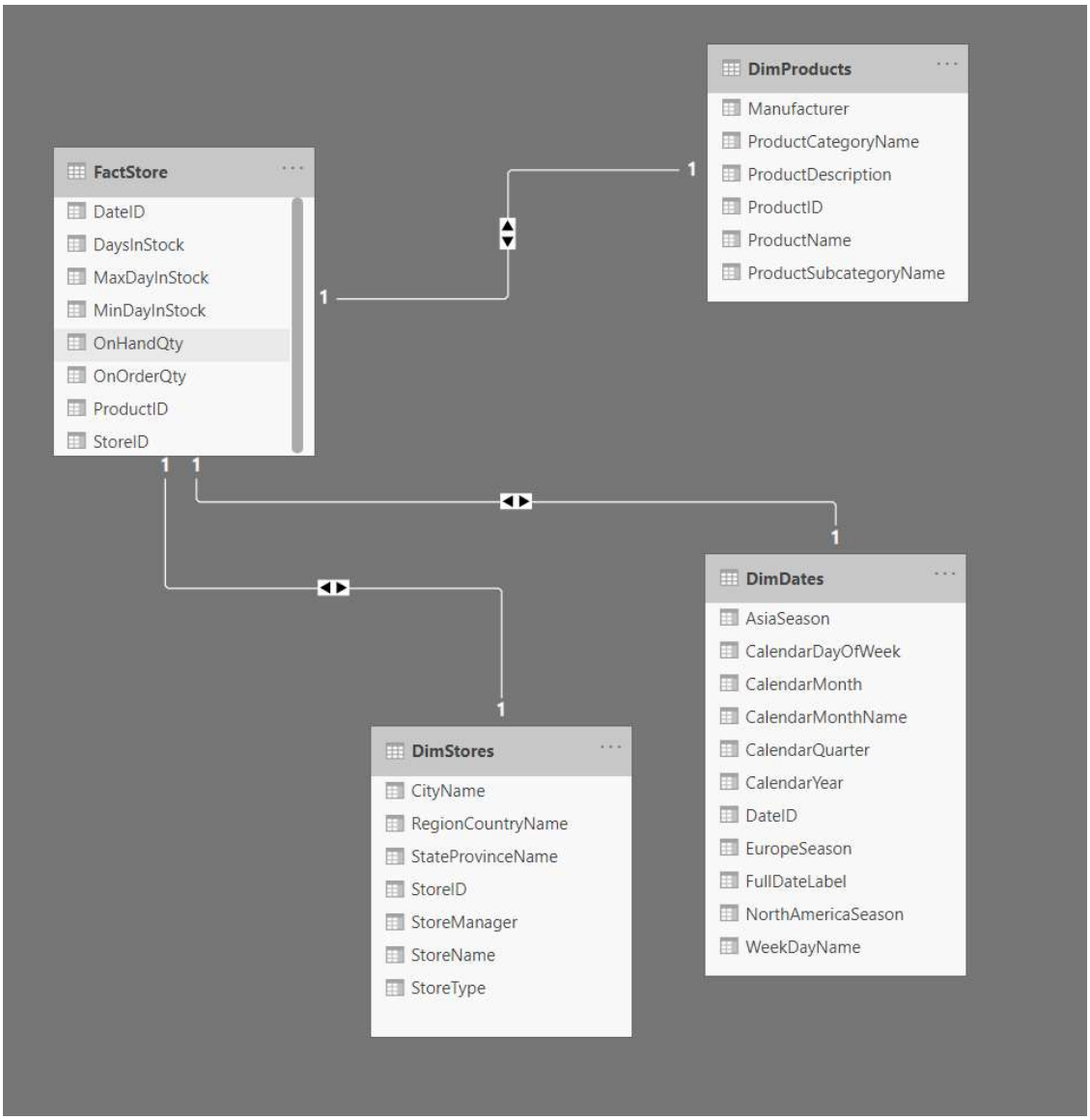
Data Analyst Interview Task Summary

Task 1: Reference DDL-DML.sql to create physical database model

Steps taken:

