# Capstone Project Template

March 26, 2020

## 1 Project Title

### 1.0.1 Data Engineering Capstone Project

**Project Summary** --describe your project at a high level--

The project follows the follow steps: * Step 1: Scope the Project and Gather Data * Step 2: Explore and Assess the Data * Step 3: Define the Data Model * Step 4: Run ETL to Model the Data * Step 5: Complete Project Write Up

```
In [2]: # Do all imports and installs here
        from pyspark.sql import SparkSession
        import os
        import configparser

        config = configparser.ConfigParser()

        #Normally this file should be in ~/.aws/credentials
        config.read_file(open('config.cfg'))

        os.environ["AWS_ACCESS_KEY_ID"] = config['AWS']['AWS_ACCESS_KEY']
        os.environ["AWS_SECRET_ACCESS_KEY"] = config['AWS']['AWS_SECRET_ACCESS_KEY']
```

### 1.0.2 Step 1: Scope the Project and Gather Data

**Scope** Explain what you plan to do in the project in more detail. What data do you use? What is your end solution look like? What tools did you use? etc>

**Describe and Gather Data** Describe the data sets you're using. Where did it come from? What type of information is included?

```
In [5]: # Read in the data here

In [6]: df.head()
```

```
Out[6]:    cicid   i94yr  i94mon  i94cit  i94res i94port  arrdate  i94mode i94addr  \
        0    6.0  2016.0     4.0   692.0   692.0     XXX  20573.0      NaN     NaN
        1    7.0  2016.0     4.0   254.0   276.0     ATL  20551.0      1.0      AL
        2   15.0  2016.0     4.0   101.0   101.0     WAS  20545.0      1.0      MI
```

```
3   16.0  2016.0     4.0   101.0   101.0     NYC  20545.0      1.0      MA
4   17.0  2016.0     4.0   101.0   101.0     NYC  20545.0      1.0      MA

   depdate   ...     entdepu  matflag  biryear   dtaddto gender insnum  \
0      NaN   ...           U      NaN   1979.0  10282016    NaN    NaN
1      NaN   ...           Y      NaN   1991.0       D/S      M    NaN
2  20691.0   ...         NaN        M   1961.0  09302016      M    NaN
3  20567.0   ...         NaN        M   1988.0  09302016    NaN    NaN
4  20567.0   ...         NaN        M   2012.0  09302016    NaN    NaN

  airline        admnum  fltno visatype
0     NaN  1.897628e+09    NaN       B2
1     NaN  3.736796e+09  00296       F1
2      OS  6.666432e+08     93       B2
3      AA  9.246846e+10  00199       B2
4      AA  9.246846e+10  00199       B2

[5 rows x 28 columns]
```

```
In [8]: # save SAS dataset to parquet files
        if True:
            from pyspark.sql import SparkSession
            spark = SparkSession.builder.\
                config("spark.jars.packages","saurfang:spark-sas7bdat:2.0.0-s_2.10")\
                .enableHiveSupport().getOrCreate()
            df_spark = spark.read.format('com.github.saurfang.sas.spark').load('../../data/18-83

            #write to parquet
        #     df_spark.write.parquet("sas_data")
        #     df_spark=spark.read.parquet("sas_data")
```

### 1.0.3 Step 2: Explore and Assess the Data

**Explore the Data** Identify data quality issues, like missing values, duplicate data, etc.

**Cleaning Steps** Document steps necessary to clean the data

```
In [7]: # Performing cleaning tasks here

        df_spark.show(2)
```

```
+-----+------+------+------+------+-------+-------+-------+-------+-------+------+-------+-----+
|cicid| i94yr|i94mon|i94cit|i94res|i94port|arrdate|i94mode|i94addr|depdate|i94bir|i94visa|count|
+-----+------+------+------+------+-------+-------+-------+-------+-------+------+-------+-----+
|  6.0|2016.0|   4.0| 692.0| 692.0|    XXX|20573.0|   null|   null|   null|  37.0|    2.0|  1.0|
|  7.0|2016.0|   4.0| 254.0| 276.0|    ATL|20551.0|    1.0|     AL|   null|  25.0|    3.0|  1.0|
+-----+------+------+------+------+-------+-------+-------+-------+-------+------+-------+-----+
only showing top 2 rows
```

### 1.0.4 Step 3: Define the Data Model

**3.1 Conceptual Data Model** Map out the conceptual data model and explain why you chose that model

**3.2 Mapping Out Data Pipelines** List the steps necessary to pipeline the data into the chosen data model

### 1.0.5 Step 4: Run Pipelines to Model the Data

**4.1 Create the data model** Build the data pipelines to create the data model.

```
In [ ]: # Write code here
```

**4.2 Data Quality Checks** Explain the data quality checks you'll perform to ensure the pipeline ran as expected. These could include: * Integrity constraints on the relational database (e.g., unique key, data type, etc.) * Unit tests for the scripts to ensure they are doing the right thing * Source/Count checks to ensure completeness

Run Quality Checks

```
In [ ]: # Perform quality checks here
```

**4.3 Data dictionary** Create a data dictionary for your data model. For each field, provide a brief description of what the data is and where it came from. You can include the data dictionary in the notebook or in a separate file.

**Step 5: Complete Project Write Up**

- Clearly state the rationale for the choice of tools and technologies for the project.
- Propose how often the data should be updated and why.
- Write a description of how you would approach the problem differently under the following scenarios:
- The data was increased by 100x.
- The data populates a dashboard that must be updated on a daily basis by 7am every day.
- The database needed to be accessed by 100+ people.

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```