Datawarehousing With IBM Cloud Db2 Warehouse

Phase 4 : Development part 2

Continue building the data warehouse by implementing ETL Processes and enabling data exploration.

ETL Process in Data warehouse

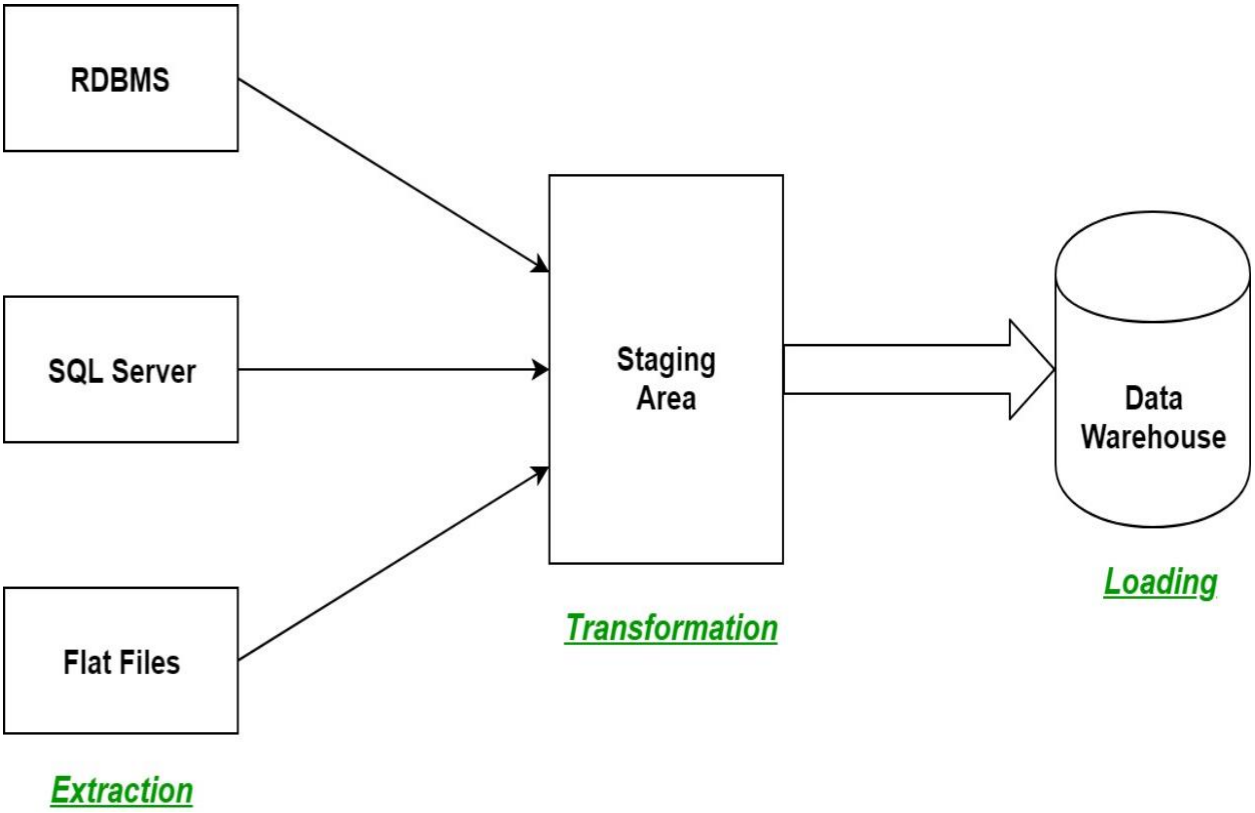
ETL stands for Extract, Transform, Load and it is a process used in data warehousing to extract data from various sources, transform it into a format suitable for loading into a data warehouse, and then load it into the warehouse.

1. **Extract:** The first stage in the ETL process is to extract data from various sources such as transactional systems, spreadsheets, and flat files. This step

involves reading data from the source systems and storing it in a staging area.

2.Transform: In this stage, the extracted data is transformed into a format that is suitable for loading into the data warehouse. This may involve cleaning and validating the data, converting data types, combining data from multiple sources, and creating new data fields.

