EMPLOYEE HIRING AND HISTORY

A DATA ANALYTICS PROJECT

PROJECT PRESENTATION

Problem Statement

Potential merger plans are in progress and the company president seeks an insightful overview of:

Employee Turnover

Diversity and Inclusivity

Important KPI's

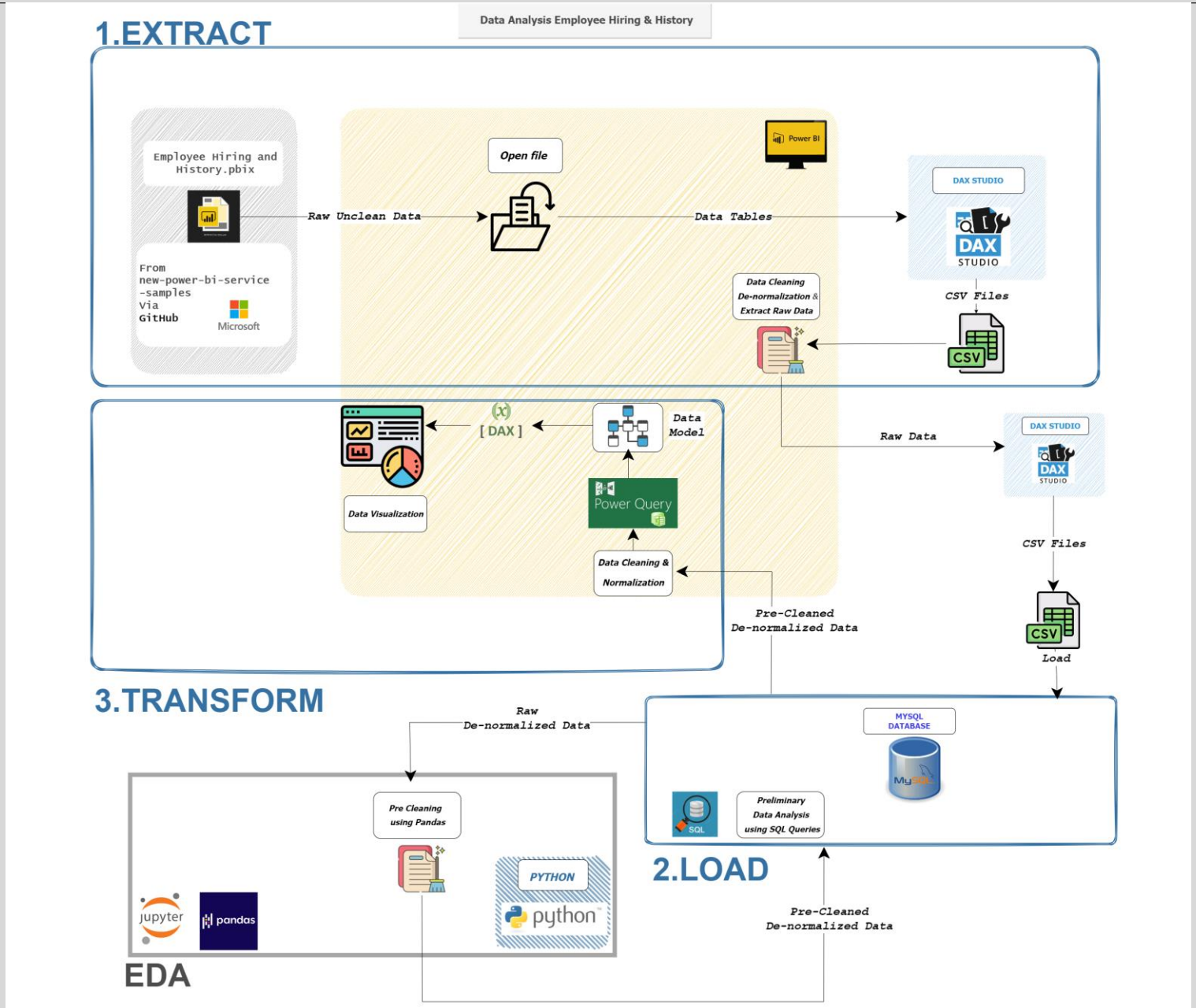
- **Turnover Rate**
- **Retention Rate**
- **Average Tenure**
- **Voluntary vs. Involuntary Separations**
- **Age Diversity Ratio**
- **Diversity Index**
- **Gender Diversity Ratio**
- **Ethnic Diversity Ratio**

