

# Step – 1: Normalization – Performed the ETL process

Extract



h1b\_datset\_BMBAN\_Fall\_2024.csv

File Origin: 1200: Unicode | Delimiter: Tab | Data Type Detection: Do not detect data types

Column1	Column2	Column3	Column4	Column5	Column6	Column7
CASE_NUMBER	CASE_STATUS	RECEIVED_DATE	DECISION_DATE	ORIGINAL_CERT_DATE	VISA_CLASS	JOB_TITLE
I-200-21270-606997	Certified	9/26/2021	10/1/2021		H-1B	APPLICATIONS SUPPORT ANALYST/ADMINISTRAT
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

Buttons: Load, Transform Data, Cancel

# Step – 1: Normalization – Performed the ETL process

worksite\_states (2) - Power Query Editor

File

Home

Transform

Add Column

View

Close & Load

Refresh Preview

Properties

Advanced Editor

Manage

Choose Columns

Remove Columns

Keep Rows

Remove Rows

Sort

Split Column

Group By

1 2 Replace Values

Merge Queries

Append Queries

Combine Files

Manage Parameters

Data source settings

New Source

Recent Sources

Enter Data

Queries [7]

cases\_2022\_1

cases\_2022\_2

cases\_2023\_1

cases\_2023\_2

employer

worksite\_states (2)

wage\_pay (2)

fx

= Table.RenameColumns(#"Reordered Columns",{{"Index", "state\_id"}})

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

Query Settings

PROPERTIES

Name

worksite\_states (2)

All Properties

APPLIED STEPS

Source

Change Type

Promoted Headers

Removed Columns

Trimmed Text

Removed Duplicates

Renamed Columns

Filtered Rows

Added Index

Reordered Columns

Renamed Columns1

2 COLUMNS, 56 ROWS

Column profiling based on top 1000 rows

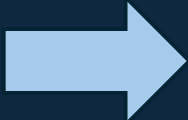
PREVIEW DOWNLOADED AT 4:14 PM

← Transform

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

# Step – 1: Normalization – Performed the ETL process

Load



Excel interface showing a table with columns A and B. The table contains data for 24 rows, with columns labeled 'state\_id' and 'state\_name'.

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

*\*\* 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_1b_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 | AL |
```

ii. Uploading the data into  
command prompt

i. Converted the loaded tabled  
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: 65200
```

# Step – 3: Query Execution – Using MySQL

Administration

Schemas

SCHEMAS

Filter objects

h\_1b\_db

Tables

cases

employer

states

wage\_pay

Views

Stored Procedures

Functions

sys

Object Info

Session

Table: wage\_pay

Columns:

WAGE\_PAY\_ID

int PK

WAGE\_PAY\_TYP

varchar(255)

E

Query 3

SQL File 3\*

SQL File 7\*

SQL File 8\*

Limit to 500 rows

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

100%

25:12

Result Grid

Filter Rows:

Search

Export:

Fetch rows:

EMPLOYER_NAME	NAICS_CODE	TOTAL_POSITIONS
amazon.com services llc	45411	130407
infosys limited	541511	119783
qualcomm technologies, inc.	334220	60871
nvidia corporation	334111	41268
grandison management, inc.	561320	39080
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

