REVOLUTION ANALYTICS

Module 2: Overview of Hadoop and R

Overview of Hadoop

Hadoop is an open-source implementation of Google's MapReduce processing framework.

- Many components - distributed file system, job scheduling, management system
- MapReduce is a programming pattern distributed computing
- Two primary steps:
  - Map - picks out identifying and subject data (key and value)
  - Reduce - aggregates values (grouped by key value)

Overview of Hadoop

Hadoop differs from former SQL databases in that information is stored by scattering it across many different machines, as opposed to keeping it all on one database.

To retrieve information, the primary node contacts the other nodes which store the desired information. An algorithm called MapReduce is then implemented to collect the information and pass it back to the user in its entirety.

There are many benefits from using this storage model: from cost effectiveness, by being able to store information on multiple cheaper machines as opposed to one expensive database, to capacity, since expanding your storage capabilities merely requires the installation of additional machines.

Overview of MapReduce

- Programming framework for distributed computing
- Provides flexibility to programmer while allowing efficient implementation

MapReduce provides a framework for users to process data in parallel on a network of compute nodes, or a cluster. There are two components to MapReduce, Map and Reduce.

Overview of MapReduce

The Map component applies an algorithm to all data in a set. For instance, sorting a subset of data, or for instance data on a single node in the Hadoop environment, Map would apply the sort function to all values of data in that subset. In the end, all values are sorted in a particular order.

The Reduce component combines the results from the Map component. Continuing our sorting example, Reduce would take the resulting sorted subsets of data obtained from the Map step, and combine them into one large set of sorted data.

Overview of MapReduce

- Map
  - One-to-one
  - Output size will be the same as Input size
- Reduce
  - Many-to-one function
  - Output size will be smaller than Input size

Overview of MapReduce - Map

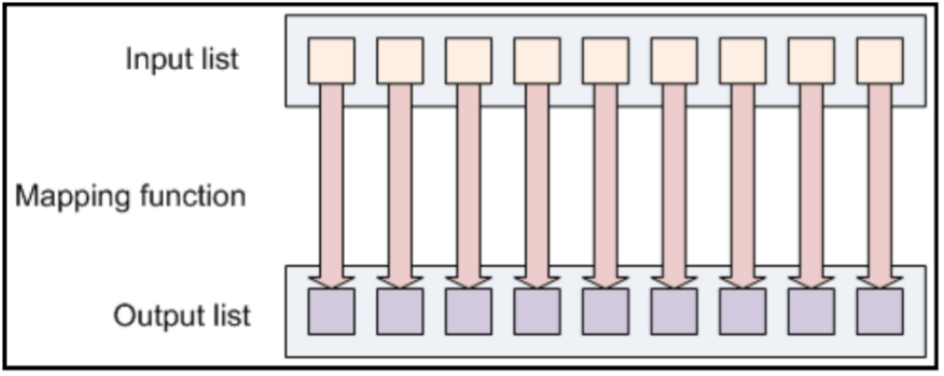

Figure: Apply the same computation to all data

Figure:

Overview of MapReduce - Reduce

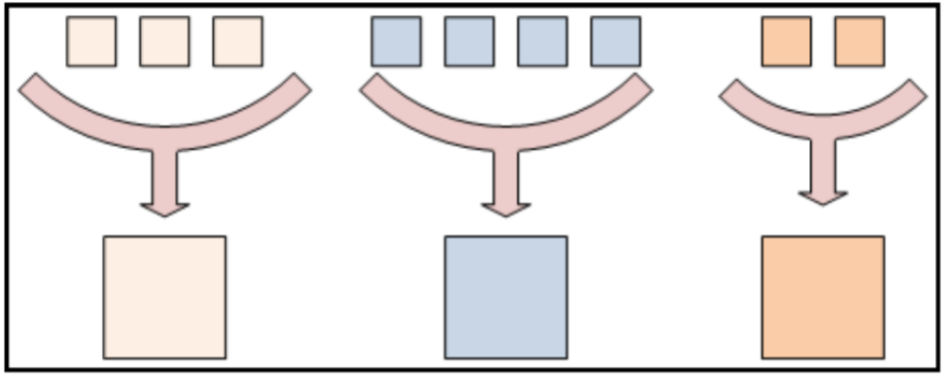

Figure: Group and Reduce data

Figure:

Overview of R

R is an open-source environment for statistical programming and analysis.

- Active, growing community
- Large library of add-on packages
- Commercial support, extensions, training
- Multi-paradigm language: functional, procedural, object-oriented
- Low-level interface with many languages (e.g. C, C++)
- Near a mathematical level of programming

Overview of R

R is a software particularly useful for statistical computing and graphics. Recently, R has become increasingly popular among statisticians and data scientists for a variety of reasons listed above. R is an open source software, and is a GNU project, meaning that users may freely run, share, study, and modify the software.

Revolution Analytics offers a proprietary version of R, referred to as ScaleR, that has several improvements over the open source software. These mainly include multithreading capabilities, whereas open source R is single threaded, and the ability to run outside of memory capacity, whereas open source R is memory bound.

Options for Using R with Hadoop

- R packages:
  - HadoopStreaming
  - Hive
  - RHIPE
  - Segue
  - Rhadoop
  - Rhdfs
  - Rhbase
  - rmr2
- RScript
  - “nacked” streaming
  - JDBC/ODBC connections to Hive
- ScaleR 7.0

Options for Using R with Hadoop

As shown above, there are several R packages which allow users to utilize the capabilities of R and Hadoop. Oftentimes, the user must write her code to import data scattered across multiple clusters (often involving knowledge of MapReduce), and parallelize the computation to run on multiple nodes.

Options also exist for users to run R scripts in collaboration with Hadoop, as in, for instance, establishing JDBC/ODBC connections to Hive. Likewise, often this requires advance knowledge of Hadoop compatible languages.

Finally, ScaleR offers a new version that is now Hadoop compatible! This is the version we will focus on for the duration of this course, and as you will come to realize, is user-friendly and does not require an advanced knowledge of Hadoop. In fact, users only require a

Recap

Let's review some of the concepts covered in this module:

- What is Hadoop, and how does it differ from older database models?
- What is MapReduce, and how does it interact with the Hadoop Environment?
  - What are the two components of MapReduce, and how do they work to complete a computation?
- Why has R become such a popular software for statistical computing?
  - What are some improvements ScaleR has over open source R?
- How can R and Hadoop be utilized together?

REVOLUTION ANALYTICS

REVOLUTION ANALYTICS

Thank you

Revolution Analytics is the leading commercial provider of software and support for the popular open source R statistics language.

www.revolutionanalytics.com, 1.855.GET.REVO, Twitter: @Revo-lutionR