CLOUD APPLICATION DEVELOPMENT (GROUP 1)

PHASE 4 : ASSIGNMENT NOTEBOOK SUBMISSION

NAME :JEMANATHAN.M

EMAIL : [chandruvasuki1@gmail.com](mailto:chandruvasuki1@gmail.com)

GitHub Repository URL : <https://github.com/hemanathan457/Hemanathan.git>

Title of the Project :-

PROJECT 7 : Data warehouse on IBM Cloud Foundary

# Data Warehousing

## **Abstract:**

Data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker to make better and faster decisions. A data warehouse is a subject-oriented, integrated, time varying, non-volatile collection of data that is used primarily in organizational decision making. Data warehouse supports on-line analytical processing, the functional and performance requirements of which are quite different from those of the on-line transaction processing applications traditionally supported by the operational databases. In this paper author suggest that for a strategic solution, the hub and spoke / centralised architecture is the more likely choice

[

import logging

import time

import mysql.connector

# Set up logger

logger = logging.getLogger(\_\_name\_\_)

logger.setLevel(logging.INFO)

formatter = logging.Formatter("%(asctime)s - %(name)s - %(levelname)s - %(message)s")

# Log to console

handler = logging.StreamHandler()

handler.setFormatter(formatter)

logger.addHandler(handler)

# Also log to a file

file\_handler = logging.FileHandler("cpy-errors.log")

file\_handler.setFormatter(formatter)

logger.addHandler(file\_handler)

def connect\_to\_mysql(config, attempts=3, delay=2):

attempt = 1

# Implement a reconnection routine

while attempt < attempts + 1:

try:

return mysql.connector.connect(\*\*config)

except (mysql.connector.Error, IOError) as err:

if (attempts is attempt):

# Attempts to reconnect failed; returning None

logger.info("Failed to connect, exiting without a connection: %s", err)

return None

logger.info(

"Connection failed: %s. Retrying (%d/%d)...",

err,

attempt,

attempts-1,

)

# progressive reconnect delay

time.sleep(delay \*\* attempt)

attempt += 1

return None

**Process of data warehousing:-**

Data warehousing is a process of collecting, integrating, and organizing data from various sources into a centralized repository for analytical purposes. Data warehousing involves data cleaning, data integration, and data consolidation. Data warehousing supports analytical reporting, structured and/or ad hoc queries, and decision making. Data warehousing also helps in improving data quality, faster queries, better business analytics, and historical insight.

Some of the basic concepts of data warehousing are:

- Data Warehouse: A data warehouse is a centralized storage system that allows for the storing, analyzing, and interpreting of data in order to facilitate better decision-making. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries, and decision making¹²³.

- Data Source: A data source is any system or application that produces or contains data that can be used for data warehousing. Data sources can be transactional databases, operational systems, external sources, files, etc¹².

- Data Integration: Data integration is the process of combining data from different sources into a consistent and unified view. Data integration involves data extraction, transformation, and loading (ETL) processes that move data from source systems to the data warehouse¹²⁴.

- Data Quality: Data quality is the degree to which data meets the expectations and requirements of the data consumers. Data quality involves ensuring the accuracy, completeness, consistency, timeliness, validity, and uniqueness of the data in the data warehouse¹⁴.

- Data Model: A data model is a logical representation of the structure and meaning of the data in the data warehouse. A data model defines the entities, attributes, relationships, constraints, and business rules of the data. A data model can be conceptual, logical, or physical¹⁴.