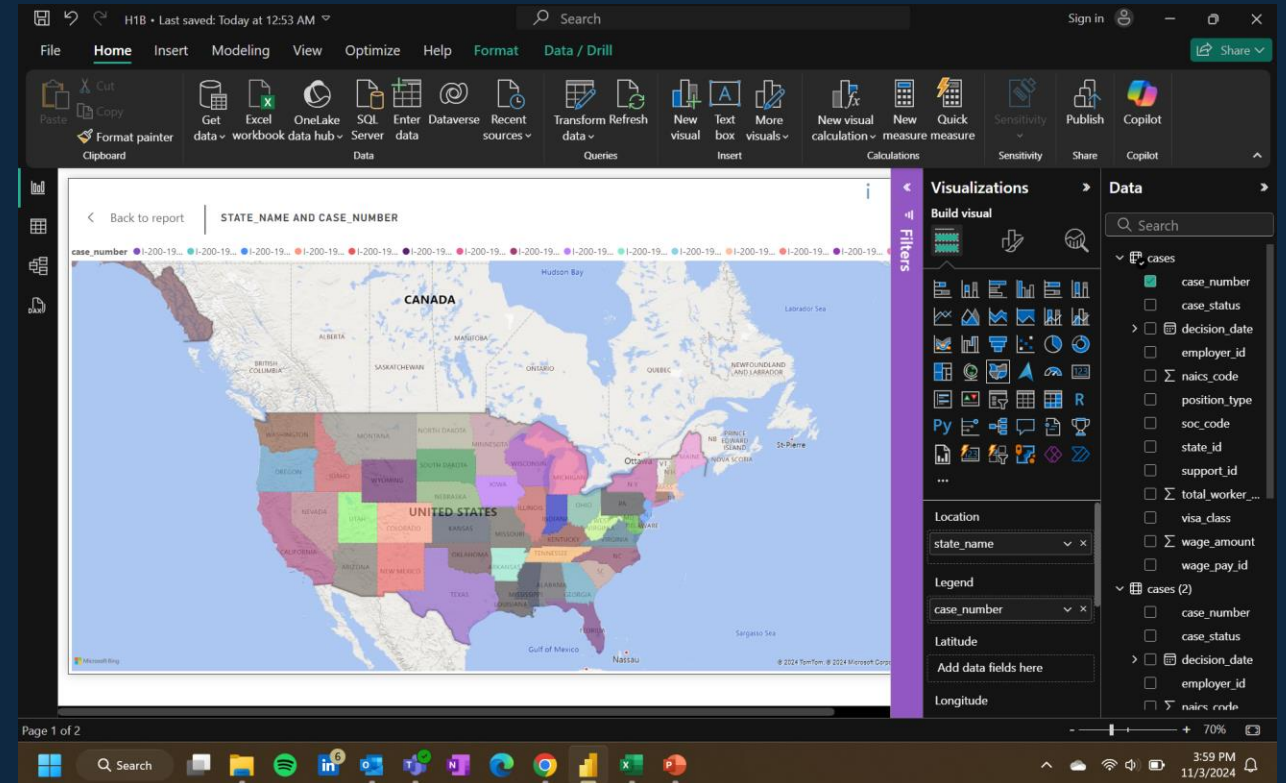
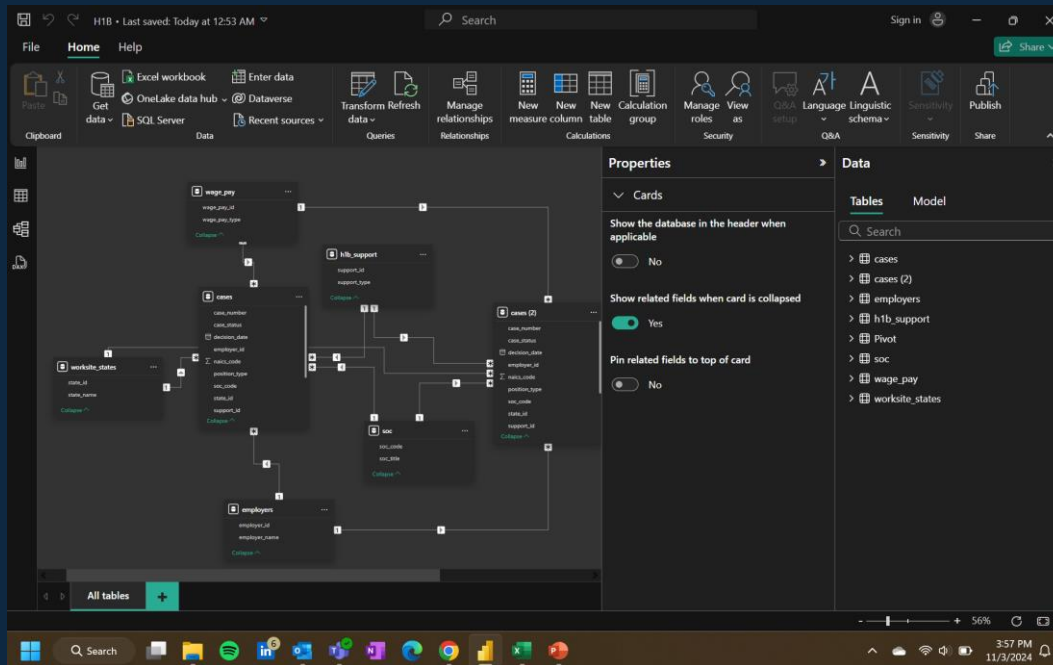
Result 7

Read Only

Query Completed

## 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.