Sri Lanka Institute of Information Technology

Data Warehousing and Business Intelligence

Assignment 1

2021

Name: Kulasekara K.M.S.B

Registration No: IT19129068

Table of Contents

Table of Figures

1. Data Set Selection

1.1 Description

Data Set – Daily Power Generation in India (2017-2020)

Description – The data set consisted of three tables. The first table which had Power Generation Data describes the region-wise power generation in India on daily basis. The table's columns were named index, Region, Date, Thermal Generation Actual, Thermal Generation Estimated, Nuclear Generation Actual, Nuclear Generation Estimated, Hydro Generation Actual and Hydro Generation Estimated. All the power generation measurements have been measured in Mega Units (MU).

The second data set consisted of Region Details which had four columns that were named Region, Area, National Share and House Hold Size.

The third and final table which was named Region Coordinates had geographical data about the region. The table's columns were Region, Latitude and Longitude.

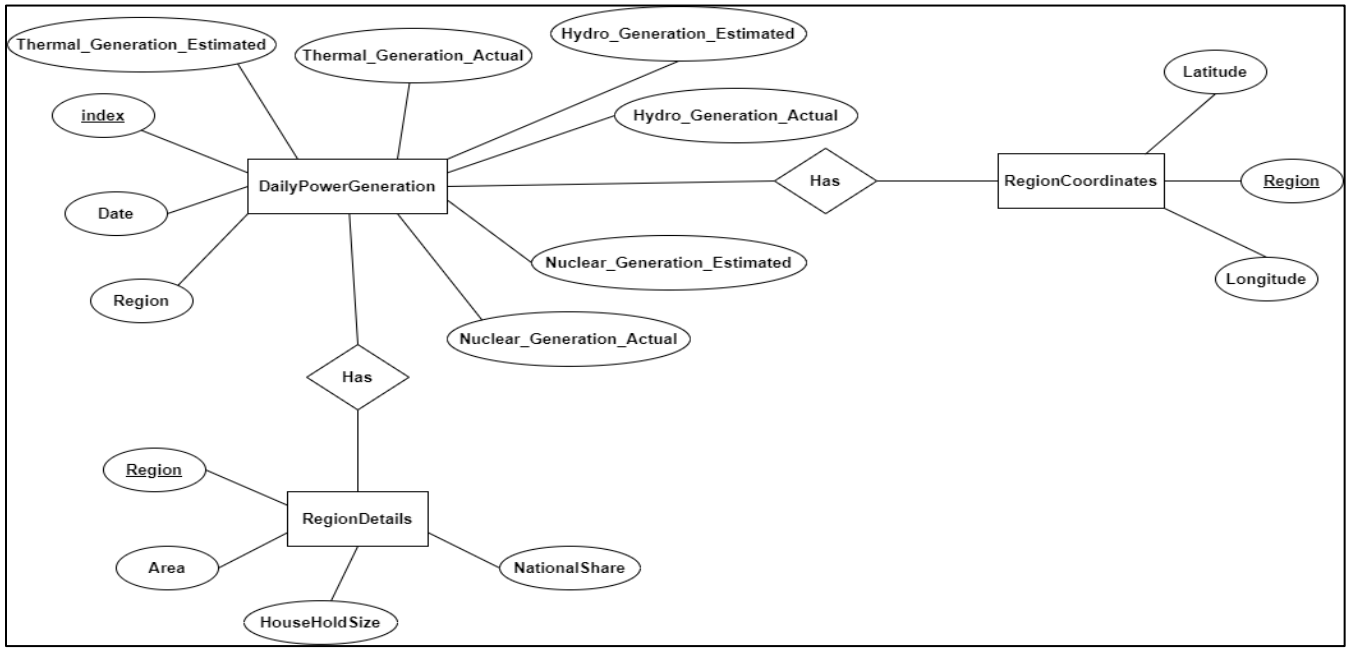

Figure 1-1 ER Diagram

2. Preparation of Data Sources

All the tables were initially in CSV format. The Region Coordinates.csv file was kept as it is and the other two were changed into text and database formats. After preparation for extraction,

1. The region_coordinates.csv file had region coordinates data.
2. The RegionDetails.txt file had region details such as area, national share and house hold size.
3. The PowerGeneration database file had daily power generation data.

Although there were details about the "Central Region" in both region coordinates and region details files, they had to be removed as there were no records related to that region in the power generation file.

