Final Project – Synapse & Power BI

Objective: To build a data engineering pipeline to ingest the existing business employment CSV file format data from an external source, transform it, and prepare it for analysis and find out the total employment over the decade & year-on-year employment.

Project outline:

1. Prerequisites
2. Data Ingestion & Preparation
3. Data Transformation & Cleaning
4. Creating a view
5. Visualizing Results
6. Conclusion & Insights

1. Prerequisites:

- Azure Synapse Workspace: Used an Azure Synapse workspace Severless SQL Pool.
- Azure Data Lake Gen 2 Storage: Used the ADLS Gen 2 account uploaded the data and used it as a source.
- Business Employment Data: Used the previously uploaded dataset called business employment CSV format for the project.

2. Data Ingestion & Preparation:

- The previously uploaded in ADLS account dataset called business_employment in CSV format for the project.

- **Created a new Database:**

```
---- Creating a new database---

USE master
GO

CREATE DATABASE bus_emp_1
GO

ALTER DATABASE bus_emp_1 COLLATE Latin1_General_100_BIN2_UTF8
GO

USE bus_emp_1
GO
```

- **Created schema in the Database as below:**

```
---Create a Schema based on Medallion Architechture----

CREATE SCHEMA bronze
GO

CREATE SCHEMA silver
GO

CREATE SCHEMA gold
GO
```

- **Created External Data Source Pointing towards our ADLS Source Storage:**

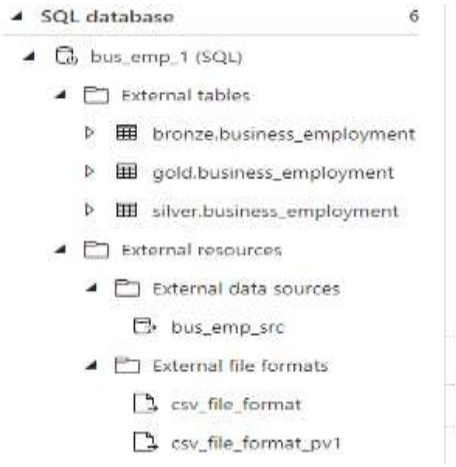
```
---- Create an External DataSource---

USE bus_emp_1;

IF NOT EXISTS (SELECT * FROM sys.external_data_sources WHERE name = 'bus_emp_src')
    CREATE EXTERNAL DATA SOURCE bus_emp_src
    WITH
    (    LOCATION   = 'https://synapsestorageadls12.dfs.core.windows.net/project'
    );
```

- **Created External File Formats for the CSV Formats:**

```
--- Creating External File Formats---

---**Creating External File Format (using parser version 2.0):**

IF NOT EXISTS (SELECT * FROM sys.external_file_formats WHERE name ='csv_file_format')
CREATE EXTERNAL FILE FORMAT csv_file_format

WITH (

FORMAT_TYPE = DELIMITEDTEXT,
FORMAT_OPTIONS (

FIELD_TERMINATOR = ','

, STRING_DELIMITER = '"'
, First_Row = 2
, USE_TYPE_DEFAULT = FALSE
, Encoding = 'UTF8'
, PARSER_VERSION = '2.0' )

);
```

```
----Creating External File Format (using parser version 1.0):

 IF NOT EXISTS (SELECT * FROM sys.external_file_formats WHERE name ='csv_file_format_pv1')
CREATE EXTERNAL FILE FORMAT csv_file_format_pv1
WITH (
    FORMAT_TYPE = DELIMITEDTEXT,
    FORMAT_OPTIONS (
      FIELD_TERMINATOR = ','
    , STRING_DELIMITER = '""'
    , First_Row = 2
    , USE_TYPE_DEFAULT = FALSE
    , Encoding = 'UTF8'
    , PARSER_VERSION = '1.0' )
    );
```

See below for the setup created on Synapse for Database, External Tables, data source, file formats:

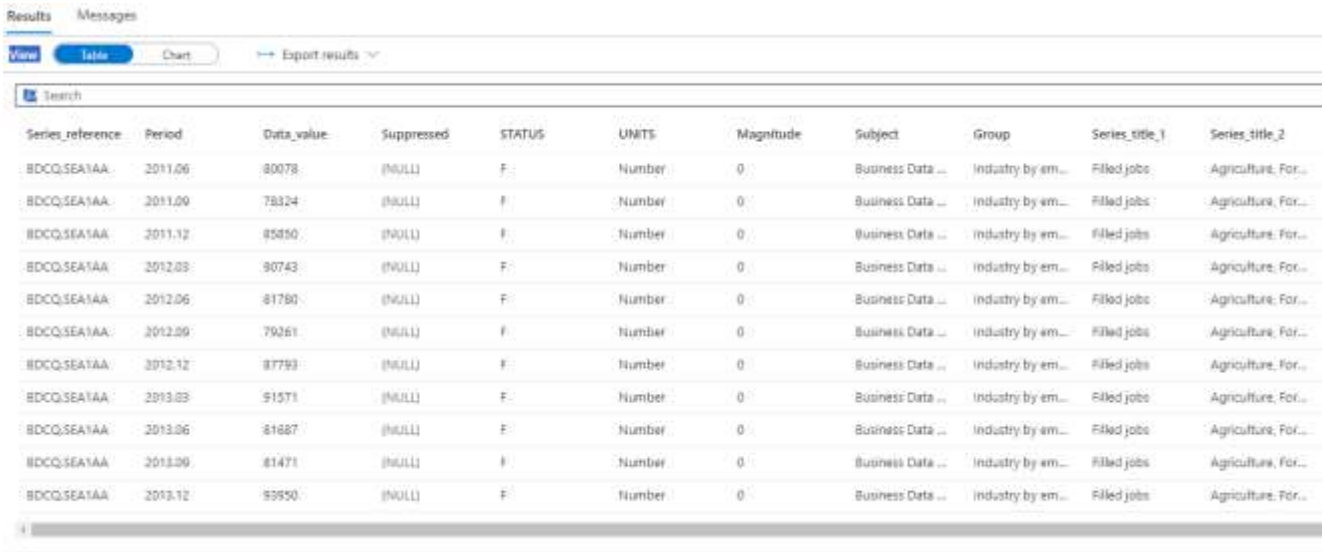


- **Created External Table for the Bronze Layer:**

```sql
--- Creating an External Table (Brozne Layer)---

IF OBJECT_ID('bronze.business_employment') IS NOT NULL
    DROP EXTERNAL TABLE bronze.business_employment;

CREATE EXTERNAL TABLE bronze.business_employment
    (
        Series_reference NVARCHAR(100),
        Period NVARCHAR(50),
        Data_value FLOAT,
        Suppressed NVARCHAR(10),
        STATUS NVARCHAR(50),
        UNITS NVARCHAR(50),
        Magnitude FLOAT,
        Subject NVARCHAR(100),
        [Group] NVARCHAR(100),
        Series_title_1 NVARCHAR(100),
        Series_title_2 NVARCHAR(100),
        Series_title_3 NVARCHAR(100),
        Series_title_4 NVARCHAR(100),
        Series_title_5 NVARCHAR(100)
        )
    WITH (
        LOCATION = 'Data/business_employment.csv',
        DATA_SOURCE = bus_emp_src,
        FILE_FORMAT = csv_file_format_pv1,
        REJECT_VALUE = 10,
        REJECTED_ROW_LOCATION = 'rejections/employment'
    );
```

The below is our source data stored in Bronze Layer:

```sql
Select* from bronze.business_employment;
```



3. Data Cleaning & Transformations:

```
---- Selecting only the desired columns needed along with the Header Row---

SELECT Period, Subject, [Group], Series_title_1 AS jobs_filled,
        Series_title_2 AS industry
FROM
    OPENROWSET(
        BULK 'Data/business_employment.csv',
        DATA_SOURCE = 'bus_emp_src',
        FORMAT = 'CSV',
        HEADER_ROW = True,
        PARSER_VERSION = '2.0'
        ) AS [result]
```

