Practical Course on Parallel Computing · SoSe 2018

Assignment Sheet 11

Assignment 1 – Scheduling in YARN

a) Which schedulers are implemented in the YARN framework? Use the ResourceManager Web Interface to determine which of them is used in our cluster!

- kinds of schedulers
    - fair scheduler
    - capacity scheduler
- The fair scheduler is used in our cluster.

b) Explain the basic functionality of the FairScheduler!

The FairScheduler is a scheduler that equally diverts the available resources to all application in such a way that in the end, on average over time, all applications have an equal share of the resources. In the basic configuration it only considers the memory consumption, but it could be extended. When extended the CPU consumption will also be considered. A single app will use the complete cluster. Additional apps will be assigned unused (free) resources. The apps are organized in queues. By default there is only one queue for all user. Queues can be limited in Memory and CPU. But the queues are also assigned a minimum share (can be assigned to other if unused). This allows smaller apps to finish fast, while big apps can work in parallel. In the queue the resources are assigned to the apps based on memory consumption, FIFO or multiple resources with Dominant Resource Fairness.

It is possible to order the queues hierarchal to split the cluster in given proportions. Moreover, the number of running apps per user/queue can be limited.

Assignment 2 - Analyzing Twitter data

a) Which Twitter ID has the most followers and how many people follow this Twitter ID?

Twitter user: 428333

Followers:2450749

b) Which Twitter ID follows the most other Twitter IDs?

Twitter user: 813286

Following 10842 users

c) Are there any bidirectional following relationships, that means are there any pairs of users A and B, such that A is a follower of B and B is a follower of A. If there are such pairs, how many of them exist in the given data?

Bidirectional Followers: 438369

Assignment 3 - Friend Finder

Ids (2322;2332) have most commons friends (258).

The number of average common friends is 12.88.

Assignment 5 – Benefits and Limitations of MapReduce

a) Why do you think MapReduce and Hadoop were such a great success?

Hadoop is a high level framework which allows the user to work with a enormous distributed amount of data in a simple way. Hadoop already stores parts of the data redundant on the nodes, so that the user does not have to deal with it. Supporting not only MapReduce, but also other algorithms also contributed to the success of Hadoop. MapReduce were such a great success, because it is a simple way to solve problems with big chunks of data paralleled. Moreover it is possible to build a Hadoop MapReduce cluster with usual consumer Hardware.

b) What are the limitations of MapReduce? Are there problem classes that can not be implemented with help of this computation model?

MapReduce is limited in some ways. By creating the Maps, sorting them internally and copying them between nodes overhead is created. Moreover it the input and the results are written to the filesystem making it use many IO operations slowing the system down. As a result, it is not build for iterative computing. Furthermore it does not support message passing. Because message passing is required to solve most graph algorithms, MapReduce cant be used to solve these. Moreover, MapReduce was designed for Batch processing.