3. Solution Architecture

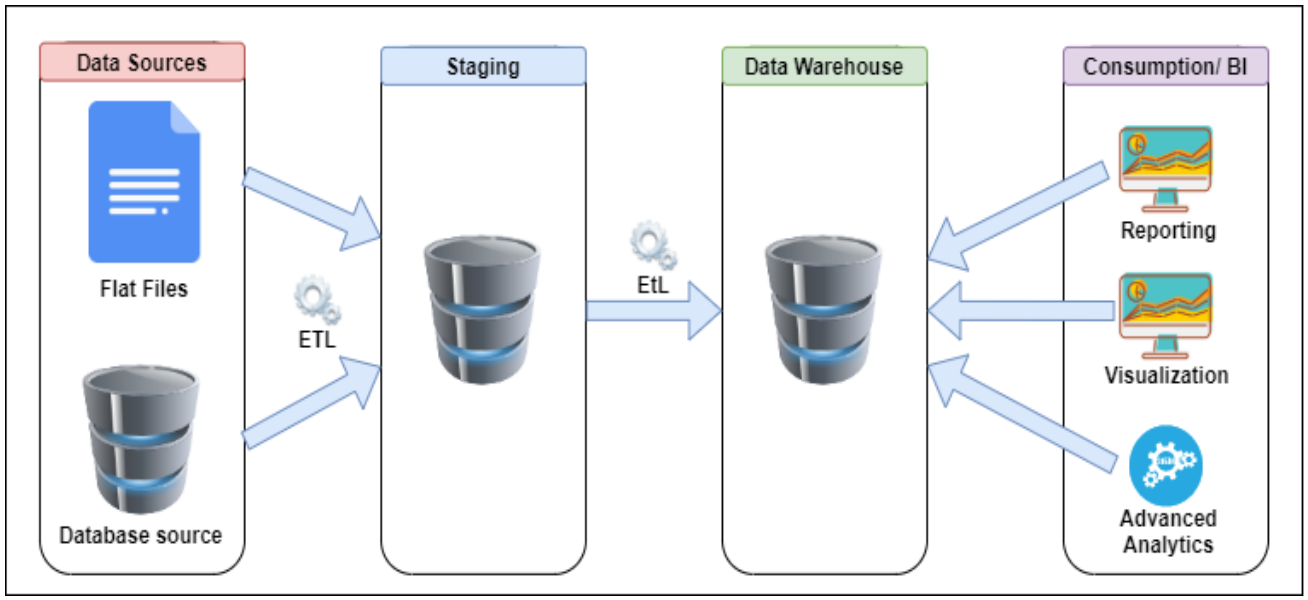

Figure 3-1 High Level Architectural diagram

As shown in the diagram there are four levels in the solution architecture. The Data Sources level consists of the three files that were prepared for data extraction. The Staging level consists of a database where the extracted data are stored in three separate database tables. The Data Warehouse level consists of three dimension tables and one fact table.

After creating the Data Warehouse, the data can be used by the consumption/ BI layer for Reporting, Visualization and Advanced Analytics such as machine learning and predictive modelling.

4. Data warehouse design & development

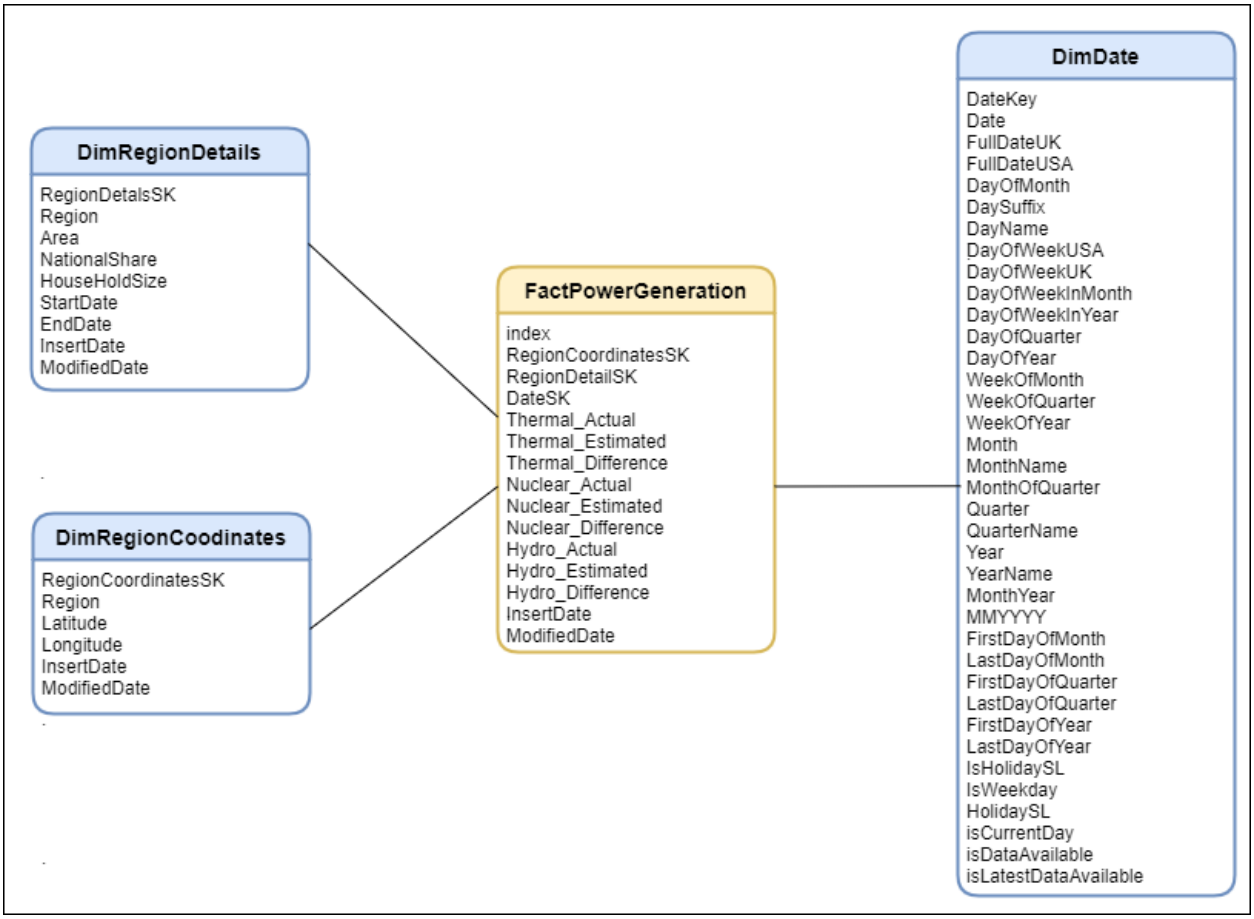

Figure 4-1 Dimensional Model

The dimensional model is designed based on the concepts of the star schema. There are three dimensions (DimRegionDetails, DemRegionCoordinates, DimDate) and one fact table (FactPowerGeneration) used to design the model.

The DimRegionDetails dimension which is considered as the Slowly Changing Dimension consists of nine attributes with RegionDetailsSK being the Surrogate Key. This has been designed as a Type Two slowly changing dimension.

The DimDate dimension consists of thirty-seven attributes. The DateKey has been used as the Surrogate Key.

The coordinates of each region have been included in the DimRegionCoordinates dimension. The RegionCoordinatesSK has been used as the Surrogate Key.

There are fifteen attributes in the FactPowerGeneration table. This table consists of the daily power generation data in India's three main power generation sources. The Thermal_Difference, Nuclear_Difference and Hydro_Difference are derived attributes that contain the difference between actual and estimated power generation.

