"WITHOUT DATA, YOU ARE JUST ANOTHER PERSON WITH AN OPINION"-**W EDWARDS DEMING**

A DATA SCIENTIST'S TOOLKIT FOR DATA WAREHOUSING

BY JASLEEN KAUR SONDHI

A DATA SCIENTIST'S TOOLKIT FOR DATA WAREHOUSING

In this book, we will be covering the core concepts of Data Warehousing that a Data Scientist must be well equipped with.

By

Jasleen Kaur Sondhi

The Definitive Guide to Dimensional Modeling.

A DATA SCIENTIST'S TOOLKIT FOR DATA WAREHOUSING

NOTICE OF OPEN SOURCING

Jasleen Kaur Sondhi curated this book as a student of Christ University, Pune, Lavasa Campus. This book is entirely open-source and can be used for reference publicly.

AUTHORED BY:
Jasleen Kaur Sondhi

EDITED BY:
Jasleen Kaur Sondhi

CREATED AT CHRIST UNIVERSITY, PUNE, LAVASA CAMPUS
Available online (Open-Sourced)

First Online Edition, 2021

PREFACE

The first version of Jasleen Kaur Sondhi's Data Scientist's Toolkit for Data Warehousing is explicitly designed for students who want to gain Data Warehousing skills while working in Data Science.

Jasleen Kaur Sondhi has done a fantastic job covering the most fundamental components of Data Warehousing, and this book serves as a great introduction to the subject. We are excited to read the other chapters of this book and witness Ms Jasleen Kaur Sondhi's expertise imparted.

By,

Mr Jolinson Richi

DEDICATION

I dedicate this book to my Database Technologies Professor, Mr KT Thomas and my parents.

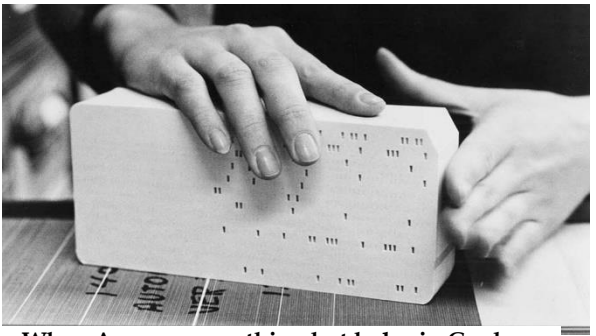
CONTENTS

Introduction

A Data Scientist cannot be a "Full Stack" Data Scientist without certain skills, and Data Warehousing is one of them. More often than not, young Data Scientists do not realise that getting raw data from various sources and making it ready for Data Modelling is 80% of the total work. And mostly, the enterprises involved have massive amounts of data stored in Data Warehouses.

Brief History of Data Warehousing

Storage of data is not a new practice. Once computers came into the picture, the curious minds of that time started thinking of ways to store the created data performing data redundancy and code reusability, even before knowing what these words meant. The first solution for storing computer-generated data was "Punch Cards", and soon enough, we evolved technologically and started using Database Management Systems.



When Apps were nothing but holes in Cards

But as the Internet was on the rise, so was the data being generated. Big enterprises realised that their data was poorly stored, linked and widely inconsistent.

That is when the concept of "Data Warehousing" came into the picture, and **Bill Inmon** was credited as the Father of Data Warehousing. So, Data Warehouses were now being created by enterprises to store the data they were taking from multiple sources in a consistent fashion.

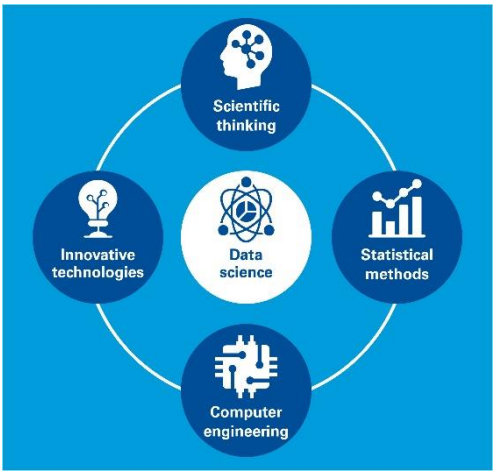
Bill Inmon- Father of Data Warehousing

CHAPTER 1: DATA SCIENCE AND DATA WAREHOUSING

Chapter 1
Data Science and Data Warehousing

What is Data Science?

