IIT School of Applied Technology

ILLINOIS INSTITUTE OF TECHNOLOGY

information technology & management

527 Data Analytics

March 22,24 2016

Week 11 Presentation

Week 11 Topic: Agenda

- Week 11 Assignment Review
- Hadoop in more detail
- Hadoop and R in SAS

Big Data:

- What is Big Data?
- What is Hadoop?
- Big Data Analytics
- Intro to R
- Installing R or RStudio

Week 11 Topic:
Assignment 1 – Cluster Analysis in R

(Worth 10 points) Due March 31st

- ◆ Read Chapter 10 of Data Smart
- ◆ Follow directions as stated and perform cluster analysis for k=3 in R
- ◆ Submit code and results in a word document.

Week 11 Topic: Assignment 2 – Exec Comp Review

(Worth 10 points) Due March 31st

- Read: https://pdfs.semanticscholar.org/52c6/d51d85e9a2c529d696cb83b6ad1d7daa45ca.pdf
- Download/Collect DEF 14A filings for 5 companies to analyze from www.sec.gov
- Summarize, for each company, its CEO name, Salary, Vested Stock Awards, and Option Awards for all available years in the filing.
- Determine total vested stock ownership of the CEO and salary trend for all available years in the filing.
- Describe (in numbered steps) how to automate the above task in SAS or R.

Week 11 Topic:
SAS and Hadoop

- The Apache Software Foundation[1] provides support for open-source software projects. One such project is Apache Hadoop[2].

- *Hadoop* is a framework for distributed processing of large data sets across a cluster of computers using simple programming models[2]

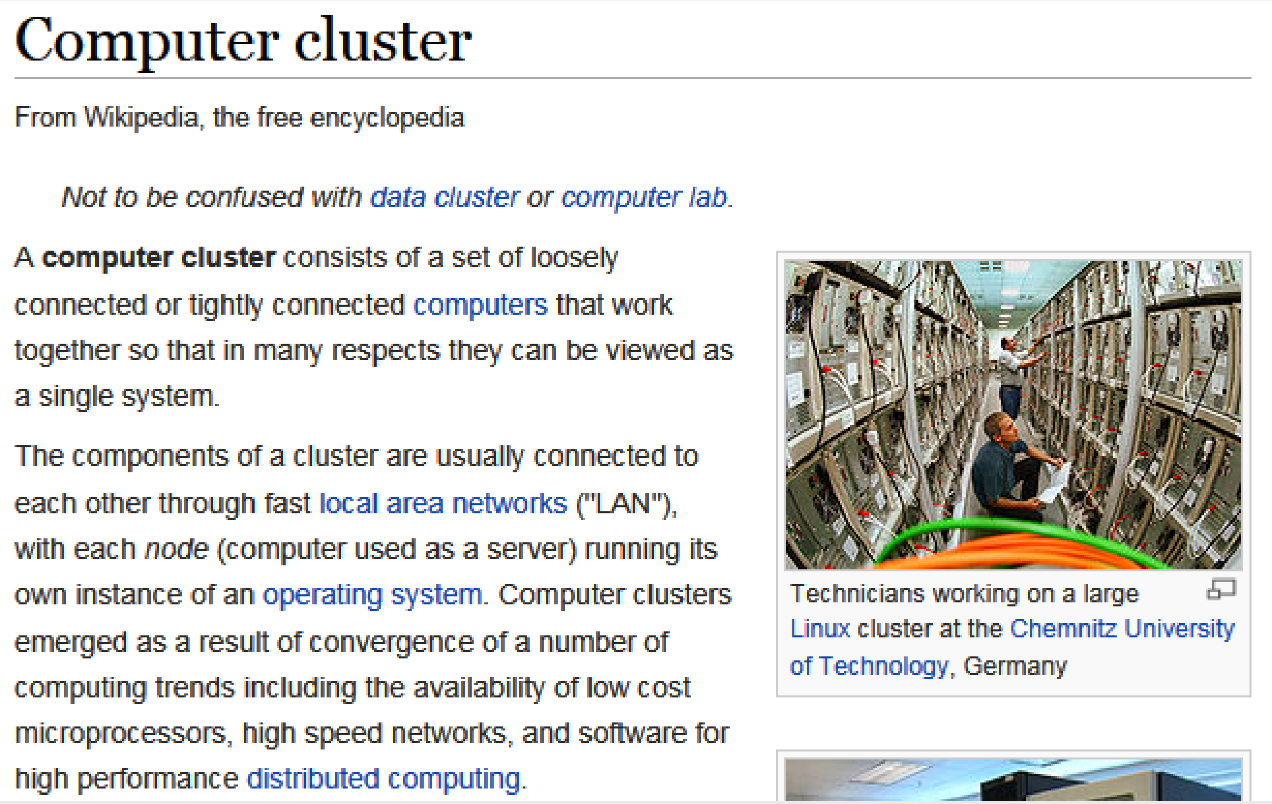
A cluster of computers:

[1] www.apache.org

[2] hadoop.apache.org

Week 11 Topic:
Cluster of Computers?

http://en.wikipedia.org/wiki/Computer_cluster



Computer cluster

From Wikipedia, the free encyclopedia

Not to be confused with data cluster or computer lab.

A **computer cluster** consists of a set of loosely connected or tightly connected computers that work together so that in many respects they can be viewed as a single system.

The components of a cluster are usually connected to each other through fast local area networks ("LAN"), with each *node* (computer used as a server) running its own instance of an operating system. Computer clusters emerged as a result of convergence of a number of computing trends including the availability of low cost microprocessors, high speed networks, and software for high performance distributed computing.

Technicians working on a large Linux cluster at the Chemnitz University of Technology, Germany

Week 11 Topic:
Revisit of Core Hadoop Modules

- Core Hadoop modules include:

HDFS (Hadoop Distributed File System)	a file system that distributes large files across the Hadoop cluster of computers
Hadoop YARN	a framework for job scheduling and cluster resource management
Hadoop MapReduce	a YARN-based system for parallel processing of large data sets

- These modules automate the process of reading, writing, and processing large files in a distributed environment, freeing programmers to write programs to process the data as if they were using a single computer.

Week 11 Topic:
File management in Hadoop

- HDFS is hierarchical with LINUX style paths and file ownership and permissions.
  - HADOOP FS commands are similar to LINUX commands.
  - HDFS in not built into the operating system.
  - Files are append-only after they are written.