Insert date and Modified date columns have been included in every table in order to keep track of the updates.

Since there were NULL values for some days in power generation data, it was assumed that whenever there is a NULL value there has not been any power generation related to that source. Hence they were converted into 0. This will be further described in the ETL development step.

5. ETL development

5.1 Load data from Source to Staging

The execution order

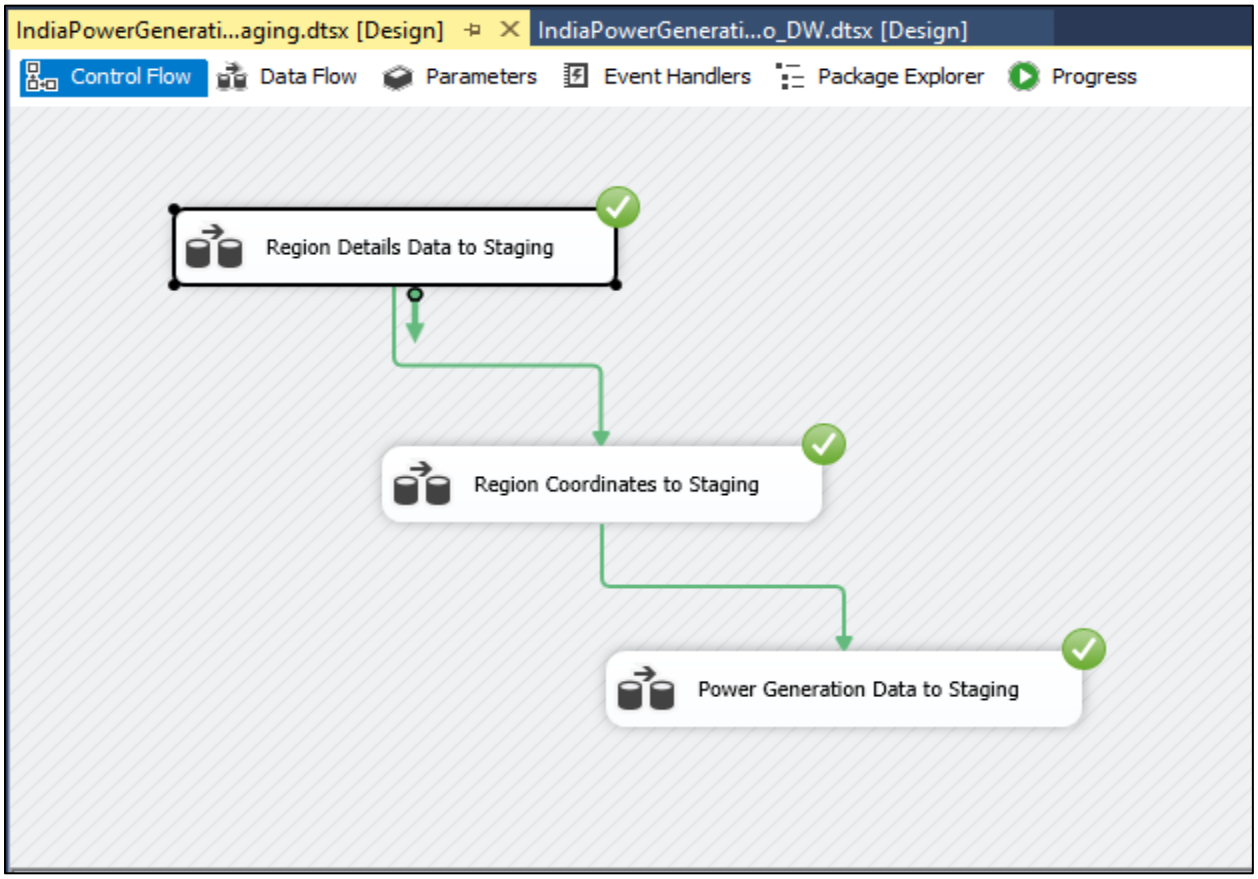

Figure 5-1 Load data from Source to Staging (Control Flow)

Step 1 - Region Detail Data to Staging

Since the region details source has been a text file all the data in the file has been in string data type. Hence they had to be converted into other data types. After conversion, the data types have been as follows,

- Region – nvarchar(50)
- Area – numeric(18,0)
- NationalShare – numeric(18,4)
- HouseHoldShare – numeric(18,4)