To put it simply, it is the Science of the Data. A formal definition of Data Science would be: "It is an interdisciplinary field that use scientific methods, processes, and systems to extract information and patterns from noisy, structured, and unstructured data, and to apply that understanding and actionable insights to a plethora of different domains."



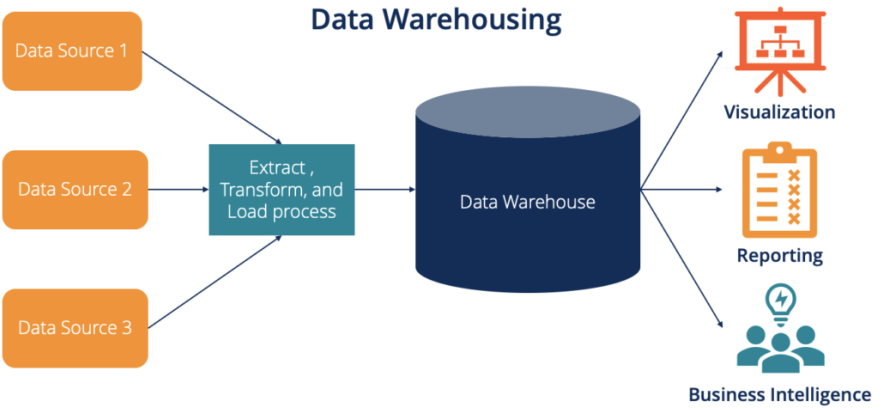
Who is a Data Scientist?

Data scientists are analytic experts that use their understanding of computers and social science to recognise trends and manage data. They uncover solutions to business challenges by integrating industry knowledge, contextual expertise, and questioning current assumptions.

Data scientists engage closely with business stakeholders to understand their goals and how data might enable them to achieve these. They create algorithms and modelling techniques to extract the data which the business needs, as well as help assess the data and share insights with peers.

What is a Data Warehouse?

A data warehouse is a type of data management system that is designed to guide and support business intelligence (BI) and analytics activities. Data warehouses are designed mainly for searching and analysis, and they typically hold vast quantities of historical data. The data warehouse's data is generally extracted from a number of sources, including application log files and transaction applications.



Why should a Data Scientist know Data Warehousing?

Data science is full of stories about data storage. In the pre-digital period, data was kept in our heads, on clay tablets, or on paper, which made data aggregation and interpretation exceedingly time-consuming. The 305 RAMAC, IBM's first commercial computer with a magnetic hard drive, was introduced in 1956. The complete unit took up 30 ft × 50 ft of physical area, weighed over a tonne, and can be leased for $3,200 per month to hold up to 5 MB of data. Prices per gigabyte of DRAM have declined dramatically in the 60 years since, from $2.64 billion in 1965 to $4.9 billion in 2017. Data storage became much denser/smaller in size, in addition to being magnitudes cheaper. The 305 RAMAC's disc held a hundred bits per square inch, compared to nearly a trillion bits every square inch in today's disc platters.

CHAPTER 2: BASICS OF DATA WAREHOUSE

Chapter 2
Basics of Data Warehouse

Features of a Data Warehouse

When the user has a shared means of explaining the patterns that are introduced as particular topics, the data warehouse can be regulated. The following are some of the important characteristics of a data warehouse:

1. **Subject-oriented-**Since it distributes information about a theme instead of an organisation's actual operations, and a data warehouse is always subject-oriented. It is feasible to do so with a certain theme. That is to add, and the data warehousing procedure is proposed to deal with a more defined theme. These themes may include sales, distribution, and marketing, for instance.