3. Load: After the data is transformed, it is loaded into the data warehouse. This step involves creating the physical data structures and loading the data into the warehouse.

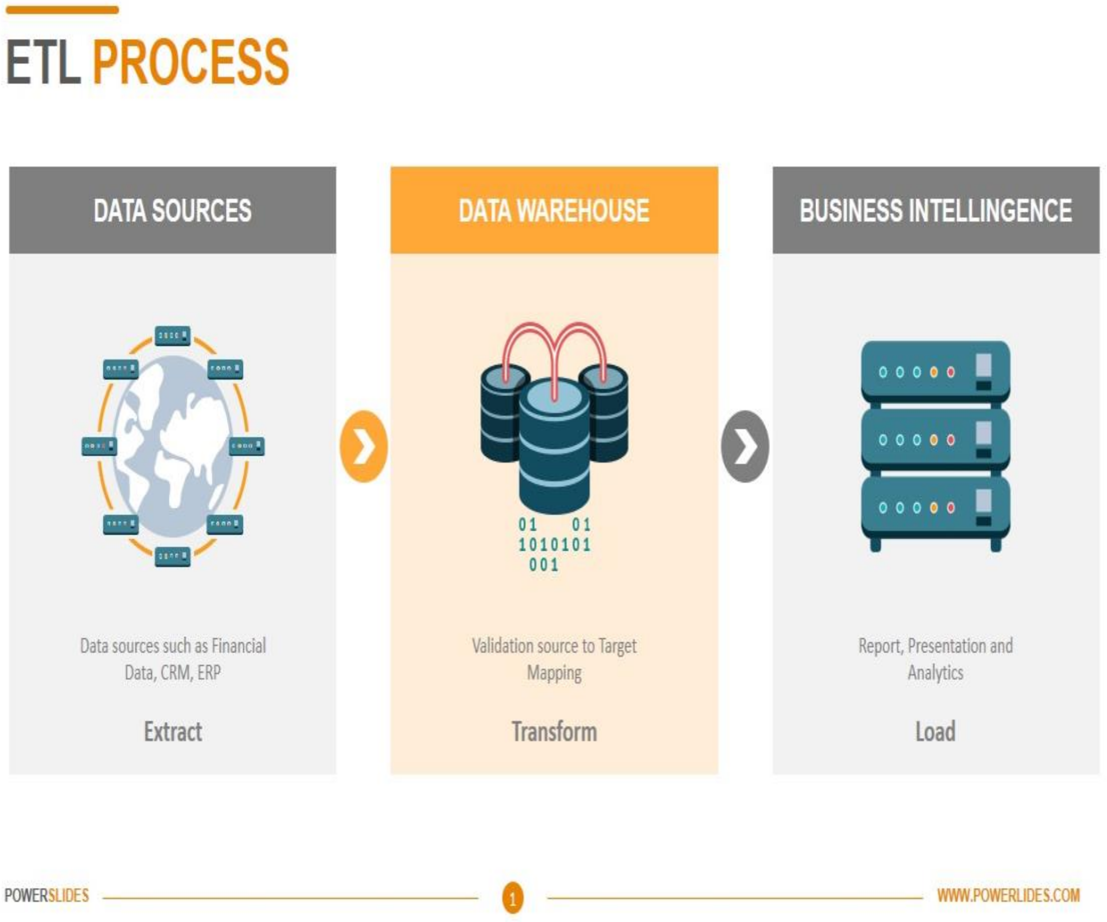


ETL Tools :

Most commonly used ETL tools are Hevo, Sybase, Oracle Warehouse builder, CloverETL, and MarkLogic.

Data Warehouse :

Most commonly used Data Warehouses are Snowflake, Redshift, BigQuery, and Firebolt.



ETL PROCESS

DATA SOURCES

Data sources such as Financial Data, CRM, ERP

Extract

DATA WAREHOUSE

Validation source to Target Mapping

Transform

BUSINESS INTELLIGENCE

Report, Presentation and Analytics

Load

POWERSLIDES 1 WWW.POWERLIDES.COM

Advantages of ETL process in data warehousing:

1. Improved data quality : ETL process ensures that the data in the data warehouse is accurate, complete, and up-to-date.