HADOOP FS commands from LINUX command prompt:

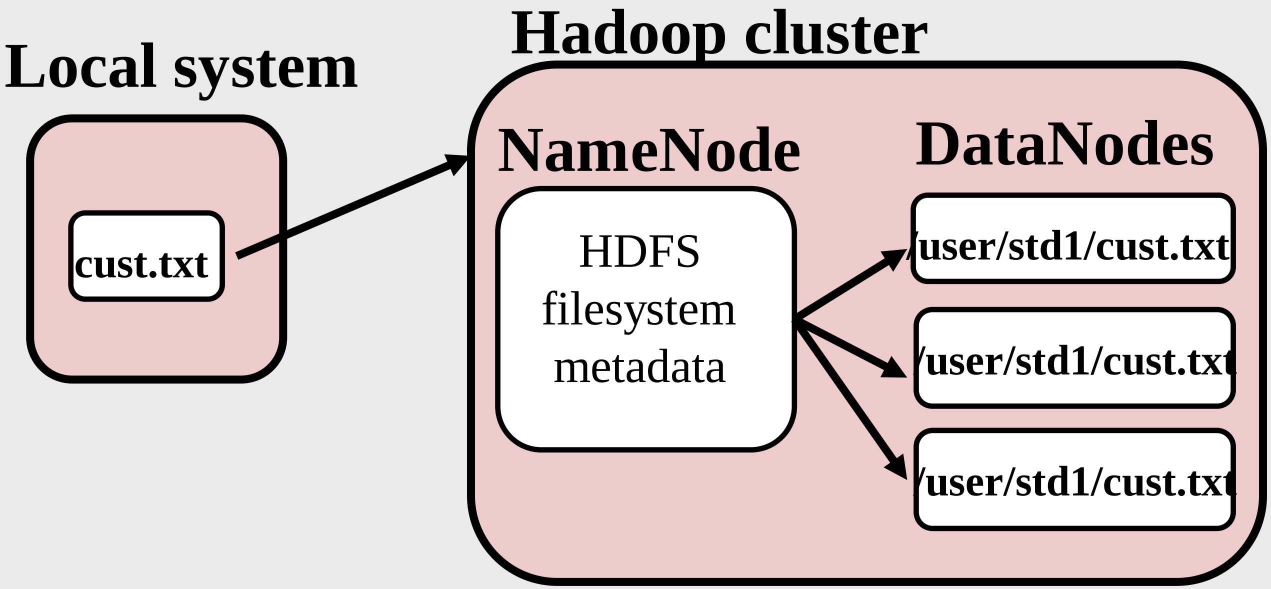
```
$ hadoop fs -ls /user/student
Found 4 items
drwxr-xr-x   - student1 sasapp 0 2014-05-30 20:00 /user/student1/.Trash
drwx------   - student1 sasapp 0 2014-05-30 10:05 /user/student1/.stage
drwxr-xr-x   - student1 sasapp 0 2014-05-28 15:25 /user/student1/data
drwxr-xr-x   - student1 sasapp 0 2014-05-28 13:59 /user/student1/users
$ hadoop fs -mkdir /user/student1/newdir
$
```

Week 11 Topic:
File management in Hadoop

- The following HDFS command moves a local file into the HDFS cluster:

```
$ hadoop fs copyfromlocal="cust.txt"  out="/user/std1"
```

- An HDFS ***NameNode*** on a root node machine distributes the file to each of the ***DataNode*** machines, and provides access to the distributed file.



Local system

cust.txt

Hadoop cluster

NameNode

HDFS filesystem metadata

DataNodes

/user/std1/cust.txt

/user/std1/cust.txt

/user/std1/cust.txt

Week 11 Topic: Characteristics of Hadoop

- Open-source
- Simple to use distributed file system
- Supports highly parallel processing
- It’s scalable, so it’s suitable for massive amounts of data
- It is designed to work on low-cost hardware
- It’s fault tolerant at the data level
  - automatic replication of data
  - automatic fail-over

Week 11 Topic: Hadoop Distribution

- Apache Hadoop is open-source technology developed by the Apache Software Foundation
  - a community of developers and users
  - a non-profit
- Several commercial vendors provide, support, and augment Apache Hadoop with their own distribution. These include:
  - Cloudera
  - Hortonworks
  - BigInsights

Week 11 Topic:
SAS Hadoop Interface

Tool	Purpose	Product
FILENAME statement	Allows the DATA step to read and write HDFS data files.	Base SAS
PROC HADOOP	Copy or move files between SAS and Hadoop. Execute MapReduce and Pig code in Hadoop. Execute Hadoop file system commands to manage files and directories.	Base SAS
SQL Pass-Through	Submit HiveQL queries and other HiveQL statements from SAS directly to Hive for Hive processing. Query results are returned to SAS.	SAS/ACCESS Interface to Hadoop
LIBNAME Statement For Hadoop	Access Hive tables such as SAS data sets using the SAS programming language. SAS/ACCESS engine translate SAS language into HiveQL and attempts do convert as much of the processing into HiveQL as possible before returning results to SAS.	SAS/ACCESS Interface to Hadoop