Results Messages

View Table Chart Export results

Search

Period	Subject	Group	jobs_filled	industry
2012.12	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing
2013.03	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing
2013.06	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing
2013.09	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing
2013.12	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing
2014.03	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing
2014.06	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing
2014.09	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing
2014.12	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing

00:00:01 Query executed successfully.

```sql
--- Correcting the Datatype for Period--

SELECT Period, Subject, [Group], Series_title_1 AS jobs_filled,
        Series_title_2 AS industry
FROM
    OPENROWSET(
        BULK 'Data/business_employment.csv',
        DATA_SOURCE = 'bus_emp_src',
        FORMAT = 'CSV',
        HEADER_ROW = True,
        PARSER_VERSION = '2.0'
        )
        WITH (
            Period FLOAT,
            Subject NVARCHAR(100),
            [Group] NVARCHAR(100),
            Series_title_1 NVARCHAR(100),
            Series_title_2 NVARCHAR(100)
              ) AS [result]
```

Results Messages

View Table Chart Export results

Search

Period	Subject	Group	jobs_filled	industry
2011.06	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing
2011.09	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing
2011.12	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing
2012.03	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing
2012.06	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing
2012.09	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing
2012.12	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing
2013.03	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing

00:00:01 Query executed successfully.

```
--- Filtering it with the Period from last 10 years i.e., 2014-01 to Current and Series Title 3 = Jobs Filled---


SELECT Period, Subject, [Group], Series_title_1 AS jobs_filled,
            Series_title_2 AS industry
FROM
    OPENROWSET(
        BULK 'Data/business_employment.csv',
        DATA_SOURCE = 'bus_emp_src',
        FORMAT = 'CSV',
        HEADER_ROW = True,
        PARSER_VERSION = '2.0'
         ) AS [result]
         WHERE Period > 2014.01 AND Series_title_1 = 'Filled jobs'
```

Results Messages

View Table Chart Export results

Search

Period	Subject	Group	jobs_filled	industry
2014.03	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing
2014.06	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing
2014.09	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing
2014.12	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing
2015.03	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing
2015.06	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing

Created Silver Layer External Table after the Transformation:

```
--- Creating External Table Silver Layer------

IF OBJECT_ID('silver.business_employment') IS NOT NULL
    DROP EXTERNAL TABLE silver.business_employment;

CREATE EXTERNAL TABLE silver.business_employment
    (
        Period FLOAT,
        [Subject] NVARCHAR(100),
        [Group] NVARCHAR(100),
        jobs_filled NVARCHAR(100),
        industry NVARCHAR(100)
        )
        WITH
        (
        LOCATION = 'Data/Silver/Silver Layer.csv',
        DATA_SOURCE = bus_emp_src,
        FILE_FORMAT = csv_file_format,
        REJECT_VALUE = 10,
        REJECTED_ROW_LOCATION = 'rejections/employment'
    );
```

Corrected the Date format from YYYY.MM to YYYY-MM-DD.

```
---- To Correct the Date Format-----

SELECT CONCAT(
        LEFT(Period, 4),    -- Get the year
        '-',                        -- Add a hyphen
        RIGHT(Period, 2),    -- Get the month
        '-01'                       -- Add '-01' for the day
    ) AS formatted_date

FROM    silver.business_employment;
```

Period	formatted_date
2014.03	2014-03-01
2014.06	2014-06-01
2014.09	2014-09-01
2014.12	2014-12-01
2015.03	2015-03-01
2015.06	2015-06-01
2015.09	2015-09-01
2015.12	2015-12-01

