Antra Assignment 2

Darshan Pradhan

1. Create a script that will read and parse the given files and remove duplicates using python, then write back into a single CSV
    - When two rows are duplicates, they have the same information but might have different separators/casing. For example
        - “1234567890” instead of “123-456-7890”
        - “JANE” instead of “Jane”
        - “ Tom” instead of “Tom”
        - …
    - Once you clean up the anomalies, two rows that are supposed to be duplicates should have the exact same information/format.
2. Split movie.json into 8 smaller JSON files.

3. A paragraph on what PaaS, SaaS and IaaS are and the differences between them.
    - https://www.bigcommerce.com/blog/saas-vs-paas-vs-iaas/#the-key-differences-between-on-premise-saas-paas-iaas
    - https://www.red74tech.com/advice/iaas-saas-paas-naas-different-flavors-of-cloud-services/
    - https://apprenda.com/library/paas/iaas-paas-saas-explained-compared/

Ans) **Infrastructure as a Service or IaaS** are self service models that provides virtualized computing resources over the inter net. It allows organizations to have an unlimited storage potential of the cloud. They are used for accessing, monitoring, and managing remote datacenter infrastructures, such as compute, storage, networking, and networking services (e.g. firewalls). IaaS platforms are accessible by multiple users, cost-effective, highly flexible and highly scalable. IaaS users are responsible for managing applications, data, runtime, middleware, and OSes.

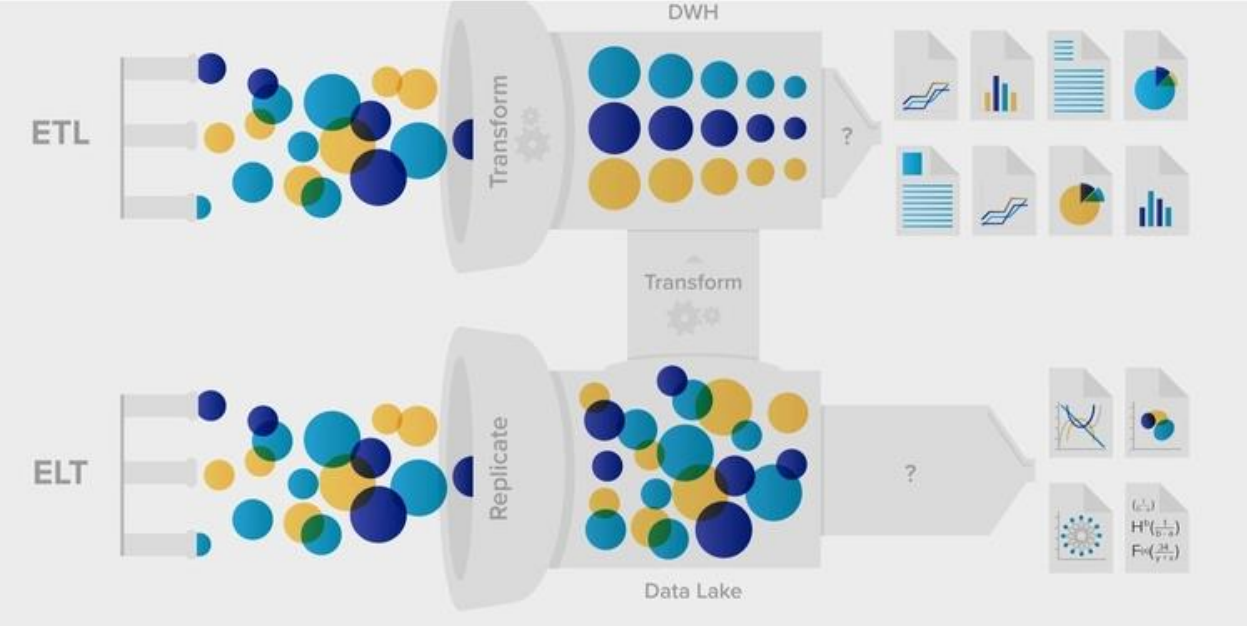
Platform as a Service or PaaS provides developers with a framework, software and tools needed to build apps and software which are all accessible through the internet. A client runs their own copies of the application, using the cloud provider’s infrastructure. Clients don’t have to worry about storage spaces and maintaining servers or hard disks. PaaS platforms are built on virtualization technology, scalable, accessible by multiple users, and easy to run without extensive system administration knowledge. PaaS clients are responsible for managing applications, data.

