REVOLUTION ANALYTICS

Introduction to Revolution R Enterprise Module 1: Overview of Big Data Analytics

Overview

Welcome to Introduction to Revolution R Enterprise!

After completing this course, you will be able to leverage Revolution R Enterprise in:

- Conducting various types of big data manipulations
- Analyzing and summarizing big data
- Constructing basic plots and graphics

The Data

Throughout this course we will be using two data sets:

- Bank Marketing data set from the Machine Learning Repository at University of California, Irvine
- an internally created, randomly generated Churn data set for exercise purposes.

[Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS (http://hdl.handle.net/1822/14838)

The Data: Bank Marketing Data

The Bank Marketing Data Set, which we will refer to as the Bank data, concerns the direct marketing campaigns, or phone calls, of a Portuguese banking institution to its clientele, and the success of those campaigns in causing customers to subscribe to a term deposit. Multiple types of data are collected on each bank client, such as information on one's age and marital status.

The Data: Bank Marketing Data

Let's examine a portion of this data to get a sense of these variables…

	age	job	marital	education	default	balance	housing
1	58	management	married	tertiary	no	2143	yes
2	44	technician	single	secondary	no	29	yes
3	33	entrepreneur	married	secondary	no	2	yes
4	47	blue-collar	married	unknown	no	1506	yes
5	33	unknown	single	unknown	no	1	no
6	35	management	married	tertiary	no	231	yes

The Data: Bank Marketing Data

Continuing the variables…

	contact	day	month	duration	campaign	pdays	previous	pout
1	unknown	5	may	261	1	-1	0	unkn
2	unknown	5	may	151	1	-1	0	unkn
3	unknown	5	may	76	1	-1	0	unkn
4	unknown	5	may	92	1	-1	0	unkn
5	unknown	5	may	198	1	-1	0	unkn
6	unknown	5	may	139	1	-1	0	unkn

REVOLUTION ANALYTICS

The Data: Bank Marketing Data

So, for instance, our third contact:

- is 33 years old
- works as an entrepreneur
- is married
- has a secondary level of education
- does not have credit in default
- has an average yearly balance of 2 euros
- has both a housing and personal loan

The Data: Bank Marketing Data

Further, our third contact in relation with the last contact of the current campaign:

- was contacted using an unknown source (i.e., not by telephone or cellular)
- was last contacted on the fifth of the month, which was may
- the duration of contact lasted 76 secs
- 1 contacts performed during this campaign for this client (including the last contact)
- was not previously contacted by a campaign (indicated by -1)
- 0 contacts performed before this campaign and for this client
- unknown outcome of previous campaign
- did not subscribe to a term deposit

The Data: Churn Data

Also, the randomly generated Churn data set is used for exercise purposes. One may imagine this data as that representing clients in a phone company, where variables such as n.family.members and n.devices refer to the number of family members and the number of devices of a particular client, respectfully. Again, in each exercise we will clarify the usage of variables and what we are attempting to gleam from the data.

What is Big Data?

Big data is a vague term used to describe an arbitrarily large amount of data that is just too difficult to handle using former computational or mathematical techniques. The data may be characterized by a variety of different attributes, such as for instance spatial data rather than conventional data. Here are a few picture examples to help clarify these concepts:

Big Data in Space

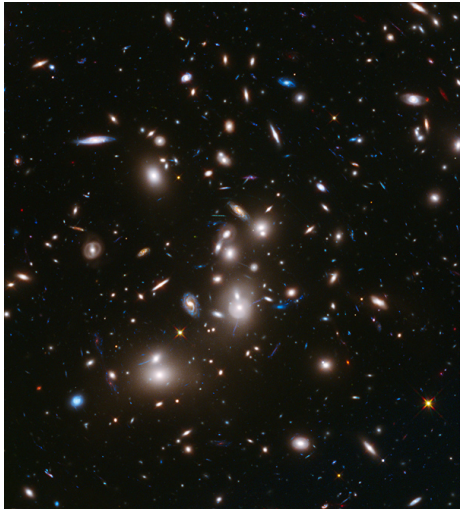


Figure: Hubble Frontier Field Abell 2744

Big Data in Space

Large amounts of digital data are collected by various sources, such as ground instrumentation and spacecraft, and must be properly stored, indexed, and processed.