2. **Integrated-** A data warehouse is constructed by combining information from a variety of sources, such as a mainframe and a relational database. It must also have dependable naming conventions, formats, and codes. The utilisation of a data warehouse allows for more effective data analysis. The consistency of name conventions, column scaling, and encoding structure, among other things, should be validated. The data warehouse integration handles a variety of subject-related warehouses.

3. **Time-Variant**- Data is kept in this system at various time intervals, such as weekly, monthly, or annually. It discovers a number of time limits that are structured between massive datasets and held in the online transaction process (OLTP). Data warehouse time

limitations are more flexible than those of operational systems. The data in the data warehouse is predictable over a set period of time and provides information from a historical standpoint. It contains explicit or implicit time elements. Another property of time variance is that data cannot be edited, altered, or updated once it has been placed in the data warehouse.

4. **Non-Volatile**- The data in a data warehouse is permanent, as the name implies. It also means that when new data is put, it is not erased or removed. It incorporates a massive amount of data that is placed into logical business alteration between the designated quantity. It assesses the analysis in the context of warehousing technologies.

Components of a Data Warehouse

A typical data warehouse has four main components: a central database, ETL (extract, transform, load) tools, metadata, and access tools. These components are designed to work quickly, allowing you to acquire findings and examine data on the fly.

The components of Data Warehouse are as follows-

1)Source Data

The data that enters the data warehouses can be divided into four categories:

Production Data: This type of information originates from the company's various operating systems. We select data segments from the various operational modes based on the data warehouse's data requirements.

Internal Data: Each organisation's "secret" spreadsheets, reports, customer profiles, and occasionally even department databases are kept by the client. This is internal information, some of which may be helpful in a data warehouse.

Data that has been archived: Operational systems are primarily designed to execute current operations. We take outdated data and put it in achieved files on a regular basis in every operating system.

External Data: For a major portion of the information they utilise, most executives rely on data from outside sources. They rely on statistics produced by an external department that are related to their industry.

2)Data Staging

1. Data Extraction: This method must work with a variety of data sources. For each data source, we must use the proper procedures.

2. Data Transformation: Data for a data warehouse originates from a variety of sources, as we all know. If data extraction for a data warehouse is a difficult task, data transformation is even more difficult. As part of data transformation, we perform a number of different tasks.

Data transformation includes a lot of standardisation of data components. Data transformation encompasses a variety of methods for merging data from various sources. Data from a single source record or comparable data components from many source records are combined.

3. Data Loading: Data loading functions are made up of two main groups of tasks. We do the loading of the information initially into the data warehouse storage once we finish designing the structure and construction of the data warehouse and go online for the first time. The first load moves large amounts of data over a long period of time.

3)Data Storing

A split repository is used to store data for data warehousing. In most cases, the data repositories for operational systems only include current data. These data repositories also contain data that is arranged in a highly standardised format for quick and efficient processing.

4)Information Delivery

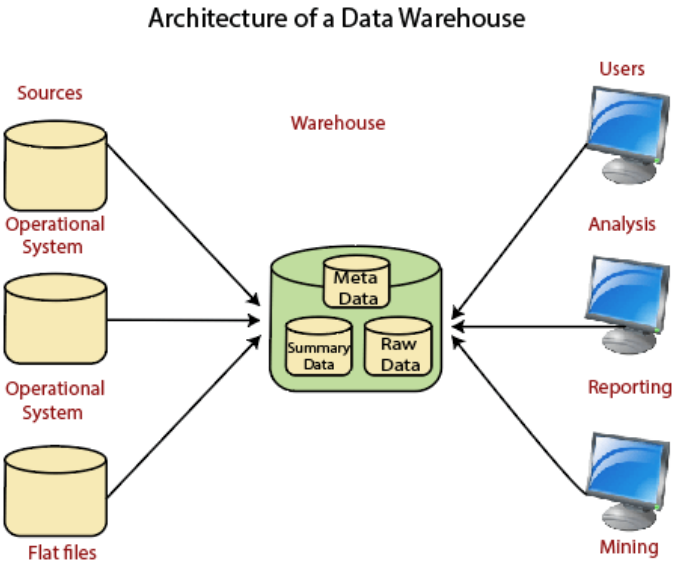
The information delivery element is used to allow customers to subscribe to data warehouse files and have them delivered to one or more destinations according to a customer-specified scheduling mechanism.