Software as a Service or SaaS is the most commonly used service. It allows users to benefit from the functionality of particular software without having to worry about storage or other issues. The software can easily be accessed by multiple users using just a web browser. SaaS platforms are available over the Internet, and are hosted on a remote server

by a third-party provider. They are ideal for small businesses or startups who cannot develop their own software applications due to their scalability and inclusivity, offering security, compliance and maintenance as part of the cost. SaaS users are only responsible for managing data.

4. A paragraph on the differences between ETL and ELT. Also, list the pros and cons of each in a chart. And specify when you'll use which.
    - https://www.xplenty.com/blog/etl-vs-elt/

Ans)



Extract, transform, and load (ETL) is the process of combining data from multiple sources into a large, central repository called a data warehouse. ETL uses a set of business rules to clean and organize raw data and prepare it for storage, data analytics, and machine learning (ML). You can address specific business intelligence needs through data analytics (such as predicting the outcome of business decisions, generating reports and dashboards, reducing operational inefficiency, and more).

Extract, load, and transform (ELT) is an extension of extract, transform, and load (ETL) that reverses the order of operations. You can load data directly into the target system before processing it. The intermediate staging area is not required because the target data warehouse has data mapping capabilities within it. ELT has become more popular with the adoption of cloud infrastructure, which gives target databases the processing power they need for transformations.

ETL	ELT
It is the Extract, Transform, and Load process for data	It is the Extract, Load, and Transform process for data
ETL only transforms and loads the data that you decide is necessary when creating the data warehouse and ETL process.	ELT can load all data immediately, and users can determine later which data to transform and analyze.
ETL is not normally a solution for data lakes. It transforms data for integration with a structured relational data warehouse system.	ELT offers a pipeline for data lakes to ingest unstructured data. Then it transforms the data on an as-needed basis for analysis.
Transformations happen within a staging area outside the data warehouse.	Transformations happen inside the data system itself, and no staging area is required.
Calculations will either replace existing columns, or you can append the dataset to push the calculation result to the target data system.	ELT adds calculated columns directly to the existing dataset.

5. **(OPTIONAL)**Create a python script that will calculate/display:
   - Names, types and sizes of blobs in a certain container

```
**************************************************************************************************************************
*  bmw-918408_1920.jpg                                                                                  BlockBlob  338362    *
*  movieshoper.png                                                                                      BlockBlob  118369    *
*  other/2020 Oct.txt                                                                                   BlockBlob  405       *
*  other/Feedback Jan 2021.txt                                                                          BlockBlob  51        *
*  other/Zhaohe Song.docx                                                                               BlockBlob  14911     *
*  smallimages/1Pager-Use-Cases.pptx                                                                    BlockBlob  49298     *
*  smallimages/getting-started-with-asynchronous-programming-in-c-using-async-and-await-slides.pdf      BlockBlob  1243673   *
*  smallimages/my-resume-5 (1).pdf                                                                      BlockBlob  46014     *
*  smallimages/other/async-and-await-advanced-topics-and-best-practices-slides.pdf                      BlockBlob  727800    *
*  smallimages/other/asynchronous-programming-deep-dive-slides.pdf                                      BlockBlob  706315    *
*  smallimages/other/test/MichaelYi-Resume.pdf                                                          BlockBlob  46014     *
*  smallimages/other/test/Nai_Chen_Chi_CV_2021.pdf                                                      BlockBlob  235702    *
*  woman-1749355_640.jpg                                                                                BlockBlob  55464     *
**************************************************************************************************************************
```

   - Names and sizes of “folders” in a certain container

```
{'other': 1731198, 'smallimages': 3054816, 'smallimages/other': 1715831, 'smallimages/other/test': 281716, 'root': 3582378}
```