COGNOS

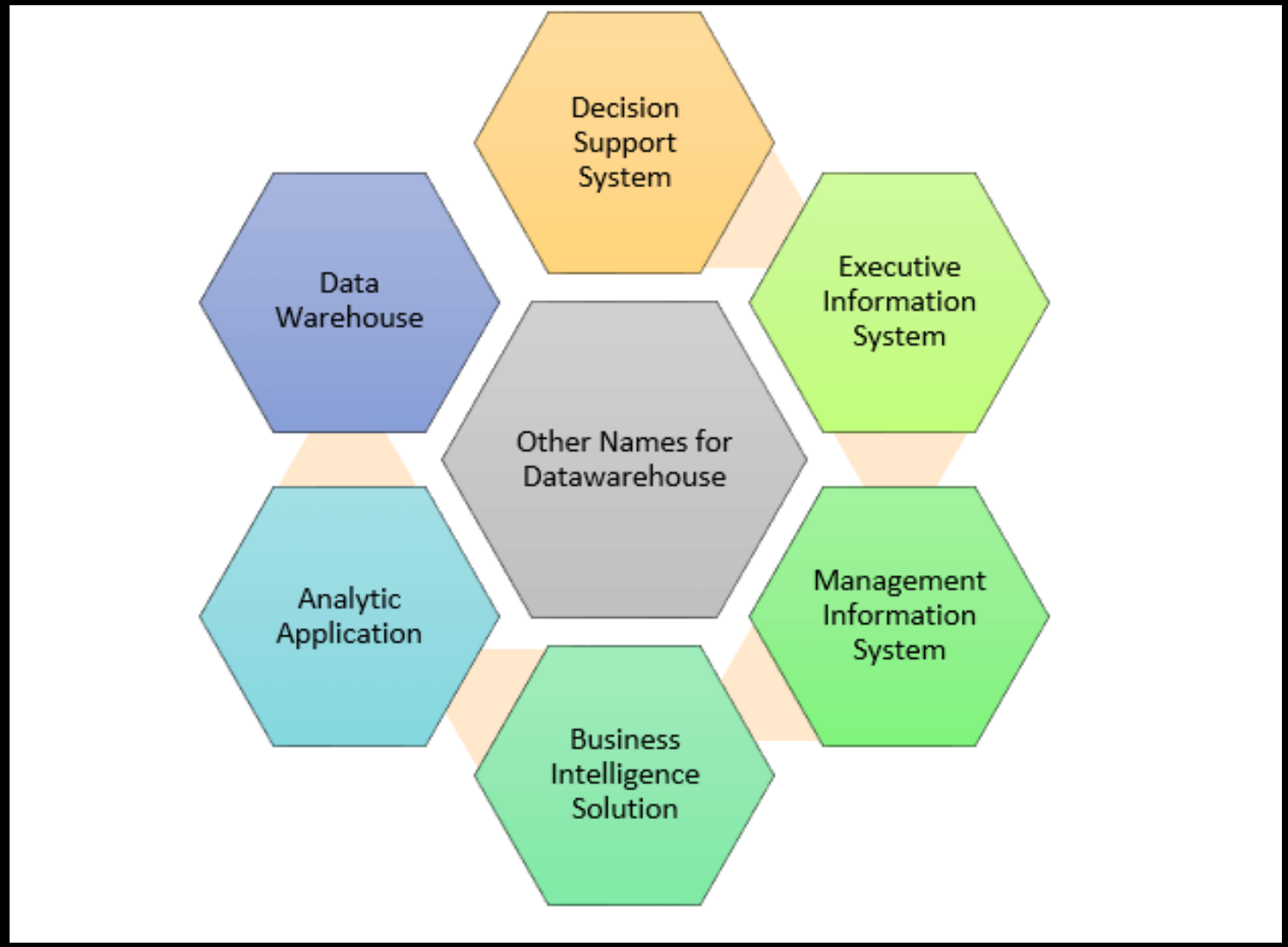
IBM

WHAT IS DATA WAREHOUSING?

- A data warehousing is a technique for collecting and managing data from varied sources to provide meaningful business insights. It is a blend of technologies and components which allows the strategic use of data.
- It is electronic storage of a large amount of information by a business which is designed for query and analysis instead of transaction processing. It is a process of transforming data into information and making it available to users in a timely manner to make a difference.

- The decision support database (Data Warehouse) is maintained separately from the organization's operational database. However, the data warehouse is not a product but an environment. It is an architectural construct of an information system which provides users with current and historical decision support information which is difficult to access or present in the traditional operational data store.
- The data warehouse is the core of the BI system which is built for data analysis and reporting.

- Data warehouse system is also known by the following name:
- Decision Support System (DSS)
- Executive Information System
- Management Information System
- Business Intelligence Solution
- Analytic Application
- Data Warehouse


Decision Support System
Data Warehouse
Executive Information System
Other Names for Datawarehouse
Analytic Application
Management Information System
Business Intelligence Solution

HOW DATA WAREHOUSE WORKS?

- Data Warehouse works as a central repository where information arrives from one or more data sources. Data flows into a data warehouse from the transactional system and other relational databases.
- Data may be:
- Structured
- Semi-structured
- Unstructured data

- The data is processed, transformed, and ingested so that users can access the processed data in the Data Warehouse through Business Intelligence tools, SQL clients, and spreadsheets. A data warehouse merges information coming from different sources into one comprehensive database.
- By merging all of this information in one place, an organization can analyze its customers more holistically. This helps to ensure that it has considered all the information available. Data warehousing makes data mining possible. Data mining is looking for patterns in the data that may lead to higher sales and profits.

TYPES OF DATA WAREHOUSE

- **Three main types of Data Warehouses are:**

1. Enterprise Data Warehouse:

Enterprise Data Warehouse is a centralized warehouse. It provides decision support service across the enterprise. It offers a unified approach for organizing and representing data. It also provide the ability to classify data according to the subject and give access according to those divisions.

2. Operational Data Store:

Operational Data Store, which is also called ODS, are nothing but data store required when neither Data warehouse nor OLTP systems support organizations reporting needs. In ODS, Data warehouse is refreshed in real time. Hence, it is widely preferred for routine activities like storing records of the Employees.

3. Data Mart:

- A data mart is a subset of the data warehouse. It specially designed for a particular line of business, such as sales, finance, sales or finance. In an independent data mart, data can collect directly from sources.

:

COMPONENTS OF DATA WAREHOUSE

- **Four components of Data Warehouses are:**
- **Load manager:** Load manager is also called the front component. It performs with all the operations associated with the extraction and load of data into the warehouse. These operations include transformations to prepare the data for entering into the Data warehouse.

- **Warehouse Manager:** Warehouse manager performs operations associated with the management of the data in the warehouse. It performs operations like analysis of data to ensure consistency, creation of indexes and views, generation of denormalization and aggregations, transformation and merging of source data and archiving and baking-up data.

- **Query Manager:** Query manager is also known as backend component. It performs all the operation operations related to the management of user queries. The operations of this Data warehouse components are direct queries to the appropriate tables for scheduling the execution of queries.

- **End-user access tools:**
- This is categorized into five different groups like 1. Data Reporting 2. Query Tools 3. Application development tools 4. EIS tools, 5. OLAP tools and data mining tools.

WHO NEEDS DATA WAREHOUSE?

- Data warehouse is needed for all types of users like:
- Decision makers who rely on mass amount of data
- Users who use customized, complex processes to obtain information from multiple data sources.
- It is also used by the people who want simple technology to access the data
- It also essential for those people who want a systematic approach for making decisions.
- If the user wants fast performance on a huge amount of data which is a necessity for reports, grids or charts, then Data warehouse proves useful.
- Data warehouse is a first step If you want to discover 'hidden patterns' of data-flows and groupings.

WHAT ARE THE MOST COMMON SECTORS WHERE DATA WAREHOUSE IS USED ?

Airline:

- In the Airline system, it is used for operation purpose like crew assignment, analyses of route profitability, frequent flyer program promotions, etc.

Banking:

- It is widely used in the banking sector to manage the resources available on desk effectively. Few banks also used for the market research, performance analysis of the product and operations.

Healthcare:

- Healthcare sector also used Data warehouse to strategize and predict outcomes, generate patient's treatment reports, share data with tie-in insurance companies, medical aid services, etc.

Public sector:

- In the public sector, data warehouse is used for intelligence gathering. It helps government agencies to maintain and analyze tax records, health policy records, for every individual.

Investment and Insurance sector:

- In this sector, the warehouses are primarily used to analyze data patterns, customer trends, and to track market movements.

Telecommunication:

- A data warehouse is used in this sector for product promotions, sales decisions and to make distribution decisions.

Hospitality Industry:

- This Industry utilizes warehouse services to design as well as estimate their advertising and promotion campaigns where they want to target clients based on their feedback and travel patterns.

ADVANTAGES & DISADVANTAGES OF DATAWAREHOUSE

- **Advantages of Data Warehouse:**
- Data warehouse allows business users to quickly access critical data from some sources all in one place.
- Data warehouse provides consistent information on various cross-functional activities. It is also supporting ad-hoc reporting and query.
- Data Warehouse helps to integrate many sources of data to reduce stress on the production system.
- Data warehouse helps to reduce total turnaround time for analysis and reporting.

- Restructuring and Integration make it easier for the user to use for reporting and analysis.
- Data warehouse allows users to access critical data from the number of sources in a single place. Therefore, it saves user's time of retrieving data from multiple sources.
- Data warehouse stores a large amount of historical data. This helps users to analyse different time periods and trends to make future predictions.

DISADVANTAGES OF DATA WAREHOUSE:

- Not an ideal option for unstructured data.
- Creation and Implementation of Data Warehouse is surely time confusing affair.
- Data Warehouse can be outdated relatively quickly
- Difficult to make changes in data types and ranges, data source schema, indexes, and queries.
- The data warehouse may seem easy, but actually, it is too complex for the average users.
- Despite best efforts at project management, data warehousing project scope will always increase.
- Sometime warehouse users will develop different business rules.
- Organisations need to spend lots of their resources for training and Implementation purpose.

DATA WAREHOUSE TOOLS

1. MarkLogic:

- MarkLogic is useful data warehousing solution that makes data integration easier and faster using an array of enterprise features. This tool helps to perform very complex search operations. It can query different types of data like documents, relationships, and metadata.
- http://developer.marklogic.com/products

2. Oracle:

- Oracle is the industry-leading database. It offers a wide range of choice of data warehouse solutions for both on-premises and in the cloud. It helps to optimize customer experiences by increasing operational efficiency.
- https://www.oracle.com/index.html

3. Amazon RedShift:

- Amazon Redshift is Data warehouse tool. It is a simple and cost-effective tool to analyze all types of data using standard SQL and existing BI tools. It also allows running complex queries against petabytes of structured data, using the technique of query optimization.
- https://aws.amazon.com/redshift/?nc2=h_m1