2. Better data integration : ETL process helps to integrate data from multiple sources and systems, making it more accessible and usable.

3. Increased data security : ETL process can help to improve data security by controlling access to the data warehouse and ensuring that only authorized users can access the data.

4. Improved scalability : ETL process can help to improve scalability by providing a way to manage and analyze large amounts of data.

5. Increased automation : ETL tools and technologies can automate and simplify the ETL process, reducing the time and effort required to load and update data in the warehouse.

Disadvantages of ETL process in data warehousing:

1. High cost : ETL process can be expensive to implement and maintain, especially for organizations with limited resources.

2. Complexity : ETL process can be complex and difficult to implement, especially for organizations that lack the necessary expertise or resources.

3. Limited flexibility : ETL process can be limited in terms of flexibility, as it may not be able to handle unstructured data or real-time data streams.

4. Limited scalability : ETL process can be limited in terms of scalability, as it may not be able to handle very large amounts of data.

5. Data privacy concerns : ETL process can raise concerns about data privacy, as large amounts of data are collected, stored, and analyzed.

```
SELECT column1, column2
FROM source_table
WHERE condition;
```

Transform: Apply transformations to the extracted data.

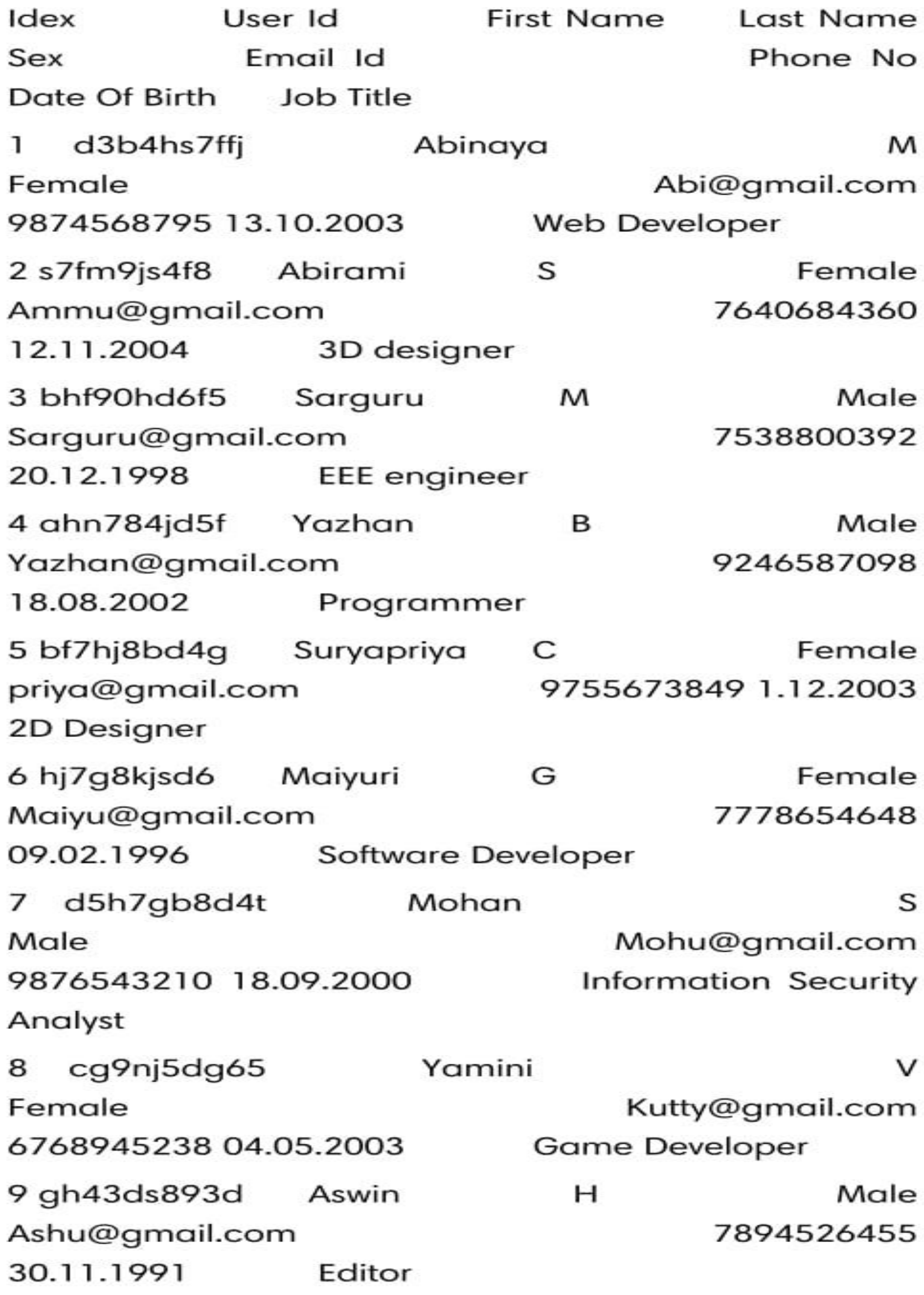
Extract: Retrieve data from the source.

```
UPDATE target_table
SET transformed_column = some_transformation_function(original_column);
```

