Google Developer Student Clubs

Data Engineering Workshop

Tawfik Y. Tawfik

LinkedIn

Google Developer Student Clubs

Data Engineering Workshop

GDSC

Session #2: Introduction to Data Warehousing

Data Warehousing

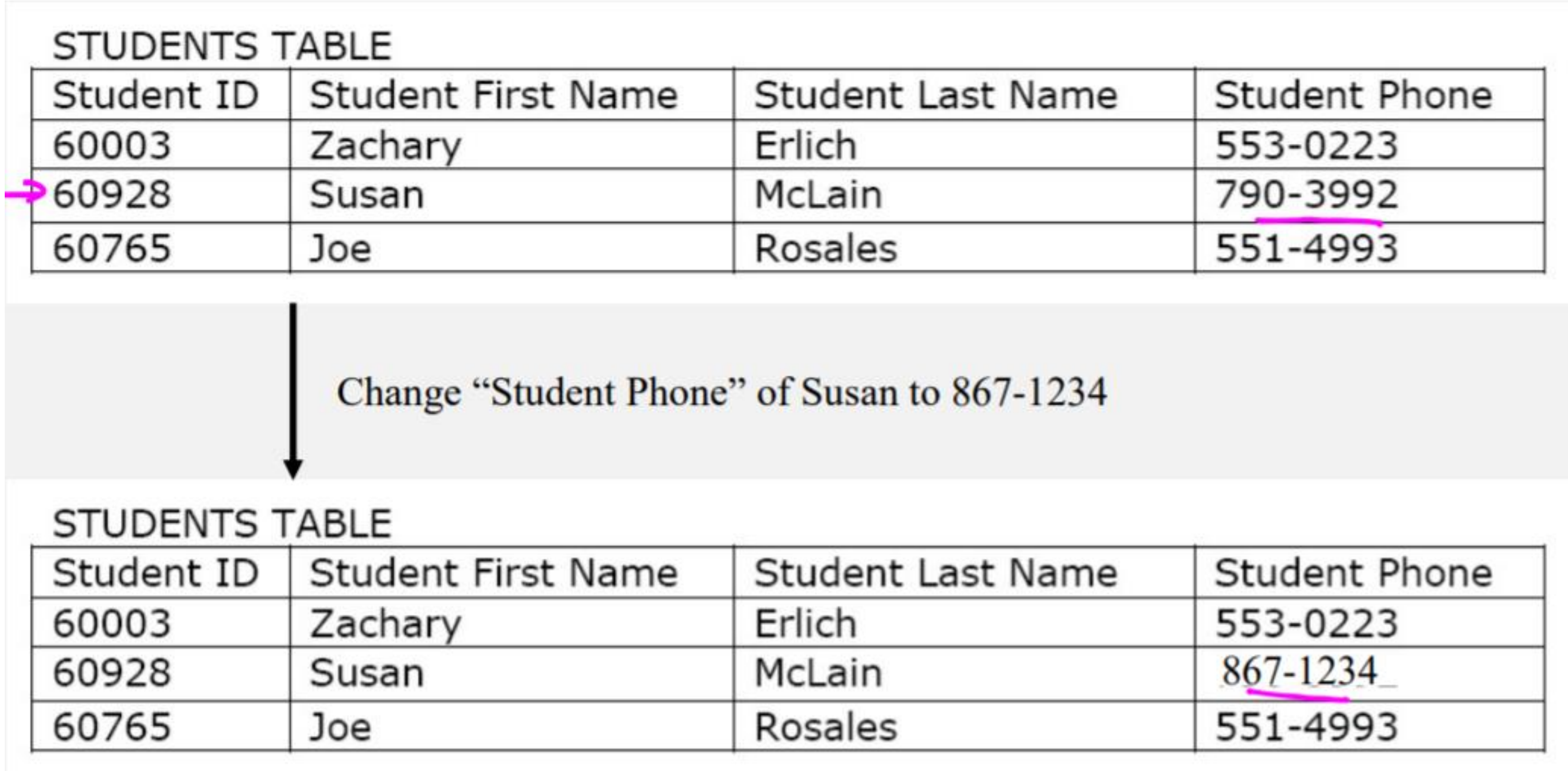
Introduction

- Data warehousing implements the process to access heterogeneous data sources; filter, transform the data; and store the data in a structure that is easy to access, understand, and use.
- The data is then ready for querying, reporting, and analysis.
- Example: The bank needs a clear view of the customers, accounts, and transactions data over time in all sites for making a correct decision.

Data Warehousing

Introduction

- In the operational systems we have no historical data.
- For decision making we need to keep the history.



STUDENTS TABLE

Student ID	Student First Name	Student Last Name	Student Phone
60003	Zachary	Erlich	553-0223
60928	Susan	McLain	790-3992
60765	Joe	Rosales	551-4993

Change "Student Phone" of Susan to 867-1234

STUDENTS TABLE

Student ID	Student First Name	Student Last Name	Student Phone
60003	Zachary	Erlich	553-0223
60928	Susan	McLain	867-1234
60765	Joe	Rosales	551-4993

