Data Cleaning And exploration

Aviation Dataset Mount & Initialization

To be able to work with data, first we have to mount the transportation dataset to our EC2 instance. I used the `lsblk` command to see available block storages mounted to my EC2. Then created a separate folder and mounted the volume to the filesystem.

```
$ lsblk
$ sudo mkdir /data
$ mount /dev/xvdb /data
```

References

- AWS User Guide: Attaching EBS volumes
- AWS User Guide: Using EBS volumes

Data exploration using `bash`

At a certain level bash is perfectly fine to discover what's included in the transportation dataset. At first with directory navigation we can investigate each folder and see what's inside. Obviously we're just interested in the `aviation` subfolder.

```
$ ls
air_carrier_employees          air_carrier_statistics_summary  airline_origin_destination    aviation_safety_reporting
...
$ cd airline_ontime
$ ls
1988  1989  1990  1991  1992  1993  1994  1995  1996  1997  1998
...
$ cd 2008
$ ls
On_Time_On_Time_Performance_2008_10.zip  On_Time_On_Time_Performance_2008_2.zip  On_Time_On_Time_Performance_2008_6.zip
...
```

We can peek inside each zip file using `bash` as well, ff we pipe the output of each file to the `gunzip` command. Directing the output to an other file allows us to save samples and test our map-reduce jobs on small portion of data.

```
cat ./On_Time_On_Time_Performance_2008_10.zip | gunzip | head -255 > ~/airline_ontime_perf.csv
```

Moving relevant data to Hadoop HDFS

We'll just work with the `airline_ontime` data, which contains on-time performance for each flight. A special `bash` script is getting all the zip archives in all subfolders, searches inside each zip file for CSV extensions and unzips only those files from the zip archive. We'll pipe each CSV output to a `hdfs put` command.

```
migration/move-ontime-perf-to-hadoop.sh /data/aviation /user/ec2-user/ontime_perf
```

References

- Migration scripts on GitHub

System Integration