Домашнее задание №1

Целью домашнего задания №1 является освоение работы с Microsoft HDInsight, технологией MapReduce, компиляцией и постановкой hadoop-задач на счет.

Срок сдачи: 28 октября, 23:59:59. Решения, присланные позже данного срока, не принимаются.

Формулировка задания: Вам необходимо проделать по инструкции, составленной преподавателями: http://bigdata.cs.msu.ru/images/c/c0/Hadoop_Introduction.pdf все шаги по созданию хранилища, созданию кластера, компиляции программы WordCount, ее запуску на кластере и просмотру результата.

После того, как Вы получите вывод статистики по частоте слов, выведите список наиболее часто встречаемых слов (с максимальной частотой встречаемости). Данный список сохраните в текстовом файле в формате «слово пробел частота встречаемости», список должен быть упорядоченным по словам.

Далее Вам необходимо модифицировать исходную программу таким образом, чтобы у Вас выводилась статистика по длинам слов.

Программу по подсчету статистики по длинам слов необходимо запустить как на наборе входных данных, указанном в инструкции, так и на наборе данных, выданном преподавателями: wasb://hometask1@bigdatamsu.blob.core.windows.net/

Далее Вам необходимо на почту курса bigdata@cs.msu.ru отправить архив в формате Task1-Фамилия.rar (фамилия на англ.), состоящий из шести файлов:

1) wordcount.java файл с Вашим кодом по выводу наиболее часто встречаемых слов
2) wordlencount.java файл с Вашим кодом по подсчету статистики длин слов
3) Сформированные wordcount.jar и wordlencount.jar файлы
4) Текстовый файл с наиболее часто встречаемыми словами (для набора данных, указанного в инструкции)
5) Вспомогательные файлы для сборки (если используются)
6) Файл readme.txt с описанием того, как Вы компилировали и запускали программы