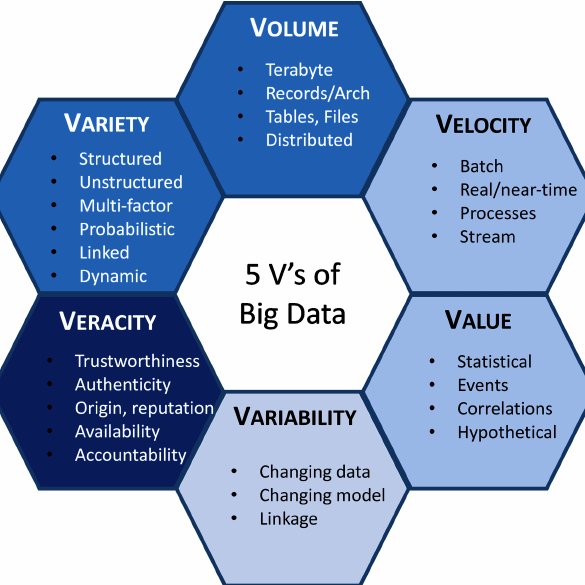
# DWH (Data Warehouse)

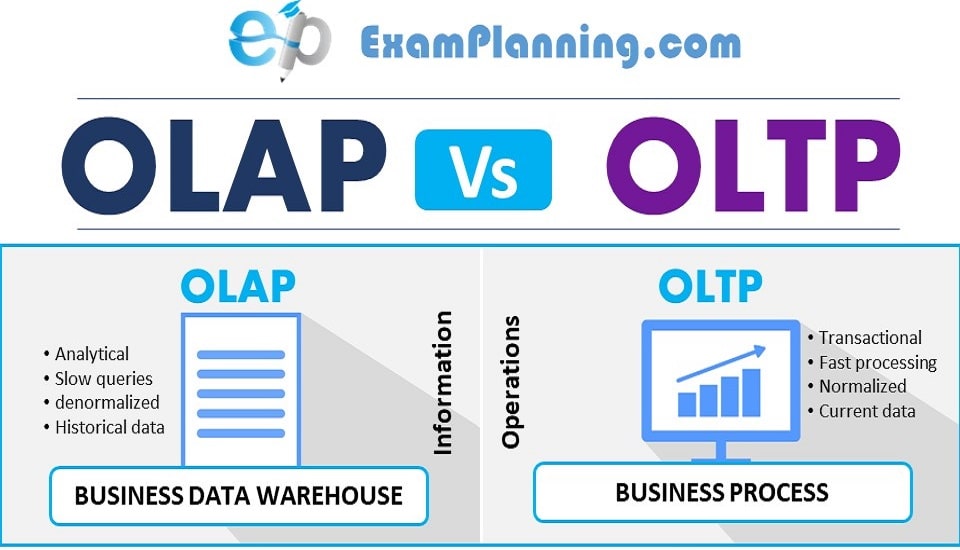
The 5Vs



## DB: e.g. SQL Server

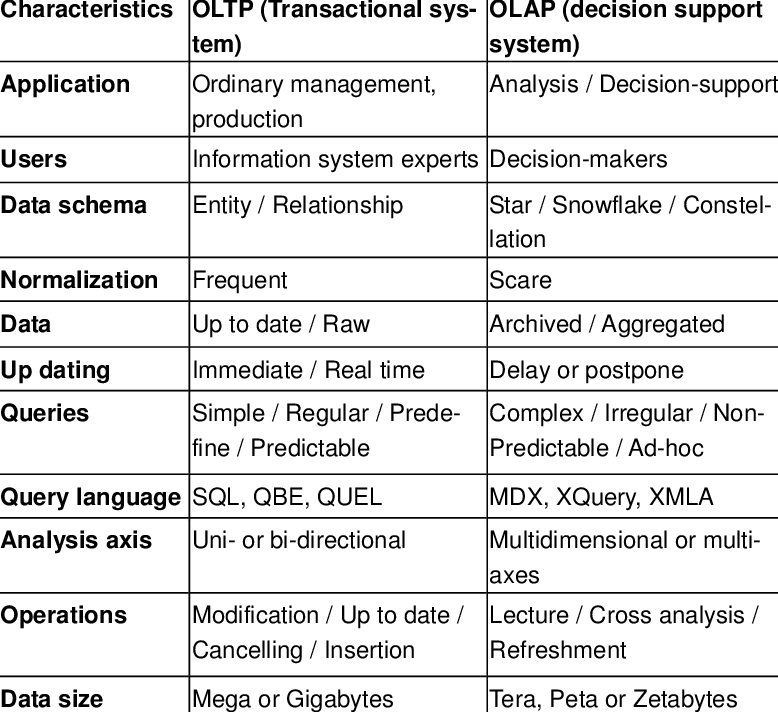
DWH: Real Time, combination of DBs, schema (structured), schema on write