Load: Load the transformed data into the target database.

```
INSERT INTO target_table (column1, column2)
VALUES (value1, value2);
```

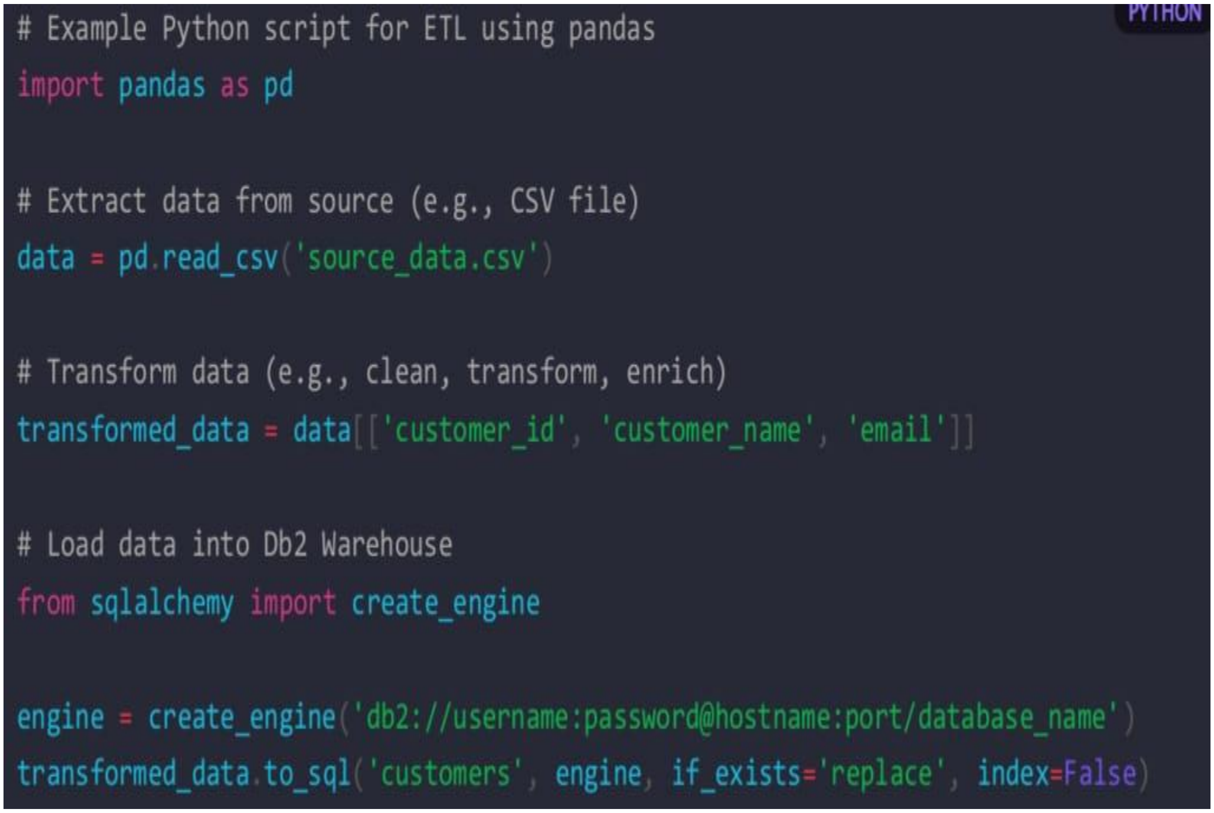
Before Dataset



Idex	User Id	First Name	Last Name	Sex	Email Id	Phone No	Date Of Birth	Job Title
1	d3b4hs7ffj	Abinaya	M	Female	Abi@gmail.com	9874568795	13.10.2003	Web Developer
2	s7fm9js4f8	Abirami	S	Female	Ammu@gmail.com	7640684360	12.11.2004	3D designer
3	bhf90hd6f5	Sarguru	M	Male	Sarguru@gmail.com	7538800392	20.12.1998	EEE engineer
4	ahn784jd5f	Yazhan	B	Male	Yazhan@gmail.com	9246587098	18.08.2002	Programmer
5	bf7hj8bd4g	Suryapriya	C	Female	priya@gmail.com	9755673849	1.12.2003	2D Designer
6	hj7g8kjsd6	Maiyuri	G	Female	Maiyu@gmail.com	7778654648	09.02.1996	Software Developer
7	d5h7gb8d4t	Mohan	S	Male	Mohu@gmail.com	9876543210	18.09.2000	Information Security Analyst
8	cg9nj5dg65	Yamini	V	Female	Kutty@gmail.com	6768945238	04.05.2003	Game Developer