Data Warehousing

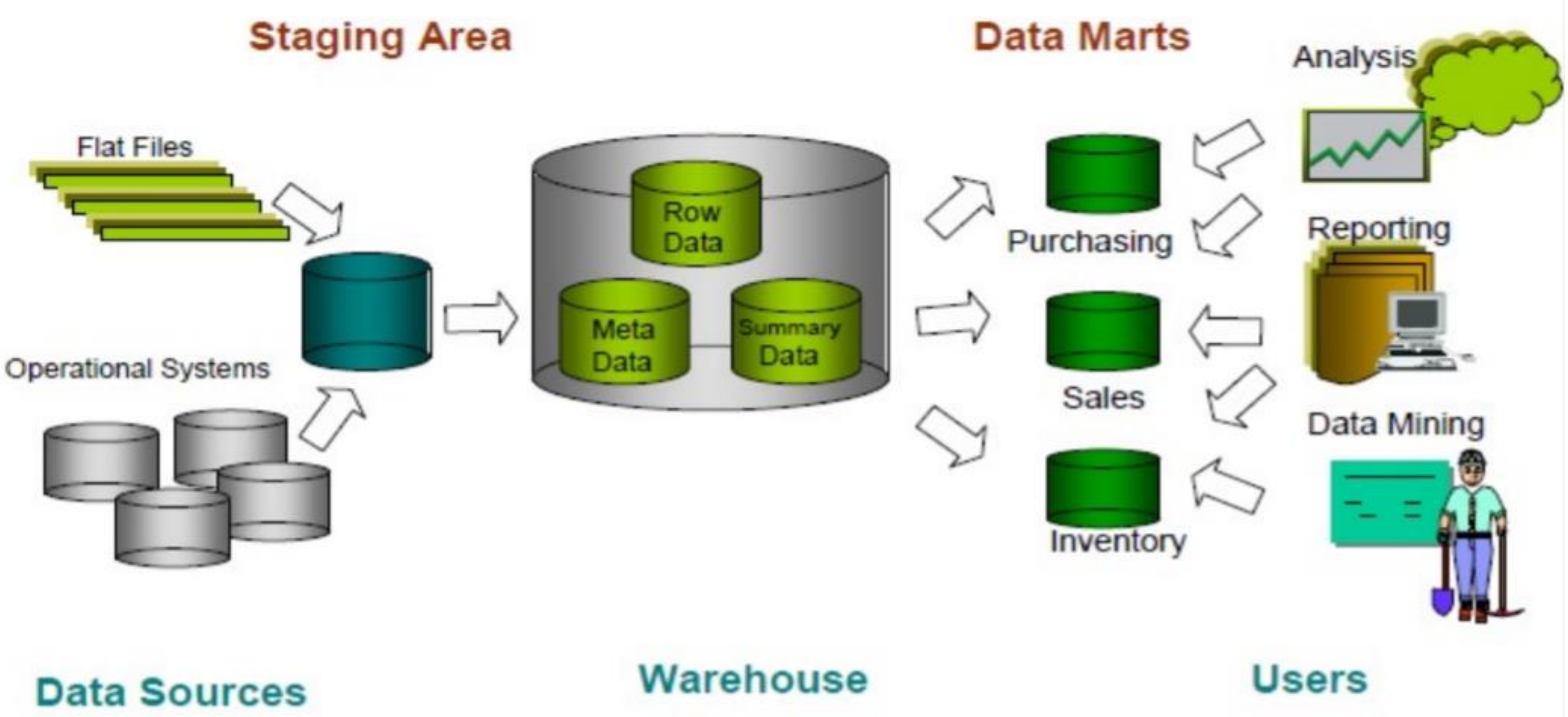
Introduction

- This database would be used for archiving, and it would be larger in size than transactional databases, but its design would make it optimal to run reports that would enable large organizations to plan and make decisions.

What is a DW (Brief Idea)?

- DW provides an excellent approach for transforming the large amounts of data that exist in organizations into useful and reliable information to support the decision-making process.
- A DW provides the base for powerful data analysis as data mining and multidimensional analysis.

Data Warehousing



Google Developer Student Clubs

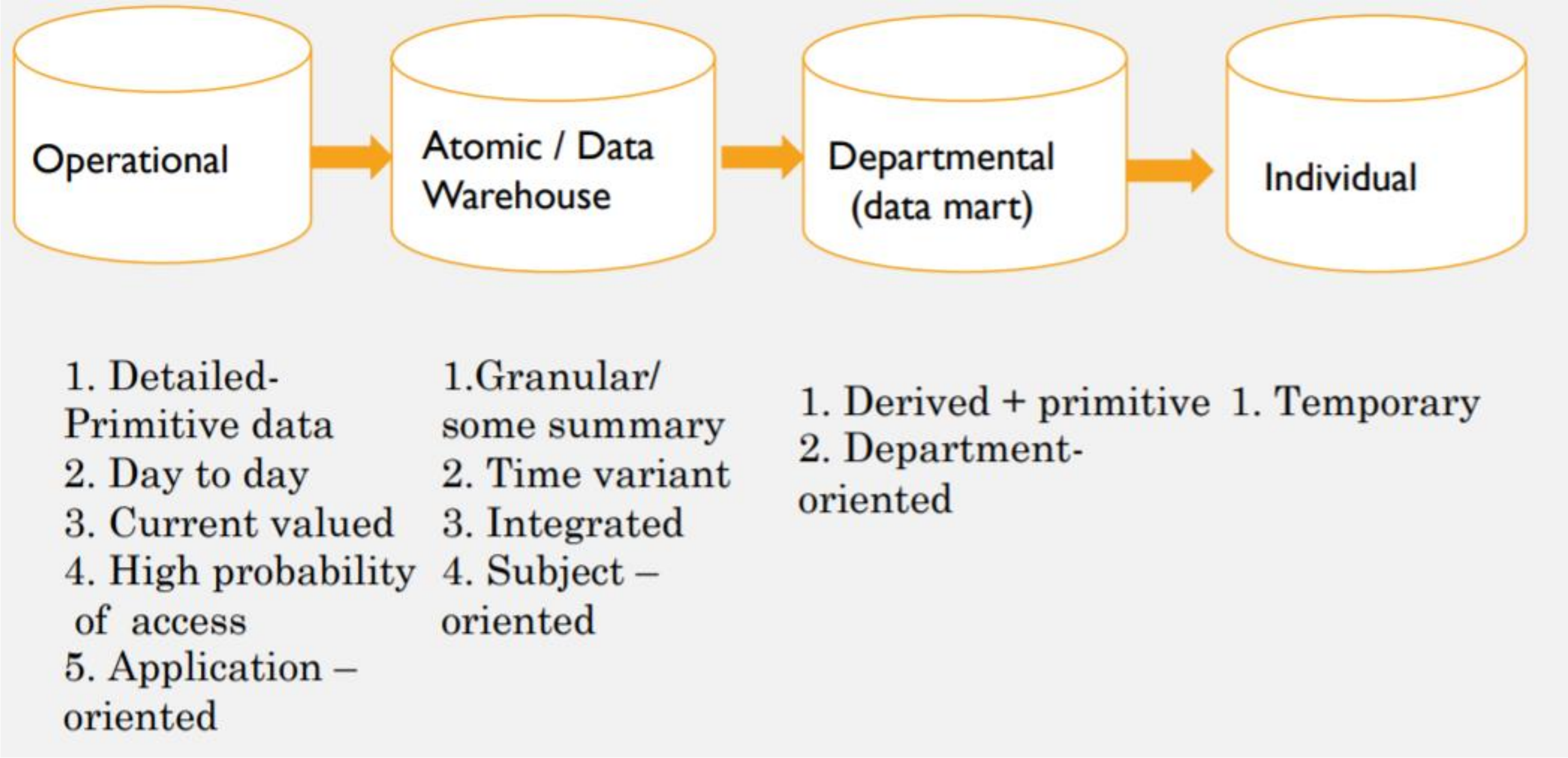
Data Warehouse

Definition

- A data warehouse is a subject oriented, integrated, nonvolatile, and time variant collection of data in support of management’s decisions.
- **Subject:** The rule of thumb is that subject areas are what the business wants to talk "about" or the "nouns" of the business. => Finance, HR, Sales, Marketing, etc.
- **Integrated:** Data in the same format and structure. **=>** Gender, Date, Address, etc.
- **Non-Volatile:** Once entered the warehouse, data should not change, loaded and accessed but not updated, which result in historical data.
- **Time Variant:** Every unit of data in the DW is accurate as of a moment of time.