Week 11 Topic:
Base SAS Tools for Hadoop

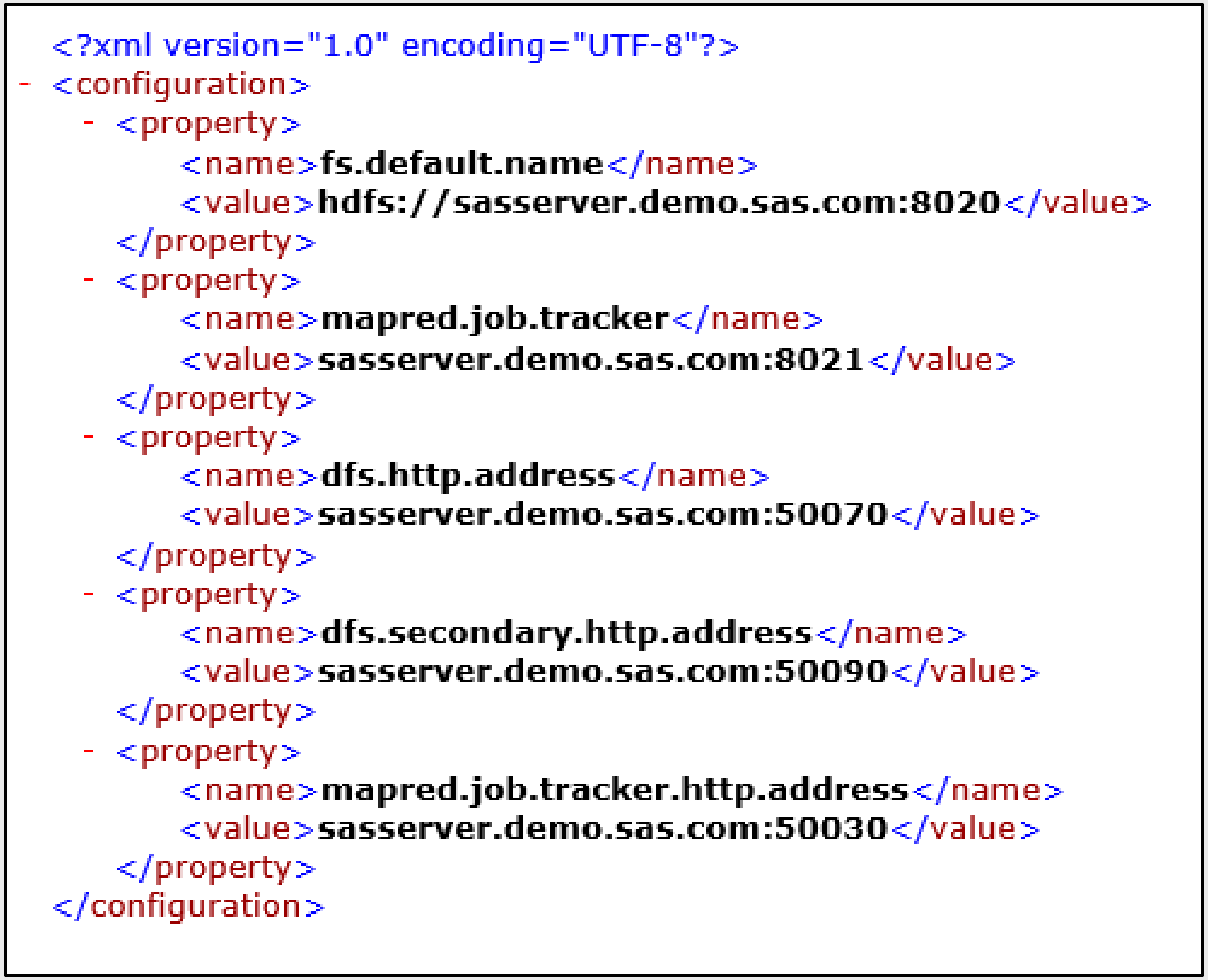
- The Hadoop FILENAME statement enables you to do the following:
  - upload local data to Hadoop using the DATA step
  - read data from Hadoop using the DATA step

- PROC HADOOP enables you to do the following:
  - submit Hadoop file system (HDFS) commands
  - submit MapReduce programs
  - submit Pig language code

Week 11 Topic: Hadoop Config.xml File

- The FILENAME statement for Hadoop and PROC HADOOP require an option that specifies a Hadoop configuration file.
  - The configuration file defines how to connect to Hadoop and other Hadoop system information.
  - The file must be accessible to the SAS client application.
  - A SAS administrator commonly manages this configuration for the SAS users.
  - This file is often referred to as the Hadoop core-site.xml file.

Week 11 Topic:
Hadoop Config.xml File

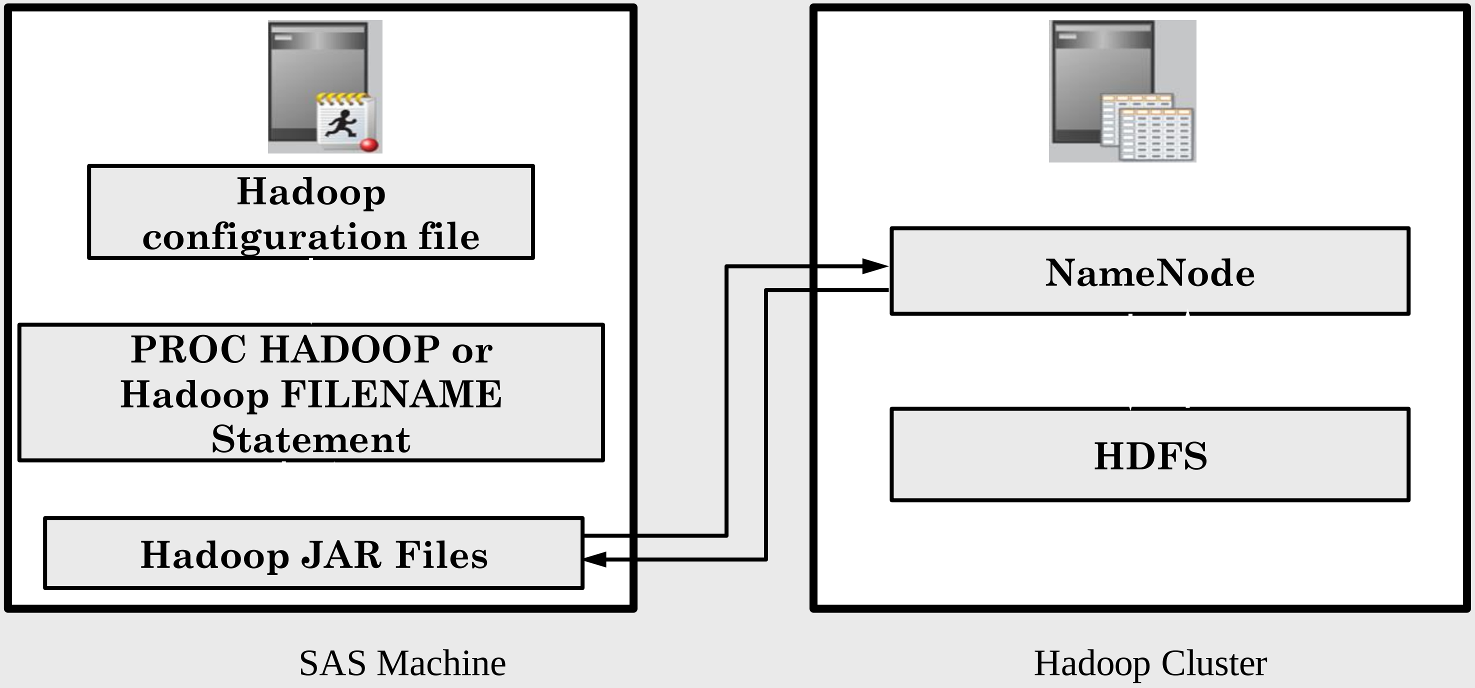


```
  <?xml version="1.0" encoding="UTF-8"?>
- <configuration>
  - <property>
        <name>fs.default.name</name>
        <value>hdfs://sasserver.demo.sas.com:8020</value>
    </property>
  - <property>
        <name>mapred.job.tracker</name>
        <value>sasserver.demo.sas.com:8021</value>
    </property>
  - <property>
        <name>dfs.http.address</name>
        <value>sasserver.demo.sas.com:50070</value>
    </property>
  - <property>
        <name>dfs.secondary.http.address</name>
        <value>sasserver.demo.sas.com:50090</value>
    </property>
  - <property>
        <name>mapred.job.tracker.http.address</name>
        <value>sasserver.demo.sas.com:50030</value>
    </property>
  </configuration>
```

