

# Step – 1: Normalization – Performed the ETL process

Extract

The screenshot shows a Microsoft Excel interface with a 'Get & Transform Data' Power Query dialog box open over a worksheet. The dialog box is titled 'h1b\_datset\_BMBAN\_Fall\_2024.csv'. It displays a preview of data with columns: Column1, Column2, Column3, Column4, Column5, Column6, and Column7. The data consists of 24 rows of H-1B visa information. The columns include fields like CASE\_NUMBER, CASE\_STATUS, RECEIVED\_DATE, DECISION\_DATE, ORIGINAL\_CERT\_DATE, VISA\_CLASS, and JOB\_TITLE. The Power Query ribbon tabs at the top are 'File', 'Home', 'Insert', 'Page Layout', 'Get & Transform Data', and 'Power Query'. The bottom of the dialog box has buttons for 'Load', 'Transform Data', and 'Cancel'. The main Excel window shows a blank sheet labeled 'Sheet1'.

Column1	Column2	Column3	Column4	Column5	Column6	Column7
I-200-21270-606997	Certified	9/26/2021	10/1/2021		H-1B	APPLICATIONS SUPPORT ANALYST/ADMINISTRATOR
I-200-21270-606867	Certified	9/26/2021	10/1/2021		H-1B	Designer
I-200-21270-606846	Certified	9/26/2021	10/1/2021		H-1B	Data Analyst
I-200-21270-606842	Certified	9/26/2021	10/1/2021		H-1B	Pharmaceutical Chemist
I-200-21270-606941	Certified	9/26/2021	10/1/2021		H-1B	Senior Systems Analyst JC60
I-200-21270-606854	Certified	9/26/2021	10/1/2021		H-1B	Regional Sales Manager
I-200-21270-606963	Certified	9/26/2021	10/1/2021		H-1B	Software Engineer - CAS-77363-R8X2M5
I-200-21270-607013	Certified	9/26/2021	10/1/2021		H-1B	Quality Assurance Analyst
I-200-21270-607015	Certified	9/26/2021	10/1/2021		H-1B	Technical Architect
I-200-21270-606898	Certified	9/26/2021	10/1/2021		H-1B	Senior Manager JC45
I-200-21270-606847	Certified	9/26/2021	10/1/2021		H-1B	Data Scientist
I-200-21270-606881	Certified	9/26/2021	10/1/2021		H-1B	Scientist 1, Analytical
I-200-21270-606834	Certified	9/26/2021	10/1/2021		H-1B	COMPUTER SYSTEMS ANALYST
I-203-21270-606930	Certified	9/26/2021	10/1/2021	E-3 Australian	SOFTWARE DEVELOPER	
I-200-21270-606832	Certified	9/26/2021	10/1/2021	H-1B	Pharmaceutical Chemist	
I-200-21270-606883	Certified	9/26/2021	10/1/2021	H-1B	Manager JC50	
I-200-21270-606908	Certified	9/26/2021	10/1/2021	H-1B	Computer Software Engineer, Applications	
I-203-21270-606902	Certified	9/26/2021	10/1/2021	E-3 Australian	Sr. Software Engineer	
I-200-21270-606879	Certified	9/26/2021	10/1/2021	H-1B	PHYSICAL THERAPIST	
I-200-21270-606934	Certified	9/26/2021	10/1/2021	H-1B	Sr. Software Engineer	

# Step – 1: Normalization – Performed the ETL process

The screenshot shows the Microsoft Power Query Editor interface. The main area displays a table with two columns: 'state\_id' and 'state\_name'. The 'state\_id' column contains integers from 1 to 25, and the 'state\_name' column lists US states. The 'Transform' ribbon tab is selected. The 'APPLIED STEPS' pane on the right shows the steps taken: Source, Change Type, Promoted Headers, Removed Columns, Trimmed Text, Removed Duplicates, Renamed Columns, Filtered Rows, Added Index, and Reordered Columns. A blue arrow points from the 'Transform' tab to the 'APPLIED STEPS' pane.

state_id	state_name
1	GA
2	MI
3	MA
4	MN
5	TN
6	CA
7	TX
8	VA
9	CT
10	MD
11	MO
12	OK
13	NY
14	WA
15	IL
16	KY
17	NJ
18	FL
19	PA
20	NC
21	AZ
22	UT
23	ND
24	AL
25	NH