9	gh43ds893d	Aswin	H	Male	Ashu@gmail.com	7894526455	30.11.1991	Editor

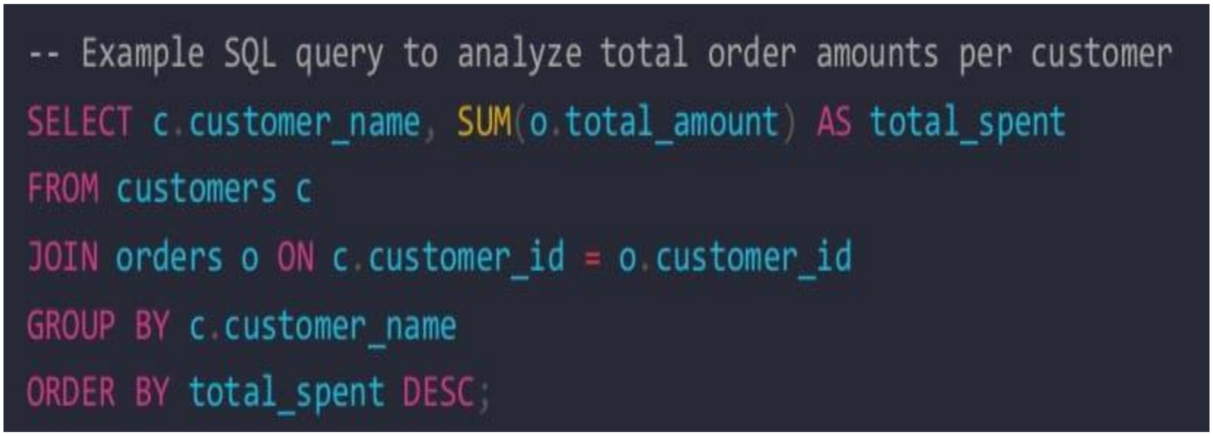
10	9jkh43risd	Midhuna	J	Female	Midhu@gmail.com	9806754356	13.05.2001	Software Developer
11	rjidkh30k7	yashika	T	Female	Yashi@gmail.com	8765432190	05.06.2000	Game Developer
12	abj9kcd74w	Rajesh	V	Male	Raj@gmail.com	7654321098	19.04.1997	Software Developer
13	cvgh45azm2	Ananya	K	Female	Ananya@gmail.com	6789012345	20.01.2003	Application Developer
14	qwpo09nmzx	Kavin	G	Male	Kavin@gmail.com	8901234567	01.01.2001	Game Designer
15	xzh45qwe80	Vijay	C	Male	Vijay@gmail.com	6786543210	10.12.2002	System Architecture
16	jk86mz34gf	pradeep	S	Male	Pradeep@gmail.com	6754839201	02.03.1996	Data Scientist
17	qw649bx93r	Dhanush	O	Male	Kavin@gmail.com	8899765876	21.07.1999	Ux Designer
18	bvg13sk25o	Kamal	A	Male	Kamal@gmail.com	9889787657	15.04.2001	Mobile App Developer
19	zm87vxeew5	Snega	W	Female	Snega@gmail.com	9788934936	18.09.2002	IT Project Manager

