Лабораторная работа №5

Цель

Произвести обработку логов NASA, используя Apache Spark в Standalone и over Yarn mode.

Задача

Необходимо:

1. Положить логи NASA на HDFS
2. Подготовить список запросов, которые закончились 5xx ошибкой, с количеством неудачных запросов
3. Подготовить временной ряд с количеством запросов по датам для всех используемых комбинаций http методов и return codes. Исключить из результирующего файла комбинации, где количество событий в сумме было меньше 10.
4. Произвести расчет скользящим окном в одну неделю количества запросов закончившихся с кодами 4xx и 5xx

Примечания

1. Все результаты должны быть сохранены на HDFS или в любом SQL / NoSQL хранилище
2. Желательно, чтобы Apache Spark был запущен в рамках docker containers (готовые варианты можно использовать отсюда)
3. Допускается использовать как RDD, так и Dataframes
4. Примеры посложнее вокруг MapReduce