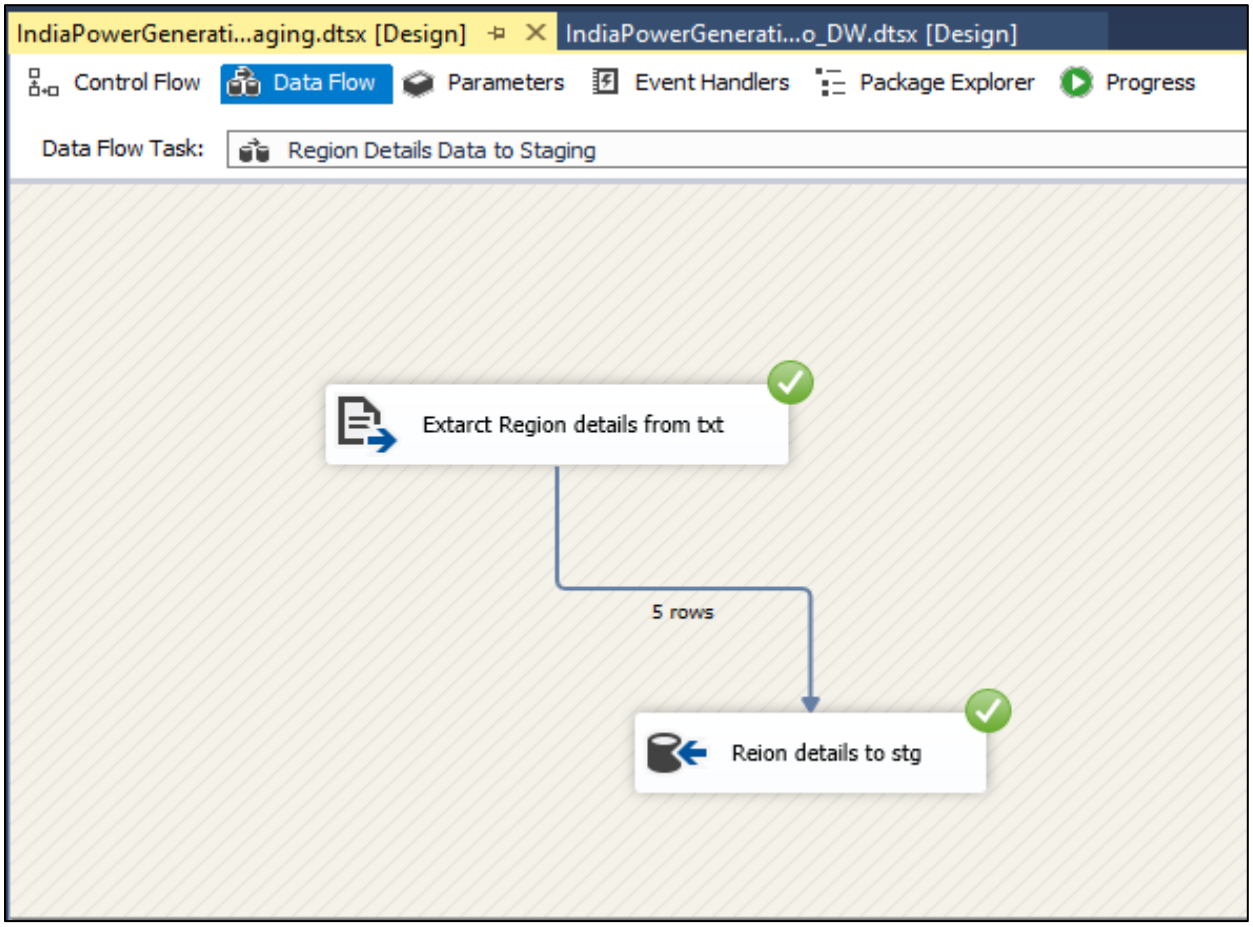

Figure 5-2 Region Details to Staging

Step 2 - Region Coordinates to Staging

The region coordinates data source has been in CSV format. Hence the data in that table were also in string data type and had to be converted into other data types. After conversion, the data types have been as follows,

- Region – nvarchar(50)
- Latitude – numeric(18,4)
- longitude – numeric(18,4)

