C&G 2nd assignment

Jacopo Urbani
<jacopo@cs.vu.nl>
C&G practical – 2nd assignment

• Your task: develop an application that processes large Web data using MapReduce
C&G practical – 2nd assignment

• Hadoop is a popular MapReduce framework designed by Yahoo!
• MapReduce is a programming model designed to process large data
• Very simple model: computation consists of two functions: map and reduce
MapReduce is a programming model designed to process large data.

Very simple: computation consists of two functions: *map* and *reduce*. 
C&G practical – 2nd assignment

- In MapReduce all information must be encoded as set of *key/value* pairs

- Computation:
  - 1) The *map* function transforms the input pairs in intermediate ones
  - 2) The framework groups the pairs (using key)
  - 3) The *reduce* function processes each group and returns output pairs
Example: consider we have in input some text files and we want to count how many times a word occurs

MapReduce program:

```java
map(null, file) {
    for (word in file)
        output(word, 1)
}
reduce(word, set<numbers>) {
    int count = 0;
    for (int value : numbers)
        count += value;
    output(word, count)
}
```
C&G practical – 2nd assignment

- The application you must develop works on Semantic Web data

- What is the Semantic Web (aka Web 3.0)?
  - “Invented” by Tim Berners Lee in the ’90
  - The semantics of the information is machine-accessible
  - Able to answer more complex queries
C&G practical – 2nd assignment

• In the Semantic Web data is encoded using the RDF format

• RDF data := set of (s,p,o) statements

• Example of statement:

<http://www.vu.nl> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/resource/University>

• Data comes from different parts (marked in blue)
C&G practical – 2nd assignment

• Write a MapReduce program that:
  - Read in input RDF statements
  - Count how many times information from different datasets was used in the same statement (report the ones with at least 1000 connections)
  - Identify what was the link that connected these datasets and report the percentage

• Output:
  - <http://www.vu.nl> <http://www.w3.org> #stats
  - <link> #stats per link
C&G practical – 2nd assignment

• Test the performance and scalability of your implementation
• Use the RDF data that I copied on the DAS
• Good analysis and scalable implementation will increase your grade
  - How does it scale?
  - How can you improve the performance?
C&G practical – 2nd assignment

• What to submit before the deadline
  - The code of your program
  - A report (max 2 pages) where you describe your findings
    • Describe your own program and justifies your choices
    • Report measurement of the performance (with 4-8-16-32-etc. reducers)
    • An analysis of the scalability (e.g. look at the min/max/avg time of mappers/reducers)
• If your program works and your analysis is correct, you pass the assignment. If you want a good grade, you must improve it.
C&G practical – deadline

- Deadline 2nd assignment: 6 weeks from now
  - 23 March 00:00 Hawaii timezone
- Submit by Blackboard

- 1st tip: don’t start late!
- 2nd tip: don’t start late!
- Good luck!