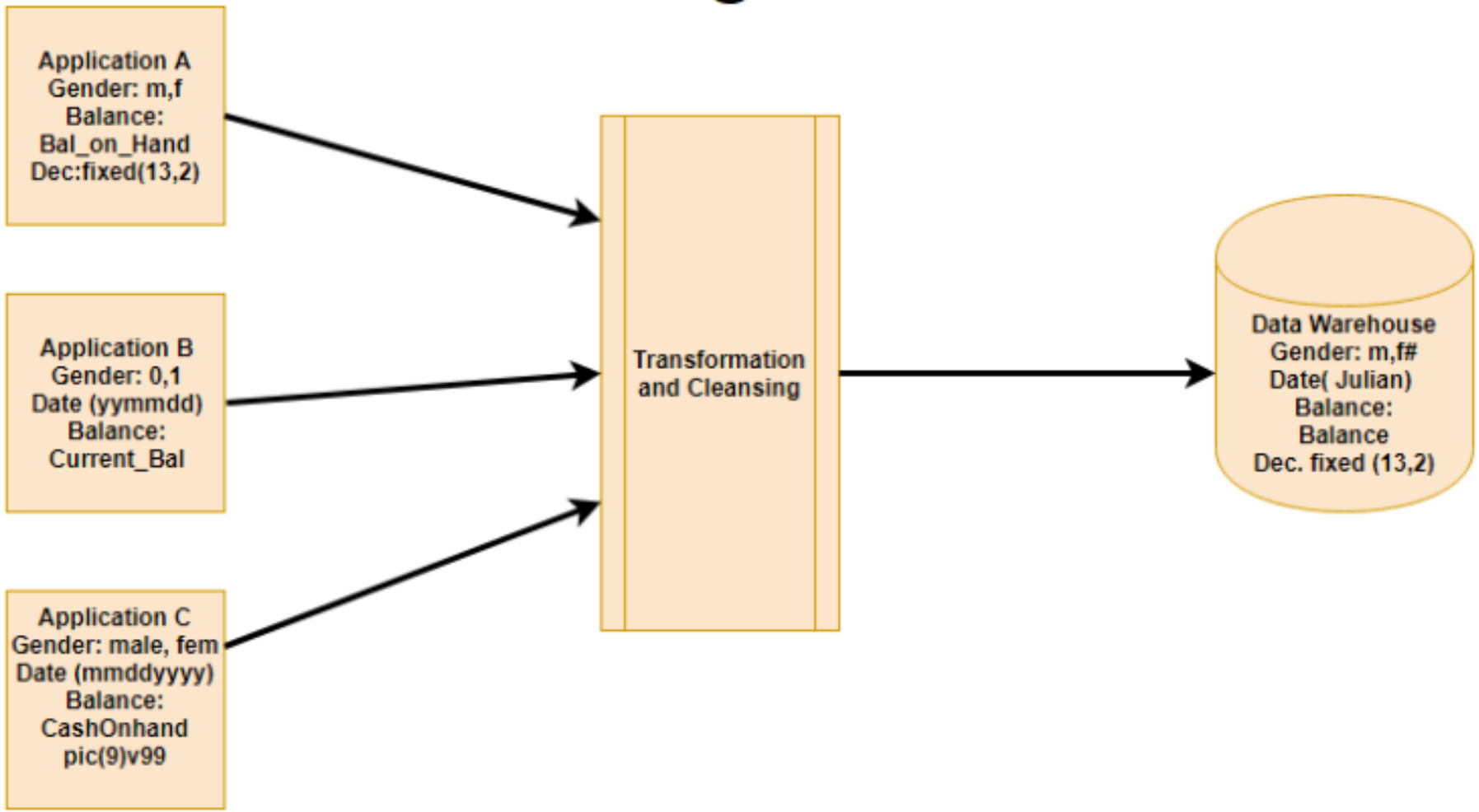
CHARACTERISTICS OF DATA WAREHOUSE

- A data warehouse has following characteristics:
- Subject-Oriented
- Integrated
- Time-variant
- Non-volatile

- **Subject-Oriented**
- A data warehouse is subject oriented as it offers information regarding a theme instead of companies' ongoing operations. These subjects can be sales, marketing, distributions, etc.
- A data warehouse never focuses on the ongoing operations. Instead, it put emphasis on modeling and analysis of data for **decision making**. It also provides a simple and concise view around the specific subject by excluding data which not helpful to support the decision process.

- **Integrated**
- In Data Warehouse, integration means the establishment of a common unit of measure for all similar data from the dissimilar database. The data also needs to be stored in the Datawarehouse in common and universally acceptable manner.
- A data warehouse is developed by integrating data from varied sources like a mainframe, relational databases, flat files, etc. Moreover, it must keep consistent naming conventions, format, and coding.
- This integration helps in effective analysis of data. Consistency in naming conventions, attribute measures, encoding structure etc. have to be ensured. Consider the following example:
-

Data Integration Issues



- In the above example, there are three different application labeled A, B and C. Information stored in these applications are Gender, Date, and Balance. However, each application's data is stored different way.
- In Application A gender field store logical values like M or F
- In Application B gender field is a numerical value,
- In Application C application, gender field stored in the form of a character value.
- Same is the case with Date and balance
- However, after transformation and cleaning process all this data is stored in common format in the Data Warehouse.

- **Time-Variant**
- The time horizon for data warehouse is quite extensive compared with operational systems. The data collected in a data warehouse is recognized with a particular period and offers information from the historical point of view. It contains an element of time, explicitly or implicitly.
- One such place where Datawarehouse data display time variance is in in the structure of the record key. Every primary key contained with the DW should have either implicitly or explicitly an element of time. Like the day, week month, etc.
- Another aspect of time variance is that once data is inserted in the warehouse, it can't be updated or changed.

- **Non-volatile**
- Data warehouse is also non-volatile means the previous data is not erased when new data is entered in it.
- Data is read-only and periodically refreshed. This also helps to analyze historical data and understand what & when happened. It does not require transaction process, recovery and concurrency control mechanisms.
- Activities like delete, update, and insert which are performed in an operational application environment are omitted in Data warehouse environment. Only two types of data operations performed in the Data Warehousing are
- Data loading
- Data access

DATA WAREHOUSE ARCHITECTURES

There are mainly three types of Datawarehouse Architectures: -

Single-tier architecture

- The objective of a single layer is to minimize the amount of data stored. This goal is to remove data redundancy. This architecture is not frequently used in practice.

Two-tier architecture

- Two-layer architecture separates physically available sources and data warehouse. This architecture is not expandable and also not supporting a large number of end-users. It also has connectivity problems because of network limitations.

Three-tier architecture

- This is the most widely used architecture.

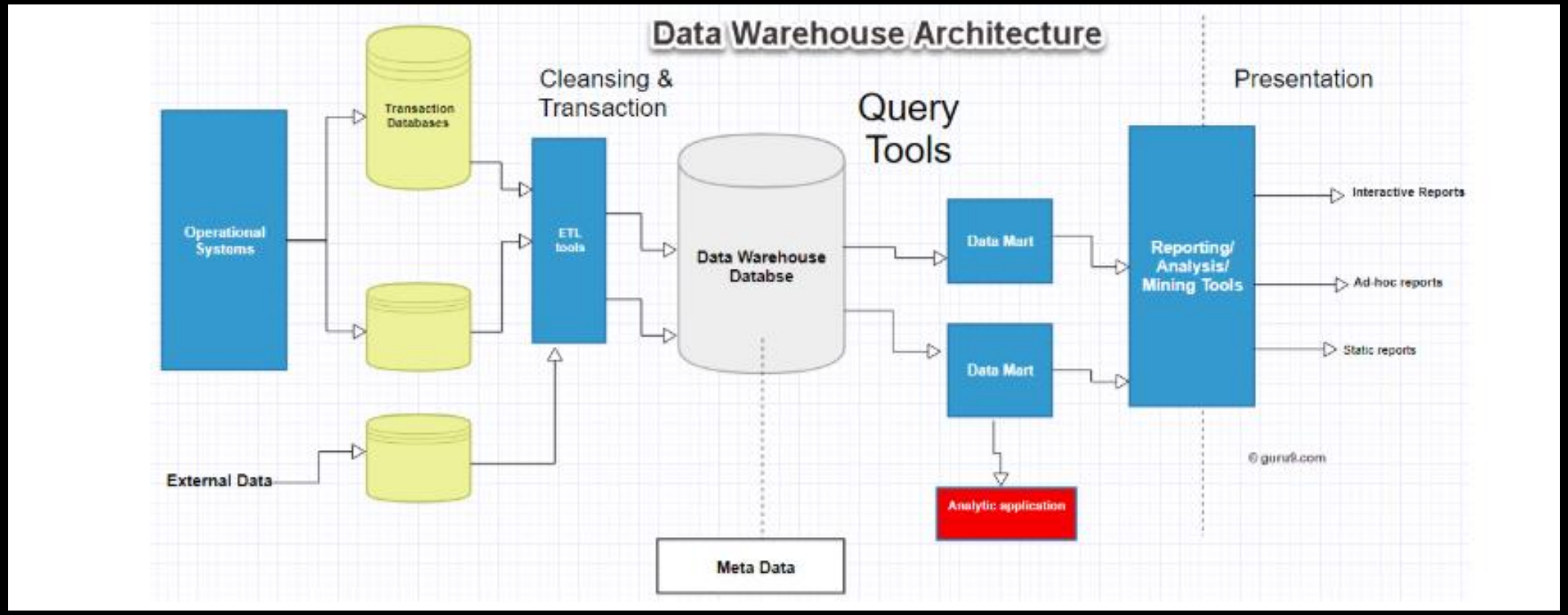
- It consists of the Top, Middle and Bottom Tier.

- **Bottom Tier:** The database of the Data warehouse servers as the bottom tier. It is usually a relational database system. Data is cleansed, transformed, and loaded into this layer using back-end tools.

- **Middle Tier:** The middle tier in Data warehouse is an OLAP server which is implemented using either ROLAP or MOLAP model. For a user, this application tier presents an abstracted view of the database. This layer also acts as a mediator between the end-user and the database.

- **Top-Tier:** The top tier is a front-end client layer. Top tier is the tools and API that you connect and get data out from the data warehouse. It could be Query tools, reporting tools, managed query tools, Analysis tools and Data mining tools.

DATAWAREHOUSE COMPONENTS



- The data warehouse is based on an RDBMS server which is a central information repository that is surrounded by some key components to make the entire environment functional, manageable and accessible.

There are mainly five components of Data Warehouse:

Data Warehouse Database

- The central database is the foundation of the data warehousing environment. This database is implemented on the RDBMS technology. Although, this kind of implementation is constrained by the fact that traditional RDBMS system is optimized for transactional database processing and not for data warehousing. For instance, ad-hoc query, multi-table joins, aggregates are resource intensive and slow down performance.

Sourcing, Acquisition, Clean-up and Transformation Tools (ETL)

- The data sourcing, transformation, and migration tools are used for performing all the conversions, summarizations, and all the changes needed to transform data into a unified format in the data warehouse. They are also called Extract, Transform and Load (ETL) Tools.

Metadata

- The name Meta Data suggests some high- level technological concept. However, it is quite simple. Metadata is data about data which defines the data warehouse. It is used for building, maintaining and managing the data warehouse.
- In the Data Warehouse Architecture, meta-data plays an important role as it specifies the source, usage, values, and features of data warehouse data. It also defines how data can be changed and processed. It is closely connected to the data warehouse.

- For example, a line in sales database may contain:

```
4030 KJ732 299.90
```

- This is a meaningless data until we consult the Meta that tell us it was
- Model number: 4030
- Sales Agent ID: KJ732
- Total sales amount of $2999.90
- Therefore, Meta Data are essential ingredients in the transformation of data into knowledge.
- Metadata helps to answer the following questions
- What tables, attributes, and keys does the Data Warehouse contain?
- Where did the data come from?
- How many times do data get reloaded?
- What transformations were applied with cleansing?

- **Query Tools**
- One of the primary objects of data warehousing is to provide information to businesses to make strategic decisions. Query tools allow users to interact with the data warehouse system.
- These tools fall into four different categories:
- Query and reporting tools
- Application Development tools
- Data mining tools
- OLAP tools

- **1. Query and reporting tools:**
- Query and reporting tools can be further divided into
- Reporting tools
- Managed query tools

- **Reporting tools:** Reporting tools can be further divided into production reporting tools and desktop report writer.
- Report writers: This kind of reporting tool are tools designed for end-users for their analysis.
- Production reporting: This kind of tools allows organizations to generate regular operational reports. It also supports high volume batch jobs like printing and calculating. Some popular reporting tools are Brio, Business Objects, Oracle, PowerSoft, SAS Institute.

Managed query tools:

- This kind of access tools helps end users to resolve snags in database and SQL and database structure by inserting meta-layer between users and database.

2. Application development tools:

- Sometimes built-in graphical and analytical tools do not satisfy the analytical needs of an organization. In such cases, custom reports are developed using Application development tools.

3. Data mining tools:

- Data mining is a process of discovering meaningful new correlation, pattens, and trends by mining large amount data. Data mining tools are used to make this process automatic.

