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/

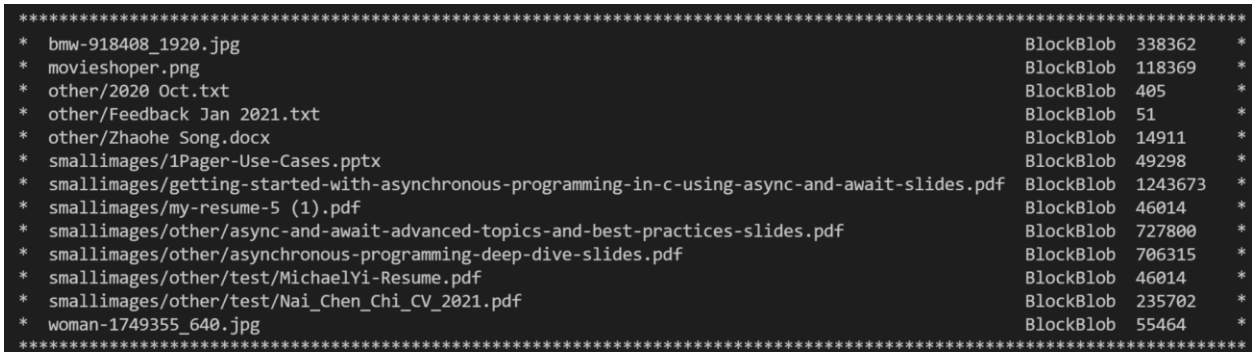
1. PaaS
    - aka platform as a service
    - provides cloud components to certain software while being used mainly for applications
    - focuses primarily on hardware and software tools available over the internet
2. SaaS
    - aka software as a service
    - involve software that is available via third-party over the Internet
    - typically ready-to-use and run from a users’ web browser, which allows businesses to skip any additional downloads or application installations
3. IaaS
    - aka infrastructure as service
    - works primarily with cloud-based and pay-as-you-go services such as storage, networking and virtualization
    - typically offers over the internet

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

    ETL is extract, transform, and load; but ELT is extract, load, and transform. Since ETL transform the data first, it allows the data analysis to be faster and more efficient. On the other hand, since ELT does not transform the data first, it is not a good choice if you need to analyze the data with fast speed. ETL is more secure than ELT because it transforms the data before storing them into the data warehouse. Also, since ETL existed earlier, it is more mature in many aspects compare to ETL. As for ELT, it is faster

in data availability, easier on maintenance, and faster when loading the data into its location.

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}
```

    - connection_string = "DefaultEndpointsProtocol=https;AccountName=antrablobstorage;AccountKey=ECVP9sDWl64Ubd6w3lGd4d4fbiZuwHWWu1q/KoS2sCR18mwwkSxf1gLC7PvqCT1jWi3IYE87ZQtJYMIztIg3vg==;EndpointSuffix=core.windows.net"
    - container_name = "imagescontainer"