ETL process



```python
# Example Python script for ETL using pandas
import pandas as pd

# Extract data from source (e.g., CSV file)
data = pd.read_csv('source_data.csv')

# Transform data (e.g., clean, transform, enrich)
transformed_data = data[['customer_id', 'customer_name', 'email']]

# Load data into Db2 Warehouse
from sqlalchemy import create_engine

engine = create_engine('db2://username:password@hostname:port/database_name')
transformed_data.to_sql('customers', engine, if_exists='replace', index=False)
```

Data Exploration

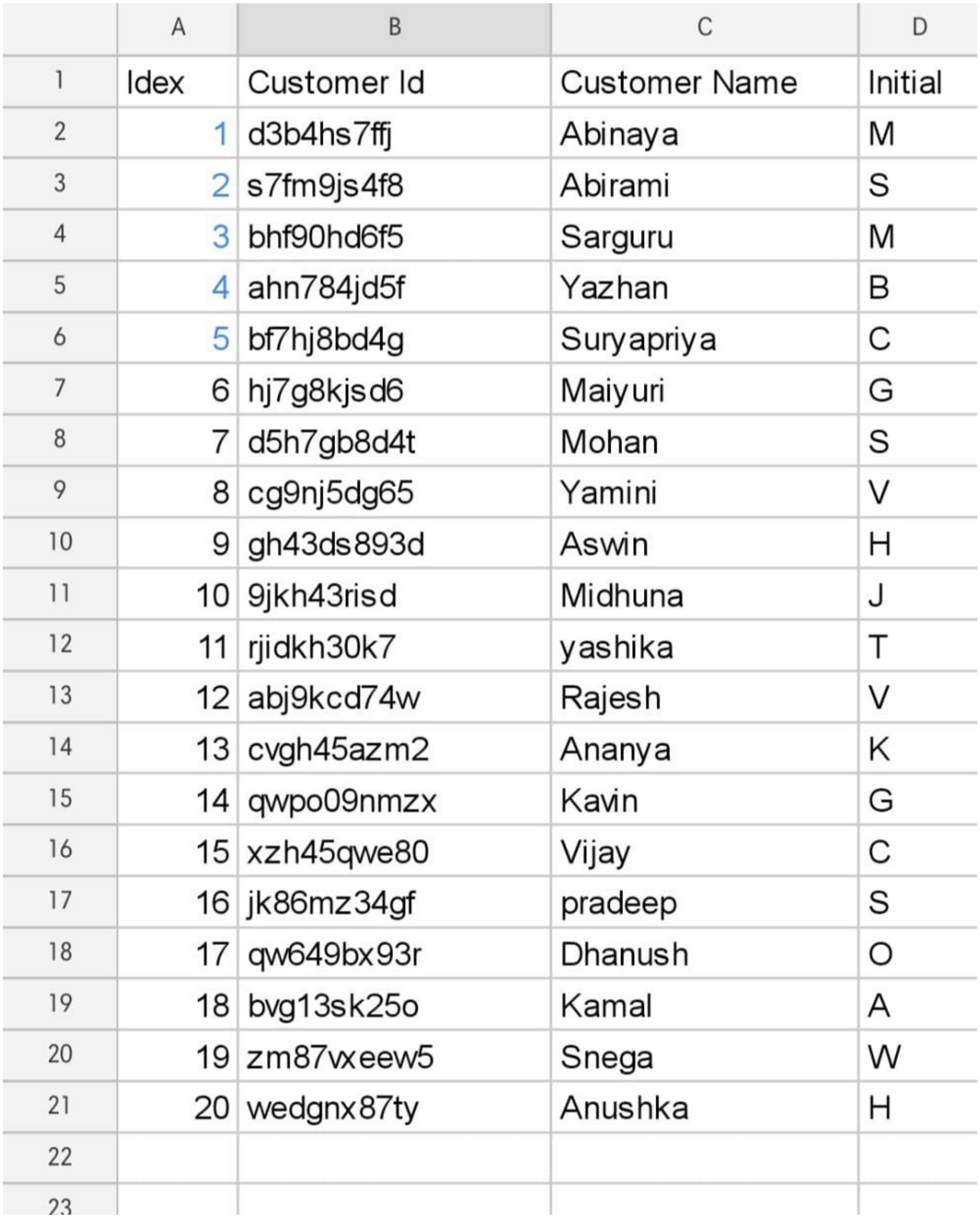


```sql
-- Example SQL query to analyze total order amounts per customer
SELECT c.customer_name, SUM(o.total_amount) AS total_spent
FROM customers c
JOIN orders o ON c.customer_id = o.customer_id
GROUP BY c.customer_name
ORDER BY total_spent DESC;
```