4. OLAP tools:

- These tools are based on concepts of a multidimensional database. It allows users to analyse the data using elaborate and complex multidimensional views.

DATA WAREHOUSE BUS ARCHITECTURE

- Data warehouse Bus determines the flow of data in your warehouse. The data flow in a data warehouse can be categorized as Inflow, Up flow, Down flow, Outflow and Meta flow.
- While designing a Data Bus, one needs to consider the shared dimensions, facts across data marts.

- **Data Marts**
- A data mart is an access layer which is used to get data out to the users. It is presented as an option for large size data warehouse as it takes less time and money to build. However, there is no standard definition of a data mart is differing from person to person.
- In a simple word Data mart is a subsidiary of a data warehouse. The data mart is used for partition of data which is created for the specific group of users.
- Data marts could be created in the same database as the Datawarehouse or a physically separate Database.

ETL (EXTRACT, TRANSFORM, AND LOAD) PROCESS

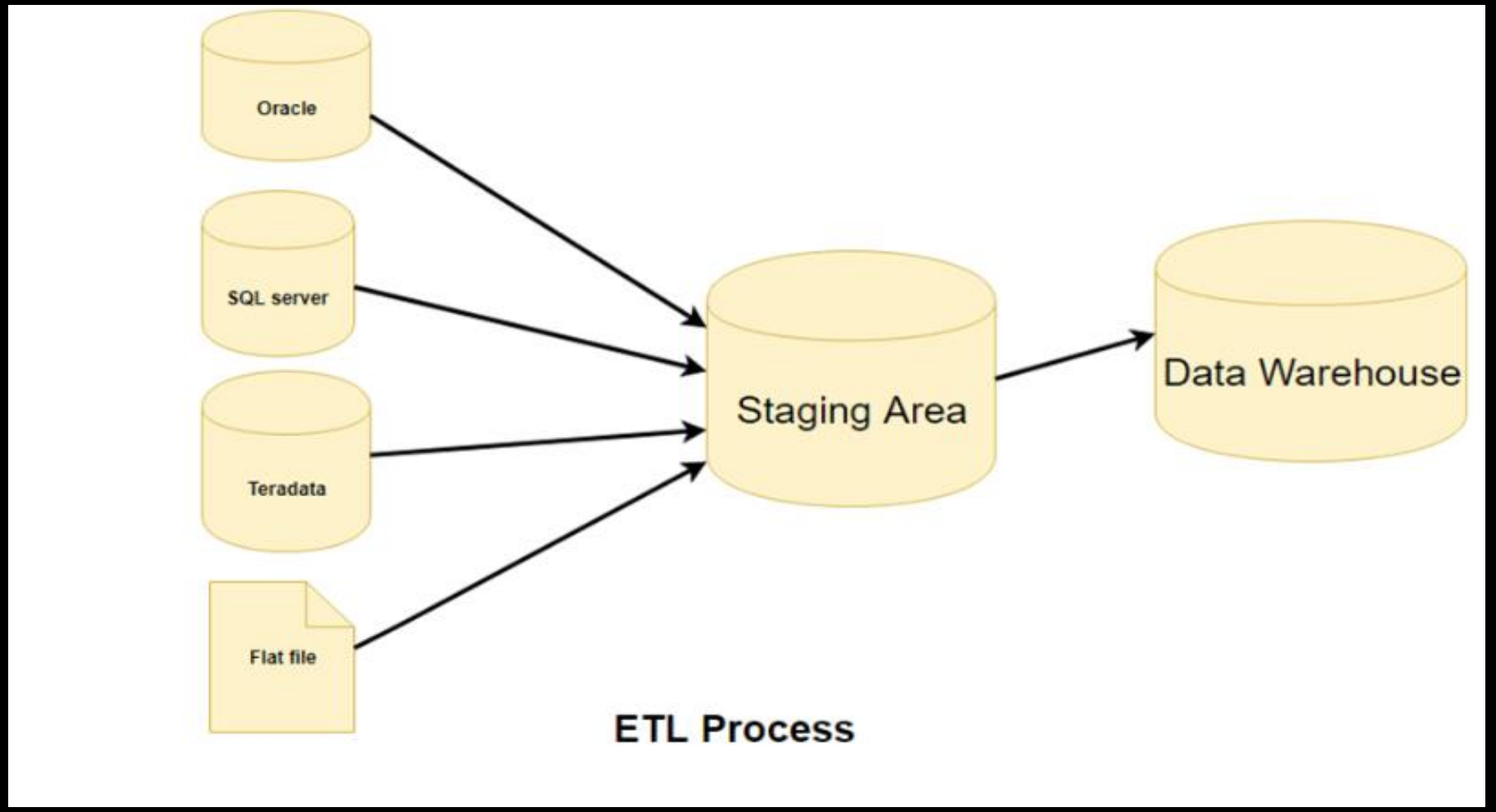
WHAT IS ETL?

- ETL is an abbreviation of Extract, Transform and Load. In this process, an ETL tool extracts the data from different RDBMS source systems then transforms the data like applying calculations, concatenations, etc. and then load the data into the Data Warehouse system.
- It's tempting to think a creating a Data warehouse is simply extracting data from multiple sources and loading into database of a Data warehouse. This is far from the truth and requires a complex ETL process. The ETL process requires active inputs from various stakeholders including developers, analysts, testers, top executives and is technically challenging.
- In order to maintain its value as a tool for decision-makers, Data warehouse system needs to change with business changes. ETL is a recurring activity (daily, weekly, monthly) of a Data warehouse system and needs to be agile, automated, and well documented.

WHY DO YOU NEED ETL?

- It helps companies to analyze their business data for taking critical business decisions.
- Transactional databases cannot answer complex business questions that can be answered by ETL.
- A Data Warehouse provides a common data repository
- ETL provides a method of moving the data from various sources into a data warehouse.
- As data sources change, the Data Warehouse will automatically update.
- Allow verification of data transformation, aggregation and calculations rules.
- ETL process allows sample data comparison between the source and the target system.
- ETL process can perform complex transformations and requires the extra area to store the data.
- ETL helps to Migrate data into a Data Warehouse. Convert to the various formats and types to adhere to one consistent system.
- ETL is a predefined process for accessing and manipulating source data into the target database.
- It helps to improve productivity because it codifies and reuses without a need for technical skills.

ETL PROCESS IN DATA WAREHOUSES



STEP 1) EXTRACTION

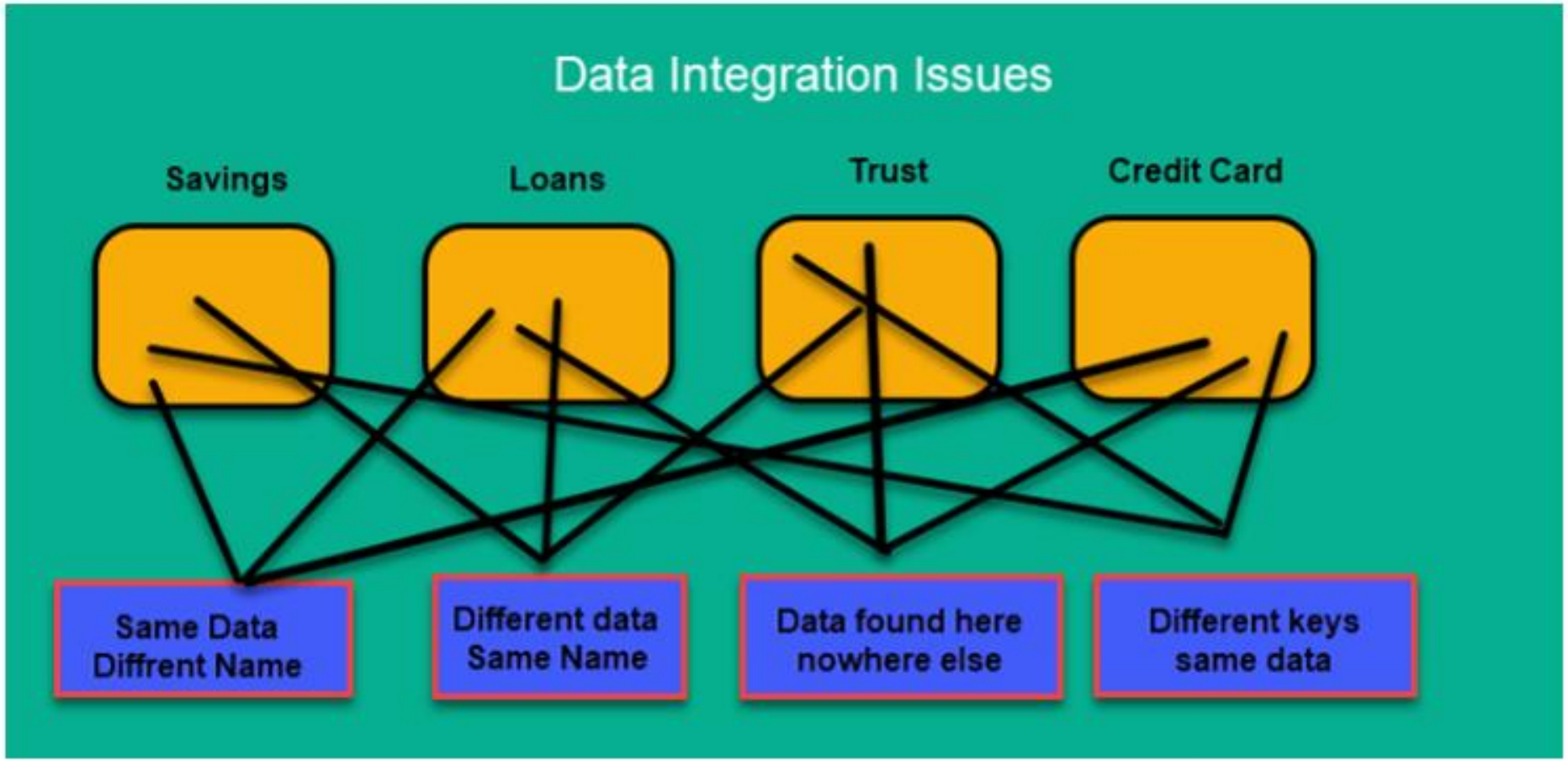
- In this step, data is extracted from the source system into the staging area. Transformations if any are done in staging area so that performance of source system in not degraded. Also, if corrupted data is copied directly from the source into Data warehouse database, rollback will be a challenge. Staging area gives an opportunity to validate extracted data before it moves into the Data warehouse.
- Data warehouse needs to integrate systems that have different
- DBMS, Hardware, Operating Systems and Communication Protocols. Sources could include legacy applications like Mainframes, customized applications, Point of contact devices like ATM, Call switches, text files, spreadsheets, ERP, data from vendors, partners amongst others.
- Hence one needs a logical data map before data is extracted and loaded physically. This data map describes the relationship between sources and target data.

Three Data Extraction methods:

- Full Extraction
- Partial Extraction- without update notification.
- Partial Extraction- with update notification
- Irrespective of the method used, extraction should not affect performance and response time of the source systems. These source systems are live production databases. Any slow down or locking could effect company's bottom line.
- **Some validations are done during Extraction:**
- Reconcile records with the source data
- Make sure that no spam/unwanted data loaded
- Data type check
- Remove all types of duplicate/fragmented data
- Check whether all the keys are in place or not

STEP 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.
- In this step, you apply a set of functions on extracted data. Data that does not require any transformation is called as **direct move** or **pass through data**.
- In transformation step, you can perform customized operations on data. For instance, if the user wants sum-of-sales revenue which is not in the database. Or if the first name and the last name in a table is in different columns. It is possible to concatenate them before loading.


Data Integration Issues
Savings
Loans
Trust
Credit Card
Same Data
Diffrent Name
Different data
Same Name
Data found here
nowhere else
Different keys
same data

Following are Data Integrity Problems:

- Different spelling of the same person like Jon, John, etc.
- There are multiple ways to denote company name like Google, Google Inc.
- Use of different names like Cleaveland, Cleveland.
- There may be a case that different account numbers are generated by various applications for the same customer.
- In some data required files remains blank
- Invalid product collected at POS as manual entry can lead to mistakes.