Big Data: schema on read



Table

Description automatically generated



## The Evolution:

3NF 🡪 WH 🡪 Big Data (CDS, RT, DL) 🡪 Hybrid 🡪 DWH

3NF (3rd Normal Form) [http://www.gitta.info/LogicModelin/en/html/DataConsiten\_Norm3NF.html]

Warehouse

Common Data Service

Real Time

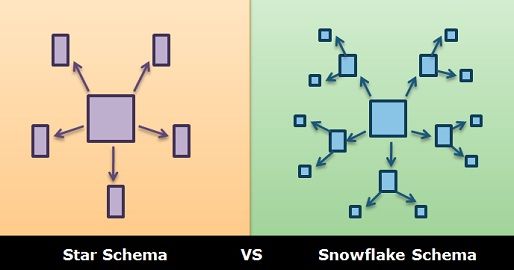
Data Lake

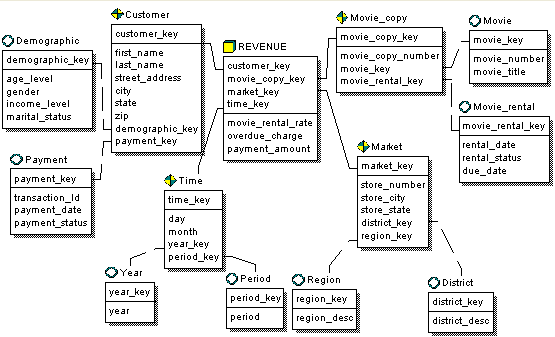
## DWH Schemas:

Dimensions and Facts together are the application for Business.

Two common schemas are:

1. Snowflake:
   1. Fact table 🡪 Dimensions 🡪 Sub-dimensions
2. Star:
   1. Fact table 🡪 Dimensions





## DWH Architecture Example:

## Example of a Fact Table:

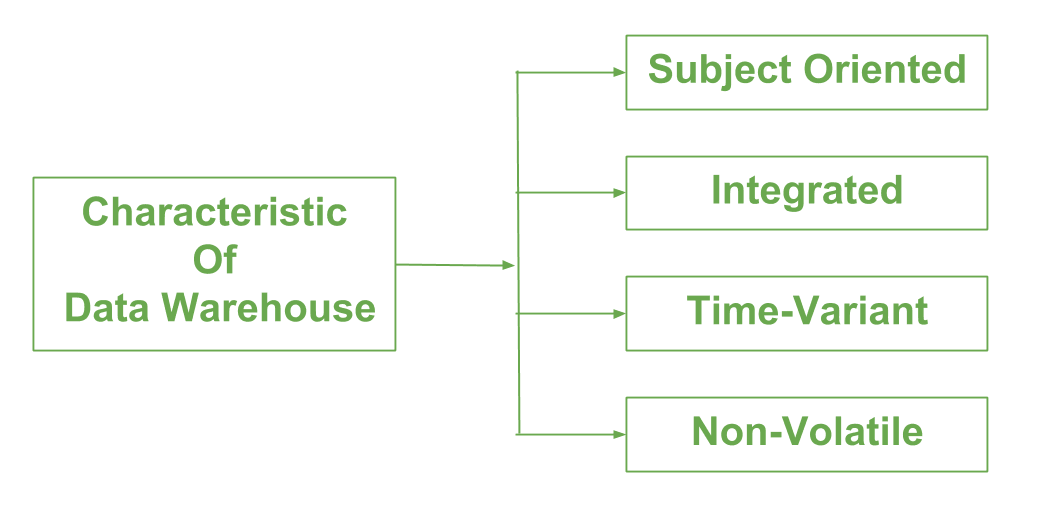
Let’s take example of a throughput table, we will the following columns:

|  |  |
| --- | --- |
| Traffic | Measure |
| Duration | Measure |
| Session Length | Measure |
| Date\_ID | Dimension (noun) |
| Location\_ID | Dimension (noun) |
| MSISDN | Dimension (noun) |
| Tech\_ID | Dimension (noun) |
| Surrogate\_ID | The Primary Key |