Data Warehouse

Architecture Explained



1. Detailed-Primitive data
2. Day to day
3. Current valued
4. High probability of access
5. Application – oriented

1.Granular/ some summary
2. Time variant
3. Integrated
4. Subject – oriented

1. Derived + primitive
2. Department-oriented

1. Temporary

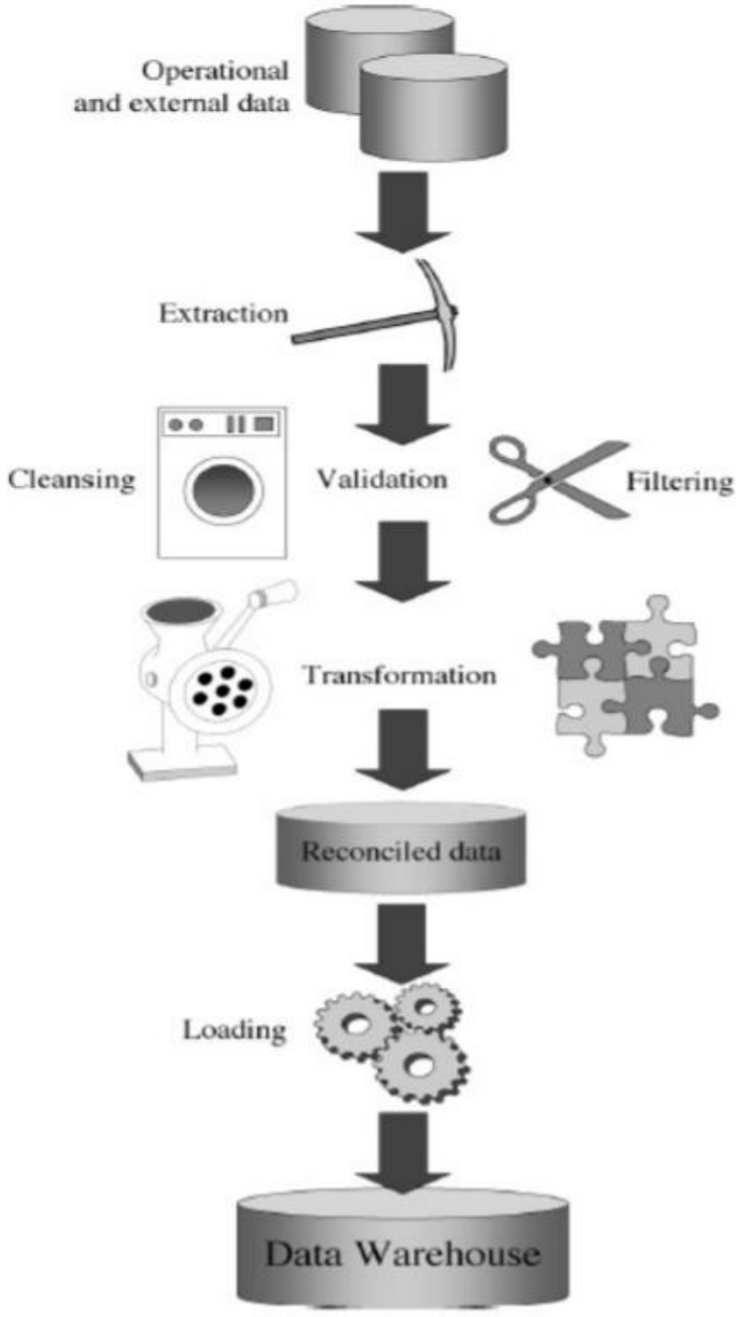
Any Questions?

Feel free to ask and interact

Data Warehouse

Staging Area (ETL)

- The data staging area is everything between the operational source system and data warehouse storage.
- ETL takes place once when a data warehouse is populated for the first time, then it occurs every time the data warehouse is regularly updated or based on an event.
- ETL consists of 3 separate phases: **extraction** (or capture), **cleansing** (or cleaning or scrubbing) and **transformation**, and **loading**.



Data Warehouse

ETL In Details

- **Extraction:** Extracting means reading and understanding the source data and copying the data needed for the data warehouse into the staging area for further manipulation.
- **Extraction methods:** 1. Get the whole table every time. 2. Get it incrementally (changes only).
- **Method #1:** SELECT * FROM table1
- **Method #2:** Based on 5 ways: 1. Using Timestamp column, 2. Using triggers, 3. Using transaction date, 4. using data partitioning, 5. Combination of the 4 steps.
- **1. Timestamp:** LSET = Last successful extract time, "SELECT * FROM table1 WHERE [date column] > LSET".

Data Warehouse

ETL In Details

- **2. Triggers:** Creating triggers for a table and when the trigger happened, data will be moved.
- **3. Transaction date:** Using the transaction time of the transaction table, to extract set of rows according to specific date.
- **4. Partitioning:** Partitioning by range over time, Extract records that fall within that time range.
- **5. May I need to combine all ways to achieve business need.**

