Hadoop and MapReduce introduction homework

Input data

Apache web-server access log available for download from KB https://kb.epam.com/download/attachments/241634481/homework_for_2nd_session.tar.gz or on shared directory \\EPBYMINSA0000.minsk.epam.com\Training Materials\EPAM Trainings\Hadoop_and_MapReduce_Introduction_by_Kostiantyn_Kudriavtsev (file homework_for_2nd_session.tar.gz)

Task

Write MR job to count average and total bytes by IP (try to use combiner).

MRUnit tests must be added.

The output is **SequenceFile** compressed with **Snappy** on **Block** level with pairs: IP - > <Average, Total>,

Where <Average, Total> is your **custom Writable**.

Input and output must be passed over configuration object, not as command line arguments.

Add custom counters to calculate number of IE, Mozzila or Other browsers were detected (parse it from UserAgent, Mozzila counts only for Mozzila UserAgent, not Firefox)

The job must run **2 reducer** instances.

Task evaluation

Task evaluation contains two steps:

1. Send zipped code and values for 3 browser counters (i.e. IE, Mozzila, Other) to your **assigned reviewer** (your team will get e-mail with his contacts) as soon as possible
2. Build JAR file and deploy on provided Hortonworks cluster (again, details will be provided). Jar file must be located in your home directory. Run on this cluster, input HDFS directory must be /user/<your_user>/hw_himr_1/input and output directory /user/<your_user>/hw_himr_1/output. Later, the online review session will be scheduled and each attendee will show own work on the cluster (for instance, reviewer might ask to show content of sequence file in output directory in readable format)