This information refers to the Udacity course "Intro to Hadoop and MapReduce".

# 1 Python

## 1.1 Mapper and Reducer Code

The mapper and reducer code discussed in the Udacity course requires some modifications to work with Python 3. An adjusted version of the code is provided via StudIP.

Note that the provided `mapper.py` and `reducer.py` are incomplete. Work through the Udacity course to finalize their implementation.

## 1.2 Making the Python files executable

To run the mapper and reducer as shown in the Udacity course, you need to make `mapper.py` and `reducer.py` executable:

```
minihive@container$ chmod +x mapper.py reducer.py
```

# 2 Input Data

The input file, `purchases.txt`, used in the Udacity course is available on GitHub. Execute the following commands in the miniHive container to clone the repository and unzip the file:

```
minihive@container$ git clone https://github.com/sdbs-uni-p/sds-artifacts
minihive@container$ cd sds-artifacts
minihive@container$ tar xzf purchases.tgz
```

For testing, we recommand to use only a subset of the data (e.g., the first 100 lines):

```
minihive@container$ head -n 100 purchases.txt > testdata.txt
```

# 3 Hadoop

## 3.1 Preparation

Before loading your data into the Hadoop Distributed File System, you may need to create the home directory (or any other path you use):

```
minihive@container$ hadoop fs -mkdir -p .
```

## 3.2 Streaming

The Udacity course introduces the `hs` command as an alias for the command-line call

```
hadoop jar /usr/lib/hadoop-0.20-mapreduce/contrib/streaming/hadoop-streaming
   -2.60-mr1-cdh4.1.1.jar
```

However, the `hs` command is not available in the miniHive Docker container. Instead, you can use the following command, which you also find in the Hadoop documentation[1]:

```
minihive@container$ mapred streaming -mapper mapper.py -reducer reducer.py \
    -file <relative-path-to-mapper.py> -file <relative-path-to-reducer.py> \
    -input <hdfs-path-to-source> -output <hdfs-path-to-output>
```

Hints:

- Your *Python* files need to be executable.
- `<hdfs-path-to-output>` must not yet exist in the Hadoop Distributed File System.

## 3.3 Job information output

Hadoop provides a web-based job information panel, which is required for Udacity's *Structural Patterns* exercise. The output seen there is not available on miniHive Docker even if you enable the needed port forwarding. You can skip this exercise but think about the Combiner problem nevertheless.

[1] `https://hadoop.apache.org/docs/r3.2.2/hadoop-streaming/HadoopStreaming.html`