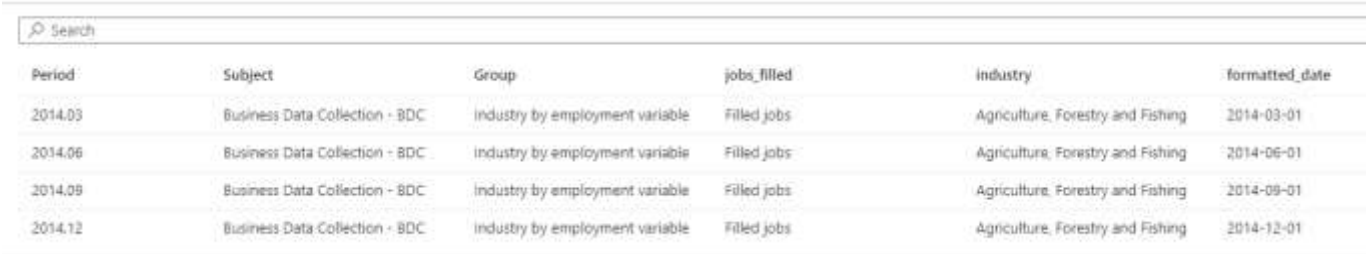
Added a new column to insert the formatted date see below:

```
---- Added a new column formatted date with the YYYY-MM-DD Format----

SELECT Period, Subject, [Group], Series_title_1 AS jobs_filled,
        Series_title_2 AS industry, CONCAT(
        LEFT(Period, 4),    -- Get the year
        '-',                        -- Add a hyphen
        RIGHT(Period, 2),    -- Get the month
        '-01' ) AS formatted_date
FROM
    OPENROWSET(
        BULK 'Data/business_employment.csv',
        DATA_SOURCE = 'bus_emp_src',
        FORMAT = 'CSV',
        HEADER_ROW = True,
        PARSER_VERSION = '2.0'
        ) AS [result]
        WHERE Period > 2014.01 AND Series_title_1 = 'Filled jobs'
```

Period	Subject	Group	jobs_filled	industry	formatted_date
2014.03	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing	2014-03-01
2014.06	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing	2014-06-01
2014.09	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing	2014-09-01
2014.12	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing	2014-12-01
2015.03	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing	2015-03-01
2015.06	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing	2015-06-01
2015.09	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing	2015-09-01
2015.12	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing	2015-12-01

Created Gold Layer External Table after the Transformation:

```
----- Created the Gold Layer External Table with updated date format---

CREATE EXTERNAL TABLE gold.business_employment (
    Period FLOAT,
    [Subject] NVARCHAR(100),
    [Group] NVARCHAR(100),
    jobs_filled NVARCHAR(100),
    industry NVARCHAR(100),
    formatted_date  DATE  --  new column here
)
WITH (
    LOCATION = 'Data/Gold/Gold Layer.csv',
    DATA_SOURCE = bus_emp_src,
    FILE_FORMAT = csv_file_format,
    REJECT_VALUE = 10,
    REJECTED_ROW_LOCATION = 'rejections/employment'
    );

Select * FROM gold.business_employment
```



Search

Period	Subject	Group	jobs_filled	industry	formatted_date
2014.03	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing	2014-03-01
2014.06	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing	2014-06-01
2014.09	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing	2014-09-01
2014.12	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing	2014-12-01

4. Created a view for ease of access with all the updates:

The below query of the dataset for key insights such as Total employment growth over the decade.

```
---- Created a view ----

CREATE VIEW my_view AS
SELECT
    Period, Subject, [Group], jobs_filled, industry, formatted_date
FROM
    gold.business_employment

SELECT * FROM my_view
```

Period	Subject	Group	jobs_filled	industry	formatted_date
2014.03	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing	2014-03-01
2014.06	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing	2014-06-01
2014.09	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing	2014-09-01
2014.12	Business Data Collection - BDC	Industry by employment variable	Filled jobs	Agriculture, Forestry and Fishing	2014-12-01

00:00:01 Query executed successfully.

