Лабораторная работа 3.

Задача связывания (join) наборов данных по ключу.

Метод — reduce side

1. Входные данные.

Файл **664600583_T_ONTIME_sample.csv** содержит данные о совершенных рейсах авиакомпаний.

Столбцы :
YEAR — год
QUARTER - квартал
MONTH - месяц
DAY_OF_MONTH — день месяца
DAY_OF_WEEK - день недели
FL_DATE - дата перелета
UNIQUE_CARRIER — ID авиакомпании
AIRLINE_ID — ID авиакомпании в классификации US DOT
CARRIER - ID авиакомпании в классификации IATA
TAIL_NUMFL_NUM — ID рейса
ORIGIN_AIRPORT_ID - ID авиапорта
ORIGIN_AIRPORT_SEQ_ID - ID авиапорта в классификации US DOT
ORIGIN_CITY_MARKET_ID - код группы аэропортов относящихся к одному городу
DEST_AIRPORT_ID — ID города авиапорта
DEST_AEROPORT_ID — Идентификатор аэропорта

WHEELS_ON — время приземления (в локальном времени hhmm)

ARR_TIME - время прибытия (в локальном времени hhmm)
ARR_DELAY — разница в минутах между расчетным временем приземления и реальным (может быть отрицательной)
ARR_DELAY_NEW - разница в минутах между расчетным временем приземления и реальным (>=0)
CANCELLED — признак отмены рейса (1 в случае отмены)
CANCELLATION_CODE — код причины отмены
AIR_TIME - время в полете в минутах
DISTANCE - расстояние в милях

Файл ***L_AIRPORT_ID.csv*** содержит список аэропортов

Столбцы :
code — идентификатор аэропорта

description — название аэропорта

2. Задача

Требуется связать наборы данных по коду аэропорта прибытия : DEST_AEROPORT_ID

Для каждого аэропорта требуется определить среднее, минимальное и максимальное время задержки для всех прибывающих рейсов.

3. Подсказки.

а. Разрабатываем Writable для каждого из входных наборов данных, который может читать данные из csv.

б. разрабатываем WritableComparable ключа имеющий два столбца :

AEROPORT_ID,индикатор набора данных (для записей с информацией об аэропорте = 0 , для рейсов =1)

сортировка по умолчанию по двум столбцам

в.разрабатываем map функцию для каждого из набора данных, которая генерирует WritableComparable ключа и Writable данных

для списка аэропортов эта функция в качестве value отправляет имя аэропорта.

для списка рейсов в качестве value эта функция отправляет время задержки (в виде строки)

также надо фильтровать только рейсы с задержкой прибытия.

в. Разрабатываем Partitioner, который учитывает только код аэропорта

г. Разрабатываем GroupingComparatorClass, который учитывает только код аэропорта

д. Разрабатываем reduce функцию, которая берет первую строку, извлекает из нее имя аэропорта, далее рассчитывает из последующих строк среднее минимальное и максимальное время задержки и печатает результат.

В случае если в аэропорт не осуществлялось рейсов, то запись в результат добавлять не надо.