



Data Warehousing Tutorial for Beginners

Introduction

Data Warehousing is like trying to catalog a library where all books contain information in various languages that are delivered at random times. It is often challenging for beginners to be able to differentiate between a normal database and a warehouse or to be confused by terms such as ETL, OLAP, and Schemas within the “alphabet soup.” Do not worry if you are having problems comprehending where to start to combine disorganized information from multiple sources to create a “single source of truth” that is not slowed down by system performance.

This data warehousing tutorial not only simplifies complex architectures but also goes through the processes involved in cleaning, transforming, and storing data for high-speed business intelligence.

Ready to Become a Data Architect? Download our full [Data Warehousing Course Syllabus](#) to understand the full journey, right from dimensional design to cloud-based implementations like Snowflake and BigQuery.

Why Students or Freshers Learn Data Warehousing?

Learning about data warehousing is one of the most influential career-changing choices for a college-going individual or a fresher.

As the world of AI & Big Data expands with each passing year, it has given rise to a requirement for a clean & well-structured data setup, which is where Data Warehousing comes into play, as it is described as the “backbone” of today’s corporate intelligence.

- **Huge Demand in the Market:** Businesses in Finance, Ecommerce, and Healthcare are transitioning to cloud-based data warehouses such as Snowflake and BigQuery, resulting in a massive demand for fresh junior ETL and DWH developers.
- **Best Compensation Packages:** In India, the average beginning salary for fresher talent in Data Warehousing ranges from 4.5 to 7 lakhs per year, which may even beat software developer rates.
- **Foundation for AI/ML:** AI is only as strong as its data. DWH allows for an environment where it can obtain “clean” data from history used for successful ML models.
- **Core Engineering Skills:** You learn skills in tech areas like SQL, Data Modeling (Star and Snowflake Data Warehouse schema) and ETL processes that can be applied anywhere in the data field.
- **Future-Proof Career:** There may be fluctuations in careers related to the tech field, but requirements in nimble storage and processing of business history are always required.

Ace Your Data Technical Round Get the jump on the competition with our comprehensive [Data Warehousing Interview Questions and Answers](#), which go in-depth on topics ranging from Slowly Changing Dimensions to Fact versus Dimension tables, as well as Cloud DWH Architecture.

Take your Knowledge
Test Report

Check your Score



Step-by-Step Data Warehousing Tutorial for Beginners

Data warehousing is an essential subject that every data analyst should consider. This tutorial is a starting point for data warehousing experts to know more about Data Warehousing (DWH) in detail. Data warehousing plays a vital role in the modern world because data warehousing is considered the 'Central Source of Truth' for any organization. It is a high-speed repository that stores data from various operational systems for decision-making.

Step 1. Database Architectures: Understanding Data Ware

A Data Warehouse is more than a large database; in fact, a Data Warehouse is a multi-level environment especially designed for performing Online Analytical Processing (OLAP) operations.

- **Bottom Tier:** The Data Warehouse Server where data will be placed.
- **Middle Tier:** OLAP engine that breaks up data into "cubes." These cubes are used for quick "slicing and dicing" analysis.
- **Top Tier:** The front-end reporting tools used for data visualization, such as Power BI or Tableau.

Step 2. Installation & Setup

For a beginner, the most convenient way to get familiarized is to either create a PostgreSQL environment yourself or use a cloud-based data warehousing solution such as Snowflake/Google BigQuery (which has a free tier available).

2.1: Local Setup with PostgreSQL

1. Install PostgreSQL.
2. Install pgAdmin 4 (the GUI for managing your database).
3. Open pgAdmin by clicking on it and then right-clicking on "Databases" and creating a new database and name it *Sales_Warehouse*.

2.2: Selection of ETL Tool

For transferring your data from your sources, whether it is in the form of CSV or Excel files, to your storage, or your data warehouse, you would require an ETL tool.

- **For Beginners:**
 - **Programming:** Python language using the 'pandas' library.
 - **GUI-based tools:** Talend Open Studio.
- **Cloud Option:** Use AWS Glue or Azure Data Factory.

Step 3. Your Schema Design (Blueprint)

Normalization is used by standard databases to conserve space. Denormalization (Star & Snow Flake Schemas) in Data Warehouses emphasizes faster query processing.