Week 11 Topic: Hadoop JAR File

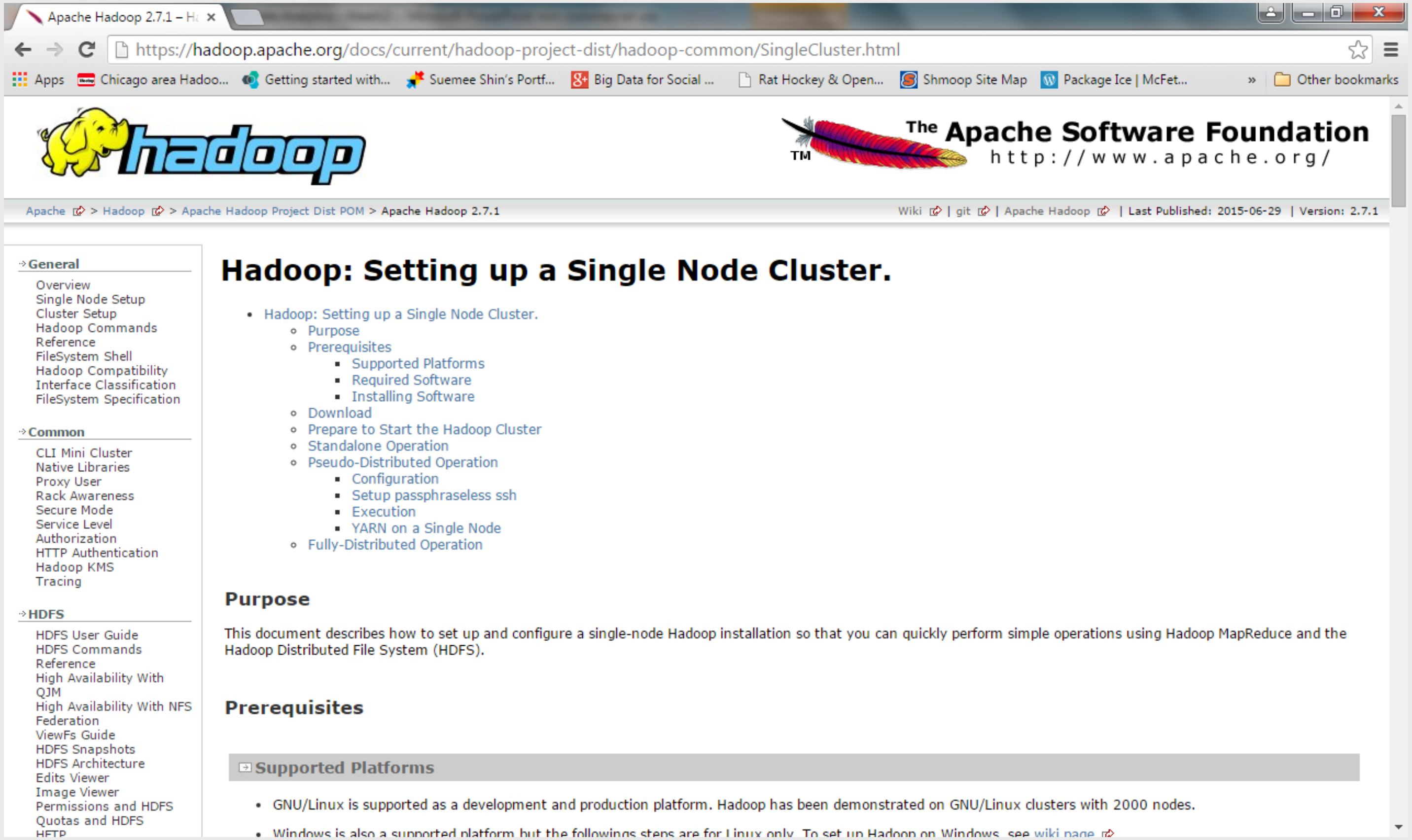
- A collection of Hadoop JAR files is also required on the SAS client machine.
  - An environment variable SAS_HADOOP_JAR_PATH on the SAS client machine defines location of the Hadoop JAR files
  - The Hadoop JAR files must be compatible with the specific Hadoop implementation and can be copied from the Hadoop server.
  - A Hadoop system administrator commonly manages the configuration of the Hadoop JAR files for the SAS users.

Week 11 Topic: Hadoop Interface to SAS


Hadoop configuration file
PROC HADOOP or Hadoop FILENAME Statement
Hadoop JAR Files
NameNode
HDFS
SAS Machine
Hadoop Cluster

Week 11 Topic: Installing Hadoop

https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SingleCluster.html (using a Linux emulator)



Apache > Hadoop > Apache Hadoop Project Dist POM > Apache Hadoop 2.7.1 | Wiki | git | Apache Hadoop | Last Published: 2015-06-29 | Version: 2.7.1

General
- Overview
- Single Node Setup
- Cluster Setup
- Hadoop Commands Reference
- FileSystem Shell
- Hadoop Compatibility
- Interface Classification
- FileSystem Specification

Common
- CLI Mini Cluster
- Native Libraries
- Proxy User
- Rack Awareness
- Secure Mode
- Service Level Authorization
- HTTP Authentication
- Hadoop KMS
- Tracing

HDFS
- HDFS User Guide
- HDFS Commands Reference
- High Availability With QJM
- High Availability With NFS
- Federation
- ViewFs Guide
- HDFS Snapshots
- HDFS Architecture
- Edits Viewer
- Image Viewer
- Permissions and HDFS
- Quotas and HDFS
- HFTP

Hadoop: Setting up a Single Node Cluster.

- Hadoop: Setting up a Single Node Cluster.
  - Purpose
  - Prerequisites
    - Supported Platforms
    - Required Software
    - Installing Software
  - Download
  - Prepare to Start the Hadoop Cluster
  - Standalone Operation
  - Pseudo-Distributed Operation
    - Configuration
    - Setup passphraseless ssh
    - Execution
    - YARN on a Single Node
  - Fully-Distributed Operation

Purpose

This document describes how to set up and configure a single-node Hadoop installation so that you can quickly perform simple operations using Hadoop MapReduce and the Hadoop Distributed File System (HDFS).

