AIM III – Large Scale Data Analysis and Data Mining

Second Assignment

Data and topic for this assignment

In this assignment we deal with bibliographic data about authors and books. The file *src/test/resources/two/authors.tsv* contains author names and ids, the file *src/test/resources/two/books.tsv* contains books, their year of publication and their author id. We will use MapReduce and Hadoop to sort and join this data.

1. Sort the books with "secondary sort"

We want to transform the book data into a list of $(century, title)$-tuples, where century just denotes the first two digits of the year of publication. Each line in the output file should have the format $century < tab > book\,title$.

The output data must be sorted ascending by century and title. You must not sort the data yourself, but must use Hadoop's "secondary sort" capabilities to have the framework do the sorting for you in the shuffle phase.

2. Join books and authors with a "map-side in-memory join"

In this task we will perform an inner join of the books and authors on the author id.

Use a "map-side in-memory join": load the smaller dataset into memory in your mapper class and perform the join before sending the tuples to the reducer over the network.

Each line in the output file must have the format:

$authorname < tab > book\,title < tab > year\,of\,publication$

3. Join books and authors with a "reduce-side join"

We will peform the same join here as in task 2, but we will use another technique called "reduce-side join". The join should be performed in the reducer class, an ideal solution would also avoid buffering more than one value in the reducer.

Deadline

Please send your solution as a patch file to *ssc@apache.org* with *isabel@apache.org* in CC until the 6th of May.