5)Meta Data

A data warehouse's metadata is the same as a database management system's data dictionary or catalogue. We retain information about logical data structures, records and addresses, indexes, and so on in the data dictionary.

The architecture of Data Warehouse

The data warehouse architecture is a way of describing the total architecture of data communication, processing, and presentation for end-user computing within an organisation. Although each data warehouse is different, they all share the same basic components.



A data warehouse's architecture can be classified into three types: single-tier, double-tier or two-tier, and three-tier approaches.

- **Single Tier Approach**- Using a single-tier method reduces storage and redundancy.
- **Double Tier Approach**- The two-tier or double-tier approach isn't employed to circumvent network limits, limited growth capabilities, and the inability to service end-users.
- **Three Tier Approach-** The three-tier technique is the most extensively utilised and is also the most user-friendly.

Operational System

An operational system is a word used in data warehousing to describe a system that processes an organisation's day-to-day transactions.

Flat Files

A flat file system is a collection of files in which transactional data is stored, with each file having a unique name.

Meta Data

A collection of data that defines and describes other data.

Metadata is utilised in the Data Warehouse for a variety of reasons, including:

Meta Data is a type of data that summarises important information about it, making it easier to discover and operate with specific instances of data. Author, data build, data altered, and file size, for example, are all instances of extremely basic document information.

A query's metadata is used to steer it to the most relevant data source.

CHAPTER 3: DIMENSIONAL MODELLING

Chapter 3
Dimensional Modelling

What is Dimensional Modelling?

Dimensional Modeling is a technique based on data structures that are specifically designed for data storage in a data warehouse. The goal of dimensional modelling is to optimise the database so that data can be retrieved more quickly. Ralph Kimball created the notion of Dimensional Modelling, which consists of "fact" and "dimension" tables.

Steps in Dimensional Modelling

Step 1: Determining the business goal

The first stage is to determine the business goal. Sales, HR, Marketing, and so on are some examples, depending on the needs of the company. Because it is the most significant step in Data Modelling, the quality of data available for that process influences the choice of business aim.

Step 2: Identifying Granularity

Granularity refers to the table's lowest level of data storage. Grain describes the level of information for a business challenge and its solution.

Step 3: Determining Dimensions and Attributes

Objects or things have dimensions. Dimensions organise and characterise data warehouse facts and metrics so that they can be used to answer

business questions. In dimension tables, descriptive attributes are organised as columns in a data warehouse. For example, the data dimension could include information such as the year, month, and weekday.

Step 4: Identifying the Fact

The fact table is where the quantifiable data is kept. The majority of the rows in the fact table are numerical values such as price or cost per unit.

Step 5: Schema Construction

In this step, we put the Dimension Model into action. A database structure is referred to as a schema. Star Schema and Snowflake Schema are two prominent systems.

Basic Fact Table Techniques

A table that contains the numeric measures produced by an operational measurement event in the real world is known as a fact table. Its's row corresponds to a measurement event and vice versa, at the lowest grain.

There are three types of numeric measures in a fact table. Facts that are fully additive are the most versatile and valuable; additive measures can be totalled across any of the fact table's dimensions.

In fact tables, null-valued measurements behave gracefully. The aggregate functions (SUM, COUNT, MIN, MAX, and AVG) all handle null facts correctly. Nulls in the fact table's foreign keys, on the other hand, must be avoided because they will create a referential integrity violation.

A record in a transaction fact table that corresponds to a transaction.