Validations are done during this stage

- Filtering – Select only certain columns to load
- Using rules and lookup tables for Data standardization
- Character Set Conversion and encoding handling
- Conversion of Units of Measurements like Date Time Conversion, currency conversions, numerical conversions, etc.
- Data threshold validation check. For example, age cannot be more than two digits.
- Data flow validation from the staging area to the intermediate tables.
- Required fields should not be left blank.
- Cleaning (for example, mapping NULL to 0 or Gender Male to "M" and Female to "F" etc.)
- Split a column into multiples and merging multiple columns into a single column.
- Transposing rows and columns,
- Use lookups to merge data
- Using any complex data validation (e.g., if the first two columns in a row are empty then it automatically reject the row from processing)

STEP 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 (nights). Hence, load process should be optimized for performance.
- In case of load failure, recover mechanisms should be configured to restart from the point of failure without data integrity loss. Data Warehouse admins need to monitor, resume, cancel loads as per prevailing server performance.

Types of Loading:

- **Initial Load** — populating all the Data Warehouse tables
- **Incremental Load** — applying ongoing changes as when needed periodically.
- **Full Refresh** —erasing the contents of one or more tables and reloading with fresh data.

Load verification

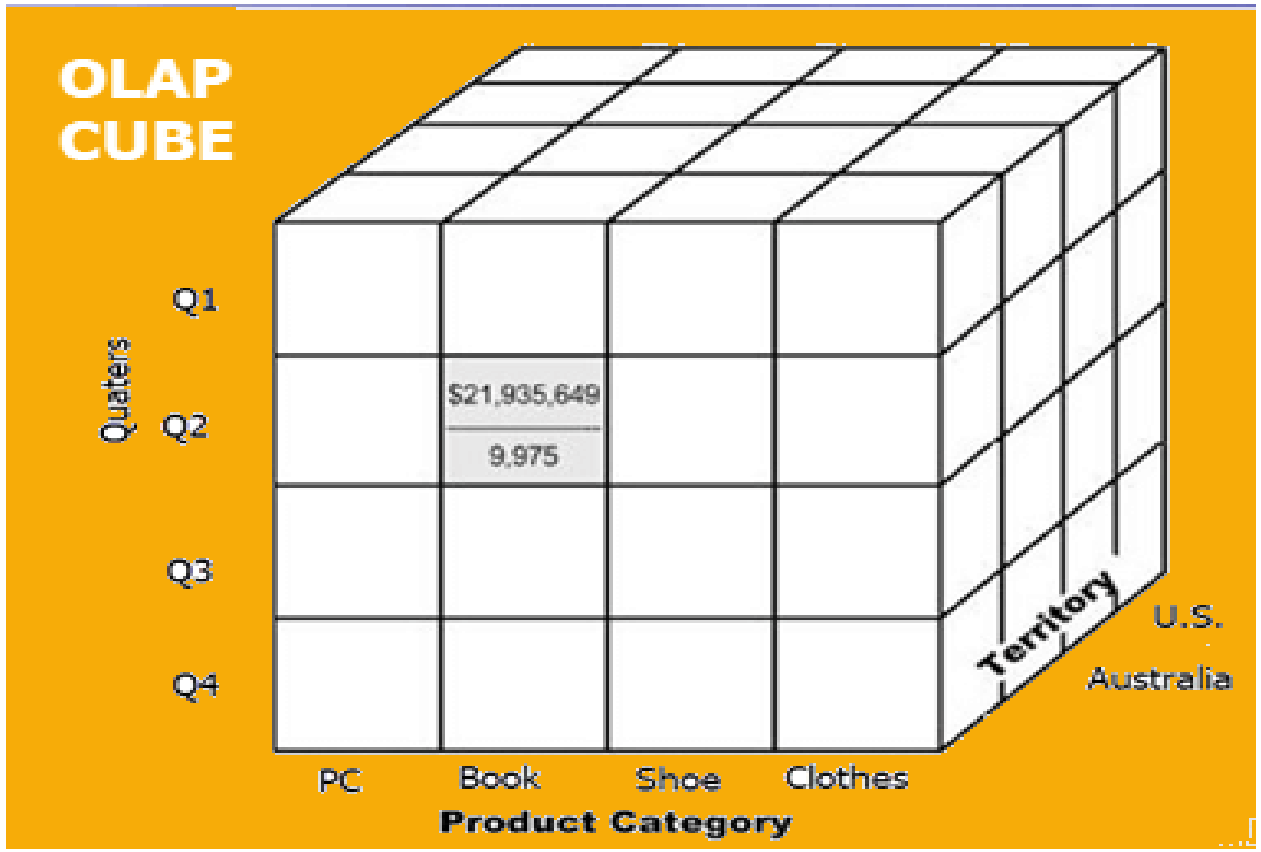
- Ensure that the key field data is neither missing nor null.
- Test modeling views based on the target tables.
- Check that combined values and calculated measures.
- Data checks in dimension table as well as history table.
- Check the BI reports on the loaded fact and dimension table.

OLAP (ONLINE ANALYTICAL PROCESSING): CUBE, OPERATIONS & TYPES

WHAT IS ONLINE ANALYTICAL PROCESSING?

- OLAP is a category of software that allows users to analyze information from multiple database systems at the same time. It is a technology that enables analysts to extract and view business data from different points of view. OLAP stands for Online Analytical Processing.
- Analysts frequently need to group, aggregate and join data. These operations in relational databases are resource intensive. With OLAP data can be pre-calculated and pre-aggregated, making analysis faster.
- OLAP databases are divided into one or more cubes. The cubes are designed in such a way that creating and viewing reports become easy.

OLAP CUBE


OLAP CUBE
Quaters
Q1
Q2
Q3
Q4
$21,935,649
9,975
Territory
U.S.
Australia
PC
Book
Shoe
Clothes
Product Category

- At the core of the OLAP, concept is an OLAP Cube. The OLAP cube is a data structure optimized for very quick data analysis.
- The OLAP Cube consists of numeric facts called measures which are categorized by dimensions. OLAP Cube is also called the **hypercube**.
- Usually, data operations and analysis are performed using the simple spreadsheet, where data values are arranged in row and column format. This is ideal for two-dimensional data. However, OLAP contains multidimensional data, with data usually obtained from a different and unrelated source. Using a spreadsheet is not an optimal option. The cube can store and analyze multidimensional data in a logical and orderly manner.

HOW DOES IT WORK?

- A Data warehouse would extract information from multiple data sources and formats like text files, excel sheet, multimedia files, etc.
- The extracted data is cleaned and transformed. Data is loaded into an OLAP server (or OLAP cube) where information is pre-calculated in advance for further analysis.

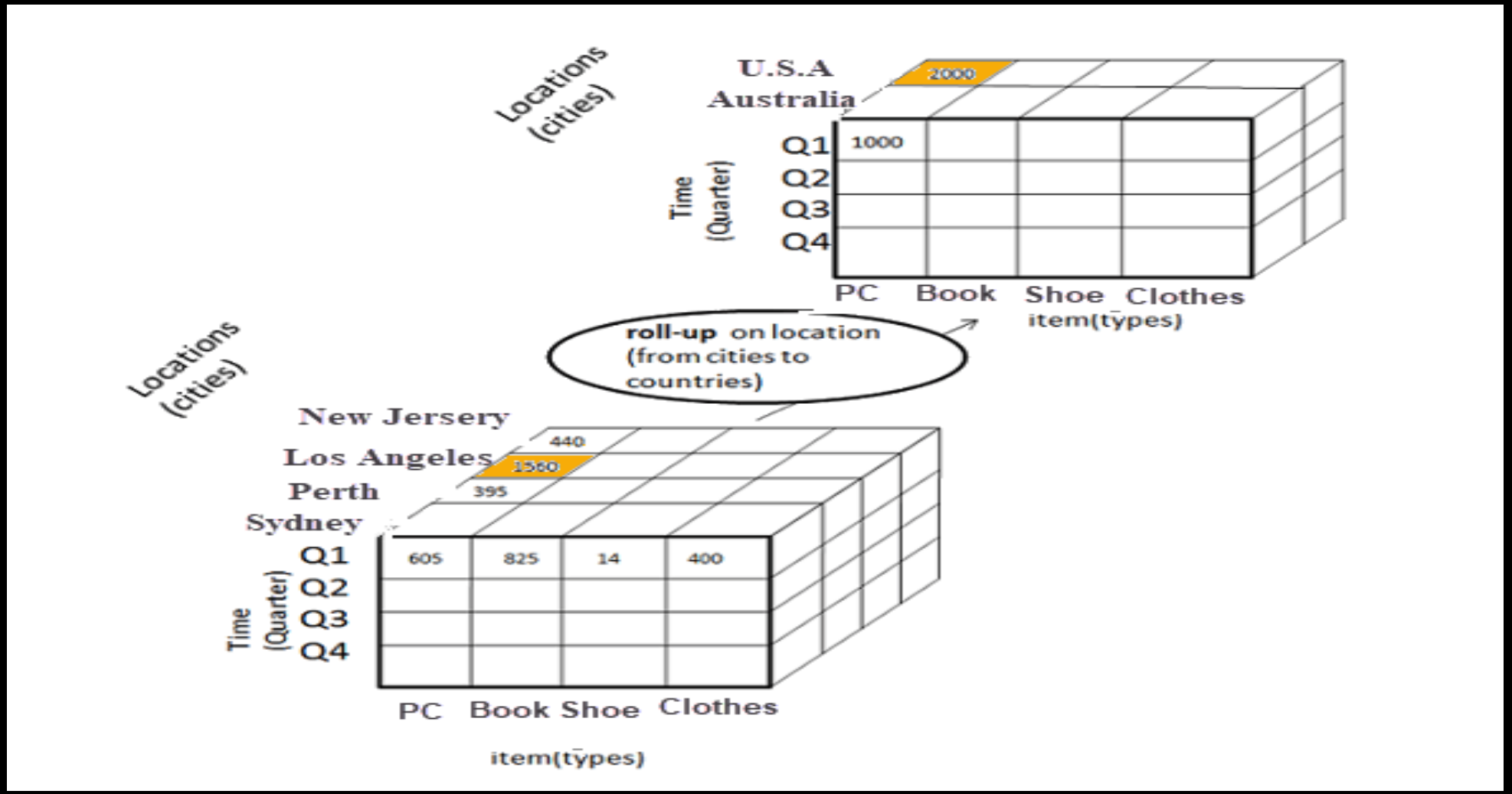
BASIC ANALYTICAL OPERATIONS OF OLAP

- Four types of analytical operations in OLAP are:
- Roll-up
- Drill-down
- Slice and dice
- Pivot (rotate)

1) ROLL-UP

- Roll-up is also known as "consolidation" or "aggregation." The Roll-up operation can be performed in 2 ways
- Reducing dimensions
- Climbing up concept hierarchy. Concept hierarchy is a system of grouping things based on their order or level.

CONSIDER THE FOLLOWING DIAGRAM23

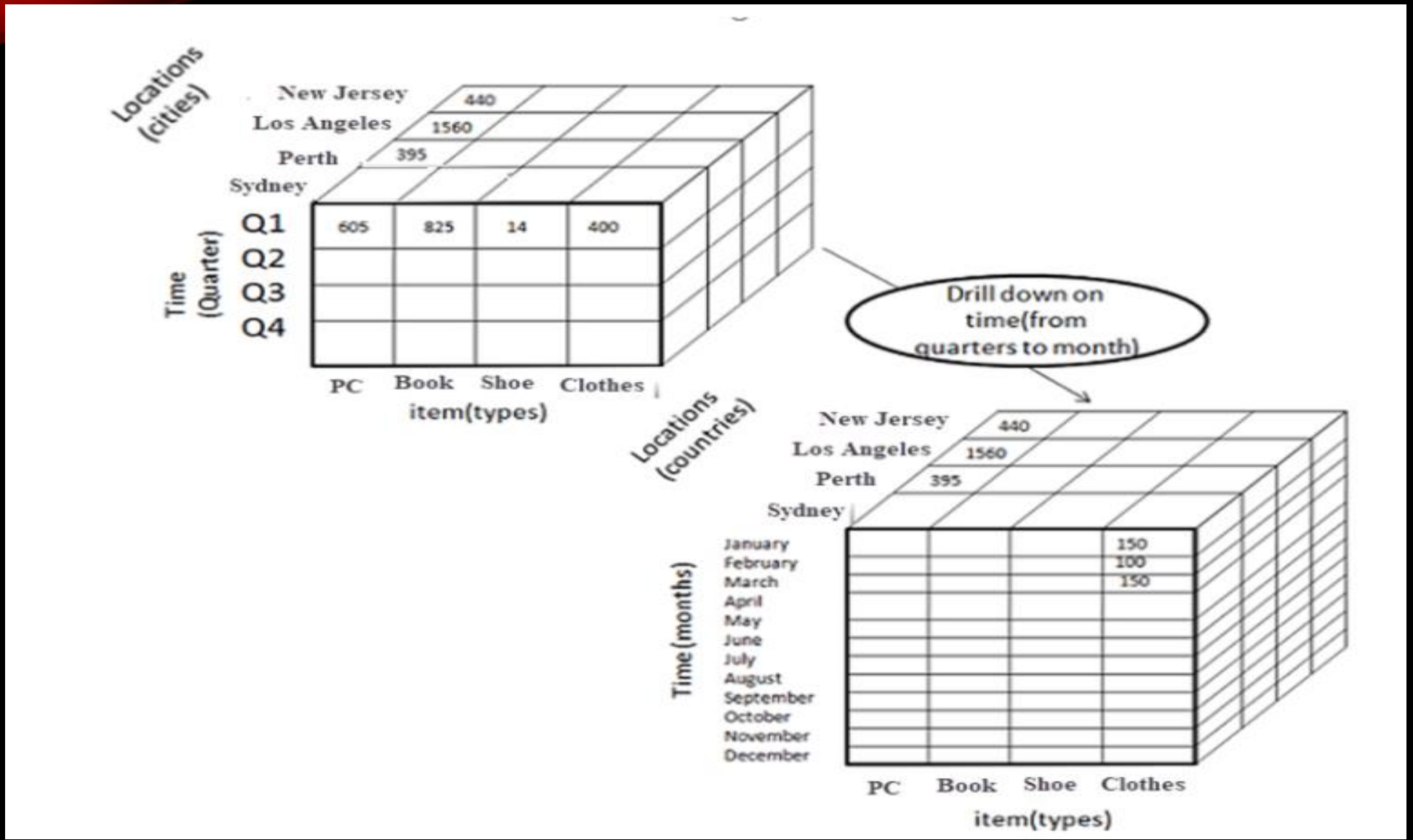


CONTINUE..

- In this example, cities New jersey and Lost Angles and rolled up into country USA
- The sales figure of New Jersey and Los Angeles are 440 and 1560 respectively. They become 2000 after roll-up
- In this aggregation process, data is location hierarchy moves up from city to the country.
- In the roll-up process at least one or more dimensions need to be removed. In this example, Quater dimension is removed.

2) DRILL-DOWN

- In drill-down data is fragmented into smaller parts. It is the opposite of the rollup process. It can be done via
- Moving down the concept hierarchy
- Increasing a dimension


Locations (cities)
New Jersey
Los Angeles
Perth
Sydney
440
1560
395
Time (Quarter)
Q1
Q2
Q3
Q4
605
825
14
400
PC
Book
Shoe
Clothes
item(types)
Drill down on time(from quarters to month)
Locations (countries)
New Jersey
Los Angeles
Perth
Sydney
440
1560
395
Time (months)
January
February
March
April
May
June
July
August
September
October
November
December
150
100
150
PC
Book
Shoe
Clothes
item(types)

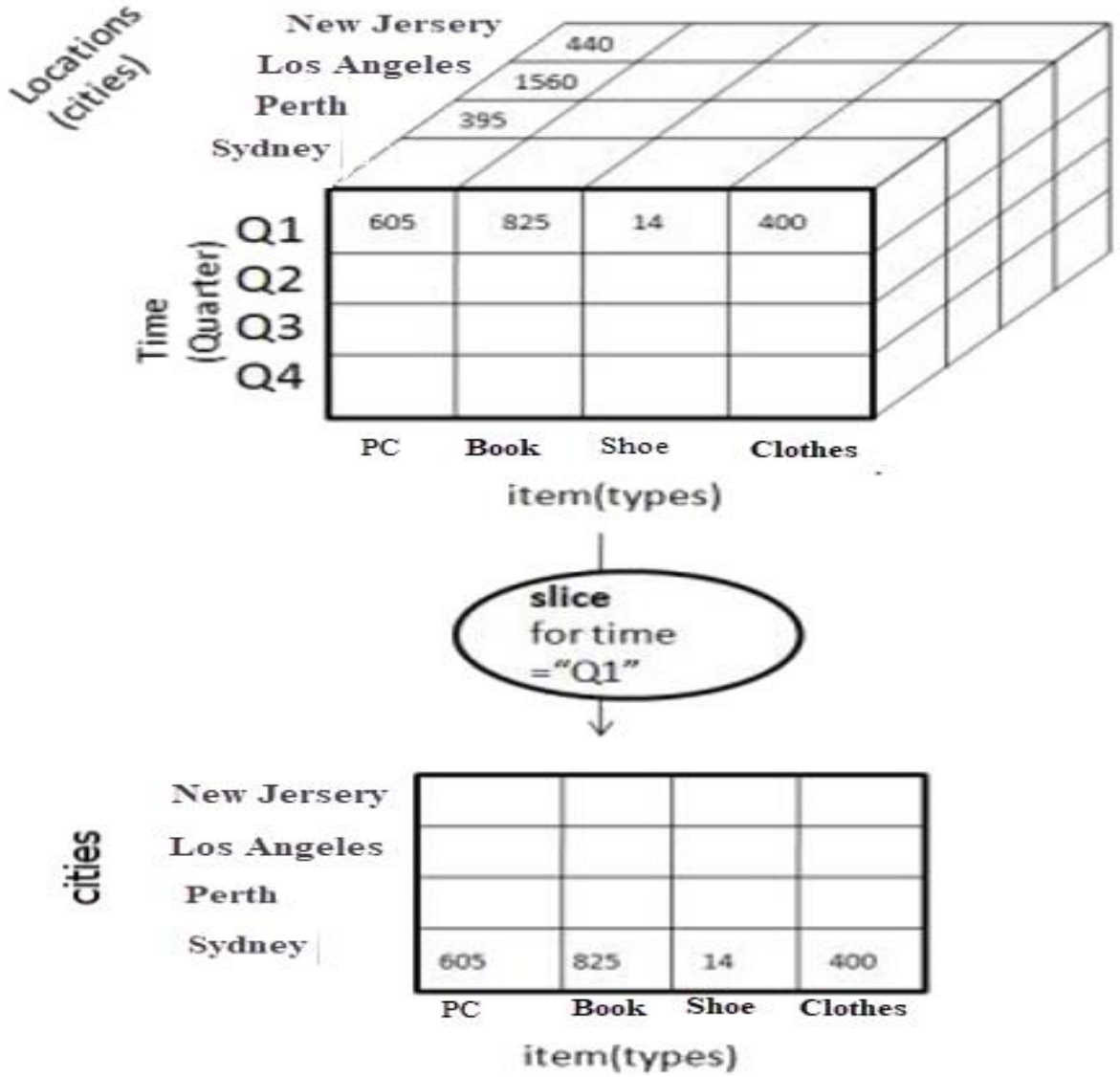
CONTINUE..

- Quater Q1 is drilled down to months January, February, and March. Corresponding sales are also registers.
- In this example, dimension months are added.

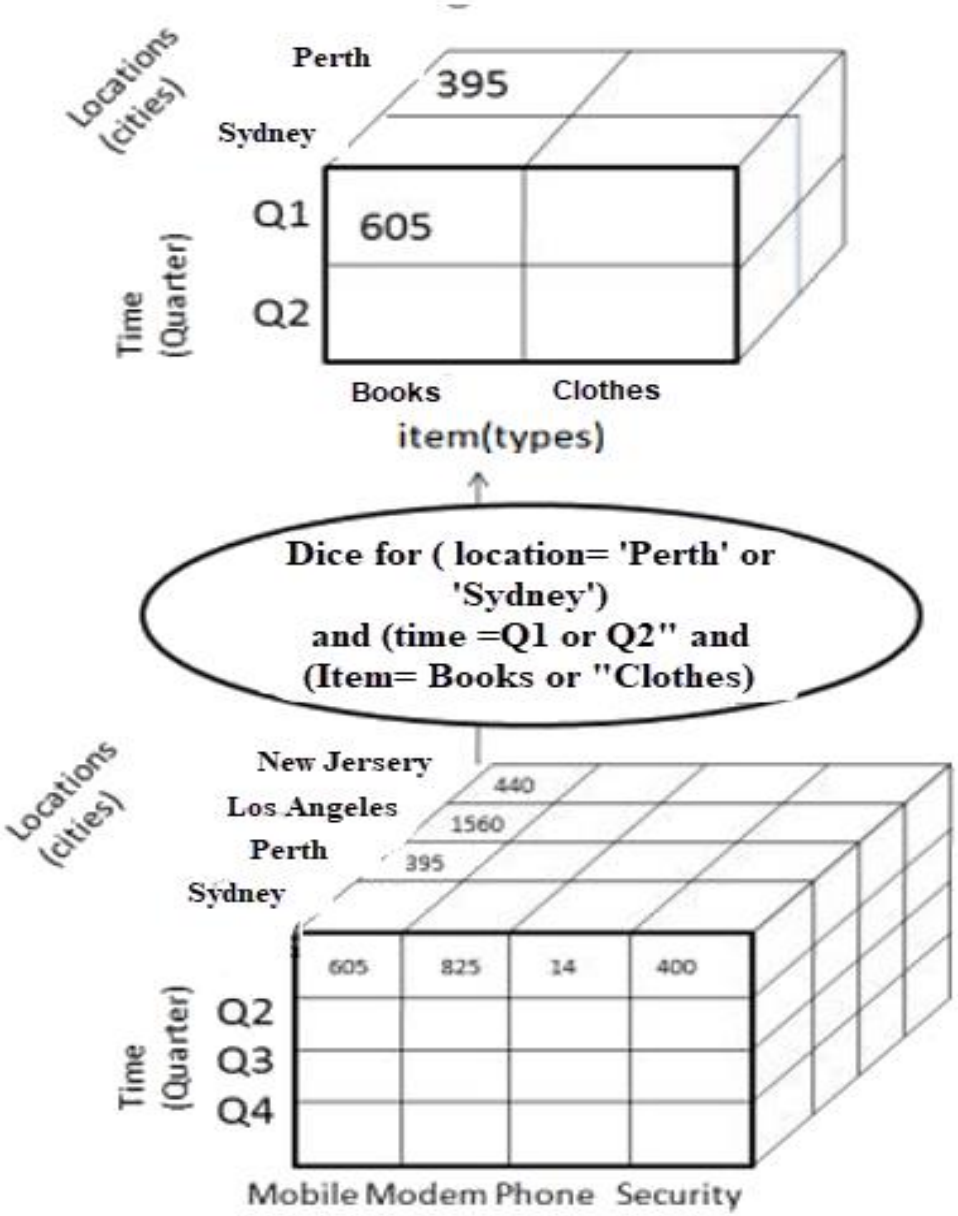
3) SLICE:

- Here, one dimension is selected, and a new sub-cube is created.
- Following diagram explain how slice operation performed:

- Dimension Time is Sliced with Q1 as the filter.
- A new cube is created altogether.

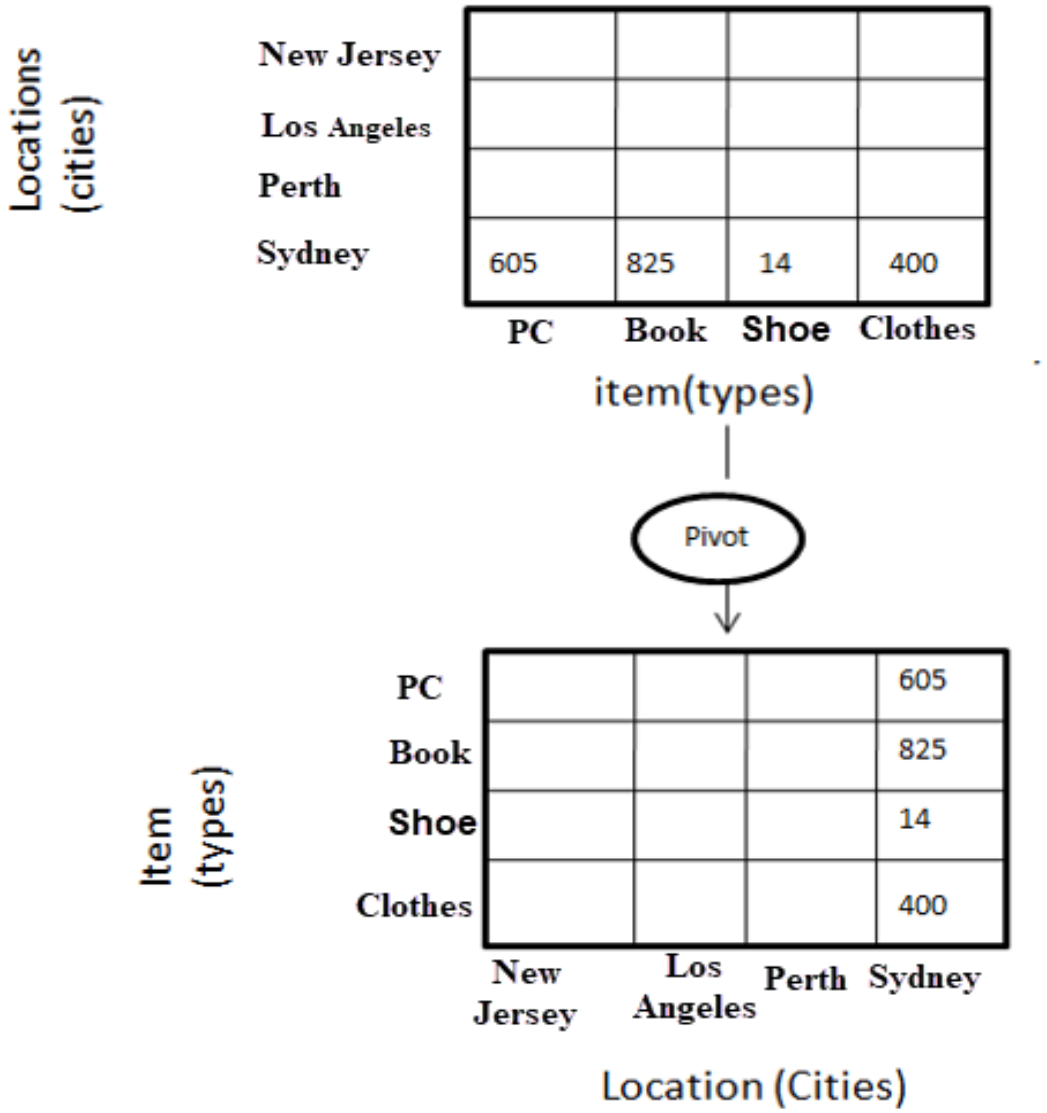


DICE:



- This operation is similar to a slice. The difference in dice is you select 2 or more dimensions that result in the creation of a sub-cube

4) PIVOT



- In Pivot, you rotate the data axes to provide a substitute presentation of data.
- In the following example, the pivot is based on item types.

Benefits of using OLAP services

- OLAP creates a single platform for all type of business analytical needs which includes planning, budgeting, forecasting, and analysis.
- The main benefit of OLAP is the consistency of information and calculations.
- Easily apply security restrictions on users and objects to comply with regulations and protect sensitive data.

Drawbacks of OLAP service

- Implementation and maintenance are dependent on IT professional because the traditional OLAP tools require a complicated modeling procedure.
- OLAP tools need cooperation between people of various departments to be effective which might always be not possible.

WHAT IS OLTP?

- **Online transaction processing** shortly known as OLTP supports transaction-oriented applications in a 3-tier architecture. OLTP administers day to day transaction of an organization.

- **The primary objective is data processing and not data analysis**

EXAMPLE OF OLTP SYSTEM

- An example of OLTP system is ATM center. Assume that a couple has a joint account with a bank. One day both simultaneously reach different ATM centers at precisely the same time and want to withdraw total amount present in their bank account.
- However, the person that completes authentication process first will be able to get money. In this case, OLTP system makes sure that withdrawn amount will be never more than the amount present in the bank. The key to note here is that OLTP systems are optimized for **transactional superiority instead data analysis.**

Other examples of OLTP system are:

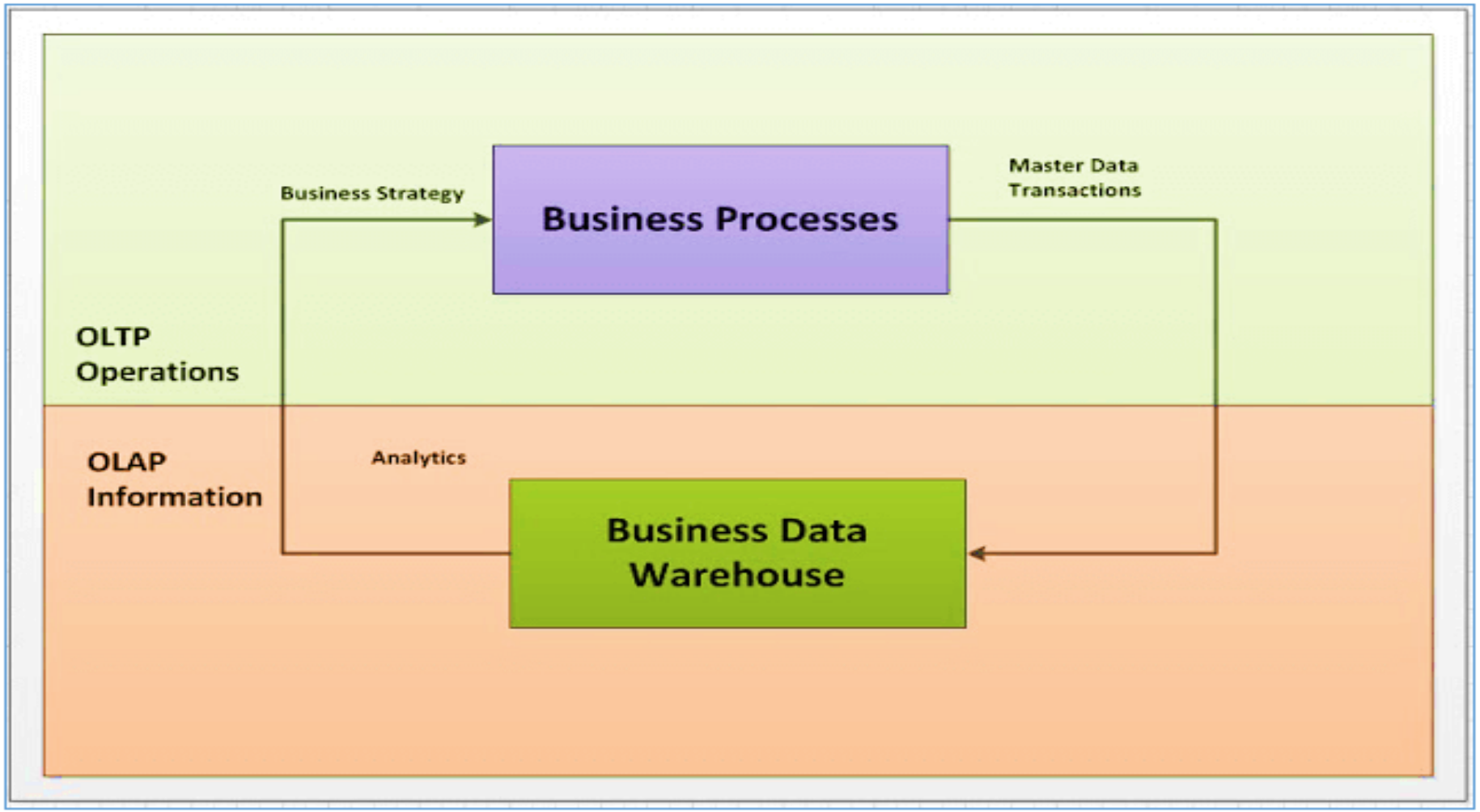
- Online banking
- Online airline ticket booking
- Sending a text message
- Order entry
- Add a book to shopping cart

Benefits of OLTP method

- It administers daily transactions of an organization.
- OLTP widens the customer base of an organization by simplifying individual processes

Drawbacks of OLTP method

- If OLTP system faces hardware failures, then online transactions get severely affected.
- OLTP systems allow multiple users to access and change the same data at the same time which many times created unprecedented situation.


Business Strategy
Business Processes
Master Data
Transactions
OLTP
Operations
OLAP
Information
Analytics
Business Data
Warehouse

DIFFERENCE BETWEEN OLTP AND OLAP

Parameters	OLTP	OLAP
Process	It is an online transactional system. It manages database modification.	OLAP is an online analysis and data retrieving process.
Characteristic	It is characterized by large numbers of short online transactions.	It is characterized by a large volume of data.
Functionality	OLTP is an online database modifying system.	OLAP is an online database query management system.
Method	OLTP uses traditional DBMS.	OLAP uses the data warehouse.

Query	Insert, Update, and Delete information from the database.	Mostly select operations
Table	Tables in OLTP database are normalized.	Tables in OLAP database are **not** normalized.
Source	OLTP and its transactions are the sources of data.	Different OLTP databases become the source of data for OLAP.
Data Integrity	OLTP database must maintain data integrity constraint.	OLAP database does not get frequently modified. Hence, data integrity is not an issue.

Response time	It's response time is in millisecond.	Response time in seconds to minutes.
Data quality	The data in the OLTP database is always detailed and organized.	The data in OLAP process might not be organized.
Usefulness	It helps to control and run fundamental business tasks.	It helps with planning, problem-solving, and decision support.
Operation	Allow read/write operations.	Only read and rarely write.
Audience	It is a market orientated process.	It is a customer orientated process.

Query Type	Queries in this process are standardized and simple.	Complex queries involving aggregations.
Back-up	Complete backup of the data combined with incremental backups.	OLAP only need a backup from time to time. Backup is not important compared to OLTP
Design	DB design is application oriented. Example: Database design changes with industry like Retail, Airline, Banking, etc.	DB design is subject oriented. Example: Database design changes with subjects like sales, marketing, purchasing, etc.
User type	It is used by Data critical users like clerk, DBA & Data Base professionals.	Used by Data knowledge users like workers, managers, and CEO.
Purpose	Designed for real time business operations.	Designed for analysis of business measures by category and attributes.

Performance metric	Transaction throughput is the performance metric	Query throughput is the performance metric.
Number of users	This kind of Database users allows thousands of users.	This kind of Database allows only hundreds of users.
Productivity	It helps to Increase user's self-service and productivity	Help to Increase productivity of the business analysts.
Challenge	Data Warehouses historically have been a development project which may prove costly to build.	An OLAP cube is not an open SQL server data warehouse. Therefore, technical knowledge and experience is essential to manage the OLAP server.

Process	It provides fast result for daily used data.	It ensures that response to the query is quicker consistently.
Characteristic	It is easy to create and maintain.	It lets the user create a view with the help of a spreadsheet.
Style	OLTP is designed to have fast response time, low data redundancy and is normalized.	A data warehouse is created uniquely so that it can integrate different data sources for building a consolidated database

WHAT IS DIMENSIONAL MODEL IN DATA WAREHOUSE?