In this case we will have Star schema as follows:

## Basics of DWH:

DWH is an Information System having data from one or multiple sources allowing users to access a unified version of truth for timely business decision-making, reporting, and forecasting.



Graphical user interface, text, application, email

Description automatically generated

ER Model: Entity – Relationship Model

### DWH Design Approaches:

1. Kimball:
   * Bottom-up Approach. Data marts are built and integrated into a DWH.
   * Advantages:
     1. Fast to construct & query.
     2. No Normalization involved.
     3. Star schema is easy to comprehend by most users.
     4. Less resources, simple management, focused on business areas & processes.
     5. Less expert resources are required.
   * Disadvantages:
     1. Data is not entirely integrated, idea of ‘single source of truth’ is lost.
     2. Redundancy makes it prone to irregularities.
     3. Alteration and Updates may cause performance issues.
     4. Cannot handle all BI reporting needs.
2. Inmon:
   * Top-Down Approach. A DWH is built having all the data, data marts are created out of it as required.
   * Advantages:
     1. Unified source of Truth with all data integrated at one place.
     2. Low redundancy makes ETL simple and less susceptible to errors.
     3. Simplifies business processes.
     4. Flexible, easier to update or alter / add.
     5. Can handle diverse BI reporting requirements.
   * Disadvantages:
     1. Complexity increases with time.
     2. More skilled resources are required.
     3. Initial setup and delivery are time consuming.
     4. Additional ETL step added, as data marts are created after DWH.
3. Comparison:
   * Reporting needs:
     1. Organization-wide and integrated reporting 🡪 Inmon
     2. Reporting focused on the business process or team 🡪 Kimball
   * Project Deadline:
     1. If delivery is time critical 🡪 Inmon
     2. If time for delivery is less 🡪 Kimball
   * Recruitment Plan:
     1. Large team of experts 🡪 Inmon
     2. Relatively smaller team 🡪 Kimball
   * Frequent changes:
     1. Reporting needs to be changed quickly 🡪 Inmon
     2. Stable reporting needs 🡪 Kimball
   * Organizational Principles:
     1. Understands the need of EDW and can bear the expenses 🡪 Inmon
     2. Just need a solution to improve reporting 🡪 Kimball

Graphical user interface, text, email

Description automatically generated

SAP Universe:

* Connecting Data Sources with BusinessObjects (Reporting, Queries, Dashboards, etc.).
* SAP Universe Designer: A tool used by BusinessObjects developers to create semantic layer between database and reporting tools.

Factless Fact Table: <https://www.zentut.com/data-warehouse/factless-fact-table/>

Dimensionality model: <https://www.guru99.com/dimensional-model-data-warehouse.html>

* A technique optimized for data storage in a Data warehouse.
* Dimensional vs Relational Models.

SCD – Slowly Changing Dimensions:

* A Slowly Changing Dimension (SCD) is a dimension that stores and manages both current and historical data over time in a data warehouse. It is considered and implemented as one of the most critical ETL tasks in tracking the history of dimension records

Table

Description automatically generated

Any System has two parts:

1. Hardware
2. Technology / Logic

Teradata = Spark

60%

Teradata Factory Training:

1. Aster
2. Hybrid
3. Normal

HDFS

HDFS

HDFS

YARN (Yet Another Resource Negotiator) – the OS of Hadoop

Hive

Impala

RAM, Storage, Processor

128MB, 1024MB

Blocks = 1024 / 128 = 8 Blocks

.orc

.parquet

Rep\_fact = 2

Node Master > App Master

Map Reduce (MR) 🡪 Pig, Sark (Sql, DF, SAS) 🡪 Impala / Hive