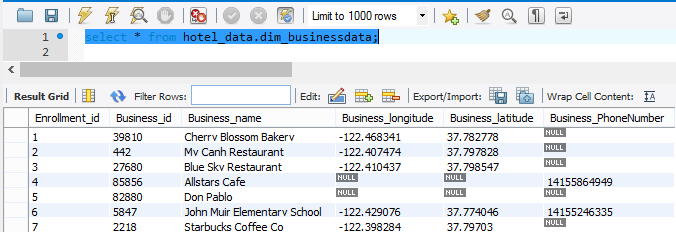
5.)Dim\_Violation





6.)Dim\_Businessdata





7.)Hoteldata\_Fact\_Build

