**Pre-requisites:**

1. Download the virtual machine disk from <https://epam-my.sharepoint.com/personal/alexey_girin_epam_com/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Falexey%5Fgirin%5Fepam%5Fcom%2FDocuments%2FBig%20Data%20QA%20stack%2FSPb%20Big%20Data%20Course%2F%5FCDH%5Fimage>
2. Import downloaded VMDK file either to VMware Workstation (<https://www.vmware.com/products/workstation-player/workstation-player-evaluation.html> or VirtualBox (<https://www.virtualbox.org/wiki/Downloads>)
3. Launch the imported Virtual Machine.  
   All cluster resources are located at [10.0.2.15](http://10.0.2.15).  
   Login and password for Cloudera services are cloudera / cloudera
4. Download airlines dataset from <https://openflights.org/data.html> and move it to the virtual machine's local file system.
5. In case you have docker installed there might be a problem with launching a virtual machine.

You need to

* Switch docker off, if it doesn’t help uninstall it.
* Turn off hyper-visor
  + Launcg cmd as administrator
  + Run command: ‘bcedit /set hypervisorlaunchtype off’

**Task 1:**

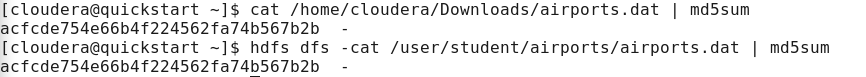
1. Create HDFS directory /user/student and nested directories airports, airlines, routes, planes, countries  
    *hdfs dfs -mkdir -p user/student*

*hdfs dfs -mkdir -p {user/student/airports, user/student/airlines,user/student/routes,user/student/planes,user/student/countries}*

Table

Description automatically generated

1. Copy each file from the local file system's airlines dataset to the corresponding HDFS directory created in step #1 (i.e. routes.dat should be stored HDFS as /user/student/routes/routes.dat)   
     
     
   hdfs dfs -put /home/cloudera/Downloads/airlines.dat /user/student/airlines  
     
   
2. Print out to the console first 9 lines of /user/student/countries/countries.dat file  
     
   Text

   Description automatically generated with medium confidence
3. Compare MD5 checksums between the local file airports.dat and HDFS file /user/student/airports/airports.dat. Is there any difference?  
     
     
     
   **no difference**
4. How much replicas does the /user/student/airports/airports.dat file have? Set number of replicas to 4 for the file  
     
     
     
   
5. Print out the total size of the /user/student dir in megabytes. Why the space consumed by the directory in HDFS doesn't equal to the space consumed by the same directory in local FS?  
   Text

   Description automatically generated  
     
   The directory size is different due to the different block size in File Systems.
6. Run the filesystem checking utility for the /user directory in HDFS. Analyze utility's output in order to answer the following questions:

* Are there any corrupted files and blocks?
* Are there any missed blocks?
* How HDFS is configured? How many Data Nodes and Racks are in place?
* Where the blocks are located? (block path)

Table

Description automatically generated

* **0 corrupted blocks**
* **3 missing replicas?**
* **1 data node, 1 rack**
* **Block path: 10.0.2.15/user**

**Expected output:**

Provide shell scripts/commands, screenshots with results, and report