Data Science II (Unit 1)

Business Intelligence

Business intelligence (BI)

Is a set of technologies and processes that allow people at all levels(from CEO to Accountant) of an organisation to access, analyse, interact with, and analyse data to manage the business, Improve performance, discover opportunities, and operate efficiently.

Business Analytics(BA)

- Is the process of transforming data into insights to improve business decisions
- Involves a set of disciplines and technologies for solving business problems using data analysis, statistical models and other quantitative methods.
- It involves an iterative, methodical exploration of an organisation's data, with an emphasis on statistical analysis, to drive decision-making.

Big Data

The term “Big Data” refer to larger data volumes, generally in the petabyte(PB) range.

Big Data is distinct and has 3 main characters

- High Volume: Big Data usually contains petabytes of data whereas in business intelligence it is just in the range of gigabytes to terabytes.
- High Velocity: The data keeps coming in at a very fast pace from IoT devices, Web Logs, E-Commerce-Sites and Social Media.
- Variety**:** It contains data in different formats like **t**extual data in the form of tweets or social comments, photos, images, and video.

Big Data Vs Business Intelligence

Business Intelligence	**Big Data**

Data is collected from traditional sources	Data is collected from multiple sources
Data is housed in a central server	Data is stored in a distributed file system
Generally data is analysed in an offline mode	Data is analysed in real time
Consits of Structured Data	Consist of Structured, Semi-Structured and Unstructured Data
Descriptive Analytics	Predictive Analytics
Diagnostics Analytics	Prescriptive Analytics

Similarities between BI(Business Intelligence) and Big Data

1. Both BI and Big Data involve large datasets
2. Both are used for deriving Value
3. Both are used to predict trends and forecasts

How BI provides value to your business

- BI touches everyone in a company and beyond to customers, suppliers, and with public data, to citizens.
- There is a correlation between the effective use of business intelligence and company performance i.e if the comapy makes the most out of BI then they can increase their profits.
- Simply having better access to data does not improve performance instead what becomes important is how the companies utiliise the data.

BI for Management and Control

- Business intelligence provides managers information to know what's going on in their ompany or business.
- Without business Intelligence, managers may talk about how they are "flying sightless" with no insight until quarterly financial numbers are published.
- With business intelligence, information is accessible on a timelier (faster) and more flexible basis to provide a view of:
  a. Sales in various regions and by various product lines,
  b. A comparision of expenses and budgets
  c. Warehouse inventory for a particular product or raw materials.
- BI provides real time reports and dashboards to help managers identify why the business is going the way it is as opposed to paper based reports or a fixed

screen transaction system which is less flexible

- BI provides early acess to expense data that allows managers to identify business unit that are over budget and then take immediate steps to reduce overtime pay, travel expenses, or to defer purchases beforehand rather than waiting until the close of the quarter to discover that excessive expenses have reduced profitability

BI for Improving performance

- Business performance is measured by a number of financial indicators, such as revenue, margin, profitability, cost to serve and so on.
- In marketing, performance gains may be achieved by improving response rates for particular campaigns by identifying characteristics of more responsive customers.
- Eliminating ineffective campaigns saves companies millions of dollars each year.

BI Scars

1. It leads to power struggles between IT and the business department when either of them loses areas of control or disagrees on the scope and approach.
2. Since BI could generate real time and effective reports, custom report developers lost their jobs
3. Many marketing manager and employees are fired when a company realized just how badly
   the manager was performing campaign management and how the employee was unproductive.
4. Software and technology that does not always work as expected, and vendors who merge, get acquired, or change strategy in ways that affect the BI deployment.

Alternate answer by ChatGPT:

While Business Intelligence (BI) offers numerous benefits, there are also potential challenges and drawbacks, often referred to as the "scars" of BI. It's important for organizations to be aware of these issues to address them effectively:

1. **Data Quality Issues:**
   - BI heavily relies on the quality of input data. Inaccurate, incomplete, or inconsistent data can lead to flawed analyses and incorrect conclusions.

Organizations need to invest in data quality assurance processes to ensure the reliability of the information used for BI.

2. **Integration Challenges:**
    - Many organizations have diverse data sources and systems that may not easily integrate with BI tools. Achieving seamless integration can be a complex and time-consuming process, requiring careful planning and implementation.
3. **Costs and Resource Intensity:**
    - Implementing and maintaining BI systems can be expensive, both in terms of software and hardware costs, as well as the need for skilled personnel. Small and mid-sized businesses may find it challenging to allocate resources for BI initiatives.
4. **Complexity and Learning Curve:**
    - BI tools often come with a learning curve, and employees may need training to use them effectively. The complexity of these tools can be a barrier to adoption, and organizations may face resistance from employees unfamiliar with analytical tools and techniques.
5. **Lack of Alignment with Business Objectives:**
    - Sometimes, organizations implement BI without a clear alignment with their business goals. This can lead to irrelevant analyses and reports, wasting resources and failing to provide meaningful insights.
6. **Security and Privacy Concerns:**
    - BI involves handling sensitive business data, and maintaining data security and privacy is paramount. Breaches or unauthorized access can have severe consequences, including reputational damage and legal issues. Organizations must implement robust security measures to protect their BI systems.
7. **Overemphasis on Technology:**
    - Organizations may fall into the trap of focusing too much on the technology aspect of BI without addressing the cultural and organizational aspects. Successful BI implementation requires a balance between technology, processes, and people.
8. **Difficulty in Change Management:**

- Implementing BI often involves changes in processes, workflows, and organizational culture. Resistance to change among employees can hinder the successful adoption of BI initiatives. Effective change management strategies are crucial to overcoming this challenge.

9. **Scalability Issues:**
    - As an organization grows, the volume of data to be processed and analyzed may increase significantly. BI systems need to be scalable to handle larger datasets and accommodate the growing analytical needs of the organization.
10. **Unrealistic Expectations:**
    - There is a risk of setting unrealistic expectations regarding the capabilities of BI tools. It's important for organizations to understand the limitations of BI and not expect it to be a panacea for all business challenges.

Understanding and addressing these scars is essential for organizations to maximize the benefits of Business Intelligence while mitigating potential risks. A well-planned BI strategy, combined with a commitment to data quality, security, and user training, can contribute to a more successful BI implementation.

LOFT in BI

In the context of Business Intelligence (BI), LOFT stands for **Luck, Opportunity, Frustration, and Threat**. It's not a technical term or framework, but rather a concept introduced by author Jim Shamlin in his book "Successful Business Intelligence: Secrets to Making BI a Killer App."

LOFT signifies the various factors that can contribute to the **success or failure of BI initiatives**. While technology and best practices play a crucial role, Shamlin argues that success often hinges on a nuanced interplay of these four external forces:

1. Luck:

- Chance events or unexpected occurrences that create favorable conditions for BI adoption and impact. For example, A change in business leadreship or a sudden competitor crisis might compel an organization to prioritize data-driven decision-making.

2. Opportunity:

- Emerging market trends, new technologies, or regulatory changes that open up avenues for utilizing BI to gain a competitive edge. For example, business doing something differently, such as entering a new market or the rise of social media

analytics presents an opportunity for companies to enhance customer understanding through sentiment analysis.

3. Frustration:

- Growing pains experienced with existing systems, rising operational costs, or recurring inefficiencies that create a burning desire for change and drive the motivation to adopt BI solutions. For example, the business feels like it is “flying blind” with its data and also persistent problems with manual data reporting might spur the investment in automated BI tools.

4. Threat:

- External pressures such as increased competition, market disruptions, or looming financial challenges that act as wake-up calls, pushing organizations to embrace data-driven approaches for survival and growth. For example, pressure from competitors or a bankruptcy and the threat of losing market share to data-driven competitors might force a reactive shift towards leveraging BI.

The LOFT concept emphasizes that **BI success is rarely a linear process solely driven by technology**. It highlights the importance of considering broader external forces and their potential to influence the adoption, utilization, and impact of BI initiatives. By understanding these dynamics, organizations can better anticipate challenges, capitalize on opportunities, and steer their BI journey towards maximizing its benefits.

Remember, LOFT is not a prescriptive framework but rather a lens for understanding the **complexities and contextual factors surrounding BI success**. It encourages organizations to adopt a holistic approach and consider the interplay of various forces in their own environment.

ERP(Enterprise Rescource Platform)

Enterprise Resource Planning (ERP) systems also kanown as Operational Systems, Transaction Processing Systems or Source Systems are integrated software solutions that organizations use to manage and streamline their business processes. ERP systems consolidate various functions across different departments into a single, unified platform, providing a holistic view of business operations. These systems aim to improve efficiency, facilitate data flow, and enhance decision-making within an organization.

Key functionalities of ERP systems include:

1. **Integrated Business Processes:**

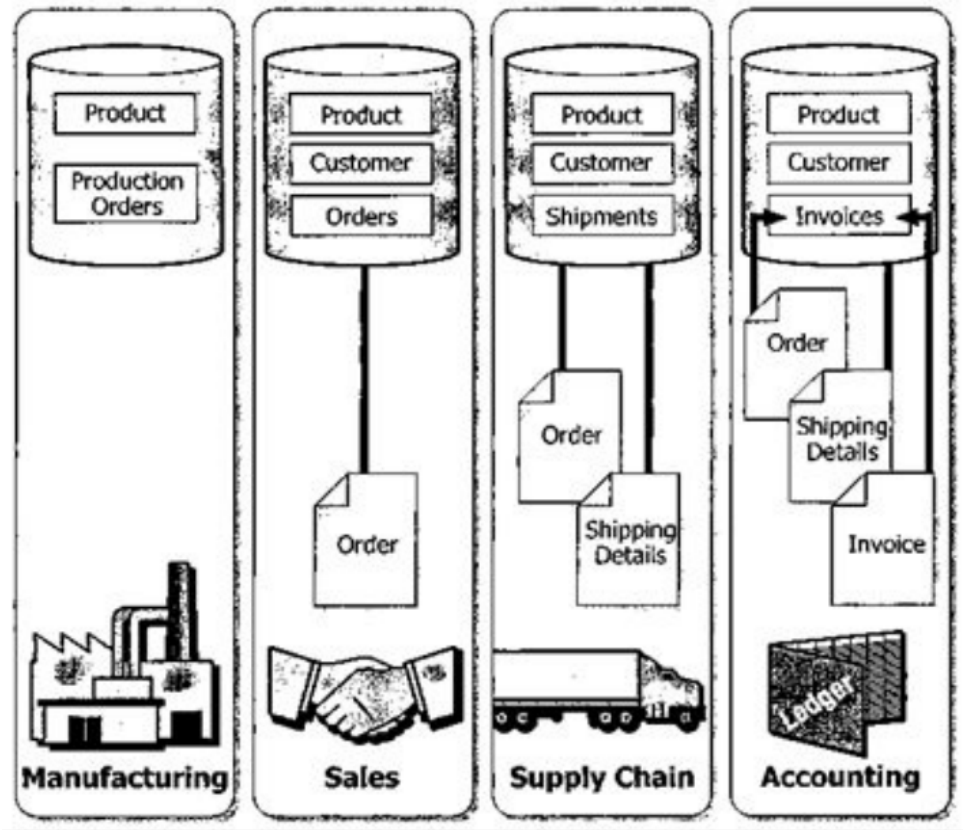
- ERP systems integrate and automate core business processes, including finance, human resources, supply chain, manufacturing, procurement, and customer relationship management. This integration eliminates data silos and ensures a seamless flow of information across the organization.

2. **Centralized Database:**
    - ERP systems maintain a centralized database that serves as a single source of truth for all relevant data. This helps in maintaining consistency and accuracy across different departments, reducing the risk of data duplication and errors.
3. **Financial Management:**
    - ERP systems provide robust financial management capabilities, including accounting, budgeting, and financial reporting. They help organizations track financial transactions, manage accounts payable and receivable, and generate financial statements.
4. **Human Resources Management:**
    - ERP systems include modules for managing human resources functions, such as payroll, employee records, benefits administration, and workforce planning. These modules help streamline HR processes and ensure compliance with regulations.
5. **Supply Chain Management:**
    - ERP systems optimize supply chain processes by managing inventory, procurement, order fulfillment, and logistics. This leads to improved coordination between suppliers, manufacturers, and distributors, reducing lead times and minimizing stockouts or overstocks.
6. **Manufacturing Management:**
    - For organizations involved in manufacturing, ERP systems offer tools to plan and control production processes. This includes features like bill of materials (BOM), work order management, quality control, and production scheduling.
7. **Customer Relationship Management (CRM):**
    - CRM modules within ERP systems help organizations manage customer interactions, sales, and marketing activities. This includes features such as lead management, opportunity tracking, and customer service support, fostering stronger customer relationships.