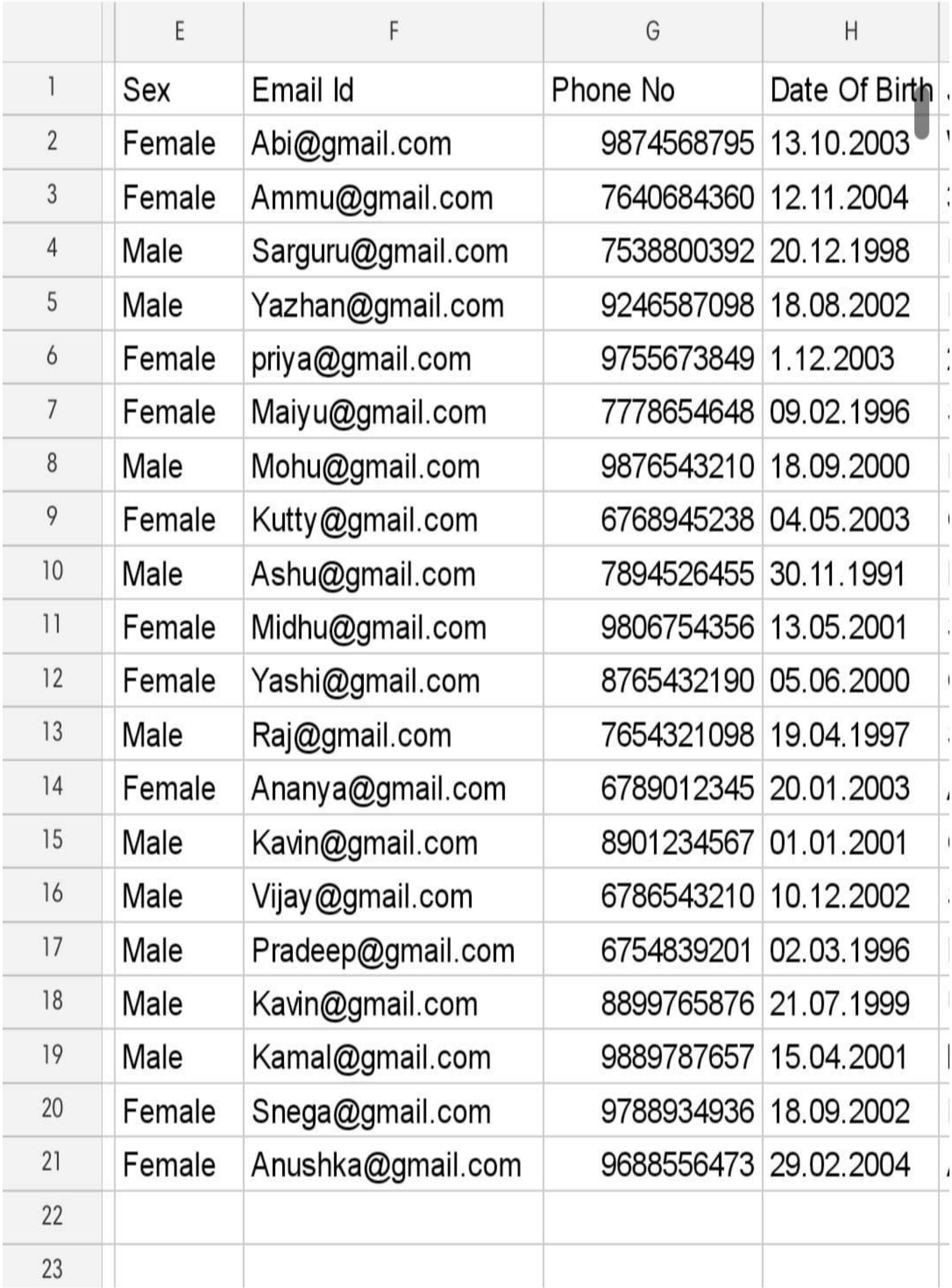
SQL Queries :

```
    SELECT Customer id , Customer Name
FROM Customer ;
   INSERT INTO Customers (CustomerName,
ContactName, Initial , Phone no, Date of
Birth, Email id)
   UPDATE Customers
SET CustomerName = ‘Afra’, Sex= ‘Female’
WHERE CustomerID = 19;
   DELETE FROM Customers WHERE
CustomerName=’Midhuna’;
   CREATE TABLE Customers(
    Customer Id Int,
    Customer Name Varchar(20),
Initial Varchar(10),
Sex Varchar (20),
Email id Varchar(20),
Phone no Number,
Date of Birth Number,
);
```