DATASET(s)

LEGEND
Raw/De-Normalized Dataset
Normalised Datasets
Common Data Issues
Cleaning Done In Python

Table Name	ColumnNames	Column Description	Data Quality Issues	TRANSFORMATIONS	
				Transformation Steps - Column	Transformation Steps - Table
Employee	~~Multiple Computed Columns~~	Multiple columns computed from the other Employee table columns	Not well computed	Delete & Recompute during Normalization	Extract Raw data & Create Table in DB
	~~Measures~~	Generated columns & not raw	Not structured properly into a measures table	Delete & Recompute during Normalization	
	Age	Employee age			
	BU	Business Unit ID	Column Name not descriptive enough	Change name to BusinessUnitID	
	date	Date of record creation		Keep	
	EmpID	Employee ID		Keep	
	EthnicGroup	Ethnic Group ID	Column Name not descriptive enough	Change name to EthnicGroupID	
	FP	Employment Type ID, full / part time	Column Name not descriptive enough	Change name to EmpTypeID	
	Gender	Gender Type ID	Column Name not descriptive enough	Change name to GenderID	
	HireDate	Date Employee Hired	NA	Keep	
	PayTypeID	Employee Pay Type ID - Hourly or Monthly	NA	Keep	
	TermDate	Termination Date	NA	Keep	
	TermReason	Tremination Reason ID	Column Name not descriptive enough	Change name to TermReasonID	
	~~AgeGroupID~~	Computed	Can be computed later	Change name to TermReasonID	
BU	BU	Business Unit	Column Name not descriptive enough	Change name to BusinessUnitID	Change table name to BusinessUnit & Extract Raw data & Create Table in DB
	~~Count of BU~~	Total Business units in each region	Measure	Delete & Recompute during Normalization	
	RegionSeq	Region id and region name	Multiple data in single column	Split the column. Create a Region table to store RegionID and corresponding Region during normalization	
	~~Region~~	Region name computed from RegionSeq	Computed	Delete & create Region table during Normalization	
	VP	Vice President Name			
~~Date~~	Date Columns	Columns computed using the data in the date column	Computed and not well structured	Delete & Recompute during Normalization	Delete & Recompute Tables During Normalization Step
~~AgeGroup~~	AgeGroupID	Age Group ID	Computed from the Employee.Age column	Change column name to AgeGroupID	
	AgeGroup	Age Group Description	NA	Keep	
~~Ethnicity~~	Ethnicity	Ethnicity ID	Column Name not descriptive enough	Change name to EthnicityID	
	Ethnic Group	Ethnic Group Name	NA	Keep	
~~FP~~	FP	Employment Type ID	Column Name not descriptive enough	Change column name to EmpTypeID	
	FPDesc	Employment Type Description , full / part time		Change name to EmpType	
~~Gender~~	ID	Gender Type ID	Column Name not descriptive enough	Change name to GenderID and Change ID from D, C to M, F	
	Gender	Gender Name	NA	Keep	
	~~Sort~~	Gender Sort Order Number	Not Needed	Delete	
~~PayType~~	PayType	PayType Description Hourly or Monthly	NA	Keep	
	PayTypeID	PayType ID	NA	Keep	
~~SeparationReasons~~	SeparationTypeID	Separation Type ID	The ID's used are not the most appropriate	Change ID from U, V to I, V (Involuntary, Voluntary)	
	SeparationReason	Separation Reason Description	NA	Keep	
	~~Sort~~	Sort Order Number	Not Needed	Delete	
Region	RegionID	Region ID	Merged together in the BU table	Create a new table to store region information	Create New Table during Normalization Step
	Region	Name of the Region			

Project Architecture



PYTHON – EDA & Pre-Cleaning

Unique / Distinct Values

Employee ID

Unique / Distinct Values in Columns

```
# Checking if EmpID's are unique for each record

# Unique EmpID's
empid_unique = len(pd.unique(df_e['EmpID']))

print(f"Unique EMPIDs : {empid_unique}")
print(f"Are the EmpID's in each row unique: {len(set(df_e['EmpID'])) == df_e['EmpID'].count()}")
```

```
Unique EMPIDs : 61843
Are the EmpID's in each row unique: False
```