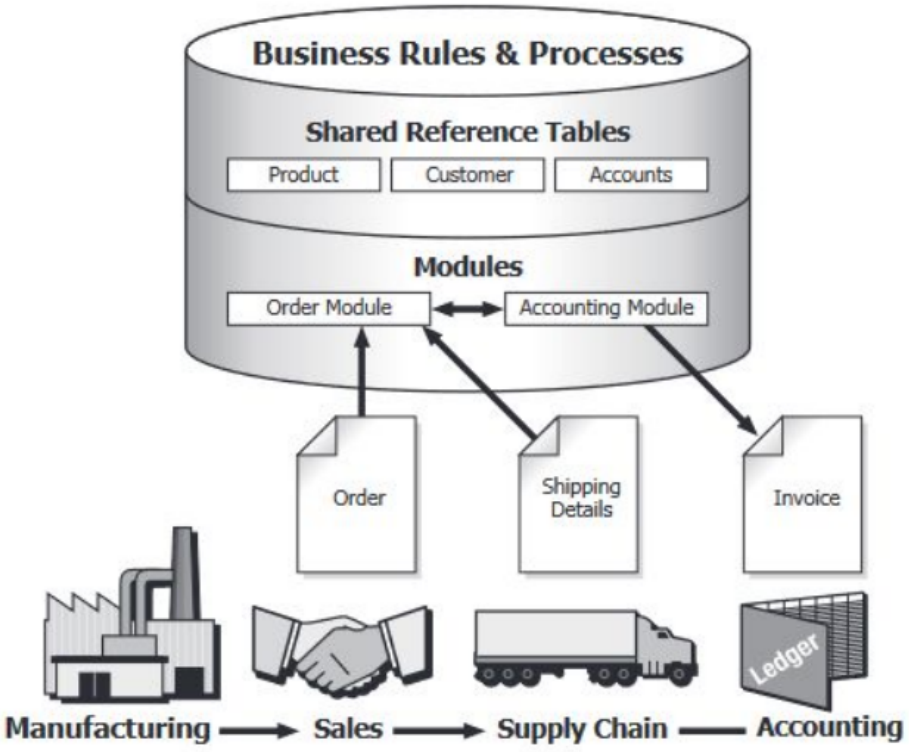
8. **Business Analytics and Reporting:**
    - ERP systems provide reporting and analytics tools to help organizations make data-driven decisions. Users can generate customizable reports, dashboards, and key performance indicators (KPIs) to gain insights into various aspects of the business.
9. **Compliance and Risk Management:**
    - ERP systems often include features to address compliance with industry regulations and manage risks. This may involve tracking changes in regulations, ensuring data security, and implementing audit trails to maintain accountability.
10. **Mobile Accessibility:**
    - Modern ERP systems often offer mobile accessibility, allowing users to access critical business information and perform tasks remotely. This facilitates real-time decision-making and enhances collaboration among employees.
11. **Workflow Automation:**
    - ERP systems automate routine business processes, reducing manual intervention and the likelihood of errors. This includes approval workflows, document management, and notifications for various tasks.
12. **Scalability:**
    - ERP systems are designed to be scalable, accommodating the growth of an organization. As the business expands, the ERP system can adapt to increased data volumes, user counts, and business complexities.

Implementing an ERP system requires careful planning, customization to fit organizational needs, and comprehensive training for users. Successful ERP adoption can result in improved operational efficiency, better resource utilization, and a more agile and responsive organization.

How ERP systems reduce duplicate data entry



Traditional Systems



ERP Systems

ETL in Data Warehousing

ETL stands for Extract, Transform, Load, and it is a crucial process in data science and data warehousing. The ETL process involves extracting data from various sources, transforming it into a suitable format, and then loading it into a target database or data warehouse for analysis. Here are the steps involved in the ETL process, along with examples:

1. **Extract (E):**
    - In this step, data is extracted from diverse source systems such as databases, spreadsheets, APIs, logs, or flat files. The extraction process can involve both structured and unstructured data.

    Example:

    - Suppose you are working with a retail company, and you want to analyze sales data. You may extract data from a transactional database that contains information about sales, customer details, and product information.
2. **Transform (T):**
    - Data transformation involves cleaning, structuring, and converting the extracted data into a format suitable for analysis. This step includes handling

missing values, standardizing formats, aggregating data, and performing any necessary calculations.

