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.

**SaaS: Software as a Service**

By the definition, software as a service is referred as getting all the benefit of a particular software without having it install locally on a users machine. The provider takes care about hosting, managing and installing and the user just pays for the service they use typically through the web browser.

**PaaS: Platform as a Service**

It is one of the cloud computing model where the platform provider or a vendor provided all the hardware and software as required to their user typically to build and deploy the web applications without need to worry about underline infrastructure. Examples include Amazon AWS and Microsoft Azure.

**IaaS: Infrastructure as a Service**

Infrastructure as a Service on the other hand is one of the cloud computing services where the vendor provides all the required infrastructure typically would be the physical hardware such as datacenters, storage devices and networks to their users. In this model users are required to install their own software and application.

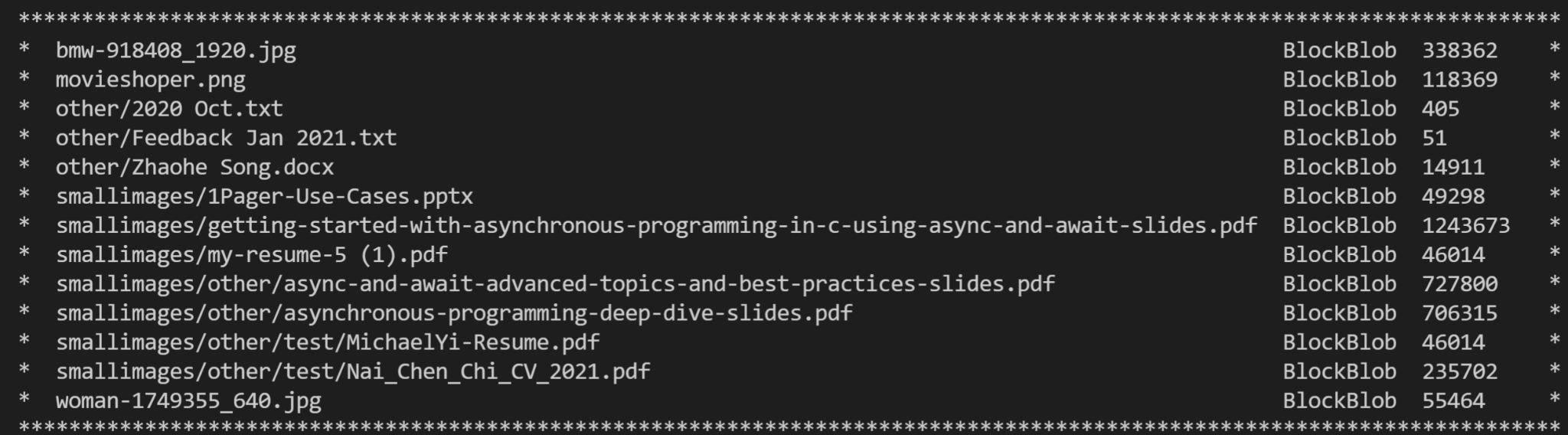
The key differences between all these three platforms are that in SaaS the vendor manages everything including Application, Data, Runtime, Middleware, OS, Virtual Machines, Storage and Networks. In PaaS, except the Application and Data, everything is managed by the vendors whereas in IaaS, typically all the hardware are managed by the vendor such as servers, storages and networks.

1. A paragraph on the differences between ETL and ELT. Also, list the pros and cons of each in a chart.

ETL and ELT are both data pipelining process typically used by data engineers in which ELT stands for Extract, Transport and Load whereas ELT stands for Extract, Load and Transfer. Either of these two process will have a source where the data has to be taken from and a destination where the data has to be loaded. The key major difference between them is the process of transforming the data before loading into the system.

Pros and Cons of ELT and ETL

|  |  |
| --- | --- |
| ETL | |
| **Pros** | **Cons** |
| Appropriate for Data Warehouse | Not appropriate for Data Lake |
| Ideal for processing smaller and relational data for analysis process | The traditional on-premise ETL process requires expensive hardware to process |
|  | Expensive to implement |

1. **(OPTIONAL)**Create a python script that will calculate/display:
   * Names, types and sizes of blobs in a certain container
   * Names and sizes of “folders” in a certain container
   * connection\_string = "DefaultEndpointsProtocol=https;AccountName=antrablobstorage;AccountKey=ECVP9sDWl64Ubd6w3lGd4d4fbiZuwHWWu1q/KoS2sCR18mwwkSxf1gLC7PvqCT1jWi3IYE87ZQtJYMIztIg3vg==;EndpointSuffix=core.windows.net"
   * container\_name = "imagescontainer"