Termination Dates

```
[12]
# Total termination dates available
len(df_e[df_e['TermDate'].notnull()])
```

```
... 29442
```

```
[13]
# Total termination dates with unique EmpID
len(pd.unique(df_e[df_e['TermDate'].notnull()]['EmpID']))
```

```
... 29442
```

Handling Null Values

BEFORE

TermReason column

- Replace empty rows with null
- Convert id from `U` to `I` to represent Involuntary Termination

```
df_e['TermReason'].value_counts()
```

```
TermReason
     1260817
V      22048
U       7394
Name: count, dtype: int64
```

AFTER

```
# Verifying if empty string is replaced by none/null
df_e['TermReason'].isnull().sum()
```

```
1260817
```

```
# Verifying Replacement
df_e['TermReason'].value_counts()
```

```
TermReason
V    22048
I     7394
Name: count, dtype: int64
```

SQL DATABASE - Loading

Before Pre-Cleaning

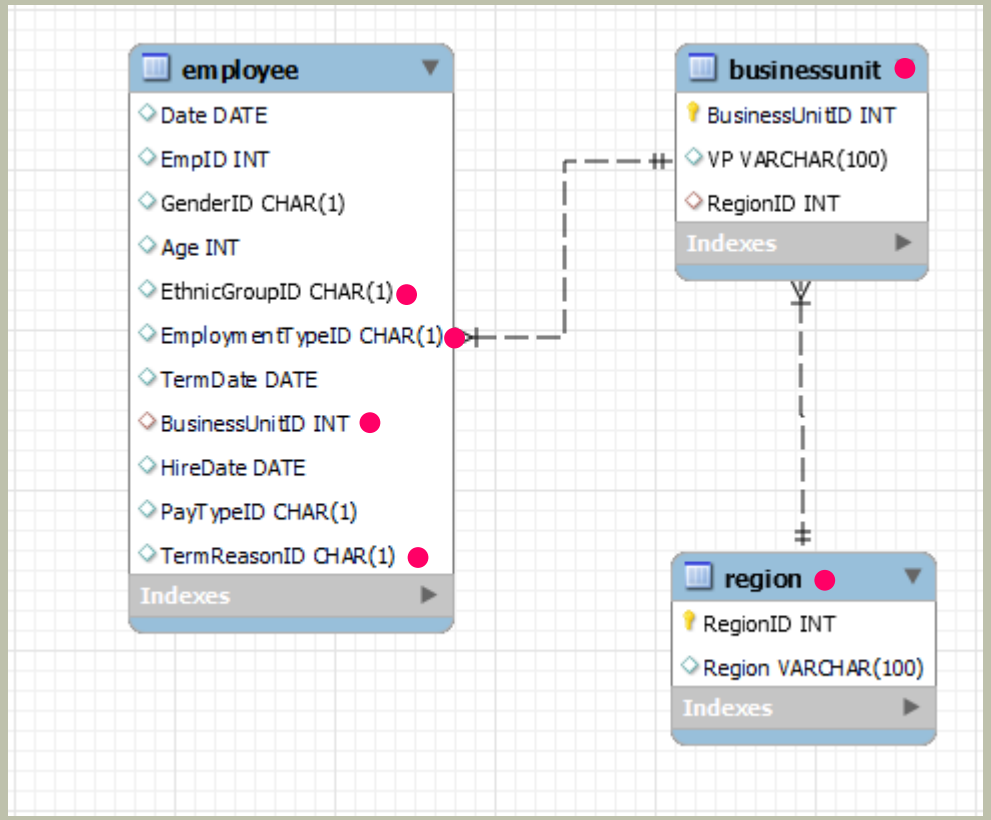
Employee

Field	Type	Null	Key	Default	Extra
date	date	YES		NULL	
EmpID	int	YES		NULL	
Gender	char(1)	YES		NULL	
Age	int	YES		NULL	
EthnicGroup	int	YES		NULL	
FP	char(1)	YES		NULL	
TermDate	date	YES		NULL	
BU	int	YES		NULL	
HireDate	date	YES		NULL	
PayTypeID	char(1)	YES		NULL	
TermReason	char(1)	YES		NULL	

