RICE DATA ANALYTICS BOOTCAMP

EXTRACT, TRANSFORM AND LOADING (ETL) PROJECT

BY

TEAM 6 MEMBERS

ORLANDO ORTEGA

CADE CULVER

JONATHAN EZEUGO

ABSTRACT

The oil and gas industry is a huge domain of data. Much of it however are proprietary and confidential by reason of strong business competition. However, there are some data made public that can provide some levels of insight into operations and stakeholders. Spanning across many countries, the oil and gas industry is one of the biggest if not the biggest industry in the world, with many countries being players. Argentina is one such country, and this project set out to explore data from the Neuquen Province in the mid-west geographical region of Argentina. The data had to be cleaned as best as possible within the project timeframe, normalized into 10 different tables and since we could establish relationships between the normalized tables, we chose to use PostgreSQL in PgAdmin to structure the tables, establish referential integrity, and upload the database as the project requirements spelt out. The primary stages of the project including CREATE, READ, UPDATE was executed, and a complete database created for any subsequent analysis.

INTRODUCTION

Our project consists of making a database from the following public website <https://hidrocarburos.energianeuquen.gob.ar/portalgis/web/> This website is used as a vehicle of information between the local government of the Neuquen Province in Argentina and the public. Entities working with these data are oil companies, market analysts, economists and others working in the energy sector. The data covers basic aspects of energy activities in the Neuquen Province. The information is provided as Geodatabases which contain various files. The files that we are using are database files which can be open by Excel. All the information is in Spanish, but some column labels have been translated to English to facilitate their identification. Our objective is to build a database that contains information about the hydrocarbon areas and the wells drilled in the Neuquen Province. Information from these tables should be translated into a database that can be queried retrieving information from either table or the two tables combined.

The information is originally contained in two separate integrated tables (Areas and Wells) that have several columns. There are multiple entities that have been integrated to produce these tables. We will identify the various entities and create indexes to connect the entities. This process will reduce the size of the original database but also will make it a relational database.

DATA MINING

Two data files - Areas and Pozos (Spanish for Wells) Excel files, were obtained from the Subsecretaria de Energia de Neuquen (which in English is: Office of the Sub-Secretary of Energy of Neuquen Province) <https://hidrocarburos.energianeuquen.gob.ar/portalgis/web/> website. The following procedures were carried out on each dataset:

1. Identified columns and row entries. This involved translating headers, checking for wrong entries, etc.
2. Established possible individual entities and designed a plan to separate entities and used an Entity Relationship Diagram (ERD) to understand respective relationships between factions of the datasets.
3. Checked for null cells and unique identifier columns. Filled null cells with 999s, unknowns and not-reported, or with values consistent with the column’s entities, and deleted redundant or duplicated rows and columns. Used find-and-replace to check for unnecessary symbols and spaces and reconciled them appropriately with the dataset.
4. All the tables have at least one column with
5. Assigned indices with unique identifiers to key columns.
6. Prepared datasets for normalization

DATASET NORMALIZATION

The first normal form produced 3 tables, drawn out by unique identifier by well names column and block information. The second normal form generated additional 3 tables by operator companies, well profiles, well types. The third normal form created additional 1 table for the fourth normal form rearranged the original datasets to make two more datasets for a total of 9.

DATA TRANSFORMATION

The datasets were transformed from excel workbooks to comma separated values (CSV) files. This is one of 3 formats that can be uploaded into PostgreSQL on pgAdmin.

TABLE GENERATION

Nine distinct tables were created using SQL with PostgreSQL with PgAdmin. By means of unique column identifiers in … tables, foreign keys were designed into parent tables to reference the key columns in mother tables. By this means referential integrity was established to create a functional database. This schema was saved in sql file formats.