- Data Mart: A data mart is a subset of a data warehouse that is tailored to meet the specific needs of a particular business unit or department. A data mart can be independent or dependent on the data warehouse. A data mart can have its own data model, ETL processes, and analytical tools¹²⁴.

from \_\_future\_\_ import print\_function

from datetime import date, datetime, timedelta

import mysql.connector

cnx = mysql.connector.connect(user='scott', database='employees')

cursor = cnx.cursor()

tomorrow = datetime.now().date() + timedelta(days=1)

add\_employee = ("INSERT INTO employees "

"(first\_name, last\_name, hire\_date, gender, birth\_date) "

"VALUES (%s, %s, %s, %s, %s)")

add\_salary = ("INSERT INTO salaries "

"(emp\_no, salary, from\_date, to\_date) "

"VALUES (%(emp\_no)s, %(salary)s, %(from\_date)s, %(to\_date)s)")

data\_employee = ('Geert', 'Vanderkelen', tomorrow, 'M', date(1977, 6, 14))

# Insert new employee

cursor.execute(add\_employee, data\_employee)

emp\_no = cursor.lastrowid

# Insert salary information

data\_salary = {

'emp\_no': emp\_no,

'salary': 50000,

'from\_date': tomorrow,

'to\_date': date(9999, 1, 1),

}

cursor.execute(add\_salary, data\_salary)

# Make sure data is committed to the database

cnx.commit()

cursor.close()

cnx.close()

- Online Analytical Processing (OLAP): OLAP is a technique for analyzing multidimensional data using various operations such as slice and dice, drill down, roll up, pivot, etc. OLAP enables users to interactively explore and manipulate large amounts of data from different perspectives and levels of detail¹²⁴.

- Data Mining: Data mining is the process of discovering hidden patterns and trends from large volumes of data using various techniques such as classification, clustering, association analysis, regression analysis, etc. Data mining helps in extracting useful information and knowledge from the data warehouse for decision making¹²⁴.

(4) Introduction to Data Warehousing: Definition, Concept, and Techniques. https://www.digitalvidya.com/blog/introduction-to-data-warehousing/.

**Background**   
A Database Management System (DBMS) stores data in the form of tables, uses ER model and the goal is [ACID properties](https://www.geeksforgeeks.org/acid-properties-in-dbms/). For example, a DBMS of college has tables for students, faculty, etc.

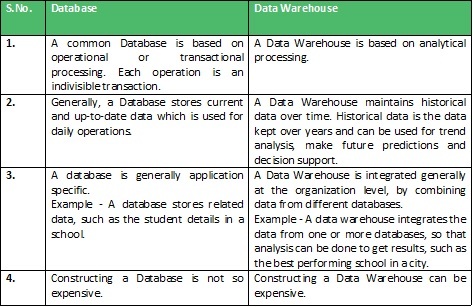
A **Data Warehouse**is separate from DBMS, it stores a huge amount of data, which is typically collected from multiple heterogeneous sources like files, DBMS, etc. The goal is to produce statistical results that may help in decision makings. For example, a college might want to see quick different results, like how the placement of CS students has improved over the last 10 years, in terms of salaries, counts, etc.

**Need for Data Warehouse**  
An ordinary Database can store MBs to GBs of data and that too for a specific purpose. For storing data of TB size, the storage shifted to Data Warehouse. Besides this, a transactional database doesn’t offer itself to analytics. To effectively perform analytics, an organization keeps a central Data Warehouse to closely study its business by organizing, understanding, and using its historic data for taking strategic decisions and analyzing trends.

**Benefits of Data Warehouse:**

1. **Better business analytics:**Data warehouse plays an important role in every business to store and analysis of all the past data and records of the company. which can further increase the understanding or analysis of data to the company.
2. **Faster Queries:**Data warehouse is designed to handle large queries that’s why it runs queries faster than the database.
3. **Improved data Quality:**In the data warehouse the data you gathered from different sources is being stored and analyzed it does not interfere with or add data by itself so your quality of data is maintained and if you get any issue regarding data quality then the data warehouse team will solve this.
4. **Historical Insight:**The warehouse stores all your historical data which contains details about the business so that one can analyze it at any time and extract insights from it

**Data Warehouse vs DBMS** 

[](https://www.geeksforgeeks.org/difference-between-database-system-and-data-warehouse/)

**Example Applications of Data Warehousing**   
Data Warehousing can be applied anywhere where we have a huge amount of data and we want to see statistical results that help in decision making.

