Felipe Gomez

04/23/2024

BIDD 330A

Module 03

GitHub BIDD 330\_Spring2024 Link: <https://github.com/Phillips094/BIDD330_Spring2024>

SQL Server Database Import/Export & Jupyter Notebook SQL Exercises Using Python

Introduction:

For module 03, we focus on developing a data warehouse from raw csv files into our localhost instance on SQL Server. The two files we use to import into our LocalHost instance for SQL Server are “bing\_covid-19\_data.csv” and “Unemployment.csv”. We utilize stored procedures to clean up our data and develop our data warehouse, and once this is developed, we focus on exporting our database (after dropping our stored procedures) into the UW Azure SQL Server database. Once our data warehouse has been imported successfully into our UW Azure SQL Server database, we create a Jupyter notebook to write Python code to run SQL scripts on our dimensions and fact tables.

Starting off our module 03 project, we download our necessary files from Microsoft’s online data source and from our Canvas site. We first create a database that is tailored to our Final project group color and our name, i.e. my database name is “Gold\_Felipe”. After creating our database we use the SQL Server Import and export Wizard to ingest our two csv files into our database. Note that we rename these two source tables by prefixing them with “Raw\_” so we recognize that these are the raw files that we use to create our data warehouse. We save a backup of this database in our C drive and then move on to importing our new database into our UW server.

During this process, it is important to note that we use “Export Data-tier Application” and “Import Data-tier Application” as our tools for importing and exporting our database. When we import our database to the UW Server, we now have our ability to create our Dimension and Fact tables. Our final data warehouse contains DimDates, DimCountry, DimStates, FactCovid and FactUnemployment. We develop stored procedures to complete this. Before moving on to the next part of the project, we must also test our data warehouse to ensure that our stored procedures worked as intended. We must perform a data quality check to make sure our data warehouse is filled up with the correct data and correct number of rows.

An example of one of our stored procedures is as below:

A screen shot of a computer program

Description automatically generated

After we have finished creating our data warehouse, we are now able to Drop our stored procedures so that we can perform another Export Data-tier Application for submitting this assignment.

A screenshot of a computer

Description automatically generated

We then move on to our second part of our assignment. We turn our attention to writing Python in a Jupyter notebook where we use python to write code to connect to our UW server. We use python to connect to our Gold\_Felipe database so that we can begin querying our data warehouse and run queries. We import our libraries that we need to utilize to write functions that will allow us to query our fact and dimension tables. We also must ensure we have our correct credentials and make sure that our conn variable in our 2nd cell runs successfully. After establishing this connection, we perform some simple Top 10 SELECT queries to make sure that our code is running smoothly and our connection works.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Our last query that we test is meant to be an interesting query where we come up with a complex query that will be interesting to analyze. Here we decide to analyze the maximum number of accumulated covid deaths in Germany by month over time from 2020 to 2023. Essentially I want to see the accumulation of covid deaths on a monthly bases over time from descending order so we see the highest number of deaths in Germany.

A screenshot of a computer program

Description automatically generated

Summary:

In summary, we demonstrated that we can quickly develop a data warehouse from scratch by initially utilizing our LocalHost instance so we can then migrate over to our cloud server and begin developing our data warehouse. It seems that this method is very powerful and achieves quick, reliable results. I personally find this method of developing a data warehouse fascinating because we do not have to go through hoops and spend a lot of time trying to architect a solution from beginning to end that encompasses an ETL solution like SSIS or complex SQL code in our final product. In the end we are able to simply utilize SQL and Python code to analyze our data.