Data Warehouse

ETL In Details

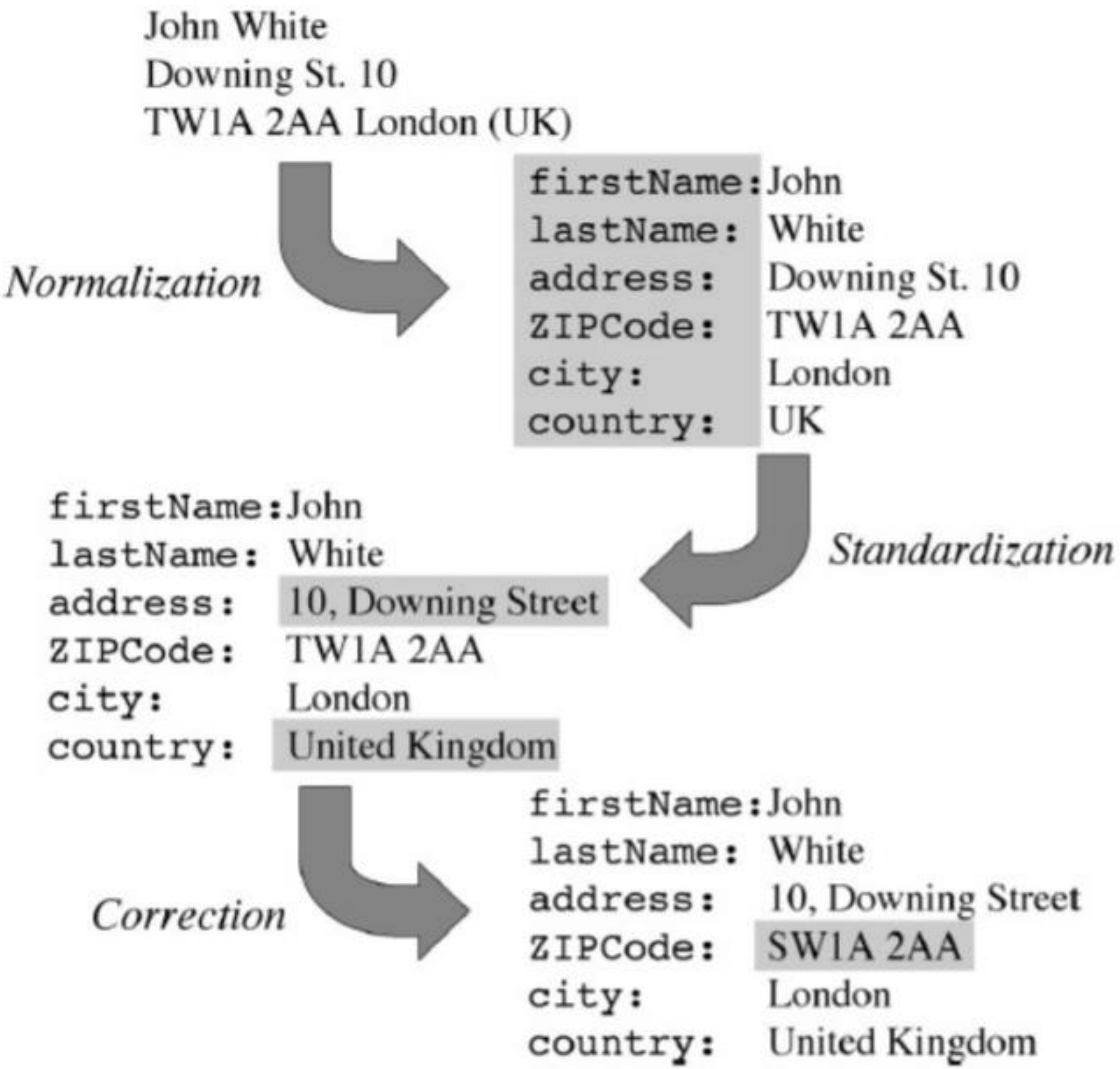
- **Data Cleaning & Transformation:** it is supposed to improve data quality.
- **Issues in Data:**
  - **Duplicate Data:** A patient is recorded many times in a hospital patient management system.
  - **Inconsistent values that are logically associated:** Such as addresses and ZIP codes.
  - **Missing data:** Such as customer's job.
  - **Unexpected use of fields:** SSN field could be used improperly to store office phone numbers.
  - **Impossible or wrong values:** 2 – 20 – 2021 (dd-MM-yyyy).
  - **Inconsistent values for a single entity because different practices were used:** Italy & I.

Data Warehouse

ETL In Details

- **Transformation:** This phase converts data from its operational source format into a specific data warehouse format.
- Standardization based on the needs.

- **Time Horizon:** Is the length of time data is represented in an environment, Time horizon of DW ~ 5-10 years. (Historical).



John White
Downing St. 10
TW1A 2AA London (UK)

Normalization

firstName: John
lastName: White
address: Downing St. 10
ZIPCode: TW1A 2AA
city: London
country: UK

Standardization

firstName: John
lastName: White
address: 10, Downing Street
ZIPCode: TW1A 2AA
city: London
country: United Kingdom

Correction

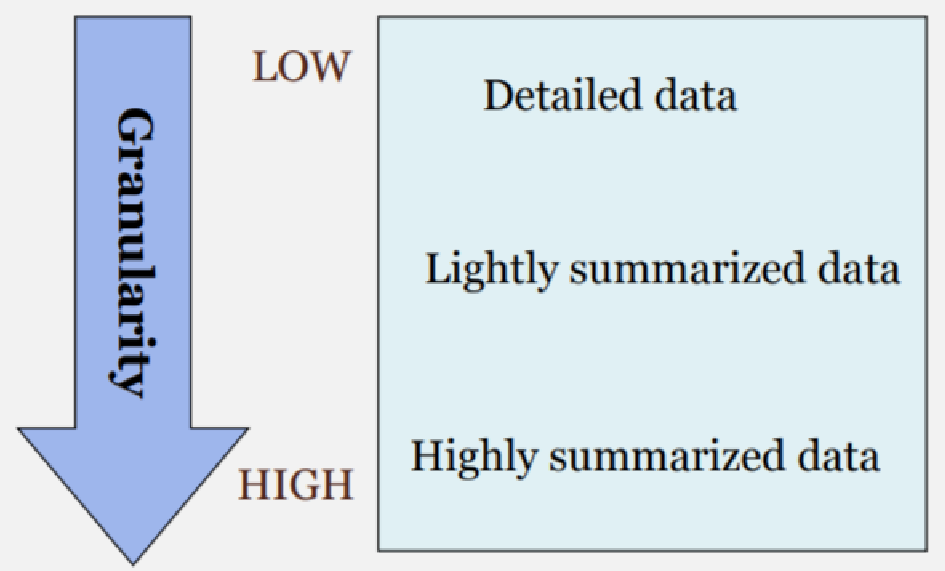
firstName: John
lastName: White
address: 10, Downing Street
ZIPCode: SW1A 2AA
city: London
country: United Kingdom

Break Time

Data Warehouse

How to Design

- We have main two concepts: **Granularity** & **Partitioning.**
- **Granularity:** The level of detail or summarization of the units of data in the data warehouse.



- **Gran** affects:

  1. Amount of data in DW.

  2. Types of queries we can ask.

Data Warehouse

Granularity

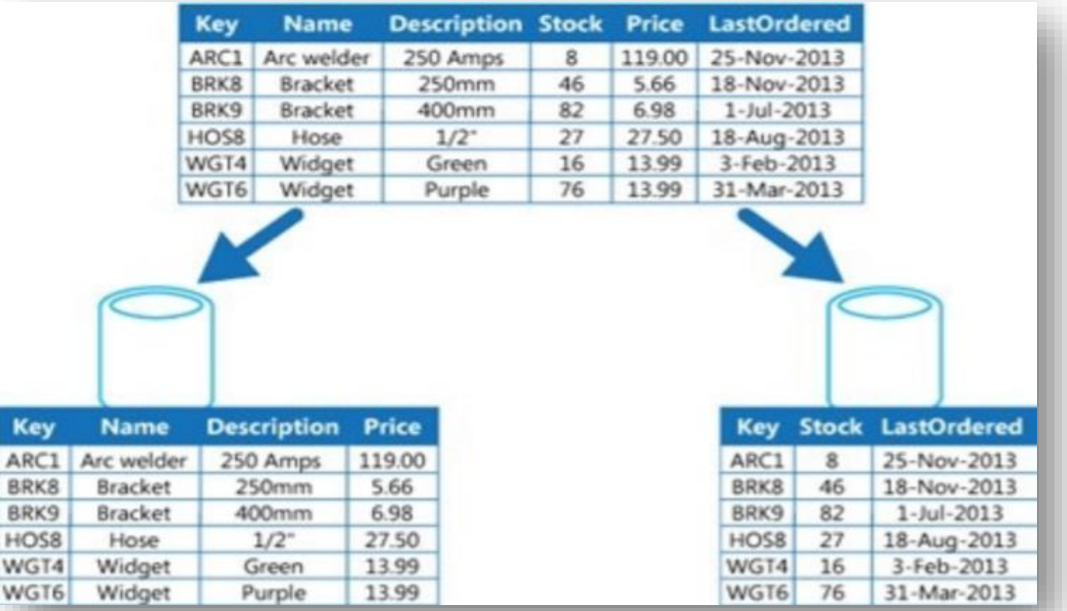
- **High Level of Detail = Low Granularity** => Answer any query, Large volume of data, More space. [Detailed Data]
- **Low Level of Detail = High Granularity** => Limited queries, Easy to manipulate, less space. [Summarized Data]
- **Example for (Detailed Data):** Details of every phone call for customers for a month,
- **Example for (Summarized Data):** Summary of phone calls for customer.

Which approach is better?

Data Warehouse

How to Design

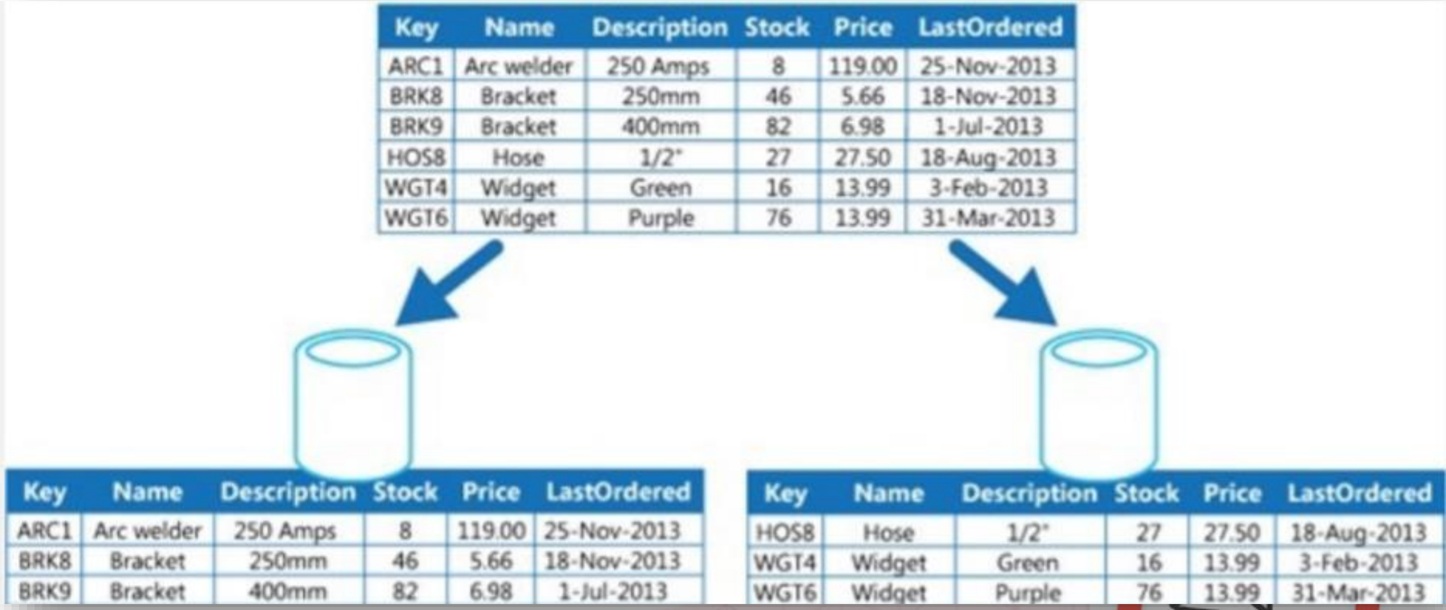
- **Partitioning:** is the breakup of data into separate physical units that can be handled independently.
- **Partitioning Types:** 1. Horizontal (Rows) – 2. Vertical (Columns).



Key	Name	Description	Stock	Price	LastOrdered
ARC1	Arc welder	250 Amps	8	119.00	25-Nov-2013
BRK8	Bracket	250mm	46	5.66	18-Nov-2013
BRK9	Bracket	400mm	82	6.98	1-Jul-2013
HOS8	Hose	1/2"	27	27.50	18-Aug-2013
WGT4	Widget	Green	16	13.99	3-Feb-2013
WGT6	Widget	Purple	76	13.99	31-Mar-2013

Key	Name	Description	Price
ARC1	Arc welder	250 Amps	119.00
BRK8	Bracket	250mm	5.66
BRK9	Bracket	400mm	6.98
HOS8	Hose	1/2"	27.50
WGT4	Widget	Green	13.99
WGT6	Widget	Purple	13.99

Key	Stock	LastOrdered
ARC1	8	25-Nov-2013
BRK8	46	18-Nov-2013
BRK9	82	1-Jul-2013
HOS8	27	18-Aug-2013
WGT4	16	3-Feb-2013
WGT6	76	31-Mar-2013

