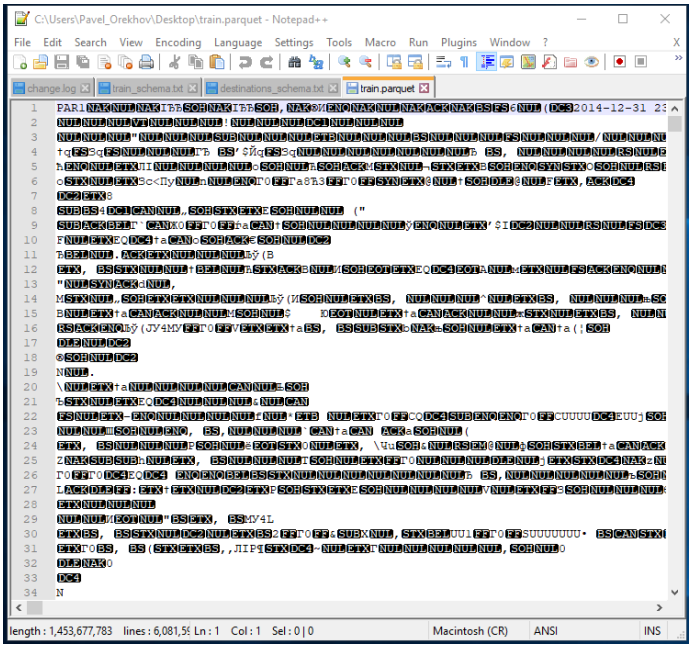
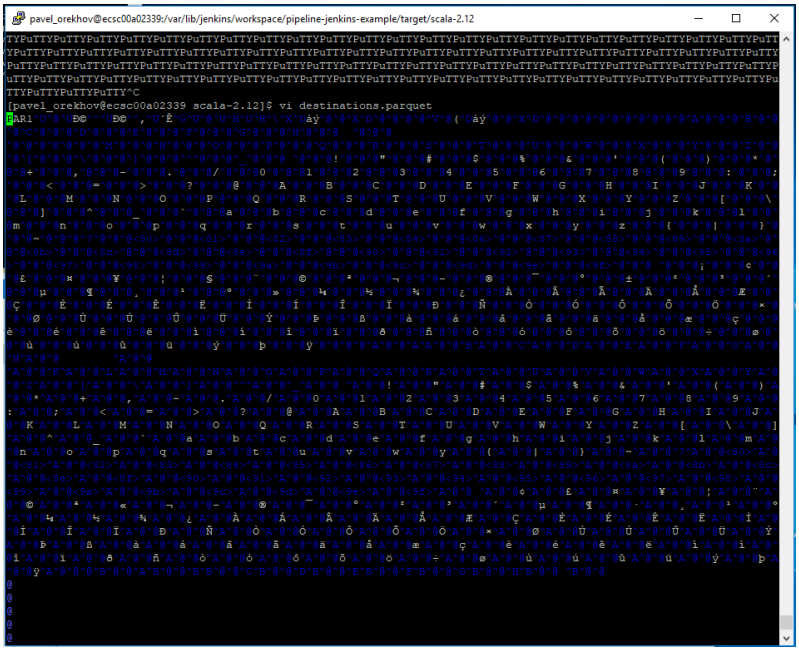
HDFS Homework.

Task 1.

Here's my screenshot of the converted parquet file:



And here's the same thing on one of the clusters' nodes:



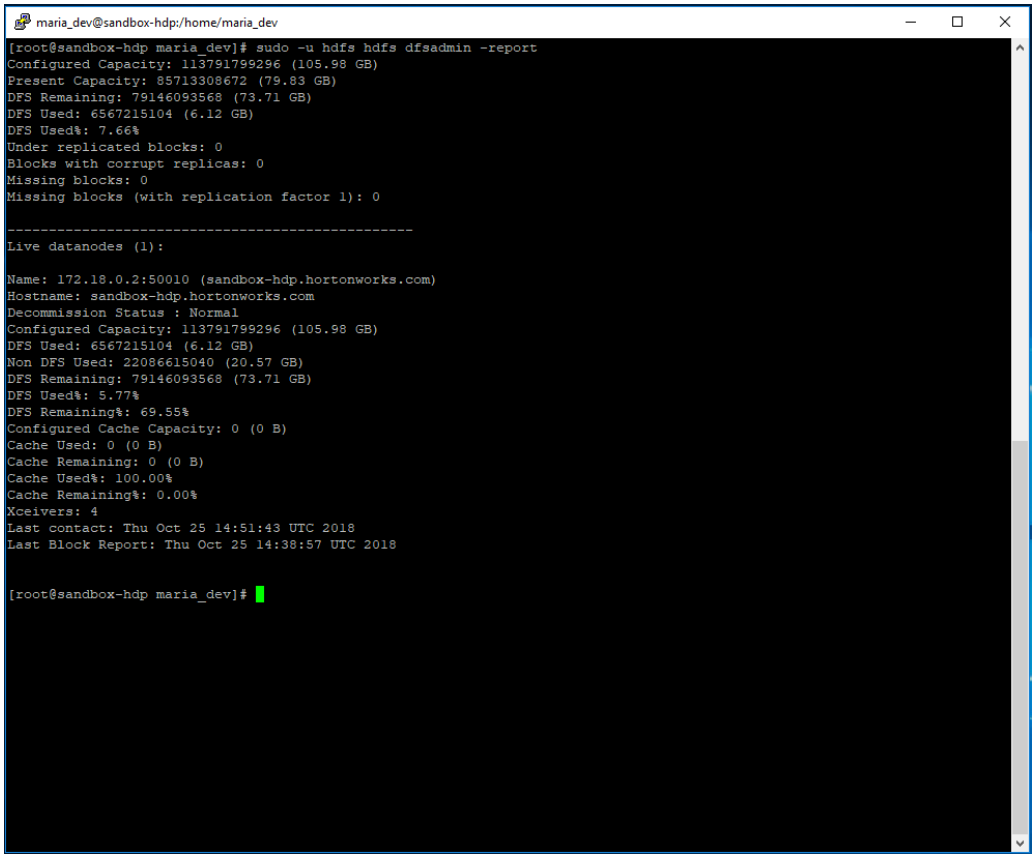
To run my jar, you need to issue the following command:

java -jar program.jar hdfs -schemaPath=schema/path -csvPath=csv/path -newFilePath=file/path -csvSeparator=,

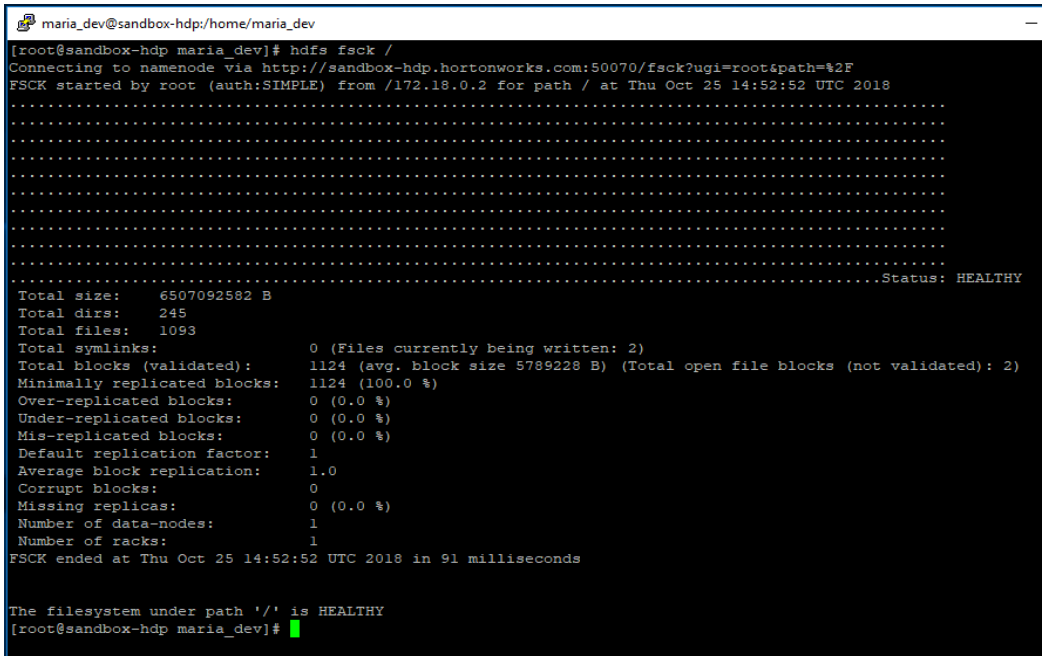
Task 2.

HDFS' health can be checked in these 2 ways:

- sudo -u hdfs hdfs dfsadmin -report



- hdfs fsck /



I will describe what the output of the first command shows.

- Configured Capacity – is the capacity that is available to HDFS for storage. It's calculated as follows: **Configured Capacity = Total Disk Space – Reserved Space**. Reserved Space is allocated for OS level operations. The parameter responsible for Reserved Space is **dfs.datanode.du.reserved**, it can be changed in **hdfs-site.xml**.
- Present Capacity – is the total amount of storage space which is actually available for storing the files after allocating some space for metadata and open-blocks (Non DFS Used space). So, the difference: **Configured Capacity – Present Capacity** is used for storing metadata and other information.
- DFS Used – is the storage space that has been used up by the HDFS. In order to get the actual size of the files stored in HDFS, you need to divide DFS Used by the replication factor. Replication factor is defined as **dfs.replication** and can be found in **hdfs-site.xml**. So, if DFS Used is 90 GB, and replication factor is 3, then the size of the actual files is 90/3=30GB.
- DFS Remaining – is the amount of storage available to HDFS to store more files. If you have 90GB of remaining space and your replication factor is 3, then you can store 90/3=30GB of files.
  Present Capacity = DFS Used + DFS Remaining.
- Non DFS Used is any data in the filesystem of the data nodes that isn't in dfs.datanode.data.dir.
  Non DFS Used = Configured Capacity – Present Capacity.
- Under-replicated blocks – are blocks whose number of replicas does not meet the target replication number. HDFS Will automatically create new replicas for under-replicated blocks until they meet the target replication. You can view the information about it by issuing the hdfs dfsadmin -metasave command.
- Blocks with corrupt replicas – are blocks who have at least one corrupt replica and at least one live replica.
- Missing blocks – blocks that have missing replicas.
- Missing blocks (with replication factor 1) – blocks that are irreparable.
- Xceivers – defines the number of server-side threads (that can be used by sockets for data connections).

And finally, this is how to move or rename a file in hdfs:

hdfs dfs -mv /path/to/file/name /new/path/to/file/new_name

Example: