Pagerank in Apache Flink

Author: Ward Schodts
Supervisor: Juan Soto

Datenbanksysteme und Informationsmanagement
Technische Universität Berlin

July 8, 2016

Agenda

Introduction

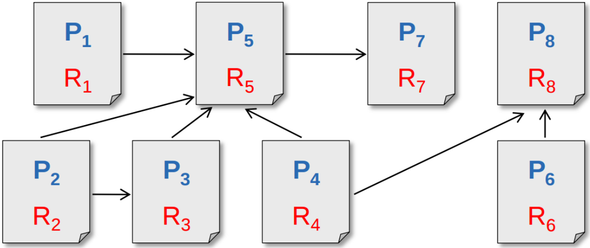
The experiment

The different algorithm implementations

Results

Conclusion

Pagerank



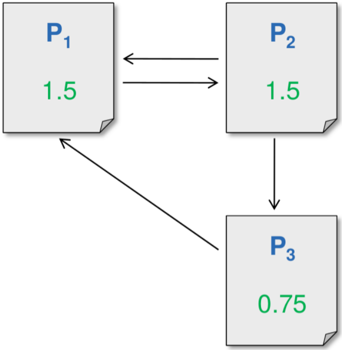
PageRank example 1 [5]

- A page has a high PageRank *R* if
  - there are many pages linking to it
  - or, if there are some pages with a high PageRank linking to it

Pagerank

$$R(P_i) = \sum_{P_j \in B_i} \frac{R(P_j)}{L_j}$$

- where
  - B_i is the set of pages that link to page P_i
  - L_j is the number of outgoing links for page P_j linking to it



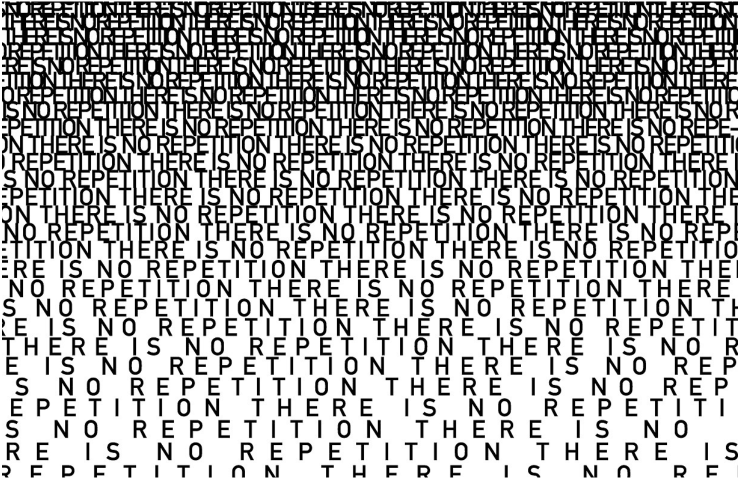
PageRank example 2 [5]

[2]

Apache Flink

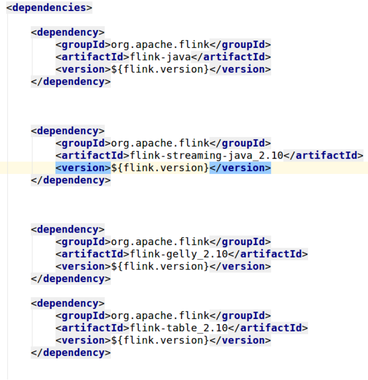
- Open source framework for distributed Big Data Analytics
- Exploits:
  - data streaming
  - in-memory processing
  - iteration operators

  to improve performance
- Formerly Stratosphere (Flink means agile)
- Developped here at TUB



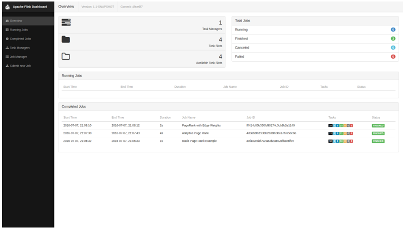
[6]

Apache Flink: 2 possible setups



```
<dependencies>

    <dependency>
        <groupId>org.apache.flink</groupId>
        <artifactId>flink-java</artifactId>
        <version>${flink.version}</version>
    </dependency>


    <dependency>
        <groupId>org.apache.flink</groupId>
        <artifactId>flink-streaming-java_2.10</artifactId>
        <version>${flink.version}</version>
    </dependency>


    <dependency>
        <groupId>org.apache.flink</groupId>
        <artifactId>flink-gelly_2.10</artifactId>
        <version>${flink.version}</version>
    </dependency>

    <dependency>
        <groupId>org.apache.flink</groupId>
        <artifactId>flink-table_2.10</artifactId>
        <version>${flink.version}</version>
    </dependency>
```

Maven



Overview Version: 1.1-SNAPSHOT Commit: 49ce6f7

1 Task Managers

4 Task Slots

4 Available Task Slots

Total Jobs	
Running	0
Finished	3
Canceled	0
Failed	0

Running Jobs

Start Time	End Time	Duration	Job Name	Job ID	Tasks	Status

Completed Jobs

Start Time	End Time	Duration	Job Name	Job ID	Tasks	Status
2016-07-07, 21:08:10	2016-07-07, 21:08:12	2s	PageRank with Edge Weights	ff414c00b530fd90174c3cb8b2e1149	14 0 0 14 0 0 0	FINISHED
2016-07-07, 21:07:38	2016-07-07, 21:07:43	4s	Adaptive Page Rank	4d3ab6f61930b23d6f630ea7f7a50e66	11 0 0 11 0 0 0	FINISHED
2016-07-07, 21:06:32	2016-07-07, 21:06:33	1s	Basic Page Rank Example	acf402ed3f702a83b2a692afb3c6ff97	10 0 0 10 0 0 0	FINISHED

Binary version (self compiled)

Agenda

Introduction

The experiment

The different algorithm implementations

Results

Conclusion

General experiment setup

1. Data file with graph and pagerank solution

Flink experiment setup

GraphLab experiment setup

Agenda

Introduction

The experiment

The different algorithm implementations

Results

Conclusion

Flink algorithm 1

Flink algorithm 2

Flink algorithm 3

GraphLab algorithm

Agenda

Introduction

The experiment

The different algorithm implementations

Results

Conclusion

Results

Agenda

Introduction

The experiment

The different algorithm implementations

Results

Conclusion

Conclusion

Thank you for your attention

Questions?

References I

- Slim Baltagi. *Overview of Apache Flink: Next-Gen Big Data Analytics Framework*. 2015. URL: %5Curl%7Bhttp://www.slideshare.net/sbaltagi/overview-of-apacheflinkbyslimbaltagi?qid=5f0b5424-d187-4c79-a600-6cae794c686e&v=&b=&from_search=3%7D.
- Apache Flink. *Apache Flink Squirrel*. URL: %5Curl%7Bhttps://flink.apache.org/img/logo/png/1000/flink_squirrel_1000.png%7D.
- Lawrence Page et al. “The PageRank citation ranking: bringing order to the web.” In: (1999).

- Lawrence Page et al. “The PageRank citation ranking: bringing order to the web.” In: (1999).

References II

- Beat signer. *Google PageRank*. 2009. URL: %5Curl%7Bhttp://www.slideshare.net/signer/google-pagerank-presentation?qid=18af8836-30e7-41cd-9edb-956bd7ca324d&v=&b=&from_search=2%7D.

- Mathias Spahlinger. *There is no repetition*. URL: %5Curl%7Bhttps://www.google.com/search?q=repeat&source=lnms&tbm=isch&sa=X&ved=0ahUKEwi4laH2tuLNAhVnB8AKHTPQCU4Q_AUICCgB&biw=1590&bih=765#tbm=isch&q=no+repetition&imgrc=h1qwLbEEezv8SM:%7D.