Distributed computing with Hadoop MapReduce

Ştefan Istrate

University of Cambridge

March 10, 2011

Hadoop MapReduce

What is MapReduce?

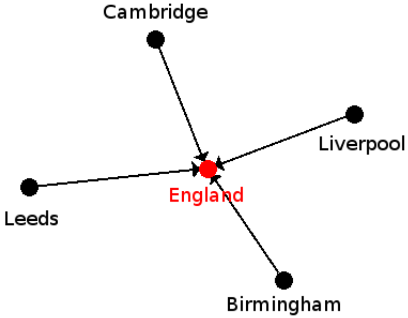
- a software framework for writing applications: *map* and *reduce*
- focus on *what* to do with data, not *how* to do it
- process vast amounts of data in parallel on large clusters
- fault-tolerant system
- introduced by Google in 2004

Why MapReduce? Why Hadoop?

- popularity
- speed
- fast development of applications
- open source

The problem: Find the reverse web-link graph

"The city of Cambridge is a university town and the administrative centre of the county of Cambridgeshire, England. It lies in East Anglia about 50 miles (80 km) north-by-east of London. Cambridge is at the heart of the high-technology centre known as Silicon Fen - a play on Silicon Valley and the fens surrounding the city." (Wikipedia)



How to solve it with MapReduce

Map:

- find every hyperlink (source -> target)
- output the pair <target, source>

Reduce:

- for a given target, concatenate the sources
- emit <target, list(source)>

What I want to do

Explore the prototype of Hadoop MapReduce:

1. investigate the architecture and the tools Hadoop provides
2. implement the reverse web-link problem
3. test on 5000 articles from Wikipedia (approx. 100MB)
4. run on a cluster (1 namenode + 2 datanodes) vs. run on a single node
5. analyse the differences in performance
6. report the results of the system (completed jobs, failed jobs, running time etc.)

Done so far:

- downloaded Wikipedia
- selected 5000 articles
- configured 2 virtual machines with 512MB of RAM (datanodes)
- copied the articles from Wikipedia into HDFS

Questions?