*\*\* refer the last slide to understand the table distribution*

# Step – 1: Normalization – Performed the ETL process

Load



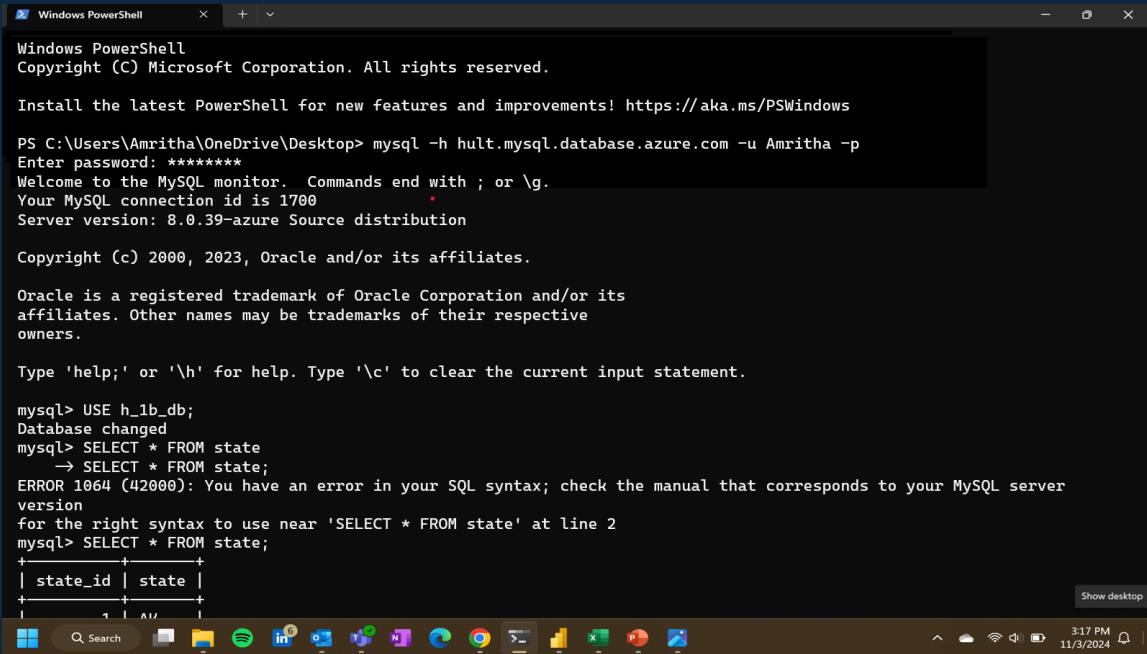
Screenshot of Microsoft Excel showing a table of state data. The table has two columns: 'state\_id' and 'state\_name'. The data consists of 24 rows, each containing a unique state ID and its corresponding name. The table is selected, and the 'Table Design' tab is active in the ribbon.

A	B
state_id	state_name
1	GA
2	MI
3	MA
4	MN
5	TN
6	CA
7	TX
8	VA
9	CT
10	MD
11	MO
12	OK
13	NY
14	WA
15	IL
16	KY
17	NJ
18	FL
19	PA
20	NC
21	AZ
22	UT
23	ND

The ribbon shows the following tabs: File, Home, Insert, Page Layout, Formulas, Data, Review, View, Automate, Help, Table Design. The 'Table Design' tab is currently selected. The status bar at the bottom right shows '4:13 PM' and '100%'. The taskbar at the bottom includes icons for File, Open, Save, New, Find, Copy, Paste, Cut, Delete, Undo, Redo, and others.