Example:

- Continuing with the retail example, the data may contain inconsistencies in product names or missing values for certain transactions. In the transformation phase, you could standardize product names, handle missing values, and calculate total sales for each product.

3. **Load (L):**
    - The transformed data is loaded into a target database, data warehouse, or data mart, making it ready for analysis by data scientists, analysts, or business intelligence tools.

    Example:

    - After cleaning and transforming the sales data, it is loaded into a data warehouse where it can be efficiently queried for analysis. The data warehouse might have specific structures and optimizations to support complex queries and reporting.

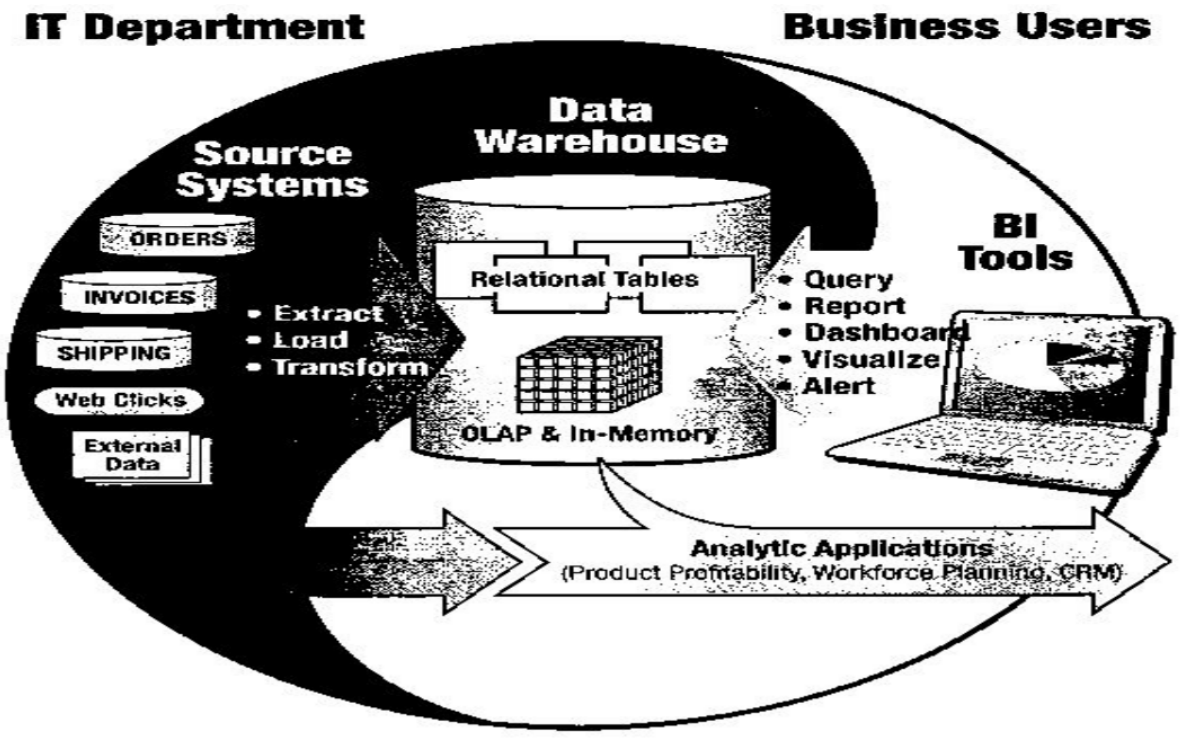
Popular solutions for ETL include Informática PowerCenter, IBM Infosphere DataStage, Oracle Data Integrator, and Microsoft Integration Services (a component of SQL Server)

Why not extract everything ?

In designing a data warehouse, requirements analysts will ask users what they need so that the ETL specialists can figure out what should be extracted from the source systems and avoid extracting everything. Some of the reasons for this is

1. The ***time window*** in which data can be ETL'd (extracted, transformed, and loaded) may be small, especially since many companies and data warehouses serve a global user base.
2. There can be a negative impact on query performance when too much detailed data is stored in the data warehouse.
3. Limited time, money, and human resources force a prioritization of what data to extract and include in the data warehouse.