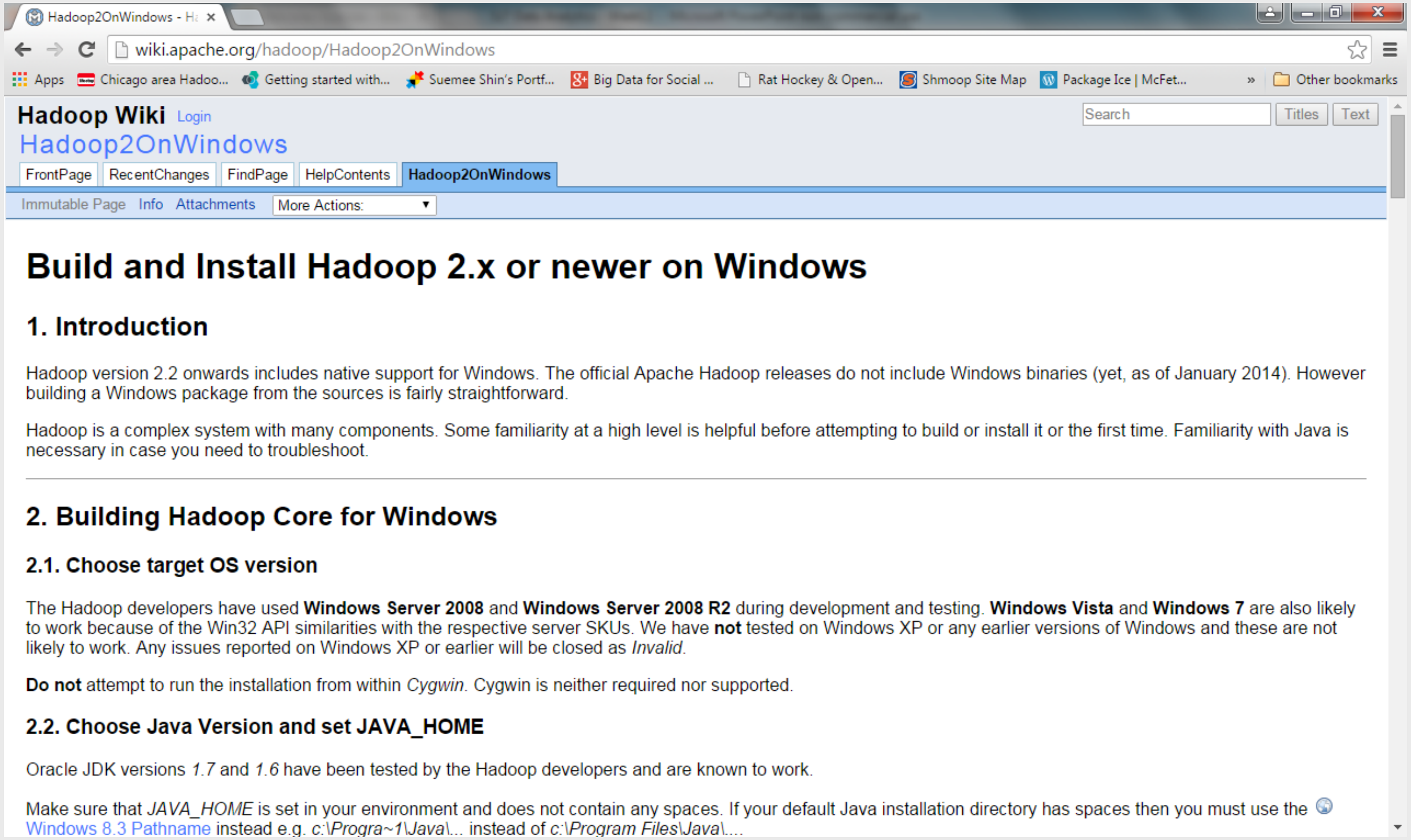
Prerequisites

Supported Platforms

- GNU/Linux is supported as a development and production platform. Hadoop has been demonstrated on GNU/Linux clusters with 2000 nodes.
- Windows is also a supported platform but the followings steps are for Linux only. To set up Hadoop on Windows, see wiki page

Week 11 Topic: Installing Hadoop

http://wiki.apache.org/hadoop/Hadoop2OnWindows



Build and Install Hadoop 2.x or newer on Windows

1. Introduction

Hadoop version 2.2 onwards includes native support for Windows. The official Apache Hadoop releases do not include Windows binaries (yet, as of January 2014). However building a Windows package from the sources is fairly straightforward.

Hadoop is a complex system with many components. Some familiarity at a high level is helpful before attempting to build or install it or the first time. Familiarity with Java is necessary in case you need to troubleshoot.

2. Building Hadoop Core for Windows

2.1. Choose target OS version

The Hadoop developers have used **Windows Server 2008** and **Windows Server 2008 R2** during development and testing. **Windows Vista** and **Windows 7** are also likely to work because of the Win32 API similarities with the respective server SKUs. We have **not** tested on Windows XP or any earlier versions of Windows and these are not likely to work. Any issues reported on Windows XP or earlier will be closed as *Invalid*.

Do not attempt to run the installation from within *Cygwin*. Cygwin is neither required nor supported.

2.2. Choose Java Version and set JAVA_HOME

Oracle JDK versions *1.7* and *1.6* have been tested by the Hadoop developers and are known to work.

Make sure that *JAVA_HOME* is set in your environment and does not contain any spaces. If your default Java installation directory has spaces then you must use the Windows 8.3 Pathname instead e.g. *c:\Progra~1\Java\...* instead of *c:\Program Files\Java\...*.

Week 11 Topic: MapReduce

- *MapReduce* is a framework written in Java that is built into Hadoop. It automates the distributed processing of data files.

map	processing of individual rows (filtering, row calculations)
shuffle and sort	grouping rows for summarization
reduce	summary calculations within groups

- The MapReduce framework coordinates multiple mapping, sorting, and reducing tasks that execute in parallel across the computer cluster.

Week 11 Topic: Executing MapReduce

A JAVA programmer can write code that calls the following ‘functions’ that are built into the MapReduce framework:

- an input reader
- a Map function
- a partition function
- a compare function
- a Reduce function
- an output writer

More recently, Pig and Hive projects have been developed for Apache Hadoop. These provide a higher-level language interface that compiles to MapReduce.

Week 11 Topic:
Pig and Hive

Pig and Hive provide less complex higher-level programming methods for parallel processing of Hadoop data files.

Pig	A platform for data analysis that includes stepwise procedural programming that converts to MapReduce.
Hive	A data warehousing framework to query and manage large data sets stored in Hadoop. Provides a mechanism to structure the data and query the data using an SQL-like language called HiveQL. Most HiveQL queries are compiled into MapReduce programs.

Week 11 Topic:
SAS and R

http://www.sas.com/content/dam/SAS/en_us/doc/conclusionpaper1/use-of-open-source-is-growing-107429.pdf:

SAS/IML – a matrix manipulation product that supports matrix-vector computation – gives users the flexibility to create custom functions. It also:

- Takes advantage of built-in functions, subroutines and SAS procedures.
- Enables interfaces with R language so users can submit R code within SAS.
- Moves between SAS and R data structures.
- Enables R functions and packages.

YouTube on how to do this:

- www.youtube.com/watch?v=rUaTTre24kI
- www.youtube.com/watch?v=nmRQ3MtkG6A

Week 11 Topic: R Reference and Exercises

SAS and R Reference:

http://sas-and-r.blogspot.com/

Sample Chapters and Exercises:

http://www3.amherst.edu/~nhorton/sasr2/

Week 11 Topic:
Profiling & Tax Strategy Submission

(Worth 20 points)

- Fill in the Profile Matrix Spreadsheet following the example in class.
- Determine your own Tax Groups. Similar to the example in class. Provide answers in the same Matrix spreadsheet.
- Determine the tax strategy per Tax Group(s) and provide the overall tax increase amount. Provide answers in the same Matrix workbook.

Week 11 Topic: What is Big Data?

Very large, distributed aggregations of loosely structured data – often incomplete and inaccessible

Before:

- ◆ Used to mainly refer to data on web behavior or social network interactions. Even 5 years ago was something to think about but nothing to set aside significant IT budget on if you were a non-internet organization.
- ◆ Used to also be a specialized capability that companies like comScore, Yahoo, Google, and Amazon developed to collect/analyze internet browsing behavior and purchase data on consumers.
- ◆ A significant application of it's technology is on fraud detection and surveillance activities.

Now:

- ◆ It has broadened to include collection and analysis of any unstructured data that are core to businesses including traditional organizations such as hospitals, banks, and supply chains.
- ◆ These traditional companies used to ETL data from systems and paper based functions into databases for analytical work. This process took a lot of time and effort. Now, the idea with Big Data is that we EL or L first then ET or T as needed for analytical requests on the fly.

Week 11 Topic: Big Data / Traditional DW

Traditional Data Warehouse:

- ◆ Records from internal transaction systems or subscription based data provider
- ◆ Data is centralized and uniform
- ◆ Batch updates of new data, mostly historical data
- ◆ Analytics designed against stable (data model) environment
- ◆ Scheduled production reports

→ Key is that the data is known and data quality is part of the processing. Long lead time in data availability but less processing to query the available data

Big Data Analytical Environments:

- ◆ Data available from various sources inside and outside of organizations, including DWs
- ◆ Distributed data in various formats
- ◆ Iteration needed to model/test data
- ◆ May require large amount of memory/resources, on the fly, to process request/query

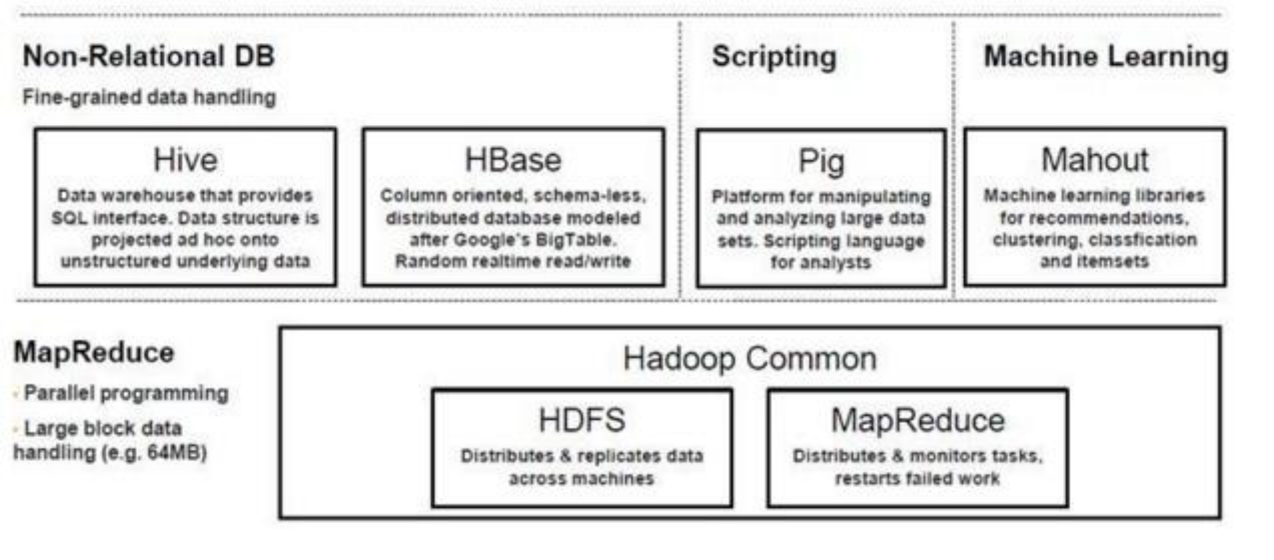
→ Key is that all data is available in its native form at any given time. Short lead time in data availability but longer processing time to query the available data if needing additional normalization steps

Week 11 Topic:
What is Hadoop?

Open-source software that enables reliable, scalable, distributed computing on clusters of inexpensive servers

- Reliable: The software is fault tolerant, it expects and handles hardware and software failures
- Scalable: Designed for massive scale of processors, memory, and local attached storage
- Distributed: Handles replication. Offers massively parallel programming model, MapReduce

Hadoop Framework:



Non-Relational DB

Fine-grained data handling

Hive

Data warehouse that provides SQL interface. Data structure is projected ad hoc onto unstructured underlying data

HBase

Column oriented, schema-less, distributed database modeled after Google's BigTable. Random realtime read/write

Scripting

Pig

Platform for manipulating and analyzing large data sets. Scripting language for analysts

Machine Learning

Mahout

Machine learning libraries for recommendations, clustering, classfication and itemsets

MapReduce

- Parallel programming
- Large block data handling (e.g. 64MB)

Hadoop Common

HDFS

Distributes & replicates data across machines

MapReduce

Distributes & monitors tasks, restarts failed work

Week 11 Topic: Hadoop Framework

Hadoop is designed to process terabytes and even petabytes of unstructured and structured data. It breaks large workloads into smaller data blocks that are distributed across a cluster of commodity hardware for faster processing. And it's part of a larger framework of related technologies:

- HDFS: Hadoop Distributed File System
- HBase: Column oriented, non-relational, schema-less, distributed database modeled after Google's BigTable. Promises "Random, real-time read/write access to Big Data"
- Hive: Data warehouse system that provides SQL interface. Data structure can be projected ad hoc onto unstructured underlying data
- Pig: A platform for manipulating and analyzing large data sets. High level language for analysts
- ZooKeeper: a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. A tool for configuring and synchronizing Hadoop clusters.

