ITEC-649 Web Technologies Assignment 2

Name: L.Dhanya Himaja

Student-ID: 45511322

In the second assignment of web technologies, we have implemented the ETL (Extract-Transform-Load) process by writing python code to extract information about jobs, people and companies from CSV files and populating them into SQL database. So far we have dealt with smaller datasets which can fit into computer memory. But challenges arise when handling large datasets which could create mistakes while implementing ETL. Not just the memory, we would require compatible driver to copy data from multiple files and extract the required data efficiently and accurately for the parse of data .During transformation of large data, curser's fetch all method takes quite a long time to iterate through a result set. In real time businesses, where there is a need for a candidate to load a large CSV files into memory in order to work with it ,spending large amount of time writing functions and ending up with dictionaries could slow down querying and it also gets difficult for transformation of data. Also, certain problems arise while importing large CSV files into SQL as no distinction exists between NULL and quotes. Few other constraints on SQL queries for large data to yield aggregate results include limitation on field names, manipulating text strings when required, sorting and replacing the data in the tables, changing the column names etc. Locking and concurrency issue arises while dealing with large chunks of data in enterprises. However, these are not only the constraints while working with large data .Certain modules and libraries should be imported to support the process of extracting and transforming data from the CSV files. When working with such bulk datasets, it's often useful to first map the data using particular attribute, filter or grouping and then reduce that using transformation mechanism. This process is called Map Reduction. Some of the popular libraries for Map Reduce with large datasets include Hadoop, Spark, pig and hive. Also, the ETL framework extracts the required information and runs the transformation to serve the business needs .however, there lays capacity limitations and a long waiting, since information access is not available until the entire ETL process has been completed. To overcome such drawbacks, ELT approach is implemented, wherein after extracting the data, we move the entire data into single, centralized data repository .This supports transformation of large data and enhances the data scalability. Another approach of dealing with large data of CSV file include using pandas sql tools to pull data from the database, as it does not hold any memory constraints.

Acknowledgements:http://www.dbta.com/Columns/DBA-Corner/What-Can-You-Do-to-Avoid-Database-Locking-Problems-100926.aspx

http://pythondata.com/working-large-csv-files-python/

https://blog.panoply.io/etl-vs-elt-the-difference-is-in-the-how