

Скачать кликстрим Википедии en за 2 месяца.

<https://dumps.wikimedia.org/other/clickstream/2018-01/clickstream-enwiki-2018-01.tsv.gz>

За первый месяц взять 10000 самых популярных пар.

За второй месяц выполнить join с самыми популярными парами за первый месяц двумя способами.

1 - через broadcast join - считать первый месяц в каждом маппере и профильтровать второй

2 - через reduce side join считать данные в мапперах, добавив метку источника, сопоставить в редьюсерах

Вывести пары и количества за второй месяц, которые были отфильтрованы за первый месяц