Bisiness Intelligence Life Cycle



BI Life Cycle Diagram

Metadata

Metadata stores information about the data itself(data about the data). It may describe things like

- When the data was extracted from the source system
- When the data was loaded into the data warehouse
- From which source system an item originated
- From which physical table and field in the source system it was extracted
- Transformation rules and logic
- How something was calculated.
  For example,
  `revenue = (price x quantitysold) - discounts`
- What the item means in a business context (revenue is based on the amount invoiced and does not include returns or bad debts)

Data Warehouse

A **data warehouse** is a type of data mangement system that is designed to enable and support business intelligence (BI) activities, especially analytics. Data warehouses are solely intended to perform queries and analysis and often contain large amounts of historical data. The data within a data warehouse is usually derived from a wide range of sources such as application log files and transaction applications.

A data warehouse centralizes and consolidates large amounts of data from multiple sources. Its analytical capabilities allow organizations to derive valuable business insights from their data to improve decision-making. Over time, it builds a historical record that can be invaluable to data scientists and business analysts. Because of these capabilities, a data warehouse can be considered an organization's "single source of truth."

A typical data warehouse often includes the following elements:

- A relational database to store and manage data
- An extraction, loading, and transformation (ELT) solution for preparing the data for analysis
  Statistical analysis, reporting, and data mining capabilities
- Client analysis tools for visualizing and presenting data to business users
- Other, more sophisticated analytical applications that generate actionable information by applying data science and artificial intelligence (AI) algorithms, or graph and spatial features that enable more kinds of analysis of data at scale

Why do we need a Data Warehouse ? (Imp Question)

Most of the organisations or comapnies will need a data warehouse separate from the transaction system in the following cases:

1. They need to perform cross-subject or cross-functional analysis, such as products ordered versus inventory on hand. Such information may exist in two different systems or different modules within an ERP system and are thus combined into the data warehouse.
2. They want to perform analysis on summary information, aggregated by time (month, quarter) or by some other hierarchy (product groupings). These hierarchies often don't exist in transaction systems, and even when they do, running such voluminous queries within a transaction system can slow it to the point of interfering with data entry.

3. They need consistently fast reporting and analysis times.Because of their different purposes and design, data warehouses allow for faster queries than operational systems.

Within the data warehouse, data is physically stored in individual tables within a relational database.

Fact Tables and Dimension Tables

Fact Table:

A fact table is a central table in a data warehouse that contains quantitative data (facts) related to a business process or event. It typically includes numerical performance metrics and is surrounded by dimension tables. Fact tables are the core of a star or snowflake schema in a data warehouse.

Example:
Consider a retail data warehouse. The fact table might be "Sales," containing quantitative data such as the number of units sold, sales revenue, and discounts. Each row in the sales fact table represents a specific sale transaction, and it includes foreign keys that link to dimension tables like "Product," "Customer," and "Time."

Attributes of a Fact Table:

1. **Foreign Keys:** Fact tables contain foreign keys that link to the primary keys of dimension tables. These links establish relationships with dimensions, providing context to the quantitative data.
2. **Measures (Facts):** The numerical data or measures in a fact table represent the key performance indicators (KPIs) of the business process or event being analyzed.
3. **Granularity:** Fact tables have a specific level of granularity, representing the level of detail at which the data is stored. For example, a sales fact table might have a granularity of individual transactions.

Dimension Table:
A dimension table in a data warehouse provides descriptive information about the business entities related to a fact table. These entities, such as customers, products, or time, are the context for the quantitative data stored in the fact table. Dimension tables help in organizing and categorizing data, providing a way to slice and dice the information in the fact table.

Example:
Continuing with the retail data warehouse example, dimension tables might include:

- **Product Dimension:** Contains information about products such as product name, category, and manufacturer.
- **Customer Dimension:** Contains details about customers, such as customer name, address, and segment.
- **Time Dimension:** Contains time-related attributes, including day, month, quarter, and year.

Attributes of a Dimension Table:

1. **Primary Key:** Each dimension table has a primary key that uniquely identifies each record within that dimension.
2. **Descriptive Attributes:** Dimension tables contain descriptive attributes that provide context to the quantitative data in the fact table. These attributes help in filtering and grouping data during analysis.
3. **Hierarchies:** Dimension tables often include hierarchical structures. For example, a time dimension might have hierarchies such as year > quarter > month > day, facilitating various levels of time analysis.
4. **Foreign Keys:** Foreign keys in the fact table link to the primary keys in dimension tables, establishing relationships between the fact and dimension tables.