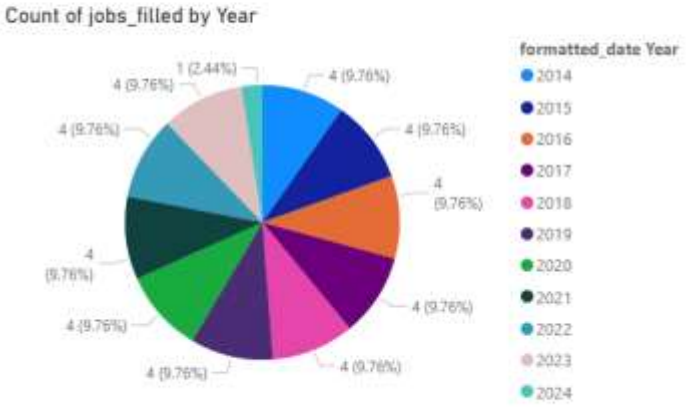
5. Visualizing Results:

Imported the Data to Power BI Desktop and created a Report indicating the jobs filled per year and per quarter in the last decade:

Pie-Chart: Showing the percentage of people employed over the last ten years.

Table-Chart: Break down of the jobs_filled as per quarter and yearly records and total records.

Multiple Card: On the top to display the Industry, Subject and jobs_filled.



Year	Quarter	Month	Day	Count of jobs_filled
2014	Qtr 1	March	1	1
2014	Qtr 2	June	1	1
2014	Qtr 3	September	1	1
2014	Qtr 4	December	1	1
2015	Qtr 1	March	1	1
2015	Qtr 2	June	1	1
2015	Qtr 3	September	1	1
2015	Qtr 4	December	1	1
2016	Qtr 1	March	1	1
2016	Qtr 2	June	1	1
2016	Qtr 3	September	1	1
2016	Qtr 4	December	1	1
2017	Qtr 1	March	1	1
2017	Qtr 2	June	1	1
Total				41

6. Conclusion & Insights:

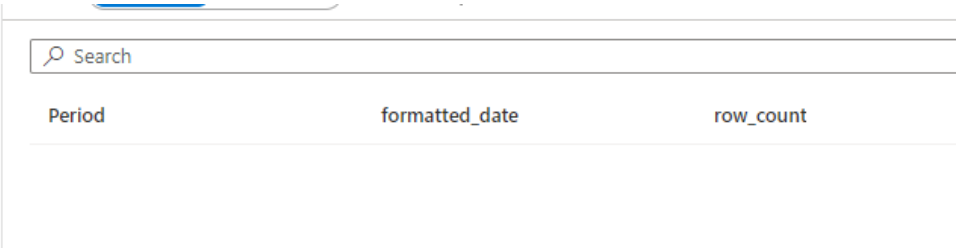
After gathering the data and visualizing trends, document key insights such as:

- Total employment over the decade.
- Year-on-year employment in the industry of Agriculture, Forest and Fishing.

For the given three tables Bronze, Silver and Gold we had ‘Period’ as a matching column.

Checked for any duplicates for Column: In order to select the key

```
--------------Checked for Duplicates ------------

SELECT Period, formatted_date,
    COUNT(*) AS row_count
FROM
    gold.business_employment
GROUP BY
    Period, formatted_date
HAVING
    COUNT(*) > 1;
```



Created a Data Model as below:

Error Faced	How it was fixed	Additional Comments

Sign-in Error with Power BI does not allow to import data to Power BI	Created the external Data Source with Master Key Encryption and Used 'SAS' - Shared access signature	CREATE MASTER KEY ENCRYPTION BY PASSWORD = 'Welcome@1'; CREATE DATABASE SCOPED CREDENTIAL bus_emp_new_credential WITH IDENTITY = 'SHARED ACCESS SIGNATURE', SECRET = 'sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupyx&se=2024-09-21T07:58:41Z&st=2024-09-20T23:58:41Z&spr=https&sig=zSLKhVY6WJYbvClmr%2BO6FwwktiWJNsuhU7JyjwzU58I%3D'; IF NOT EXISTS (SELECT * FROM sys.external_data_sources WHERE name = 'bus_emp_src_new') CREATE EXTERNAL DATA SOURCE bus_emp_src_new WITH (LOCATION = 'https://synapsestorageadls12.dfs.core.windows.net/project', CREDENTIAL = bus_emp_new_credential);
Couldn’t transform the external tables directly	Worked on the data present in ADLS to clean and transform	N/A