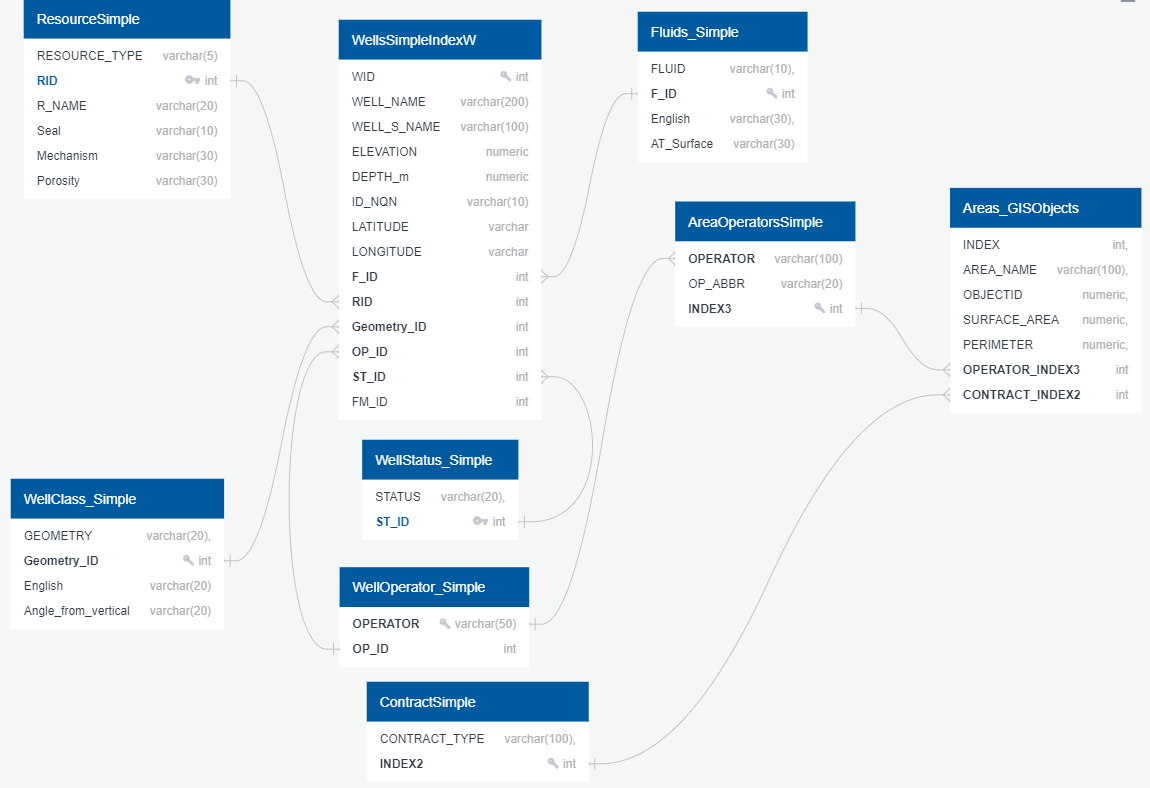
The order of loading data into the tables by position is as follows:

|  |  |
| --- | --- |
| Order of table data loading | Respective Tables |
| 1 | Fluids\_Simple |
| 2 | ResourceSimple |
| 3 | WellClass\_Simple |
| 4 | ContractSimple |
| 5 | WellOperator\_Simple |
| 6 | WellStatus\_Simple |
| 7 | AreaOperatorsSimple |
| 8 | WellsSimpleIndexW |
| 9 | Areas\_GISObjects |

*Figure 1. Table showing the order of loading in data into the tables.*

ERD GENERATION

Using the free Entity Relationship Diagram (ERD) on <https://app.quickdatabasediagrams.com/#/d/LsC3Ko> an ERD was generated for this Team6\_ETL\_Proj2 database structure. The ERD can be seen in figure 1.



*Figure 1. Entity Relationship Diagram for the Team6\_ETL\_Proj2 database*

DATA LOADING

Each csv file was uploaded into a corresponding table with matching table names to file names. This also ensured that the data structure in each csv file, matched the code in the SQL schema. An initial query of each table was done to determine data loading capability. It was successful.

CONCLUSION

Data mining, cleaning, normalizing, relationship identification is necessary for preparing data for analysis in order to obtain key insights from the data. This project afforded Team 6 the capacity to source a functional data set which had to be cleaned, normalized, and prepared for first level analysis. From this analysis, it was obvious that a relational database could be created or generated hence the choice was made to use SQL on PostgreSQL. Referential integrity was established using an ERD between entities as well as primary keys and foreign keys. This project thus produced an incredible knowledge enhancement opportunity which further inculcated the concepts learned in class regarding data modeling and data engineering.