*\*\* refer the last slide to understand the table distribution*

## Step – 2: Data Uploading – Using Command Prompt



```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\Users\Amritha\OneDrive\Desktop> mysql -h hult.mysql.database.azure.com -u Amritha -p
Enter password: *****
Welcome to the MySQL monitor. Commands end with ; or \g.
Your MySQL connection id is 1700
Server version: 8.0.39-azure Source distribution

Copyright (c) 2000, 2023, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

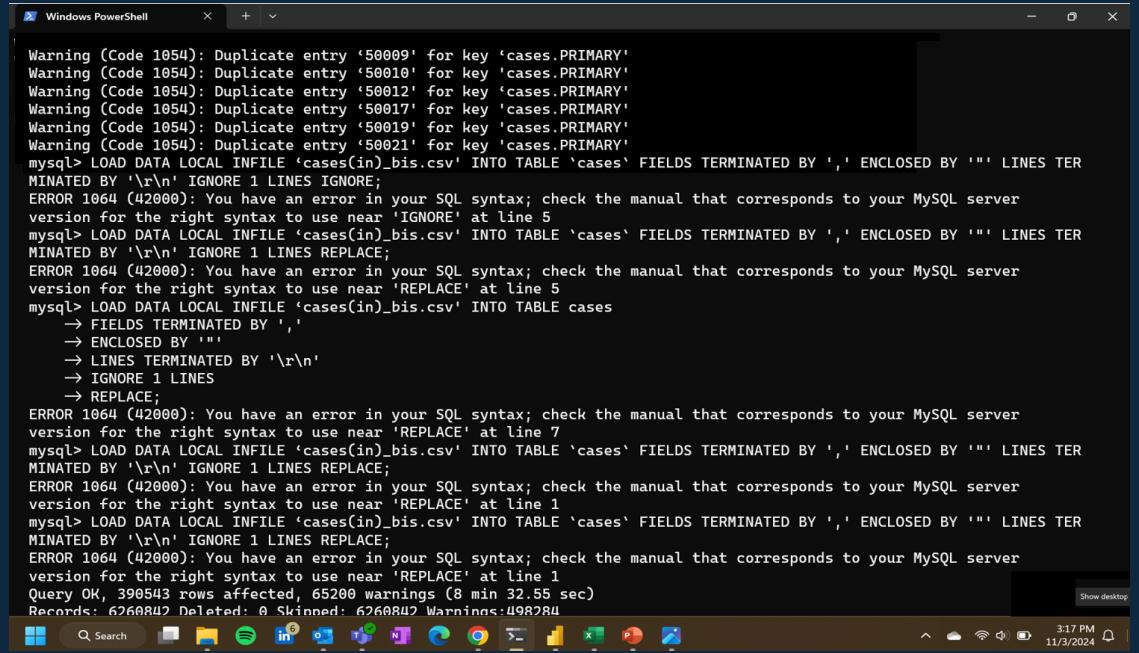
Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> USE h_lb_db;
Database changed
mysql> SELECT * FROM state
    → SELECT * FROM state;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server
version
for the right syntax to use near 'SELECT * FROM state' at line 2
mysql> SELECT * FROM state;
+-----+-----+
| state_id | state |
+-----+-----+
|       1 | AU   |
+-----+-----+
1 row in set (0.00 sec)

Show desktop
3:17 PM 11/3/2024
```

ii. Uploading the data into command prompt

i. Converted the loaded table in .csv file type



```
Warning (Code 1054): Duplicate entry '50009' for key 'cases.PRIMARY'
Warning (Code 1054): Duplicate entry '50010' for key 'cases.PRIMARY'
Warning (Code 1054): Duplicate entry '50012' for key 'cases.PRIMARY'
Warning (Code 1054): Duplicate entry '50017' for key 'cases.PRIMARY'
Warning (Code 1054): Duplicate entry '50019' for key 'cases.PRIMARY'
Warning (Code 1054): Duplicate entry '50021' for key 'cases.PRIMARY'
mysql> LOAD DATA LOCAL INFILE 'cases(in)_bis.csv' INTO TABLE 'cases' FIELDS TERMINATED BY ',' ENCLOSED BY '\"' LINES TER
MINATED BY '\r\n' IGNORE 1 LINES IGNORE;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server
version for the right syntax to use near 'IGNORE' at line 5
mysql> LOAD DATA LOCAL INFILE 'cases(in)_bis.csv' INTO TABLE 'cases' FIELDS TERMINATED BY ',' ENCLOSED BY '\"' LINES TER
MINATED BY '\r\n' IGNORE 1 LINES REPLACE;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server
version for the right syntax to use near 'REPLACE' at line 5
mysql> LOAD DATA LOCAL INFILE 'cases(in)_bis.csv' INTO TABLE cases
    → FIELDS TERMINATED BY ','
    → ENCLOSED BY '\"'
    → LINES TERMINATED BY '\r\n'
    → IGNORE 1 LINES
    → REPLACE;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server
version for the right syntax to use near 'REPLACE' at line 7
mysql> LOAD DATA LOCAL INFILE 'cases(in)_bis.csv' INTO TABLE 'cases' FIELDS TERMINATED BY ',' ENCLOSED BY '\"' LINES TER
MINATED BY '\r\n' IGNORE 1 LINES REPLACE;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server
version for the right syntax to use near 'REPLACE' at line 1
mysql> LOAD DATA LOCAL INFILE 'cases(in)_bis.csv' INTO TABLE 'cases' FIELDS TERMINATED BY ',' ENCLOSED BY '\"' LINES TER
MINATED BY '\r\n' IGNORE 1 LINES REPLACE;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server
version for the right syntax to use near 'REPLACE' at line 1
Query OK, 390543 rows affected, 65200 warnings (8 min 32.55 sec)
Records: 6260842 Deleted: 0 Skipped: 6260842 Warnings: 498284
Show desktop
3:17 PM 11/3/2024
```