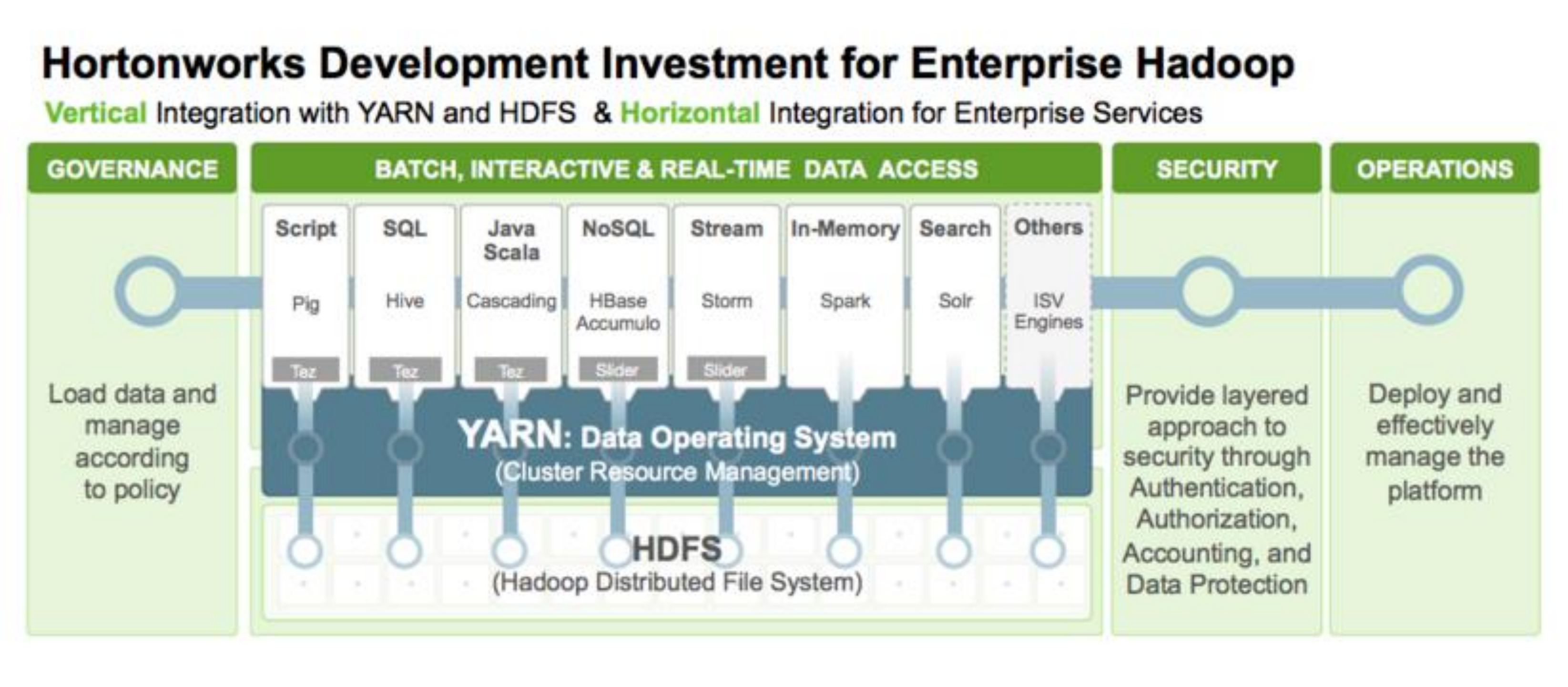
Week 11 Topic: When to use Hadoop?

- Complex but parallelizable algorithms needed, such as geo-spatial analysis or genome sequencing
  - Complex information processing is needed
  - Unstructured data needs to be turned into structured data
  - Queries can't be reasonably expressed using SQL
  - Heavily recursive algorithms e.g., need for parallel processing like machine learning
- Data sets are too large to fit into database RAM, discs, or require too many cores (10's of TB up to PB). Fault tolerance and scale is a challenge.
- Data value does not justify expense of constant real-time availability, such as archives or special interest info, which can be moved to Hadoop and remain available at lower cost and lower latency. Results are not needed in real time

Week 11 Topic:
Blueprint for Hadoop

http://hortonworks.com/blog/announcing-hdp-2-2/

(https://hortonworks.com/wp-content/uploads/2013/10/Big-Data-Reference-Architecture_4AA5-6136ENW.pdf):


Hortonworks Development Investment for Enterprise Hadoop
Vertical Integration with YARN and HDFS & Horizontal Integration for Enterprise Services
GOVERNANCE
BATCH, INTERACTIVE & REAL-TIME DATA ACCESS
SECURITY
OPERATIONS
Script
Pig
Tez
SQL
Hive
Tez
Java Scala
Cascading
Tez
NoSQL
HBase Accumulo
Slider
Stream
Storm
Slider
In-Memory
Spark
Search
Solr
Others
ISV Engines
YARN: Data Operating System
(Cluster Resource Management)
HDFS
(Hadoop Distributed File System)
Load data and manage according to policy
Provide layered approach to security through Authentication, Authorization, Accounting, and Data Protection
Deploy and effectively manage the platform

Week 11 Topic: Blueprint for Hadoop

(https://hortonworks.com/wp-content/uploads/2013/10/Big-Data-Reference-Architecture_4AA5-6136ENW.pdf):

Key highlights of HDP 2.2 include the following:

- ◆ Batch and interactive SQL queries via Apache Hive and Apache Tez, along with a cost-based optimizer powered by Apache Calcite
- ◆ High-performance ETL via Pig and Tez
- ◆ Stream processing via Apache Storm and Apache Kafka
- ◆ YARN labels
- ◆ Search via Apache Solr
- ◆ Streamlined cluster operations via Apache Ambari
- ◆ Data lifecycle management via Apache Falcon
- ◆ Perimeter security via Apache Knox
- ◆ Centralized security administration for HDFS, Hive, HBase, Storm, and Knox via Apache Ranger