For NASA, according to the Jet Propulsion Laboratory in Pasadena, Calif, “hundreds of terabytes are collected every hour. Just one terabyte is equivalent to the information printed on 50,000 trees of paper (Jet Propulsion Laboratory).”

http://www.jpl.nasa.gov/news/news.php?release=2013-299

Big Data in Climate

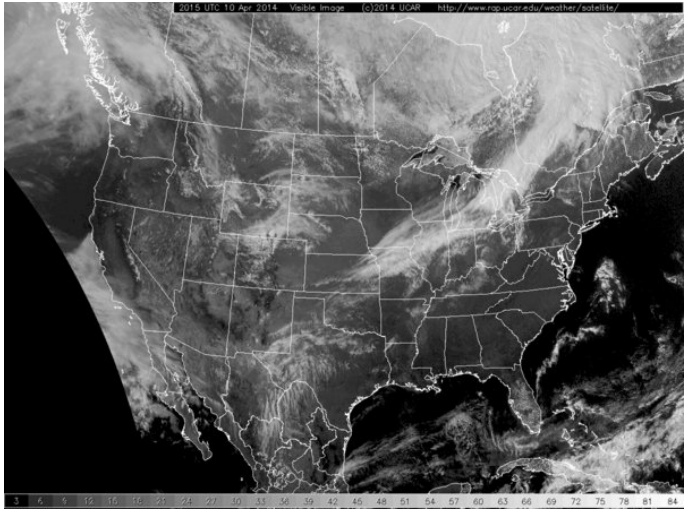


Figure: Satellite Image

Big Data in Climate

At NCAR, the National Center for Atmospheric Research located in Boulder, Colorado, big data is a common topic as well. Climate data can come in many various formats and is often very massive.

http://weather.rap.ucar.edu/satellite/

Big Data in Health

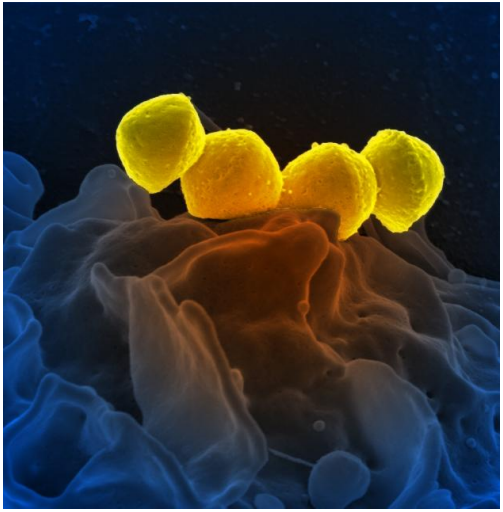


Figure: National Institute of Allergy and Infectious Diseases (NIAID), Streptococcus pyogenes Bacteria

Big Data in Health

Recent developments in medicine also focus on big data and modern statistical and mathematical methods. At the National Institutes of Health in Washington D.C., much focus has been placed recently on developing better methods to utilize these methods in dealing with big data.

http://bd2k.nih.gov/#sthash.6flWqr6f.dpbs

Key Challenges in Analyzing Big Data

- Accessing
- Moving
- Merging
- Manage
- Munge

Accessing and Moving Data

- Accessing data refers to reaching out and grabbing data in memory or, if ‘big’ data, usually in a database such as Hadoop.
- Moving data refers to changing directories where data is stored, for instance moving data from memory to hard drive space (such as the technique ScaleR uses) or moving data within or outside databases.

Merging, Managing, and Munging

- Merging data refers to combining data from two or more separate sources. For instance, perhaps we have two data sets measuring different variables on the same set of subjects, and we might wish to combine these sets to produce a comprehensive data set.
- Managing data refers to refining stored data. For instance, perhaps we want to make a change in data row, remove NA values, make a data transformation, or other such actions.
- Munging data refers to a list of changes made on a part of data which are usually reversible but transform the original data to something unrecognizable. The result of these actions could be entirely destructive, or could, for instance, hide data from outside sources.

Big Data challenges for R

- Memory Bound

In open source R, computations are limited to the amount of data that can be stored in memory. Consequently, problems involving large amounts of data may be limited or entirely impossible to handle using the open source software.

Revolution Analytics ScaleR stored data outside of memory by utilizing external sources, such as hard drive space on a local computer or a database such as Hadoop on a compute cluster, and therefore is not memory bound. Accordingly, ScaleR can handle big data computations quite well without maxing its storage capabilities.

Big Data challenges for R

- Single Threaded

Single threading refers to sequential processing, or processing a series of commands one at a time. For instance, when sorting a group of numbers, perhaps you take the first number and put it to the side, take the second and compare it to the first, and place it in front or behind according to its value. You repeat this for each number, and after a brutally long time you have a pile of sorted numbers.

