Proposal: RHadoop

Title

- Working with R in Hadoop, using the RHadoop project

Presenters

- Andrie de Vries
  - Affiliation: Revolution Analytics
  - Email: andrie@revolutionanalytics.com
- Simon Field
  - Affiliation: Revolution Analytics
  - Email: simon.field@revolutionanalytics.com

Goals

- Understand the elements of a Hadoop system
- Write an algorithm in R that calls a MapReduce job
- Introduction to some machine learning libraries, e.g. RevoScaleR and Spark

Detailed outline

- A brief introduction to Hadoop
  - File systems (e.g. HDFS)
  - Job schedulers (e.g. MapReduce)
  - Databases, e.g. Hive
- Running R in Hadoop
  - Getting to the data
  - Running a simple job
- A first example using R
  - From lapply() to mapreduce
- Hello world
  - Doing a word count (the prototypical Hadoop example)
- Thinking functionally
  - Mappers
  - Combiners
  - Reducers
- Writing a simple parallel job
  - Writing k-means clustering in Hadoop
- Doing distributed matrix algebra in Hadoop using the `rmr2` package
  - Writing a simple least squares regression
  - Distributed matrix operations in the mappers
  - Combining and reducing the results

    - Inverting the matrix on the master node
- Working with data in Hive using the RHive package
- Conclusion

Justification

- Using Hadoop for big data is one of the most hyped technology terms. The technology is widely in use in companies with web-scale data, and is increasingly being evaluated by IT departments in many other industries.
- The R user needs to know how to modify algorithms to make use of the map-reduce paradigm
- Fortunately, R has many features of functional language, for example lapply() which is a simple example of map-reduce
- This tutorial is an introduction to RHadoop for people who have not used Hadoop before

Background knowledge required

- This is a dummies guide to RHadoop and we assume very little prior knowledge. We will distribute a virtual machine image (running on Ubuntu linux) in advance of the tutorial. Attendees will need to be able to download and install the virtual machine in advance of the session.

Potential attendees

- This is an excellent opportunity to get familiar with the big data technology of Hadoop. Any R user who is curious to know more about interfacing with Hadoop will find this session useful.

Additional notes

- Prior to the conference, we will make available a virtual machine containing a Hadoop distribution (e.g. Cloudera, Hortonworks or MapR) as well as R, the RHadoop packages and a pre-configured RStudio environment
- This virtual machine allows attendees to work on their laptops and follow the examples