# Step – 3: Query Execution – Using MySQL

The screenshot shows the MySQL Workbench interface. On the left, the 'Schemas' tree view displays the 'h\_1b\_db' schema, with the 'wage\_pay' table selected and highlighted in blue. The main area contains a SQL query editor with the following code:

```
1 SELECT
2     e.EMPLOYER_NAME,
3     c.NAICS_CODE,
4     SUM(c.TOTAL_WORKER_POSITIONS) AS TOTAL_POSITIONS
5 FROM
6     employer e
7     JOIN
8         cases c ON e.EMPLOYER_ID = c.EMPLOYER_ID
9 GROUP BY
10    e.EMPLOYER_NAME, c.NAICS_CODE
11 ORDER BY
12    TOTAL_POSITIONS DESC
13 LIMIT 10;
14
```

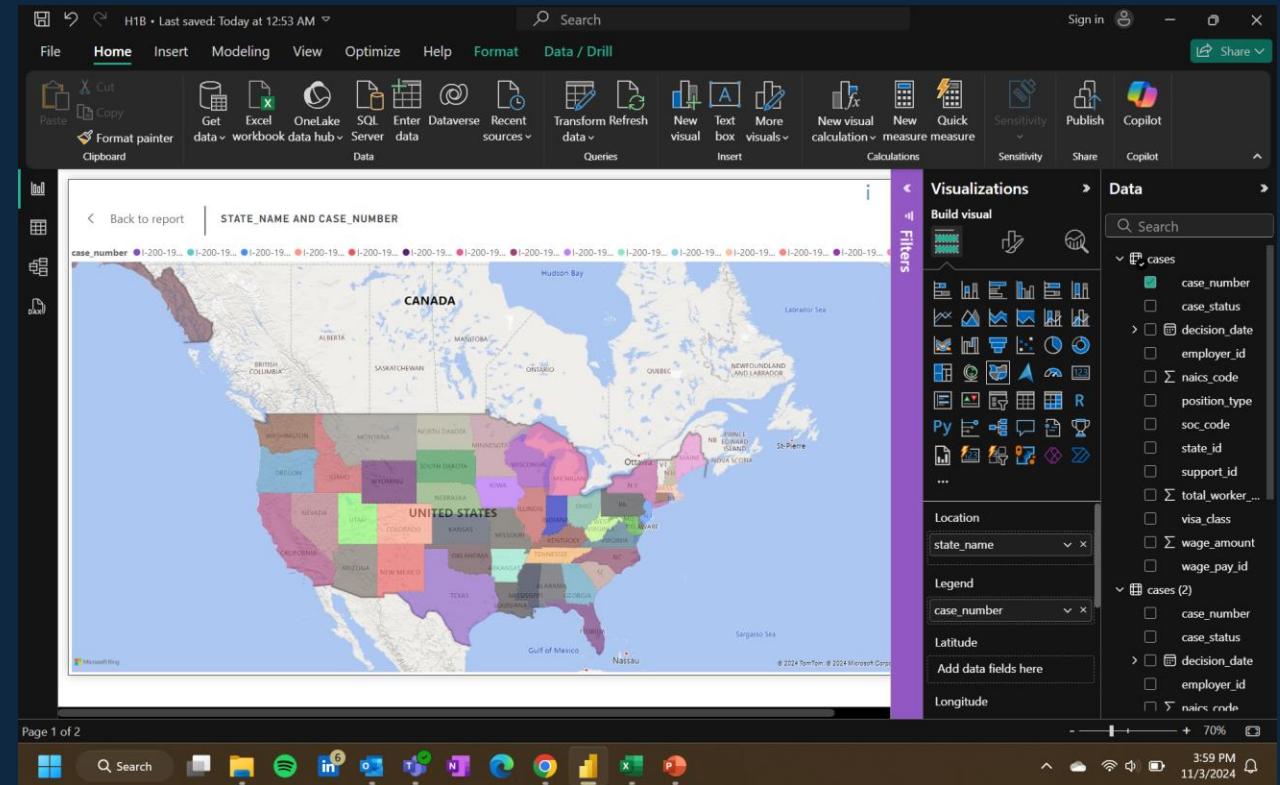
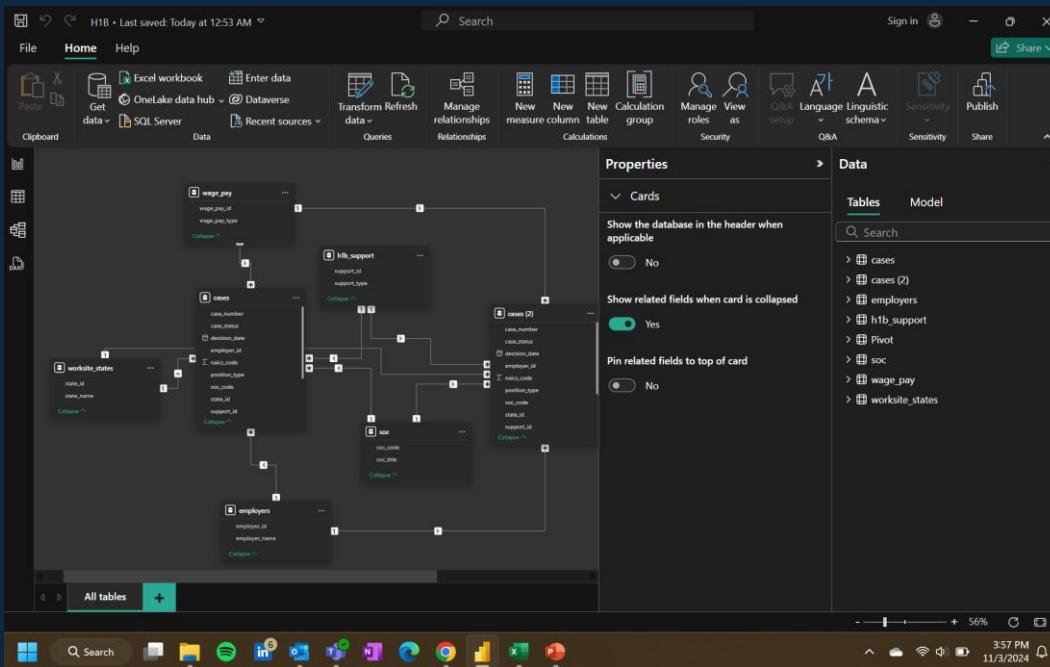
Below the query editor is a 'Result Grid' showing the results of the executed query. The columns are 'EMPLOYER\_NAME', 'NAICS\_CODE', and 'TOTAL\_POSITIONS'. The data includes:

EMPLOYER_NAME	NAICS_CODE	TOTAL_POSITIONS
amazon.com services llc	45411	139407
infosys limited	541511	119783
qualcomm technologies, inc.	334220	60871
nvidia corporation	334111	41268
grandison management, inc.	561320	39980
amazon web services, inc.	518210	33974
cisco systems, inc.	334111	30758
servicenow, inc.	541511	24306
cognizant technology solutions us corp	541512	22594
tata consultancy services limited	541511	20381

The bottom status bar indicates 'Query Completed' and 'Read Only'.

## Step – 4: PowerBI - Visualization

i. Uploaded and merged queries into PowerBI model



ii. Analyzed and created graphs using PowerBI visualization tools

*\*\* refer the last slide to understand the table distribution*

## Notes

**In the dataset - Total number of records = 1,655,836 & Total number of columns = 96**

In 2022,  
Number of records from (Q1 to Q4) = 626,084  
Number of records (Q1 to Q3) = 507,439

In 2024,  
Number of records (Q1 to Q3) = 1,029,752  
Out of which 587,378 records are NULL

- Hence, in order to perform fair analysis, we have considered **YTD (Year-to-Date)** data from both the years.
- Considering only Q1, Q2 and Q3 data from the respective years 2022 and 2024.

**\*\***

- It can be observed in the previous images that the cases table is mentioned as cases 2022\_1, cases 2022\_2, cases 2023\_1 and cases 2023\_2.
- This is because of the huge data records; we have divided the table into two equal halves and worked on it to avoid crashing the laptop.