In summary, fact tables store quantitative data related to business events, while dimension tables provide the descriptive context for that data. Together, they form a dimensional model, which is a popular schema design in data warehousing for organizing and querying large volumes of data efficiently.

What are facts and dimensions?

Data warehousing terminology includes facts and dimensions. A fact is a piece of information with a specific numerical value, like a sale or a download. Facts are kept in fact tables, which are linked to several dimension tables by a foreign key. Facts are accompanied by dimensions, which describe the items in a fact table.

Why is the fact table larger than the dimension table?

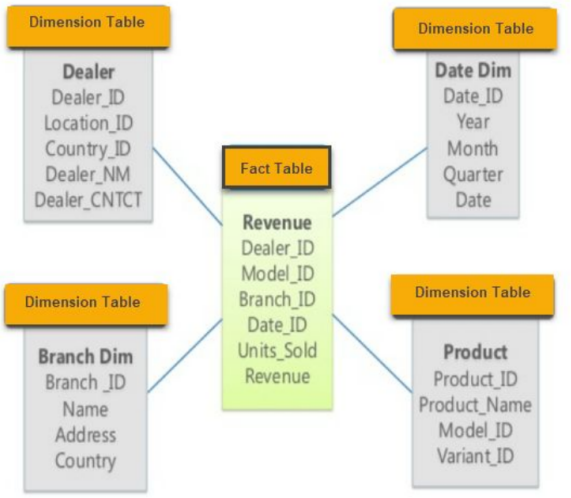
Fact tables have more records and fewer attributes, while dimension tables have more attributes and fewer records. While the dimension table expands horizontally,

the fact table expands vertically. While the dimension table has a primary key, the table has a concatenated key.

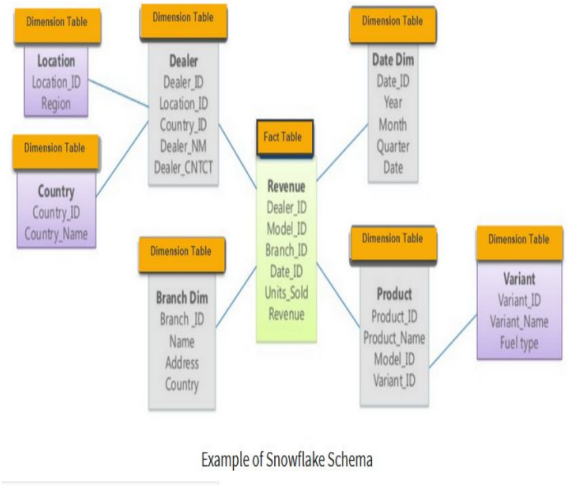
What is the relationship between facts and dimensions?

A single-dimension membership can be connected to many facts in the majority of dimensions, where each fact links to one and only one member of the dimension. This is known as each relationship in relational database jargon. However, connecting a single piece of evidence to several dimension members is frequently helpful.

Star Schema Vs Snowflake Schema



Example of Star Schema Diagram



Example of Snowflake Schema

In **Star Schema** a single fact table will be connected and surrounded by multiple dimension tables and only single join defines the relationship between the fact table and any dimension tables.

Snowflake Schema a single fact table is surrounded by dimension table which are in turn surrounded by subdimensional table hence snowflake schema requires multiple joins to fetch the data.

Star Schema	**Snowflake Schema**
Hierarchies for the dimensions are stored in the dimensional table.	Hierarchies are divided into separate tables.

It contains a fact table surrounded by dimension tables.	One fact table surrounded by dimension table which are in turn surrounded by dimension table
In a star schema, only single join creates the relationship between the fact table and any dimension tables.	A snowflake schema requires many joins to fetch the data.
Simple DB Design.	Very Complex DB Design.
Denormalized Data structure and query also run faster.	Normalized Data Structure.
High level of Data redundancy	Very low-level data redundancy
Single Dimension table contains aggregated data.	Data Split into different Dimension Tables.
Cube processing is faster.	Cube processing might be slow because of the complex join.
Offers higher performing queries using Star Join Query Optimization.Tables may be connected with multiple dimensions.	The Snowflake schema is represented by centralized fact table which unlikely connected with multiple dimensions.