WHAT IS DIMENSIONAL MODEL?

- A dimensional model is a data structure technique optimized for Data warehousing tools. The concept of Dimensional Modelling was developed by Ralph Kimball and is comprised of "fact" and "dimension" tables.
- A Dimensional model is designed to read, summarize, analyze numeric information like values, balances, counts, weights, etc. in a data warehouse. In contrast, relation models are optimized for addition, updating and deletion of data in a real-time Online Transaction System.

- These dimensional and relational models have their unique way of data storage that has specific advantages.
- For instance, in the relational mode, normalization and ER models reduce redundancy in data. On the contrary, dimensional model arranges data in such a way that it is easier to retrieve information and generate reports.
- Hence, Dimensional models are used in data warehouse systems and not a good fit for relational systems.

ELEMENTS OF DIMENSIONAL DATA MODEL

Fact

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

Dimension

- 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 – Customer Names
- Where – Location
- What – Product Name

In other words, a dimension is a window to view information in the facts

Attributes

- The Attributes are the various characteristics of the dimension.
- In the Location dimension, the attributes can be
- State
- Country
- Zipcode etc.

Attributes are used to search, filter, or classify facts. Dimension Tables contain Attributes

Fact Table

- A fact table is a primary table in a dimensional model.
- A Fact Table contains
- Measurements/facts
- Foreign key to dimension table

Dimension table

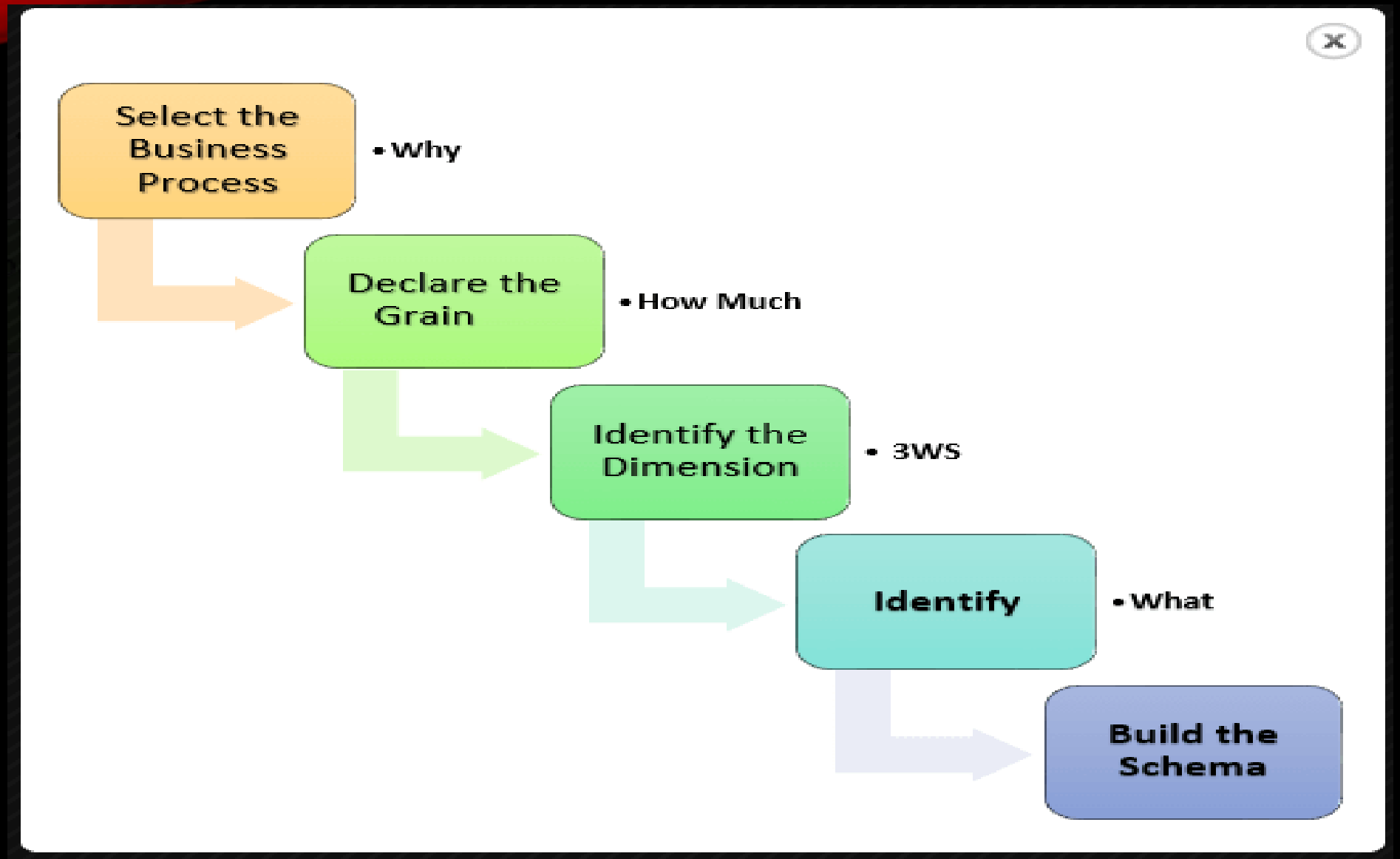
- A dimension table contains dimensions of a fact.
- They are joined to fact table via a foreign key.
- Dimension tables are de-normalized tables.
- The Dimension Attributes are the various columns in a dimension table
- Dimensions offers descriptive characteristics of the facts with the help of their attributes
- No set limit set for given for number of dimensions
- The dimension can also contain one or more hierarchical relationships

STEPS OF DIMENSIONAL MODELLING

The accuracy in creating your Dimensional modeling determines the success of your data warehouse implementation. Here are the steps to create Dimension Model

- Identify Business Process
- Identify Grain (level of detail)
- Identify Dimensions
- Identify Facts
- Build Star

The model should describe the Why, How much, When/Where/Who and What of your business process


Select the Business Process
• Why
Declare the Grain
• How Much
Identify the Dimension
• 3WS
Identify
• What
Build the Schema

- **Step 1) Identify the business process**
- Identifying the actual business process a datarehouse should cover. This could be Marketing, Sales, HR, etc. as per the data analysis needs of the organization. The selection of the Business process also depends on the quality of data available for that process. It is the most important step of the Data Modelling process, and a failure here would have cascading and irreparable defects.

- **Step 2) Identify the grain**
- The Grain describes the level of detail for the business problem/solution. It is the process of identifying the lowest level of information for any table in your data warehouse. If a table contains sales data for every day, then it should be daily granularity. If a table contains total sales data for each month, then it has monthly granularity.
- During this stage, you answer questions like
- Do we need to store all the available products or just a few types of products? This decision is based on the business processes selected for Datawarehouse
- Do we store the product sale information on a monthly, weekly, daily or hourly basis? This decision depends on the nature of reports requested by executives
- How do the above two choices affect the database size?

- **Example of Grain:**
- The CEO at an MNC wants to find the sales for specific products in different locations on a daily basis.
- So, the grain is "product sale information by location by the day.“

- **Step 3) Identify the dimensions**
- Dimensions are nouns like date, store, inventory, etc. These dimensions are where all the data should be stored. For example, the date dimension may contain data like a year, month and weekday.
- **Example of Dimensions:**
- The CEO at an MNC wants to find the sales for specific products in different locations on a daily basis.

  Dimensions: Product, Location and Time

  Attributes: For Product: Product key (Foreign Key), Name, Type, Specifications

  Hierarchies: For Location: Country, State, City, Street Address, Name

- **Step 4) Identify the Fact**
- This step is co-associated with the business users of the system because this is where they get access to data stored in the data warehouse. Most of the fact table rows are numerical values like price or cost per unit, etc.
- **Example of Facts:**
- The CEO at an MNC wants to find the sales for specific products in different locations on a daily basis.
- The fact here is Sum of Sales by product by location by time.

- **Step 5) Build Schema**
- In this step, you implement the Dimension Model. A schema is nothing but the database structure (arrangement of tables). There are two popular schemas

Star Schema

- The star schema architecture is easy to design. It is called a star schema because diagram resembles a star, with points radiating from a center. The center of the star consists of the fact table, and the points of the star is dimension tables.The fact tables in a star schema which is third normal form whereas dimensional tables are de-normalized.

Snowflake Schema

- The snowflake schema is an extension of the star schema. In a star schema, each dimension are normalized and connected to more dimension tables.

STAR AND SNOWFLAKE SCHEMA IN DATA WAREHOUSING

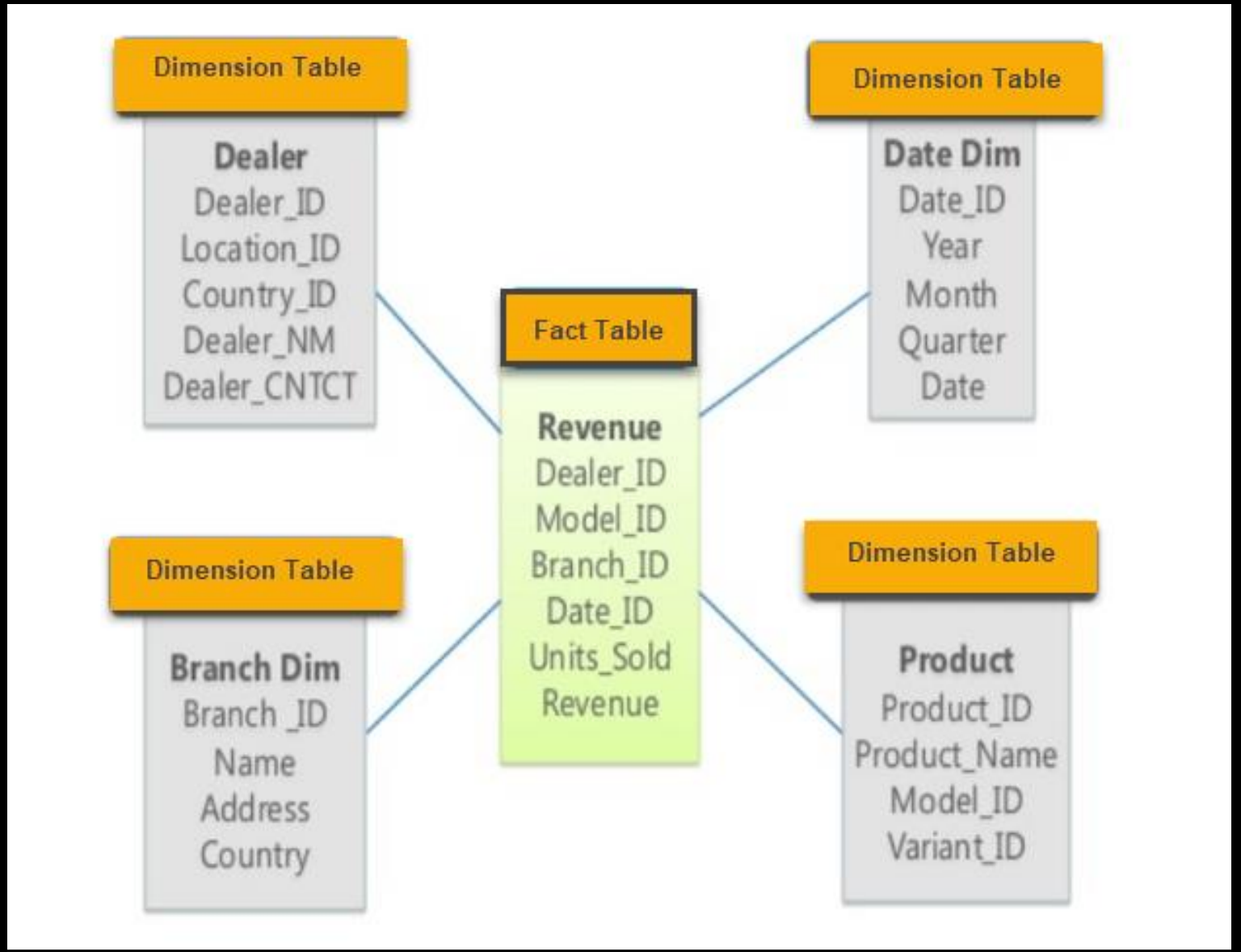
- **What is Multidimensional schemas?**

- Multidimensional schema is especially designed to model data warehouse systems. The schemas are designed to address the unique needs of very large databases designed for the analytical purpose (OLAP).
- Types of Data Warehouse Schema:
- Following are 3 chief types of multidimensional schemas each having its unique advantages.
- **Star Schema**
- **Snowflake Schema**
- **Galaxy Schema**

WHAT IS A 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.