A row in a transaction fact table corresponds to a measurement event at a point in space and time. In a periodic snapshot fact table, a row represents a collection of measurement events that occurred during a particular time period, such as a day, week, or month. The era, not the specific transaction, is the grain. Because any measurement event compliant with the fact table

grain is acceptable, periodic snapshot fact tables frequently contain a large number of facts.

Basic Dimension Table Techniques

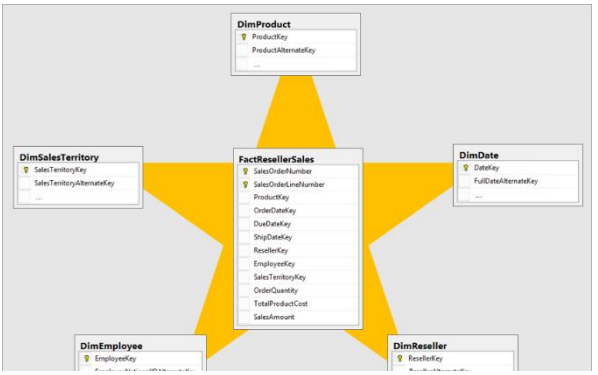
A single primary key column exists in each dimension table. This primary key is used as a foreign key in any associated fact table where the descriptive context of the dimension row matches that of the fact table row.

A dimension table has a single column that serves as the primary key. When changes are monitored over time, this main key cannot be the operational system's natural key because there will be numerous dimension rows for that natural key.

Drilling down is the most fundamental way data is analysed by business users.

Star Schema

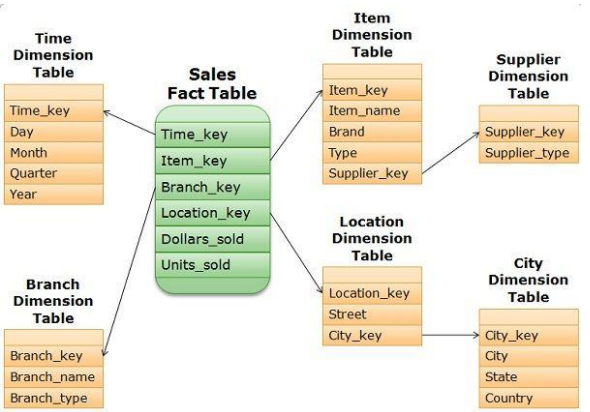
The star schema divides business process data into facts, which contain measurable, quantitative information about a company, and dimensions, which are descriptive features associated with fact data. Sales price, amount sold, and time, distance, speed, and weight measurements are all examples of fact data. Product models, product colours, product sizes, geographic locations, and salesperson names are all examples of related dimension attributes.



Example of Star Schema

Snowflake Schema

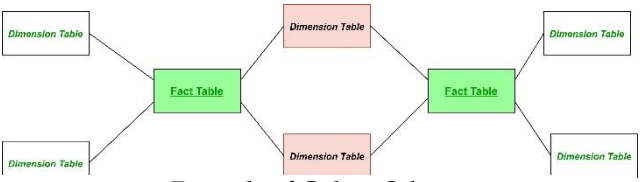
A snowflake schema in a multidimensional database is a logical arrangement of tables that resembles a snowflake form on an entity-relationship diagram. The snowflake schema is made up of centralised fact tables that are linked to a variety of dimensions. A method of standardising the dimension tables in a star schema is known as "snowflaking." When all of the dimension tables are normalised, the resultant structure resembles a snowflake with the fact table in the middle. Snowflaking is the process of normalising dimension tables by deleting low cardinality features and separating them into distinct tables.



Example of Snowflake Schema

Galaxy Schema

Fact constellation is a metric for online analytical processing that is made up of several fact tables that share dimension tables and is visualised as a constellation of stars. It can be considered to be a continuation of the star schema. There are several fact tables in a fact constellation structure. Galaxy schema is another name for it.



Example of Galaxy Schema

THANKS FOR READING!