Eric Dee

Module 3 Introduction to ETL

10/1/2022

***1, 2. Initial data***

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

***4.***

First data pull

Diagram, schematic

Description automatically generated

Data extraction process

Diagram

Description automatically generated

Documentation

This database includes the schemas named Assignment, and Extract. The assignment schema is the defacto schema for current data, and the extract schema is where to load data needed from a CSV file. The only stored procedure that exists is called ExtractEmployeePay which is the procedure used to pull data from the extraction table into the defacto table. The columns on the 2012StateEmployeePay table are: Calendar\_Year, Agency\_Name, Position\_Title, Employee\_Name, YTD\_Gross\_Pay, and are in this order. These conventions must be utilized in order to load valid data into this database.

***5. Stored procedure extraction on Extract schema***

Proof that data is not being replicated….truncate first

Graphical user interface, text, application

Description automatically generated

No truncation/no rows affected

Graphical user interface, text, application

Description automatically generated

Final procedure

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, website

Description automatically generated

***Transform***

Transformation is intended to generate alternations of extractable data in the way that you might abstract classes away from each other into models, or other tables. This allows more user control for employees who don’t have time to either separate their concerns into different excel files, or teams that choose to just use one large file with lots of data.

***Load***

Load takes the definitions that the team would define for the transform process, and adds any new data to the database.

***6.***

If I needed to test these outputs for validity on no duplicates, I would expect either an error to be thrown due to unique columns, or the query to return 0 rows affected on test runs. I have found that unique identifiers are not necessarily needed in all cases, but I would probably use them in 100% of my databases if they were to be built for teams, as this would ensure that the database is strongly typed, and would reduce the potential for “hacky” changes, or accidental insertions due to lack in knowledge of SQL, or badly formatted extraction files.