BU

Field	Type	Null	Key	Default	Extra
BU	int	YES		NULL	
RegionSeq	varchar(100)	YES		NULL	
VP	varchar(100)	YES		NULL	

After Pre-Cleaning using Pandas



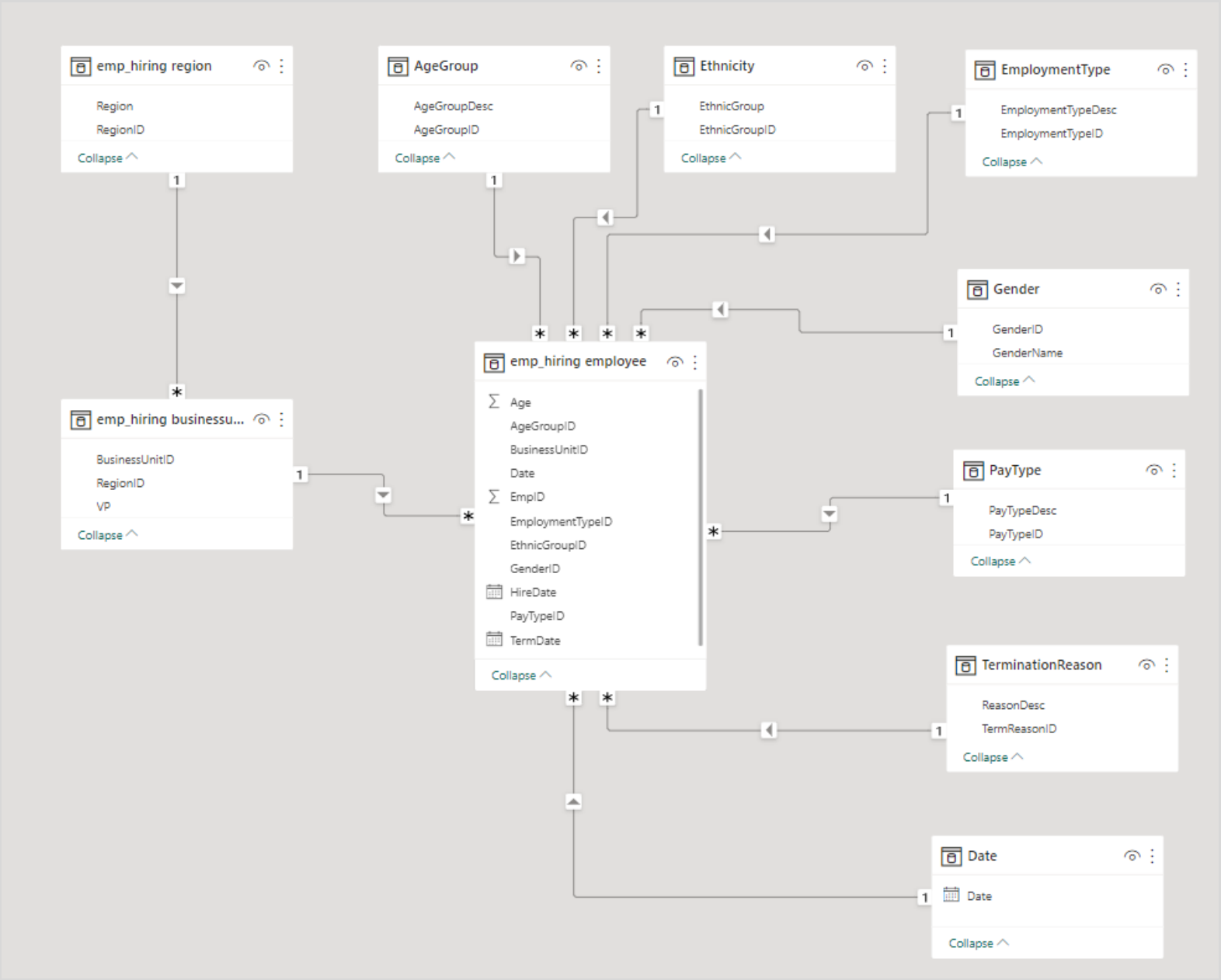
POWER BI Data Model

Dimension Tables Created

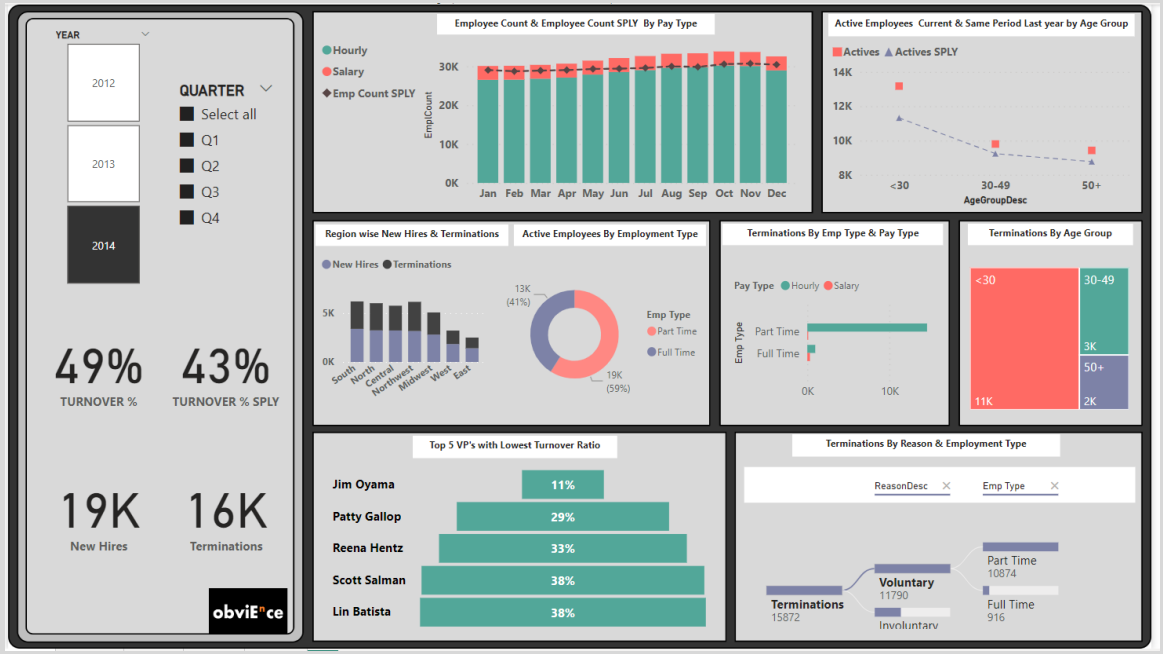
Connecting Primary Keys

Normalized Structure

Snowflake Schema



Visualization



Recommendations

- **Employee Turnover and Retention**
- **Diversity and Inclusivity**
- **Gender Disparities**
- **Seasonal Analysis**
- **Ongoing Monitoring & Communication**
- **Legal and Ethical Considerations**