After writing the SQL Queries the final output is



Idex	Customer Id	Customer Name	Initial
1	d3b4hs7ffj	Abinaya	M
2	s7fm9js4f8	Abirami	S
3	bhf90hd6f5	Sarguru	M
4	ahn784jd5f	Yazhan	B
5	bf7hj8bd4g	Suryapriya	C
6	hj7g8kjsd6	Maiyuri	G
7	d5h7gb8d4t	Mohan	S
8	cg9nj5dg65	Yamini	V
9	gh43ds893d	Aswin	H
10	9jkh43risd	Midhuna	J
11	rjidkh30k7	yashika	T
12	abj9kcd74w	Rajesh	V
13	cvgh45azm2	Ananya	K
14	qwpo09nmzx	Kavin	G
15	xzh45qwe80	Vijay	C
16	jk86mz34gf	pradeep	S
17	qw649bx93r	Dhanush	O
18	bvg13sk25o	Kamal	A
19	zm87vxeew5	Snega	W
20	wedgnx87ty	Anushka	H



	E	F	G	H
1	Sex	Email Id	Phone No	Date Of Birth
2	Female	Abi@gmail.com	9874568795	13.10.2003
3	Female	Ammu@gmail.com	7640684360	12.11.2004
4	Male	Sarguru@gmail.com	7538800392	20.12.1998
5	Male	Yazhan@gmail.com	9246587098	18.08.2002
6	Female	priya@gmail.com	9755673849	1.12.2003
7	Female	Maiyu@gmail.com	7778654648	09.02.1996
8	Male	Mohu@gmail.com	9876543210	18.09.2000
9	Female	Kutty@gmail.com	6768945238	04.05.2003
10	Male	Ashu@gmail.com	7894526455	30.11.1991
11	Female	Midhu@gmail.com	9806754356	13.05.2001
12	Female	Yashi@gmail.com	8765432190	05.06.2000
13	Male	Raj@gmail.com	7654321098	19.04.1997
14	Female	Ananya@gmail.com	6789012345	20.01.2003
15	Male	Kavin@gmail.com	8901234567	01.01.2001
16	Male	Vijay@gmail.com	6786543210	10.12.2002
17	Male	Pradeep@gmail.com	6754839201	02.03.1996
18	Male	Kavin@gmail.com	8899765876	21.07.1999
19	Male	Kamal@gmail.com	9889787657	15.04.2001
20	Female	Snega@gmail.com	9788934936	18.09.2002
21	Female	Anushka@gmail.com	9688556473	29.02.2004
22				
23				