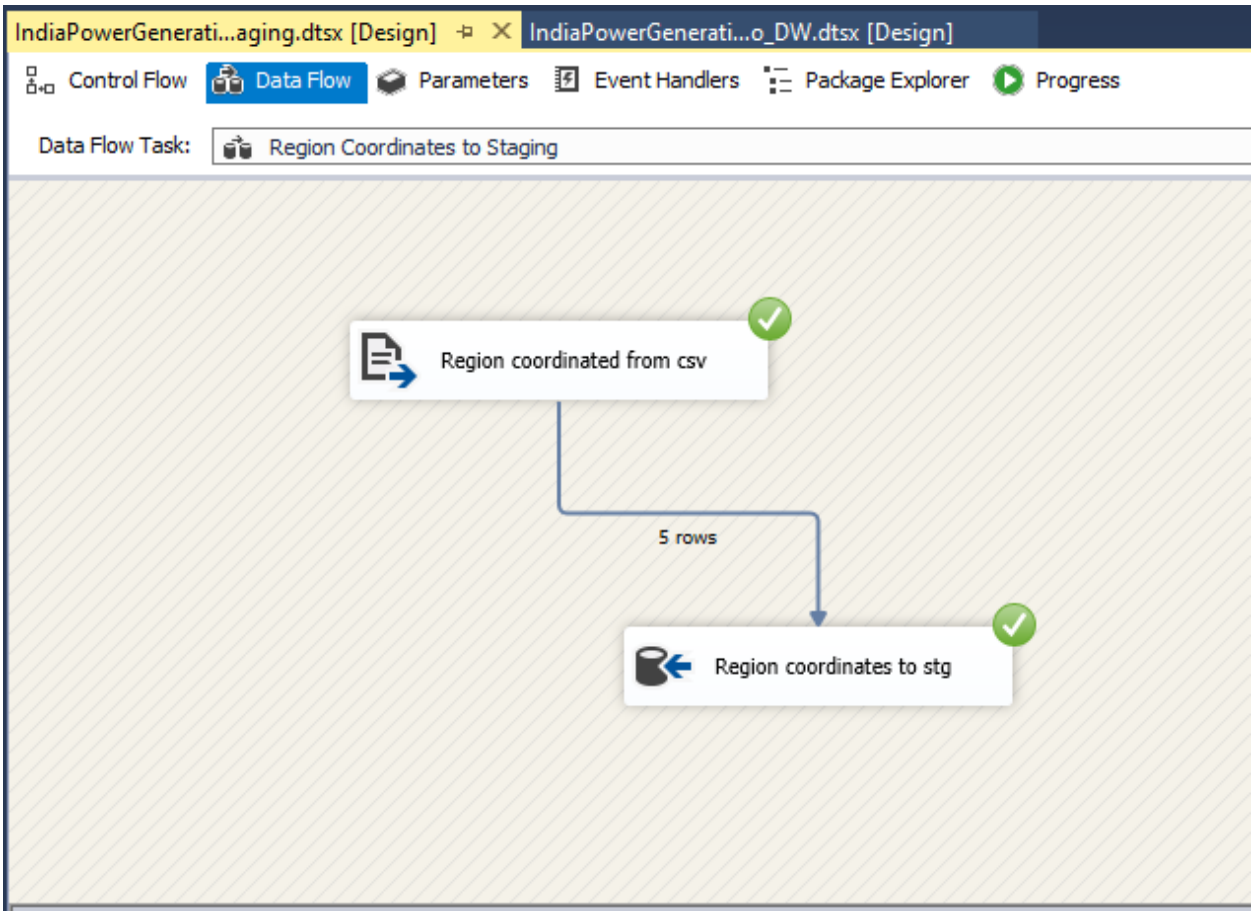

Figure 5-3 Region Coordinates to Staging

Step 3 - Power Generation data to staging

Since the power generation source has been a database table there was no need to convert any data types. However, since there were some NULL values in the source table as mentioned in the previous step, a stored procedure has been used to convert the null values into 0 s.

```
CREATE PROCEDURE [dbo].[insertPowerGeneraionStg]
@index int,
@Date date,
@Region nvarchar(20),
@Thermal_Actual NUMERIC(6,2),
@Thermal_Estimated NUMERIC(6,2),
@Nuclear_Actual NUMERIC(6,2),
@Nuclear_Estimated NUMERIC(6,2),
@Hydro_Actual NUMERIC(6,2),
@Hydro_Estimated NUMERIC(6,2)
AS
BEGIN
if(@Thermal_Actual IS NULL)
    BEGIN
        SET @Thermal_Actual = 0;
    END;

if(@Thermal_Estimated IS NULL)
    BEGIN
        SET @Thermal_Estimated = 0;
    END;

if (@Nuclear_Actual IS NULL)
    BEGIN
        SET @Nuclear_Actual = 0;
    END;

if(@Nuclear_Estimated IS NULL)
    BEGIN
        SET @Nuclear_Estimated = 0;
    END;

if(@Hydro_Actual IS NULL)
    BEGIN
        SET @Hydro_Actual = 0;
    END;

if(@Hydro_Estimated IS NULL)
    BEGIN
        SET @Hydro_Estimated = 0;
    END;

INSERT INTO dbo.PowerGenerationStg
VALUES (@index,@Date,@Region,@Thermal_Actual,@Thermal_Estimated,@Nuclear_Actual,@Nuclear_Estimated,@Hydro_Actual,@Hydro_Estimated)
END;
```

Figure 5-4 Power Generation to Staging – Stored Procedure to Convert NULLs into 0s