The Star Schema:

It is the most popular model. It includes:

- **Fact Table:** Stores numerical information such as *Sales_Amount*, *Quantity*.
- **Dimension Tables:** These store the descriptive attributes (for example, *Product_Name*, *Store_Location*, *Date*.)

Step 4. Building Your First Warehouse (Code Example)

So, let's design a simplified Star Schema for a retail store using SQL.

4.1: Create Dimension Tables

— *Create Product Dimension*

```
CREATE TABLE Dim_Product (  
  
    Product_Key SERIAL PRIMARY KEY,  
  
    Product_Name VARCHAR(100),  
  
    Category VARCHAR(50),
```

Brand VARCHAR(50)

);

— *Create Date Dimension*

CREATE TABLE Dim_Date (

Date_Key INT PRIMARY KEY,

Full_Date DATE,

Year INT,

Month INT,

Day_Of_Week VARCHAR(15)

);

4.2: Create the Fact Table

CREATE TABLE Fact_Sales (

Sales_ID SERIAL PRIMARY KEY,

Product_Key INT REFERENCES Dim_Product(Product_Key),

Date_Key INT REFERENCES Dim_Date(Date_Key),

Quantity_Sold INT,

Total_Revenue DECIMAL(10,2)

);

Step 5. ETL Procedure: Raw Data to Refined

This is where the 'magic' takes place. Below is a simplified logic of how one could transform extracted data into meaningful data using Python programming language.

```
import pandas as pd
```

```
# 1. EXTRACT: Load raw CSV data
```

```
raw_data = pd.read_csv("daily_sales.csv")
```

```
# 2. TRANSFORM: Clean and format data
```

```
# Convert date string to actual datetime objects
```

```
raw_data['date'] = pd.to_datetime(raw_data['date'])
```

```
# Handle missing values by filling with 0
```

```
raw_data['revenue'] = raw_data['revenue'].fillna(0)
```

```
# 3. LOAD: (Conceptual) Send to your SQL Warehouse
```

```
# raw_data.to_sql('Fact_Sales', engine, if_exists='append')
```

Step 6. Common Challenges and Solutions

- **Data Quality Issues:** Duplicates or incomplete records.
 - **Solution:** Data Profiling has to be implemented in ETL for rejecting bad rows even before they reach the data warehouse.
- **Slow Queries:** Joins become weighty as data size increases.
 - **Solution:** Index on primary/foreign keys and establish materialized queries for frequently run queries.
- **Cost of Scaling:** High Cost of Storing History.
 - **Solution:** Maintain Cold Storage when data is over 5 years old, with "Hot" data housed in our central warehouse.

Ready to unlock industry-level data puzzles? Take a look at our [Data Warehousing Challenges and Solutions](#) resource that delves into detail on processing CDC (Change Data Capture), slowly changing dimensions (SCD Type 2), and petabyte cloud optimizing.

Real Time Examples for Data Warehousing Tutorial for Learners

To fully grasp the skills involved with Data Warehousing, understanding the way that this technology works as the “brain” of the corporate world, taking millions of disconnected “data points” and using them to inform business decisions, can be very helpful. Below are three ways that this technology is being used:

E-Commerce: Hyper-Personalized Buying (The “Amazon Effect”)

The Scenario: A retail company wishes to provide personalized “Buy it Again” recommendations and discounts the moment the user logs in with considering the browsing history of many years.

- **The DWH Approach:** The raw data available through clicks on the websites, mobile apps, and payment transactions is received through an ETL approach in an environment such as Snowflake.
- **The Purpose:** With the help of Star Schemas, the ability to instantly combine ‘User Dimensions’ and ‘Purchase Facts’ means that the system could easily determine a consumer’s ‘Lifetime Value’ and send a unique discount code to the consumer within milliseconds to avoid ‘Cart Abandonment.’

Healthcare: Predictive Patient Care

The Scenario: A hospital network needs to identify which diabetic patients are at the highest risk of readmission before they even leave the facility.

- **DWH Approach:** It combines Electronic Health Records (EHR), prescriptions from pharmacy services, and even IoT wearables (such as heart rate monitors) into one system.
- **The Goal:** Through Historical Trend Analysis, the warehouse compares the current patient’s vital statistics with millions of historical patients. If the pattern matches the “high risk” profile that emerged five years earlier, it prompts the

doctor to make an instant change in the treatment plan, which improves patient health dramatically.