* **Social Media Websites:** The social networking websites like Facebook, Twitter, Linkedin, etc. are based on analyzing large data sets. These sites gather data related to members, groups, locations, etc., and store it in a single central repository. Being a large amount of data, Data Warehouse is needed for implementing the same.
* **Banking:**Most of the banks these days use warehouses to see the spending patterns of account/cardholders. They use this to provide them with special offers, deals, etc.
* **Government:** Government uses a data warehouse to store and analyze tax payments which are used to detect tax thefts.

|  |
| --- |
| # importing required libraries  **import** mysql.connector    dataBase **=** mysql.connector.connect(    host **=**"localhost",    user **=**"user",    passwd **=**"password",    database **=** "gfg"  )    # preparing a cursor object  cursorObject **=** dataBase.cursor()    query **=** "SELECT \* FROM STUDENT LIMIT 2 OFFSET 1"  cursorObject.execute(query)    myresult **=** cursorObject.fetchall()    **for** x **in** myresult:      print(x)    # disconnecting from server  dataBase.close() |

**Output:**

**('Nikhil', 'CSE', 98, 'A', 18)**

**('Nisha', 'CSE', 99, 'A' , 18)**

### Features :

**Centralized Data Repository:** Data warehousing provides a centralized repository for all enterprise data from various sources, such as transactional databases, operational systems, and external sources. This enables organizations to have a comprehensive view of their data, which can help in making informed business decisions.

**Data Integration:**Data warehousing integrates data from different sources into a single, unified view, which can help in eliminating data silos and reducing data inconsistencies.

**Historical Data Storage:** Data warehousing stores historical data, which enables organizations to analyze data trends over time. This can help in identifying patterns and anomalies in the data, which can be used to improve business performance.

**Query and Analysis:**Data warehousing provides powerful query and analysis capabilities that enable users to explore and analyze data in different ways. This can help in identifying patterns and trends, and can also help in making informed business decisions.

**Data Transformation:**Data warehousing includes a process of data transformation, which involves cleaning, filtering, and formatting data from various sources to make it consistent and usable. This can help in improving data quality and reducing data inconsistencies.

**Data Mining:** Data warehousing provides data mining capabilities, which enable organizations to discover hidden patterns and relationships in their data. This can help in identifying new opportunities, predicting future trends, and mitigating risks.

**Data Security:**Data warehousing provides robust data security features, such as access controls, data encryption, and data backups, which ensure that the data is secure and protected from unauthorized access.

### Advantages:

**Improved data quality:**Data warehousing can help improve data quality by consolidating data from various sources into a single, consistent view.

**Faster access to information:** Data warehousing enables quick access to information, allowing businesses to make better, more informed decisions faster.

**Better decision-making:** With a data warehouse, businesses can analyze data and gain insights into trends and patterns that can inform better decision-making.

**Reduced data redundancy:** By consolidating data from various sources, data warehousing can reduce data redundancy and inconsistencies.

**Scalability:**Data warehousing is highly scalable and can handle large amounts of data from different sources.

### Disadvantages:

**Cost:**Building a data warehouse can be expensive, requiring significant investments in hardware, software, and personnel.

**Complexity:**Data warehousing can be complex, and businesses may need to hire specialized personnel to manage the system.

**Time-consuming:**Building a data warehouse can take a significant amount of time, requiring businesses to be patient and committed to the process.

**Data integration challenges:**Data from different sources can be challenging to integrate, requiring significant effort to ensure consistency and accuracy.

**Data security:**Data warehousing can pose data security risks, and businesses must take measures to protect sensitive data from unauthorized access or breaches.

There can be many more applications in different sectors like E-Commerce, telecommunications, Transportation Services, Marketing and Distribution, Healthcare, and Retail.