**Lab 2 Advanced - Spark with Hadoop**

1. Скачайте Docker container (<https://hub.docker.com/r/uhopper/hadoop-spark/>) или одну из виртуальных машин с Hadoop или

* <http://hortonworks.com/downloads/>
* <http://www.cloudera.com/downloads/quickstart_vms/5-8.html>
* другую VM

1. Скачайте файл с наборов HTTP access logs (http://ita.ee.lbl.gov/html/contrib/Sask-HTTP.html) и запишите их на HDFS из командной строки
2. Используя любой из языков (Scala, Java, Python):
   * Найти чаще всего запрашиваемые ресурсы, исключив картинки и корень сайта
   * Найдите топ N клиентов потребляющих наибольший трафик (скачавших больше всего данных)
   * Сколько данных вернули /cgi-bin/ запросы и прямые обращения
   * Какие были самые популярные запросы и от кого - см. GET /cgi-bin/cusi?query
3. Результаты работы сохраняем в файлы на HDFS

Для выполнения задания можно использовать *spark-shell* или *pyspark*

**Quick Start**

<http://spark.apache.org/docs/latest/quick-start.html>

Video about Spark:

<https://www.youtube.com/watch?v=kZSfIfjFB3A>

<https://www.youtube.com/watch?v=5OWqAjBBDBA>