1. Create 4 sheets in excel corresponding to the different tables
2. Import the excel sheets into power BI
3. Create relationship between tables using primary keys



Task 2: Process the raw data in zip folder and load the processed data model into a database for querying the results

Steps taken:

1. Using python, append a list of directories of all files ending with ".txt.gz"

```
# initiate an empty list
list=[]

# loop through a list of file names
for filename in glob.iglob(root_dir + '**/**/***.txt.gz', recursive=False):
    list.append(filename)
```

2. Unzip each file and load into a pandas dataframe

```
# initiate an empty list
df_list = []

# unzip all txt files and convert into pandas df
for i in list:
    df = pd.read_csv(i, compression='gzip', header=0, sep=None, quotechar='"', error_bad_lines=False,
                     engine = 'python', index_col=False, encoding = "utf_8")
    df_list.append(df)
```

3. There are two approaches to process the data model:
    a. Python - concatenate all dataframes into one single dataframe. Extract the columns and set the column type to the appropriate format. Then export a complete csv file

```
# concatenate to a single df
df_con = pd.concat(df_list)
```

```
# create a new df with selected columns
df_new = df_con.iloc[:, [0, 1, 3, 8, 9, 12, 13, 14]]
```

```
# rename columns
df_new.columns = ["DateID", "StoreID", "ProductID", "OnHandQty", "OnOrderQty", "DaysInStock", "MinDayInStock", "MaxDayInStock"]
```

```
pd.options.mode.chained_assignment = None # suppress warning

# update "Date" column to type date
df_new["DateID"] = pd.to_datetime(df_new["DateID"], format="%Y-%m-%d")
```

```
df_new.head()
```

	DateID	StoreID	ProductID	OnHandQty	OnOrderQty	DaysInStock	MinDayInStock	MaxDayInStock	month
0	2009-01-03	1	6	19	0	60	43	76	1
1	2009-01-03	1	29	19	0	58	11	97	1
2	2009-01-03	1	31	19	0	63	17	111	1
3	2009-01-03	1	51	19	0	79	12	86	1
4	2009-01-03	1	54	23	4	30	30	71	1

    b. Power BI – print each unzipped dataframe into separate csv files, load into power BI and concatenate into one single dataframe.

```
# print all csv to dirctory
for index, dataset in enumerate(df_list):

    # Export to CSV
    filepath = os.path.join(destination, 'dataset_'+str(index)+'.csv')
    dataset.to_csv(filepath)
```

DateID	StoreID	ProductID	OnHandQuantity	OnOrderQuantity	DaysInStock	MinDayInStock	MaxDayInStock	AvailableQty	Month	Day	Quarter
10/01/2009	4	181	19	0	34	19	60	19	1	10	1
10/01/2009	4	372	19	0	34	10	63	19	1	10	1
10/01/2009	5	707	19	0	34	25	78	19	1	10	1
10/01/2009	10	2012	19	0	34	56	61	19	1	10	1
10/01/2009	11	305	19	0	34	42	103	19	1	10	1
10/01/2009	11	2046	19	0	34	30	86	19	1	10	1
10/01/2009	13	970	19	0	34	12	108	19	1	10	1
10/01/2009	13	2328	19	0	34	18	98	19	1	10	1

4. Data Transformation is processed as follows:

a. Check column types according to the required data model
b. New column “Availability” = OnHandQuantity – OnOrderQuantity
c. New column “Month”, “Day”, “Quarter” → extracted from DateID column

Task 3: Apply analytics

Steps taken:

1. Define data type dimensions
   a. Time - quarter, month, year
   b. Inventory level– by storeID, by productID
   c. Data dimensions – stock availability (OnHandQty, OnOrderQty), stock days (DaysInStock, MaxDayInStock, MinDayInStock)
2. Questions to ask:
   a. What’s the stock situation like over time by store and product?
      - Uncover trends and patterns over time dimensions
   b. How many products are there in each store?
      - Identify stores with too few or too many products (overstock or underutilization)
   c. What is the distribution of inventory like over time dimensions
      - Uncover trends in patterns
      - Identify products with excess stock

Task 4: Create a visual story behind the data using power bi