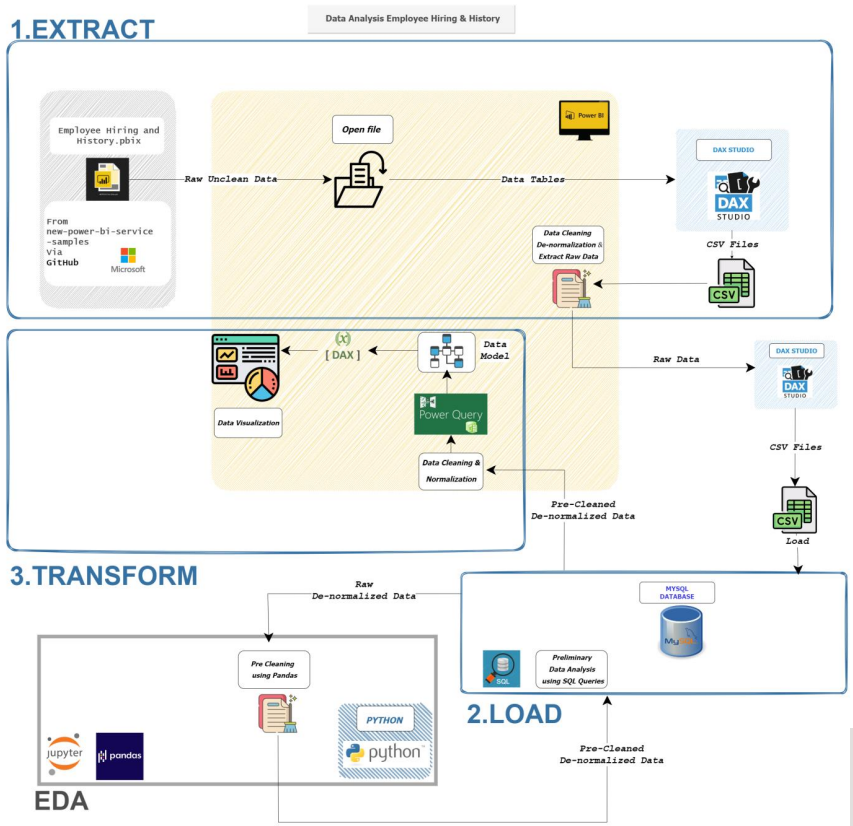


Table Name	ColumnNames	Column Description	Data Quality Issues	TRANSFORMATIONS	
				Transformation Steps - Column	Transformation Steps - Table
Employee	~~Multiple Computed Columns~~	Multiple columns computed from the other Employee table columns	Not well computed	Delete & Recompute during Normalization	Extract Raw data & Create Table in DB
	~~Measures~~	Generated columns & not raw	Not structured properly into a measures table	Delete & Recompute during Normalization	
	Age	Employee age			
	BU	Business Unit ID	Column Name not descriptive enough	Change name to BusinessUnitID	
	date	Date of record creation		Keep	
	EmpID	Employee ID		Keep	
	EthnicGroup	Ethnic Group ID	Column Name not descriptive enough	Change name to EthnicGroupID	
	FP	Employment Type ID, full / part time	Column Name not descriptive enough	Change name to EmpTypeID	
	Gender	Gender Type ID	Column Name not descriptive enough	Change name to GenderID	
	HireDate	Date Employee Hired	NA	Keep	
	PayTypeID	Employee Pay Type ID - Hourly or Monthly	NA	Keep	
	TermDate	Termination Date	NA	Keep	
	TermReason	Tremination Reason ID	Column Name not descriptive enough	Change name to TermReasonID	
	~~AgeGroupID~~	Computed	Can be computed later	Change name to TermReasonID	
BU	BU	Business Unit	Column Name not descriptive enough	Change name to BusinessUnitID	Change table name to BusinessUnit & Extract Raw data & Create Table in DB
	~~Count of BU~~	Total Business units in each region	Measure	Delete & Recompute during Normalization	
	RegionSeq	Region id and region name	Multiple data in single column	Split the column. Create a Region table to store RegionID and coresponding Region during normalization	
	~~Region~~	Region name computed from RegionSeq	Computed	Delete & create Region table during Normalization	
	VP	Vice President Name			
~~Date~~	Date Columns	Columns computed using the data in the date column	Computed and not well structured	Delete & Recompute during Normalization	Delete & Recompute Tables During Normalization Step
~~AgeGroup~~	AgeGroupID	Age Group ID	Computed from the Employee Age column	Change column name to AgeGroupID	
	AgeGroup	Age Group Description	NA	Keep	
~~Ethnicity~~	Ethnicity	Ethnicity ID	Column Name not descriptive enough	Change name to EthnicityID	
	Ethnic Group	Ethnic Group Name	NA	Keep	
~~FP~~	FP	Employment Type ID	Column Name not descriptive enough	Change column name to EmpTypeID	
	FPDesc	Employment Type Description , full / part time		Change name to EmpType	
~~Gender~~	ID	Gender Type ID	Column Name not descriptive enough	Change name to GenderID and Change ID from D, C to M, F	
	Gender	Gender Name	NA	Keep	
~~Sort~~	PayType	PayType Description Hourly or Monthly	Not Needed	Delete	
~~PayType~~	PayTypeID	PayType ID	NA	Keep	
	SeparationTypeID	Separation Type ID	The ID's used are not the most appropriate	Change ID from U, V to I, V (Involuntary, Voluntary)	
~~SeparationReasons~~	SeparationReason	Separation Reason Description	NA	Keep	
	~~Sort~~	Sort Order Number	Not Needed	Delete	
Region	RegionID	Region ID	Merged together in the BU table	Create a new table to store region information	Create New Table during Normalization Step
	Region	Name of the Region