Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Matthias Schubert
Julian Busch, Daniyal Kazempour

Munich, 06.12.2017

Big Data Management and Analytics

WS 2017/18

Tutorial 6: Apache Flink, Stream Analytics

Assignment 6-1 *Stream Processing with Apache Flink - WordCount*

In this assignment we are going to implement the wordcount example using Apache Flinks streaming API. For this purpose please download (from: https://flink.apache.org/downloads.html) and setup Apache Flink. It is recommended to consult the following manual for Flink setup https://ci.apache.org/projects/flink/flink-docs-stable/ and the streaming API documentation https://ci.apache.org/projects/flink/flink-docs-release-1.3/dev/datastream_api.html

(a) Write a wordcount program using the DataStream API.

(b) Write a wordcount program using the DataSet API.

Assignment 6-2 *Matrix-Matrix multiplication with Flink*

(a) Download the code template mm_flink_template.java and become familiar with it.

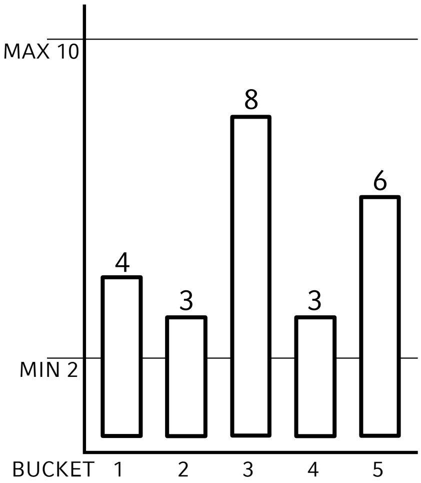
(b) Implement the method *map* in the MapToProduct class and implement the ellipses ... with your code.

(c) Test your implementation by checking the result for multiplying the matrices
$A := \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$ and $B := \begin{pmatrix} 7 & 8 \\ 9 & 10 \\ 11 & 12 \end{pmatrix}$

Assignment 6-3 *K-Buckets*

Given the histogram as seen below, execute the K-Buckets Histogram algorithm for inserts and deletes, assuming the following rules:

- The histogram consists of constantly $k = 5$ buckets.
- The upper threshold (MAX) per bucket is 10, the lower threshold (MIN) is 2.
- For split-and-merge operations: a split occurs when the size of a bucket would otherwise **exceed** MAX; a merge occurs between the two consecutive buckets that were not product of the preceding split with the lowest overall sum of sizes.
- For merge-and-split operations: a merge occurs with the neighbour bucket that has the smallest size, when the size of a bucket would otherwise be below MIN.



INSERTING Insert the items of the given sequence into the histogram, until the first overflow occurs. Execute the resulting split-and-merge and move on to the next section (deleting). Each item is denoted as the index of its respective bucket.

Sequence = 3,1,3,5,2,3,4,1,5,3

DELETING Starting with the resulting histogram of the insert section, remove the items of the given sequence from the histogram, until the first underflow occurs. Execute the resulting merge-and-split. Each item is denoted as the index of its respective bucket.

Sequence = 1,3,4,5,4,3,2,5,1,2

Assignment 6-4 *CUSUM – Change Detection*

Given a mean value $\omega = 3$ and a threshold value $\alpha = 8$, execute the Cumulative Sum algorithm for change detection on the following sequence:

Sequence = 2,3,7,4,0,2,5,6,8,7

n	x$_n$ - ω	G$_n$
0		
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		