For example, as you can see in the above-given image that fact table is at the center which contains keys to every dimension table like Deal_ID, Model ID, Date_ID, Product_ID, Branch_ID & other attributes like Units sold and revenue.

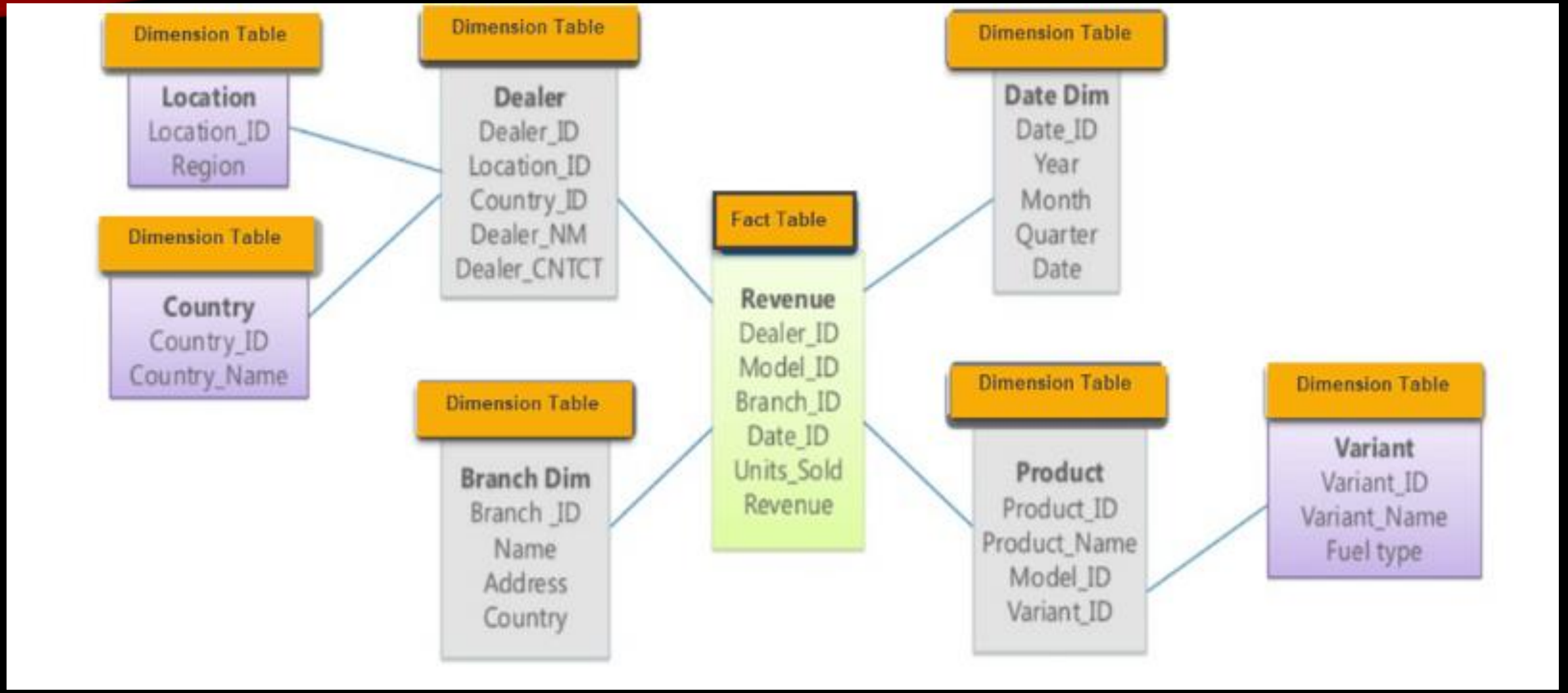


Characteristics of Star Schema:

- Every dimension in a star schema is represented with the only one-dimension table.
- The dimension table should contain the set of attributes.
- The dimension table is joined to the fact table using a foreign key
- The dimension table are not joined to each other
- Fact table would contain key and measure
- The Star schema is easy to understand and provides optimal disk usage.
- The dimension tables are **not normalized**. For instance, in the above figure, Country_ID does not have Country lookup table as an OLTP design would have.
- The schema is widely supported by BI Tools

WHAT IS A 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.
- The dimension tables are **normalized** which splits data into additional tables. In the following example, Country is further normalized into an individual table.

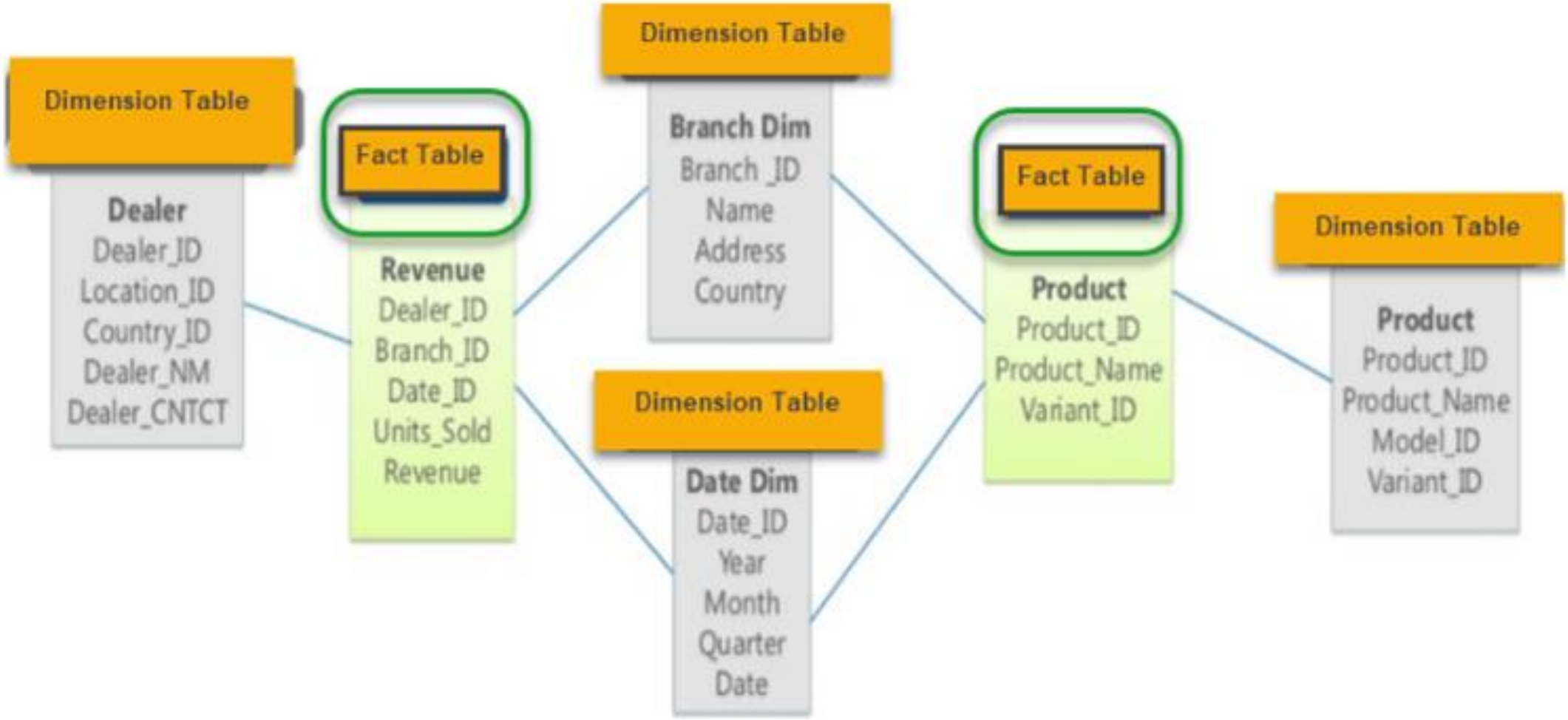

Dimension Table
Location
Location_ID
Region
Dimension Table
Country
Country_ID
Country_Name
Dimension Table
Dealer
Dealer_ID
Location_ID
Country_ID
Dealer_NM
Dealer_CNTCT
Dimension Table
Branch Dim
Branch _ID
Name
Address
Country
Fact Table
Revenue
Dealer_ID
Model_ID
Branch_ID
Date_ID
Units_Sold
Revenue
Dimension Table
Date Dim
Date_ID
Year
Month
Quarter
Date
Dimension Table
Product
Product_ID
Product_Name
Model_ID
Variant_ID
Dimension Table
Variant
Variant_ID
Variant_Name
Fuel type

Characteristics of Snowflake Schema:

- The main benefit of the snowflake schema it uses smaller disk space.
- Easier to implement a dimension is added to the Schema
- Due to multiple tables query performance is reduced
- The primary challenge that you will face while using the snowflake Schema is that you need to perform more maintenance efforts because of the more lookup tables.

WHAT IS A GALAXY SCHEMA(FACT CONSTELLATION)?

- 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.


Dimension Table
Dealer
Dealer_ID
Location_ID
Country_ID
Dealer_NM
Dealer_CNTCT
Fact Table
Revenue
Dealer_ID
Branch_ID
Date_ID
Units_Sold
Revenue
Dimension Table
Branch Dim
Branch _ID
Name
Address
Country
Dimension Table
Date Dim
Date_ID
Year
Month
Quarter
Date
Fact Table
Product
Product_ID
Product_Name
Variant_ID
Dimension Table
Product
Product_ID
Product_Name
Model_ID
Variant_ID

- As you can see in above figure, there are two facts table
- Revenue
- Product.

In Galaxy schema shares dimensions are called Conformed Dimensions.

Characteristics of Galaxy Schema:

- The dimensions in this schema are separated into separate dimensions based on the various levels of hierarchy.
- For example, if geography has four levels of hierarchy like region, country, state, and city then Galaxy schema should have four dimensions.
- Moreover, it is possible to build this type of schema by splitting the one-star schema into more Star schemes.
- The dimensions are large in this schema which is needed to build based on the levels of hierarchy.
- This schema is helpful for aggregating fact tables for better understanding.

WHAT IS DATA MART?

- A data mart is focused on a single functional area of an organization and contains a subset of data stored in a Data Warehouse.
- A data mart is a condensed version of Data Warehouse and is designed for use by a specific department, unit or set of users in an organization. E.g., Marketing, Sales, HR or finance. It is often controlled by a single department in an organization.
- Data Mart usually draws data from only a few sources compared to a Data warehouse. Data marts are small in size and are more flexible compared to a Datawarehouse.

WHY DO WE NEED DATA MART?

- Data Mart helps to enhance user's response time due to reduction in volume of data
- It provides easy access to frequently requested data.
- Data mart are simpler to implement when compared to corporate Datawarehouse. At the same time, the cost of implementing Data Mart is certainly lower compared with implementing a full data warehouse.
- Compared to Data Warehouse, a datamart is agile. In case of change in model, datamart can be built quicker due to a smaller size.
- A Datamart is defined by a single Subject Matter Expert. On the contrary data warehouse is defined by interdisciplinary SME from a variety of domains. Hence, Data mart is more open to change compared to Datawarehouse.
- Data is partitioned and allows very granular access control privileges.
- Data can be segmented and stored on different hardware/software platforms.