Banking: Real-Time Fraud Detection

The Scenario: A credit card transaction occurs in London, followed by a subsequent transaction in New York, 10 minutes later. The bank must make an immediate decision about which transaction to flag.

- **The DWH Approach:** Banks employ their Active Data Warehouse (ADW) system that consumes worldwide streaming data from both ATMs as well as Point-of-Sale Terminals.
- **The Purpose:** The warehouse retains a 'Normal Behavior Profile' for each user. As soon as the New York transaction occurs, the high-speed query is executed against the 'history' information for that particular user. Upon recognizing the impossibility in terms of geography, the transaction is labeled 'Fraudulent' along with an SMS notification sent to the customer.

Planning to put theory into practice? Check out our [Data Warehousing Project Ideas](#) for Beginners, including tutorials on how to create a COVID-19 Global Tracker, a Movie Recommendation Engine based on the IMDb dataset, and a Retail Inventory Management System.

FAQs About Data Warehousing Tutorial for Beginners

1.What is meant by data warehousing?

Data Warehousing is a process where all data is collected and consolidated from multiple sources and stored in a single location. When compared to other normal databases used for day-to-day operations, a Data Warehouse is designed to provide analytical support to businesses and help organizations recognize any trends that emerge over time.

2.What is a data warehouse and an example?

The Data Warehouse is the electronic system used in this process. For instance, a multi-national chain like Walmart uses a data warehouse system where data is coupled

in order to determine what items to replenish based on clicks and supply chain data among thousands of stores.

3.What are the three types of data warehousing?

Enterprise Data Warehouse (EDW): It is an organizational warehouse that stores data for the entire organization. Enterprise Data Warehouse is used to:

Operational Data Store (ODS): It is employed for real-time reporting for current, or daily, data.

Data Mart: Typically a mini-warehouse with a focus on a specific department, like finance or marketing.

4.What are the 4 components of data warehouse?

Load Manager: It does the ETL function (Extract, Transform, Load).

Warehouse Manager: Manages data storage, indexing, and backup.

Query Manager: Handles requests submitted by users and accelerates the retrieval process of data.

End-User Access Tools: Dashboards/SQL tools for query/analysis of the data.

5.What are data warehouse tools?

Some popular tools are: Cloud-based solutions include Snowflake, Amazon Redshift, Google BigQuery On-Premise: Oracle Data Warehouse and IBM Db2. Open Source: PostgreSQL (can be optimized for warehouse applications).

6.What is ETL in data warehousing?

ETL is an acronym for Extract, Transform, and Load. This is what is meant by “pipeline”:

Extract: Getting data from sources.

Transform: Cleaning and formatting it, for instance, converting all currencies to USD.

Load: Moving the polished data to the warehouse.

7.Why is data warehousing?

Its purpose is to offer “Single Source of Truth.” If it did not exist, it would be possible that different departments could have different information. It enables executives to perform complicated queries on historical data without hindering the applications that its customers were using.

8.Is Amazon a data warehouse?

Amazon.com (the store) is a data warehousing user, while the provider in this relationship is Amazon Web Services (AWS). Its product, Amazon Redshift, is the most widely used cloud data warehouse globally. Thousands of other companies use it.

9.What skills are needed for data warehousing?

Advanced SQL: Query and manage big datasets.

Data Modeling: Star and Snowflake

Schemas: For writing automation scripts in Python/Java.

Cloud Platforms: Familiarity with AWS, Azure, Google Cloud.

10.Which company uses a data warehouse?

Almost all large-scale organizations rely on one of them. Some of its applications are Netflix (for analyzing viewing patterns), Uber (for analyzing patterns of trips), American Express (for analyzing fraud patterns), and Coca-Cola (for managing global inventory).

Explore [Data Warehousing Salary for Freshers](#).

Conclusion

Skill in Data Warehousing involves more than handling big data, it involves building a high performance architecture with strategic decision-making capabilities. By gaining an ability to work with ETL processes, optimize a Star Schema design, and integrate cloud solutions such as Snowflake or Big Query into this system, one closes the gap between data and decision-making capabilities. This set of skills represents the foundation of Business Intelligence today and represents fuel for successful machine learning projects. Ready to Learn the Infrastructure of the Future? To gain experience with real-world data, complex models, and cloud warehouse management, sign up for our Professional [Data Warehousing Course in Chennai](#) to receive our Data Warehousing Certification.