While the above process works, perhaps you realize that you can accomplish the same task by splitting the large pile of numbers up into smaller groups, inviting some friends over, and sorting them all separately in individual groups. Then, you combine those groups and sort the numbers respectfully, use the saved computational time to hang out with your friends! This method is referred to as

Open Source R Architecture

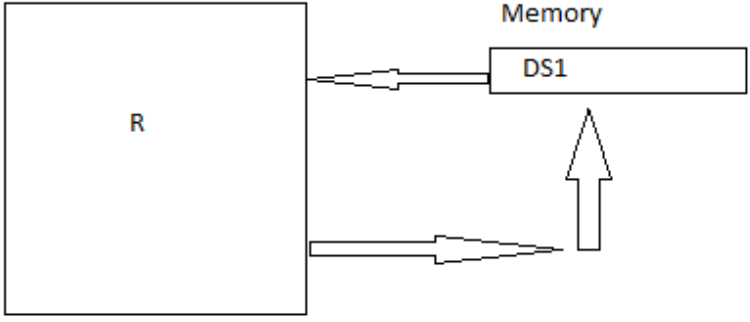


Figure:

Revolution R Enterprise Architecture

ScaleR improves this process of handling big data by storing data in an XDF format in hard drive space rather than memory. This improves big data computations on a single computer and across a distributed network.

Revolution R Enterprise Architecture

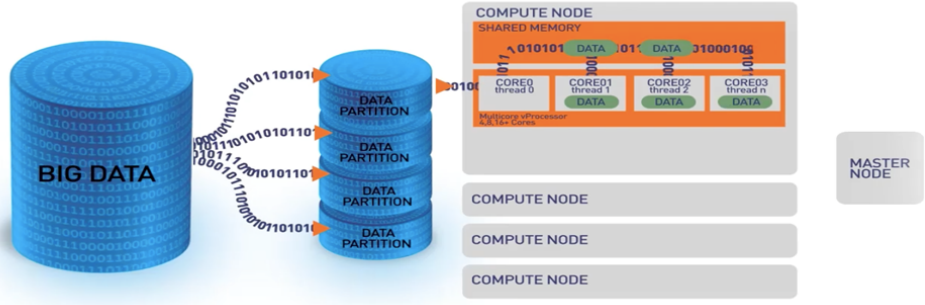
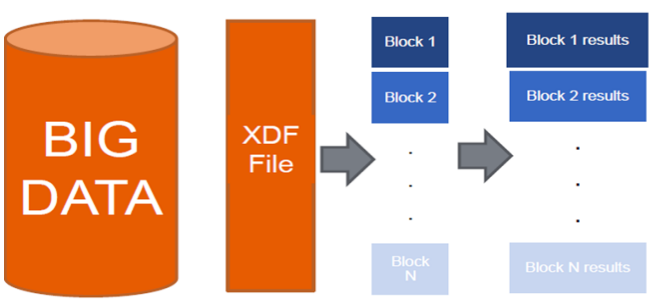
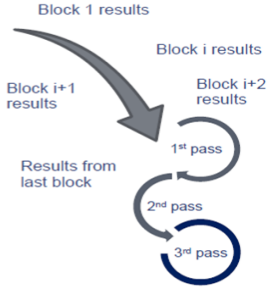


Figure:

Parallel External Memory Algorithms





Read Blocks, Compute Intermediate Results in Parallel, Iterating as Necessary

Figure:

Best Practices in Dealing with Big Data

For maximizing scalability when working with large data, keep the following points in mind:

- When handling big data in ScaleR, it is often best to subset the important data and leave the unnecessary data not applicable to your analysis out.
- To most efficiently subset data, it is best to do this outside of statistical functions. Instead, use hadoop commands in UNIX, or more perhaps even more easily use the compatible ScaleR commands in hadoop.
- Remember to create the proper HDFS directories to store your big data. When conducting analyses, remember that ScaleR works by assigning a pointer, called a data source, to the location of your data on the HDFS. Creating this data source is

Recap

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

- What is "Big Data"?
  - What are some examples you can think of?
- What are the five key challenges in analyzing big data?
  - What do these terms mean and how do they apply to your data?
- How does memory bound open source R differ from Revolution Analytics ScaleR?
- What is the difference between single threading and multithreading?
- What are some best practices for handling big data?

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