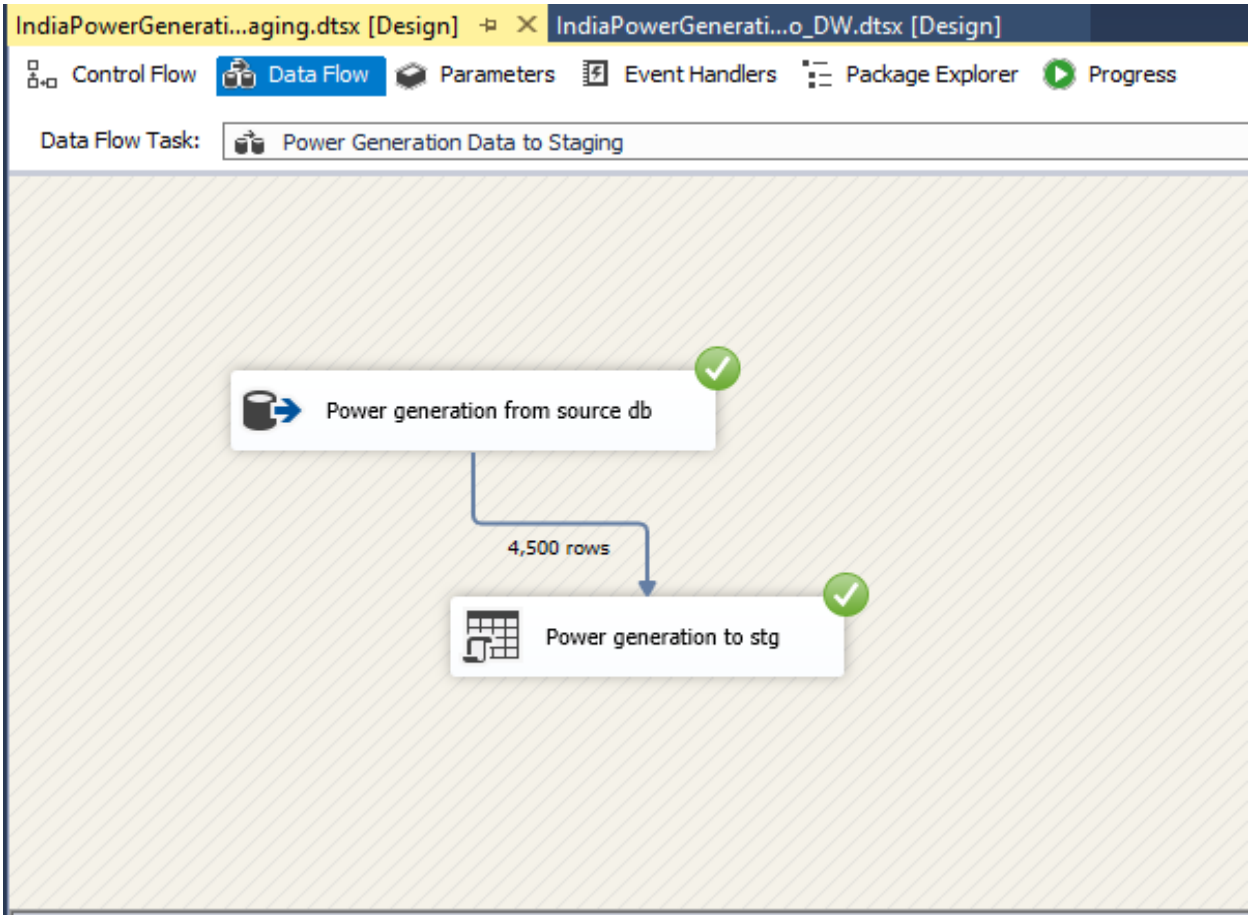

Figure 5-5 Power Generation data to Staging

5.2 Load data from Staging to Data Warehouse

The execution order

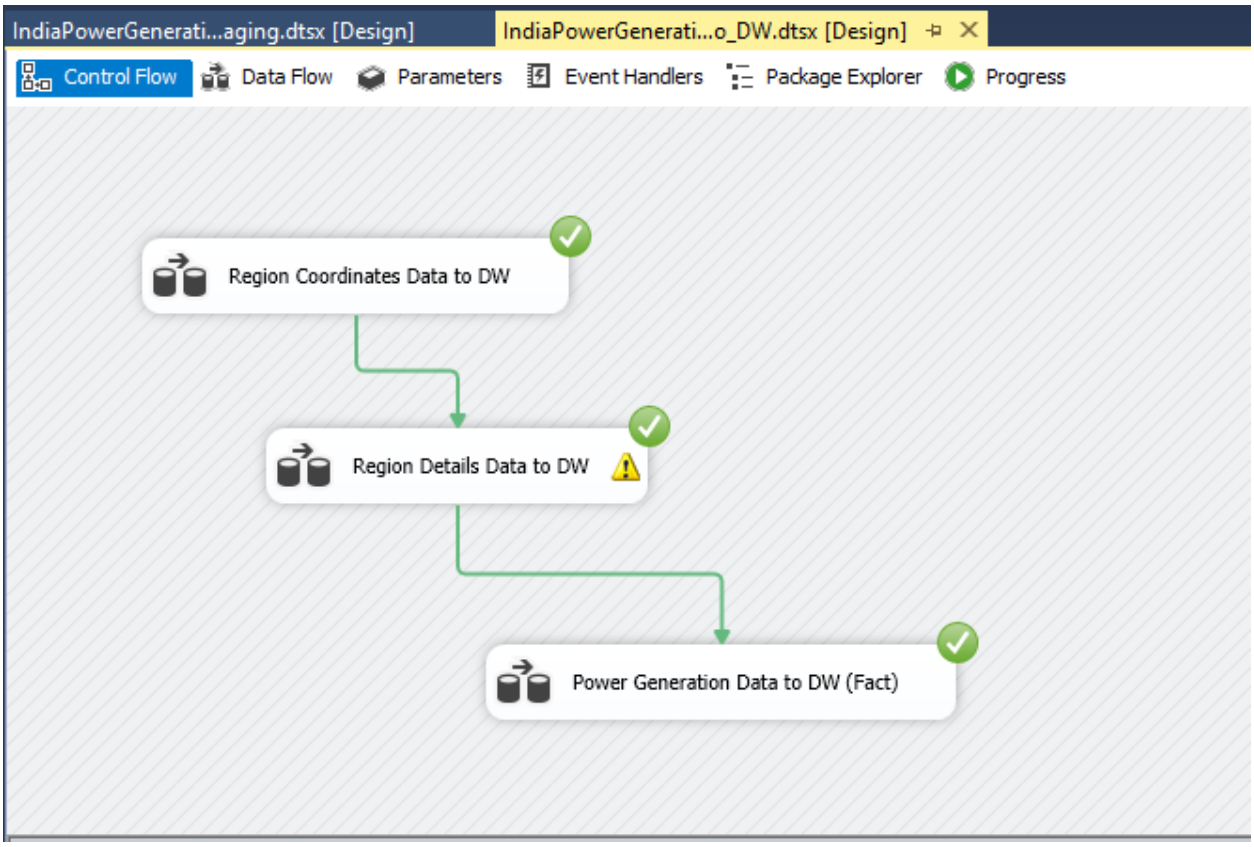

Figure 5-6 Load data from Staging to Data Warehouse (Control Flow)

Step 1 - Region Coordinates Data to DW

A stored procedure has been used to maintain the Insert Date and the Modified Date columns.

```
CREATE PROCEDURE [dbo].[updateDimRegionCoordiates]
@Region nvarchar(50),
@Latitude numeric(18,4),
@Longitude numeric(18,4)
AS
BEGIN

if not exists(SELECT [RegionCoordinatesSK]
FROM [dbo].[DimRegionCoordinates]
WHERE [Region] = @Region)
BEGIN
INSERT INTO [dbo].[DimRegionCoordinates]
([Region],[Latitude],[Longitude],[InsertDate],[ModifiedDate])
VALUES
(@Region,@Latitude,@Longitude,GETDATE(),GETDATE())
END;

if exists(SELECT [RegionCoordinatesSK]
FROM [dbo].[DimRegionCoordinates]
WHERE [Region] = @Region)
BEGIN
UPDATE [dbo].[DimRegionCoordinates]
SET [Latitude] = @Latitude,
[Longitude] = @Longitude,
[ModifiedDate] = GETDATE()
WHERE [Region] = @Region
END;
END;
```

Figure 5-7 Stored Procedure for loading Region Coordinates to DW

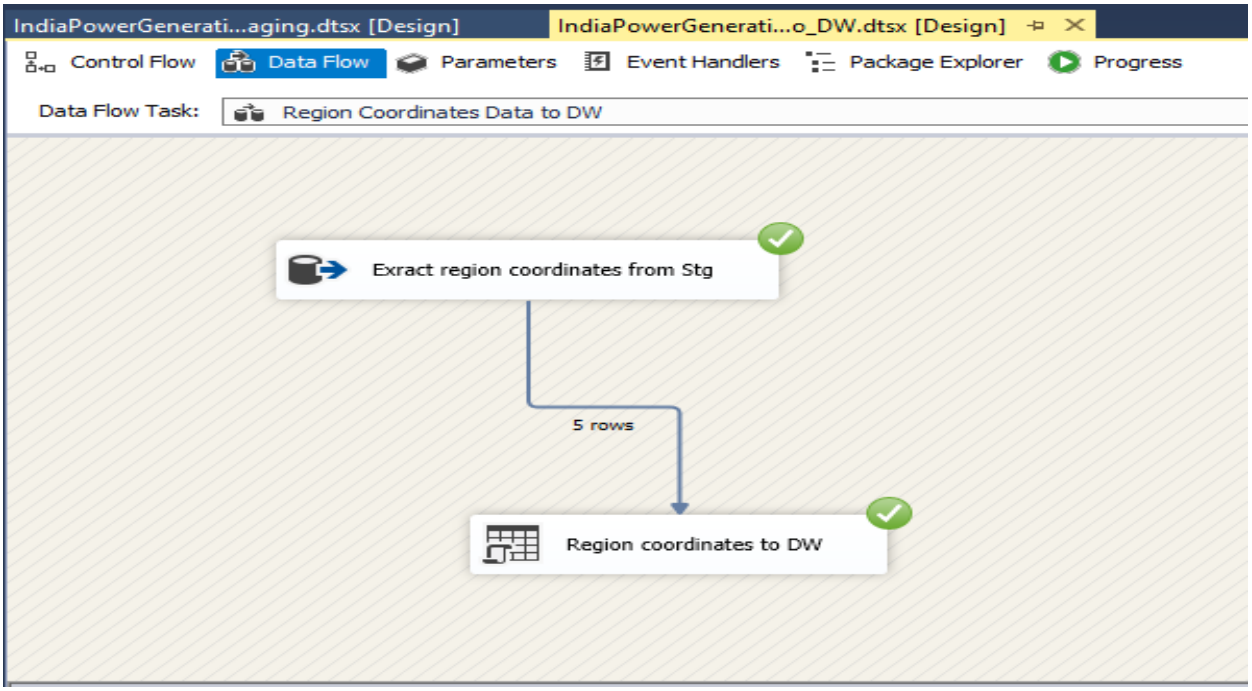

Figure 5-8 Region Coordinates data to DW

Step 2 - Region Details Data to DW (Slowly Changing Dimension)

In the dimension table, the HouseHoldSize attribute has been considered as the slowly changing attribute as it can change from time to time.

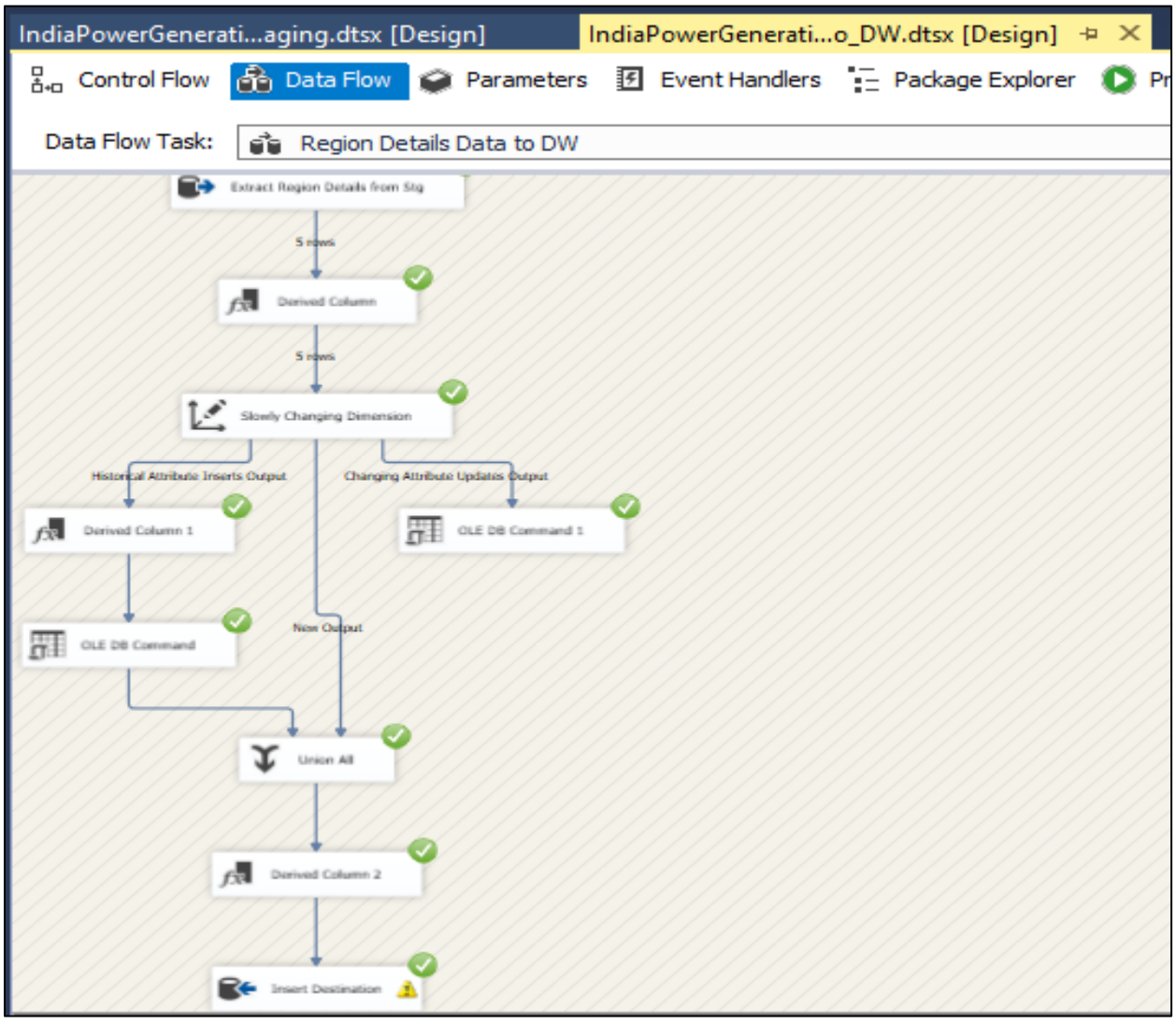

Figure 5-9 Region Details Data to DW

Step 3 - Power Generation Data to DW (Fact table)

As the final step, three lookups has been used in the Power Generation Data to DW task to get the surrogate keys from the dimension tables. After that, a stored procedure has been used to insert the data since the insert date and the modified data have to be maintained.

