Raymond Yang

15 August 2022

Python Assignment

1. See PeopleDataCleaner.py
2. See JSON_splitter.py
3. Platform as a Service (PaaS), Software as a Service (SaaS), and Infrastructure as a Service (IaaS) refer to different models of cloud services. Software as a Service is probably the most common model, which everyone uses on a daily basis. It uses the internet to deliver software services to customers, with google drive and gmail being two common examples. SaaS tends to be managed from a central location, hosted on a remote server and accessed over the internet. As a result, users are not responsible for updating hardware or software with regards to accessing the service. On the flip side, since the hardware and software are controlled by the vendor, the user has little control with regards to the application, including customizations and when updates are rolled out. There is also no guarantee that the applicaiton will integrate well with other software you may be using, especially if the other software is from a different vendor. PaaS generalizes from SaaS and provides a platform where developers can create their own cloud based software. It gives the user more control. In this model, the hardware is managed by the vendor, but the maintenance of the software is controlled by the developer or user. Windows Azure and Amazon Web Services both have PaaS services. One benefit for PaaS is that the hardware is virtualized, so resources can be allocated and scaled easily as demand changes. IaaS gives even more control to the user, allowing them to control the infrastructure, including servers, network, OS, and storage. AWS and Microsoft Azure both also have IaaS services. The hardware is still maintained by the vendor, but the storage and servers are controlled by the user, acting like a virtual data center. This is the most flexible cloud computing model, allowing for great scalability.
4. ETL and ELT are methods for integrating data into a system. The letters are the same between them, with E for Extract, T for Transform, and L for Load, with the order determining the different between the methodologies. Extract is the process of pulling the data of interest from a source. Transform is the process of formatting the data to match that of the target system. Load is the process of adding the data to the target system. The big difference between ETL and ELT is when the data transformation occurs. ETL transforms the data in a staging area before loading. Practically, this means that it can take a while for the data to be transformed, especially for vast amounts of data, but once the data is loaded, analyses of the data should be much quicker. On the flip side, ELT transforms the data after it is loaded. Practically, this means that the data is transformed as it is called upon for analysis. Since the transform is only performed on the data as it is called, the transform does not need to be performed on uncalled data, saving that processing time. More details are provided in the charts below.

ETL	
Pros	Cons
Fast Analysis	Slow Loading to target system
Allows compliance with security and privacy	Rigid Data Structure
Good for data predetermined to be of importance	Not good for large volumes of data

ELT	
Pros	Cons
Much faster loading to target system	Slow Analysis
Any Data structure can be accepted (raw data)	Since Raw data is stored, may not comply with security and privacy
Good for large volumes of Data	Newer methodology, less tools around it