Wikipedia author count using Apache Flink

Introduction

The most prominent MapReduce program that corresponds to the classical “Hello World” example in regular programming languages is the so called word count example.

See for example

- http://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Example:_WordCount_v1.0
- https://github.com/apache/flink/blob/master/flink-examples/flink-java-examples/src/main/java/org/apache/flink/examples/java/wordcount/WordCount.java
- https://github.com/apache/flink/blob/master/flink-examples/flink-scala-examples/src/main/scala/org/apache/flink/examples/scala/wordcount/WordCount.scala
- http://stackoverflow.com/questions/15487413/scala-beginners-simplest-way-to-count-words-in-file
- http://hortonworks.com/hadoop-tutorial/word-counting-with-apache-pig/

This task is inspired by that and will already be related to the outlined thesis topic.

Summary

Please write a Flink-JAVA-program that reads a Wikipedia XML Dump file as input and outputs the author name (i.e. mediawiki/page/revision/contributor/username) and edit counts tuples for all pages in namespace 0.

Details

A description of the Wikipedia XML Dump format is available here

https://www.mediawiki.org/wiki/Help:Export#Export_format

Please test your program with small real data dumps (available from here: https://dumps.wikimedia.org/backup-index.html)

and publish it as Github open source project with documentation how to use it.

The output format might be a simple CSV file using like that:

```
User, edits
alfons, 500
howardcohl, 23
physikerwelt, 39
```

The headline is not required. The list must not be sorted

See also

Some links to past and current projects you might find useful:

https://ci.apache.org/projects/flink/flink-docs-release-0.9/quickstart/run_example_quickstart.html

https://github.com/alexeygrigorev/project-mlp

https://github.com/TU-Berlin/mathosphere