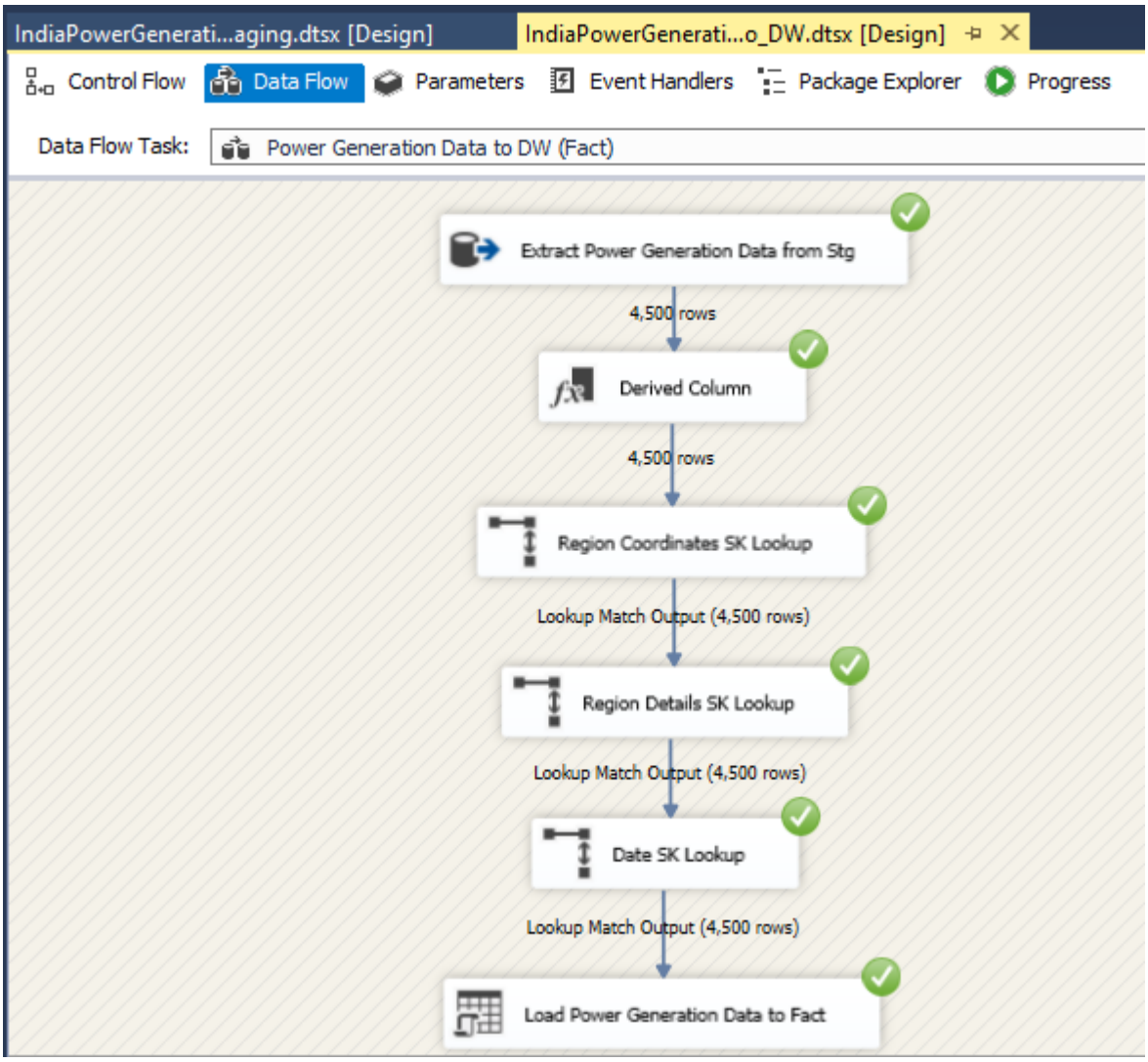

Figure 5-10 Power Generation Data to DW

```
CREATE PROCEDURE [dbo].[insertPowerGeneraionToFACT]
@index int,
@RegionCoodinatesSK int,
@RegionDetailsSK int,
@DateSK int,
@Thermal_Actual numeric(6,2),
@Thermal_Estimated numeric(6,2),
@Nuclear_Actual numeric(6,2),
@Nuclear_Estimated numeric(6,2),
@Hydro_Actual numeric(6,2),
@Hydro_Estimated numeric(6,2)
AS
BEGIN

if not exists(SELECT [index]
FROM [dbo].[FactPowerGeneration]
WHERE [RegionDetailsSK] = @RegionDetailsSK AND [DateSK] = @DateSK)
BEGIN
INSERT INTO [dbo].[FactPowerGeneration]
([index],[RegionCoordinateSK],[RegionDetailsSK],[DateSK],[Thermal_Actual],[Thermal_Estimated],[Nuclear_Actual],[Nuclear_Estimated],[Hydro_Actual],[Hydro_Estimated],[InsertDate],[ModifiedDate])
VALUES
(@index,@RegionCoodinatesSK,@RegionDetailsSK,@DateSK,@Thermal_Actual,@Thermal_Estimated,@Nuclear_Actual,@Nuclear_Estimated,@Hydro_Actual,@Hydro_Estimated, GETDATE(),GETDATE())
END;

if exists (SELECT [index]
FROM [dbo].[FactPowerGeneration]
WHERE [RegionDetailsSK] = @RegionDetailsSK AND [DateSK] = @DateSK)
BEGIN
UPDATE [dbo].[FactPowerGeneration]
SET [RegionCoordinateSK] = @RegionCoodinatesSK,
[RegionDetailsSK] = @RegionDetailsSK,
[DateSK] = @DateSK,
[Thermal_Actual] = @Thermal_Actual,
[Thermal_Estimated] = @Thermal_Estimated,
[Nuclear_Actual] = @Nuclear_Actual,
[Nuclear_Estimated] = @Nuclear_Estimated,
[Hydro_Actual] = @Hydro_Actual,
[Hydro_Estimated] = @Hydro_Estimated,
[ModifiedDate] = GETDATE()
WHERE [index] = @index
END;
END;
```

Figure 5-11 Stored Procedure for loading Power Generation data to DW

6. References

- https://www.kaggle.com/
- https://courseweb.sliit.lk/