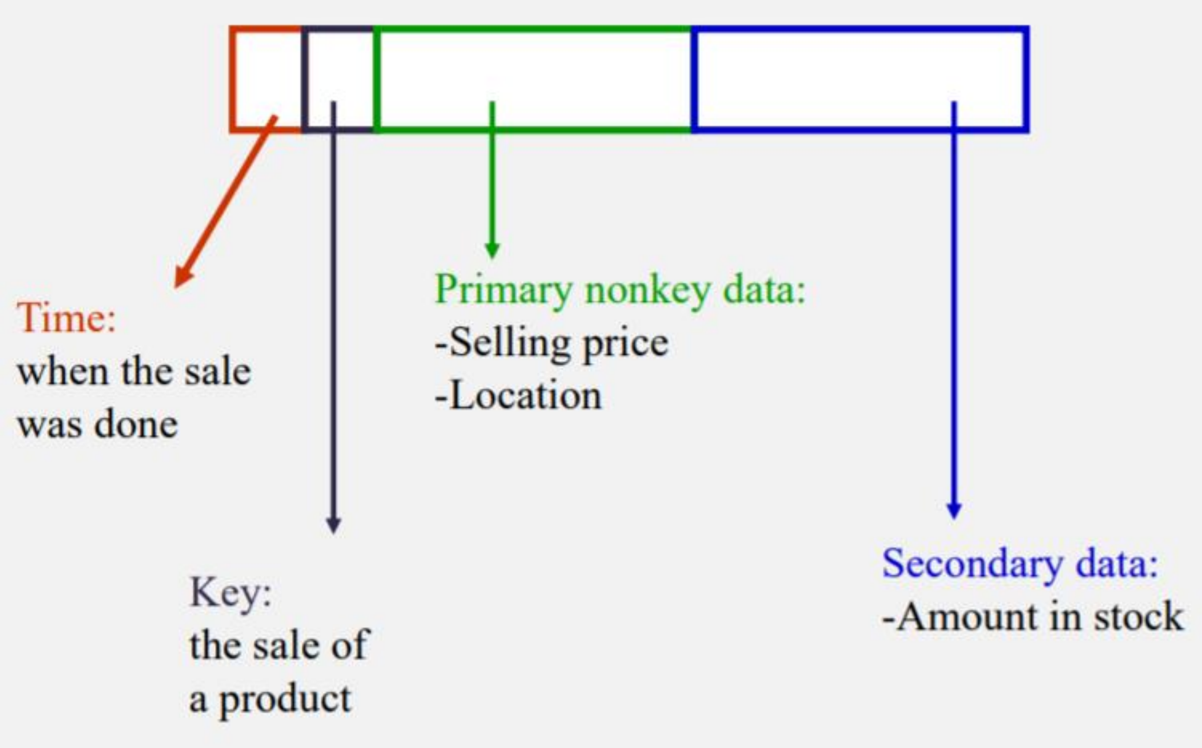


Key	Name	Description	Stock	Price	LastOrdered
ARC1	Arc welder	250 Amps	8	119.00	25-Nov-2013
BRK8	Bracket	250mm	46	5.66	18-Nov-2013
BRK9	Bracket	400mm	82	6.98	1-Jul-2013
HOS8	Hose	1/2"	27	27.50	18-Aug-2013
WGT4	Widget	Green	16	13.99	3-Feb-2013
WGT6	Widget	Purple	76	13.99	31-Mar-2013

Key	Name	Description	Stock	Price	LastOrdered
ARC1	Arc welder	250 Amps	8	119.00	25-Nov-2013
BRK8	Bracket	250mm	46	5.66	18-Nov-2013
BRK9	Bracket	400mm	82	6.98	1-Jul-2013

Key	Name	Description	Stock	Price	LastOrdered
HOS8	Hose	1/2"	27	27.50	18-Aug-2013
WGT4	Widget	Green	16	13.99	3-Feb-2013
WGT6	Widget	Purple	76	13.99	31-Mar-2013

Data Warehouse



Snapshots

- Snapshots are a result of **event occurring.**
- DW is a series of **Snapshots.**
- **Types of Triggering Events:**

  1. Activity Generated Events: Ex: Reaching a certain amount of balance.

  2. Time Generated Events: Ex: End of Day, End of Week, etc.

- **Snapshot Components:** Look at the image above.

Data Warehouse

Modeling

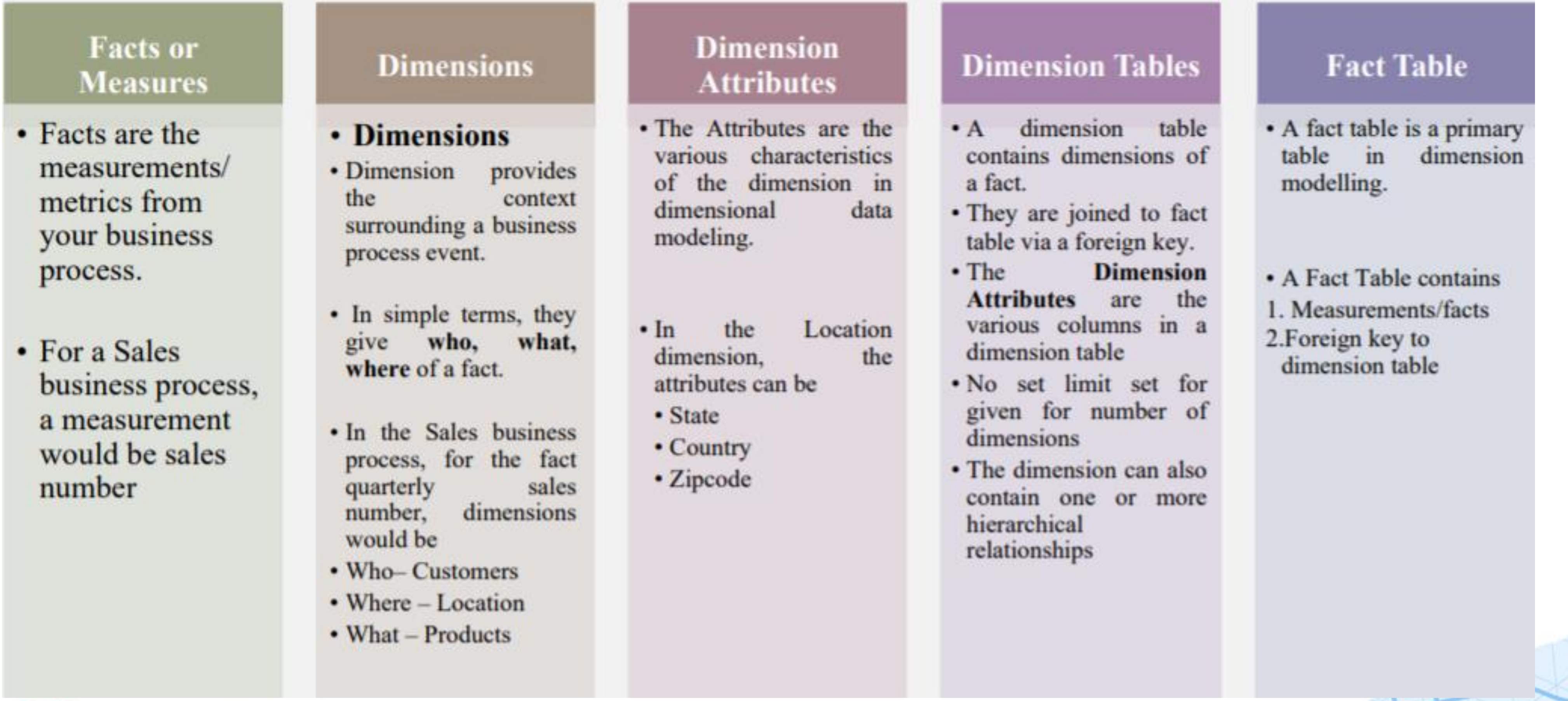
- A model is an abstraction and reflection of the real world.
- Modeling gives us the ability to visualize what we cannot yet realize.
- The data model plays the role of a guideline, or plan, to implement the DW.

Dimensional Modeling Technique

- Is especially useful for summarizing, rearranging the data and presenting views of the data to support data analysis.
- Dimensional modeling focuses on numeric data, such as values, counts, weights, balances, and occurrences.
- Has 3 main concepts: Facts, Dimensions, and Measures.

Data Warehouse

Elements of the Dimensional Model



Facts or Measures

- Facts are the measurements/ metrics from your business process.
- For a Sales business process, a measurement would be sales number

Dimensions

- **Dimensions**
- Dimension provides the context surrounding a business process event.
- In simple terms, they give **who, what, where** of a fact.
- In the Sales business process, for the fact quarterly sales number, dimensions would be
- Who– Customers
- Where – Location
- What – Products

Dimension Attributes

- The Attributes are the various characteristics of the dimension in dimensional data modeling.
- In the Location dimension, the attributes can be
- State
- Country
- Zipcode

Dimension Tables

- A dimension table contains dimensions of a fact.
- They are joined to fact table via a foreign key.
- The **Dimension Attributes** are the various columns in a dimension table
- No set limit set for given for number of dimensions
- The dimension can also contain one or more hierarchical relationships

Fact Table

- A fact table is a primary table in dimension modelling.
- A Fact Table contains

1. Measurements/facts
2. Foreign key to dimension table

Any Questions?

Feel free to ask and interact

Data Warehouse

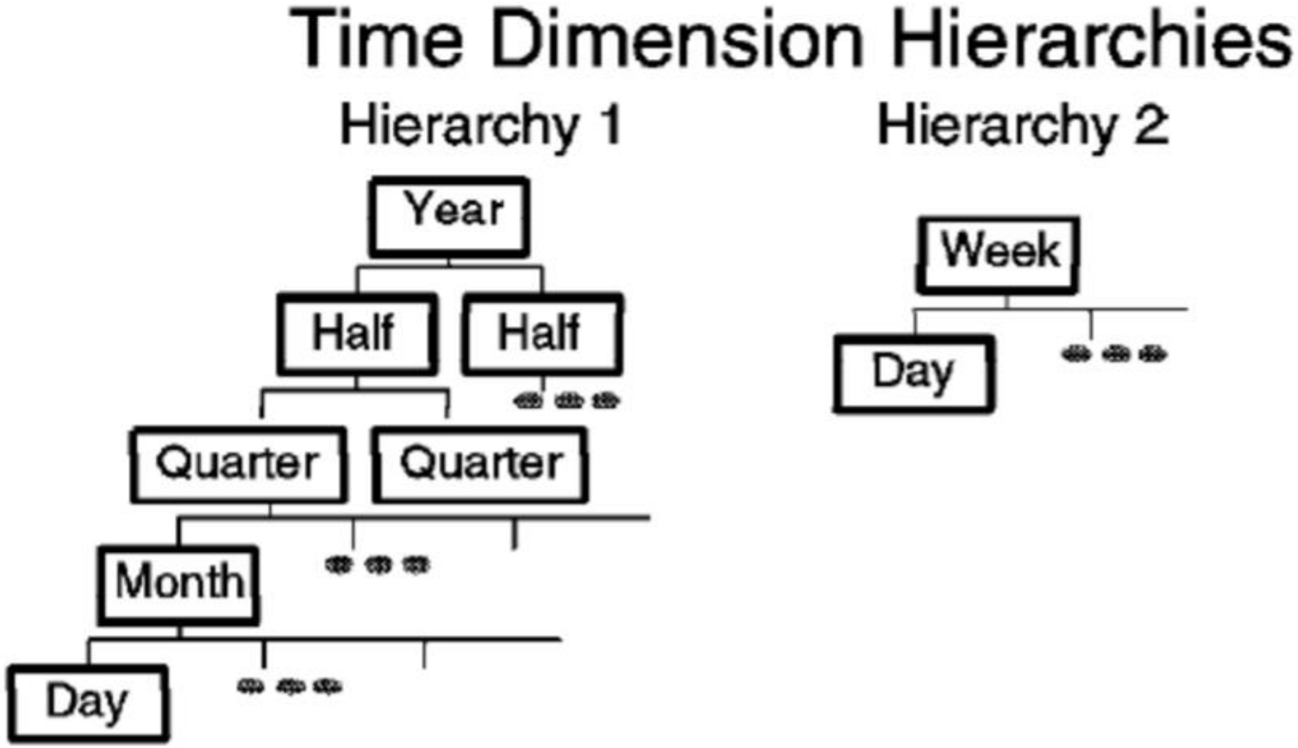
Dimensional Model

- **1. Fact:** is a collection of related data items, consisting of measures, each fact typically represents a business item, a business transaction, or an event that can be used in analyzing the business or business processes,
- **2. Dimension:** is a collection of members or units of the same type of views, In a dimensional model, every data point in the fact table is associated with one and only one member from each of the multiple dimensions.
- **For example**, in a database for analyzing all sales of products, common dimensions could be Time - Location/region – Customers -Salesperson

Data Warehouse

Continue Dimensional Model

- **Dimension Members (Attributes):** All cities, regions, and countries make up a geography dimension.
- **Dimension Hierarchies:** We can arrange the members of a dimension into one or more hierarchies, Each hierarchy can also have multiple hierarchy levels.



Data Warehouse

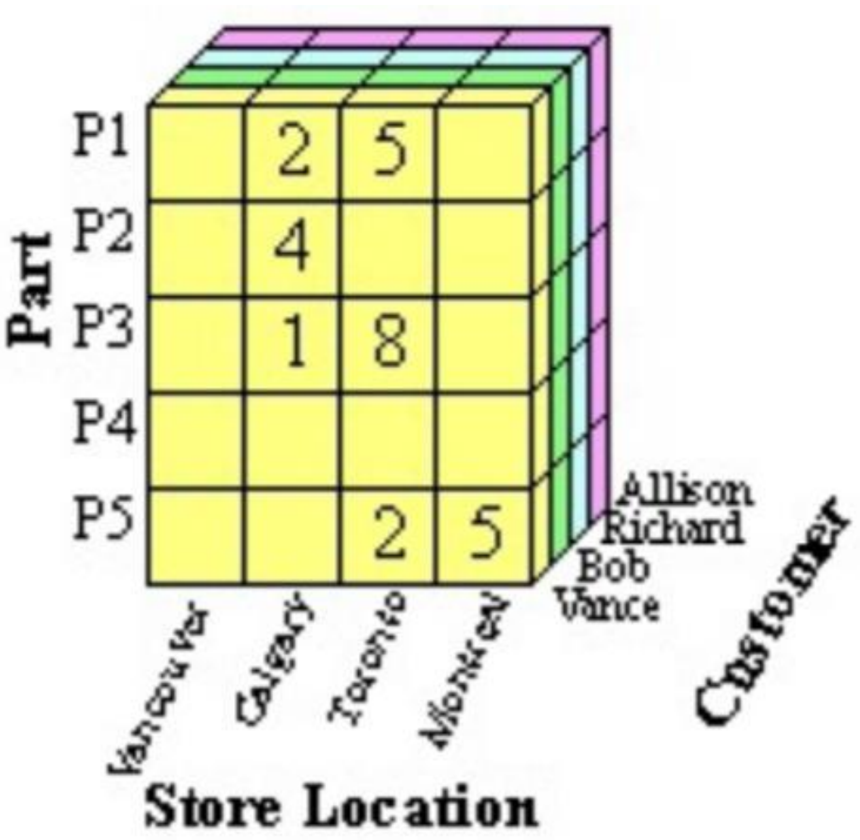
Continue Dimensional Model

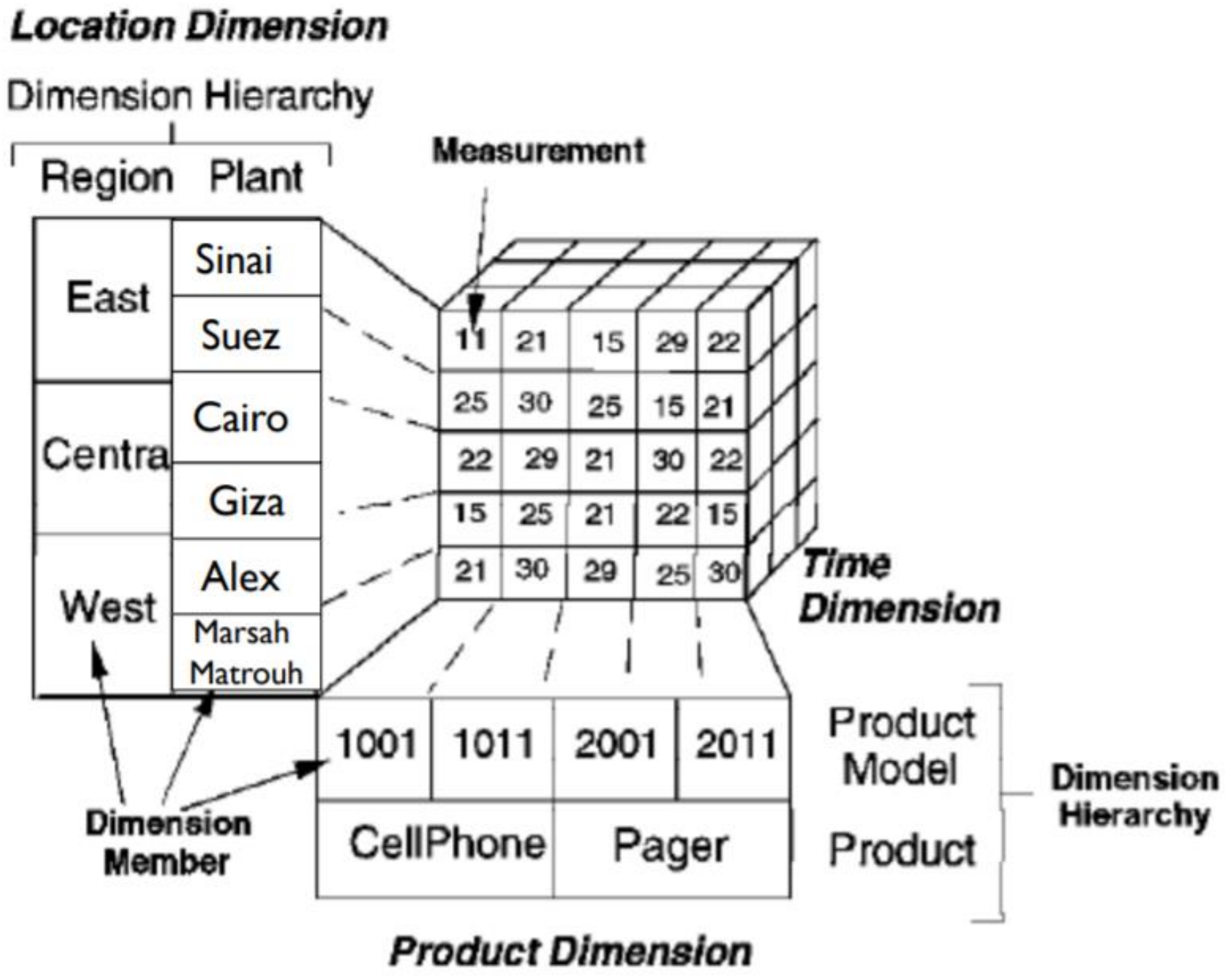
- **Measure:** is a numeric attribute of a fact, representing the performance or behavior of the business relative to the dimensions, for example, measures are the sales in money, the sales
- volume, the quantity supplied, the supply cost, the transaction amount, etc.

- **To Visualize the Dimensional Model:** The most popular way of visualizing a dimensional model is to draw a cube, we can represent a three-dimensional model using a cube, Usually a dimensional model consists of more than three dimensions and is referred to as a hypercube.

Data Warehouse

Examples of the Dimensional Model





Data Warehouse

Design the DW using the Star Schema (Multi-Dimensional Models)

- **4-Step Dimension Design Process:**
  - Select the business process to model.
  - Declare the grain of the business process.
  - Choose the dimensions that apply to each fact table row.
  - Identify the numeric facts that will populate each fact table row.

- We'll apply those steps in designing phase. (Session #4)

Practice and Fun!

We're almost done

I wish you all enjoyed.

We’ll have a feedback form, and questions form now.

Thanks

GDSC

Data Engineering Workshop