Week 11 Topic:
DB vs Hadoop

	Database	Hadoop (HDFS+MapReduce)
Function	- Defined as a database - Built to support structured data	- Defined as a distributed file system – NoSQL (Not only SQL) - Built to support unstructured as well as structured data
Schema	- Stored in rows and columns - Required on write	- Stored as key/value pairs - Required on read
Processing	- Unlimited read and write - Pooled processing across nodes - Configuration performed during install	- Write once, process on read - Recommend processing per node - Use YARN to manage resources
Query	- Returns one answer across all data - Immediate results, real time	- Returns all answers to process on - Requires time to process raw files or can build a database on top of Hadoop (Hbase)
API	SQL, Native, ODBC, JDBC	MapReduce
Scale	- Scaling is a challenge - Vertical scaling, boost server for performance	- Built to scale with ease - Horizontal scaling, add more nodes
TCO	- Typically, licensed software on dedicated servers - Functions built out of the box	- Open source software with commodity servers - Need to build traditionally available functions e.g., security, logging, ETL, server brokering, etc.
When to use	Workloads are constant and predictable	Challenged by increasing data demands

Week 11 Topic: Big Data Playbook

This may counter what we just discussed in the previous slide but does offer a good perspective.

http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/non-geeks-big-data-playbook-106947.pdf

Week 11 Topic: Big Data Analytics

- Big Data Analytics is going to be different than what was offered via traditional architectures. Data being evaluated is different. Technology involved is different. The massive amount of data also makes you think of different ways to perform analytics on the data
- Some highlights of what to expect:
  - For batch oriented data analysis / prototyping with lower processing power required use:

    → R, Python, Matlab
  - For low latency / real time data analysis / prototyping with lower processing power required use:

    → Spark (Scala)
  - For low latency / real time data analysis / prototyping with lots of processing power required use:

    → C/C++, Fortran
  - Other incumbents to the low latency, fast processing arena: Julia, Python, Spark
  - SAS, SPSS, and other analytic applications have Hadoop APIs that can be leveraged

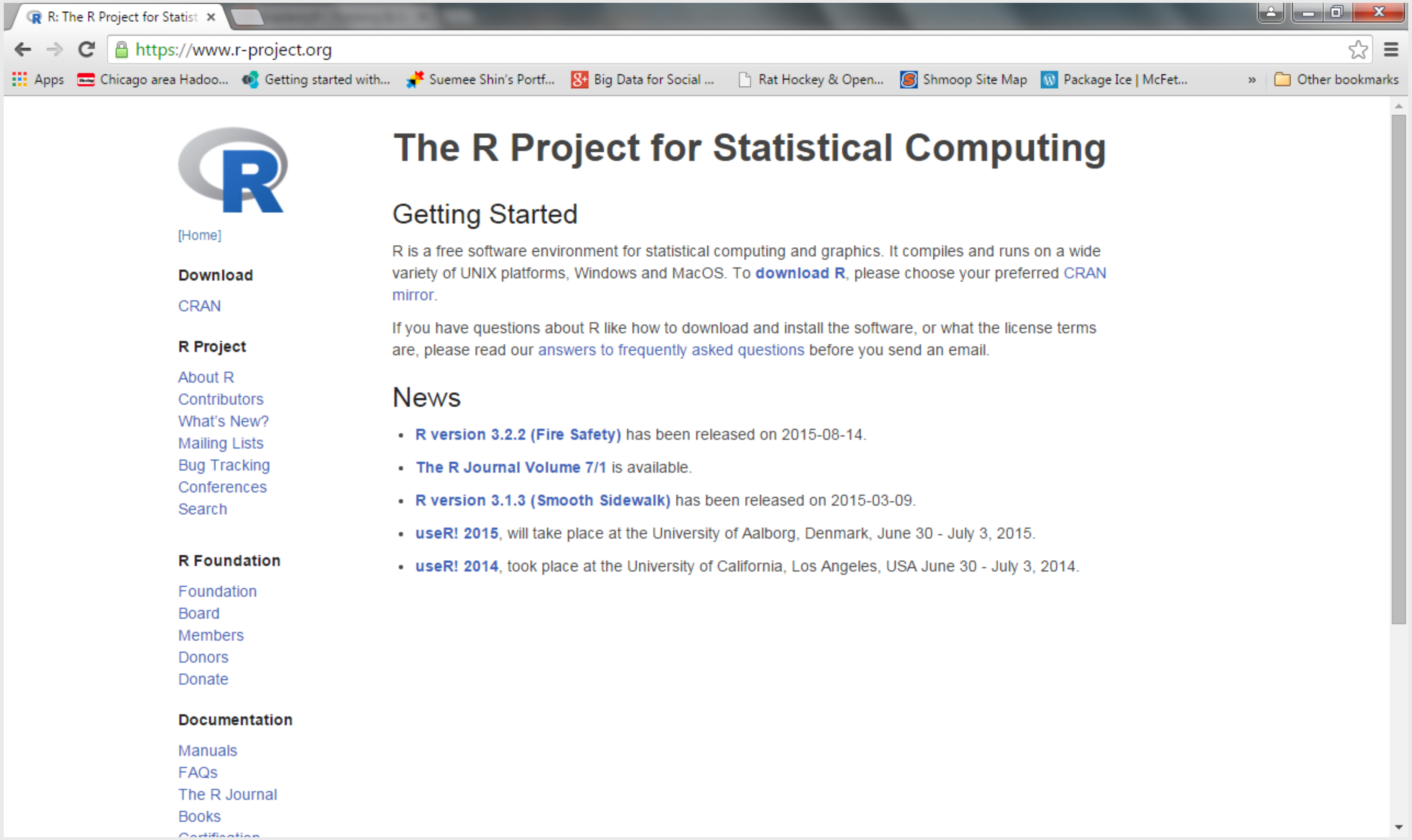
Week 11 Topic:
R

http://en.wikipedia.org/wiki/R_(programming_language):

- *Ross Ihaka and Robert Gentleman* created the open-source language R in 1995 as an implementation of the S programming language. The purpose was to develop a language that focused on delivering a better and more user-friendly way to do data analysis, statistics and graphical models. At first, R was primarily used in academics and research, but lately the enterprise world is discovering R as well. This makes R one of the fastest growing statistical languages in the corporate world.
- One of the main strengths of R is its huge community that provides support through mailing lists, user-contributed documentation and a very active Stack Overflow group. There is also CRAN (http://cran.r-project.org/), a huge repository of curated R packages to which users can easily contribute. These packages are a collection of R functions and data that make it easy to immediately get access to the latest techniques and functionalities without needing to develop everything from scratch yourself.

Week 11 Topic: Introducing R Project

https://www.r-project.org/:



[Home]

Download

CRAN

R Project

About R
Contributors
What's New?
Mailing Lists
Bug Tracking
Conferences
Search

R Foundation

Foundation
Board
Members
Donors
Donate

Documentation

Manuals
FAQs
The R Journal
Books
Certification

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

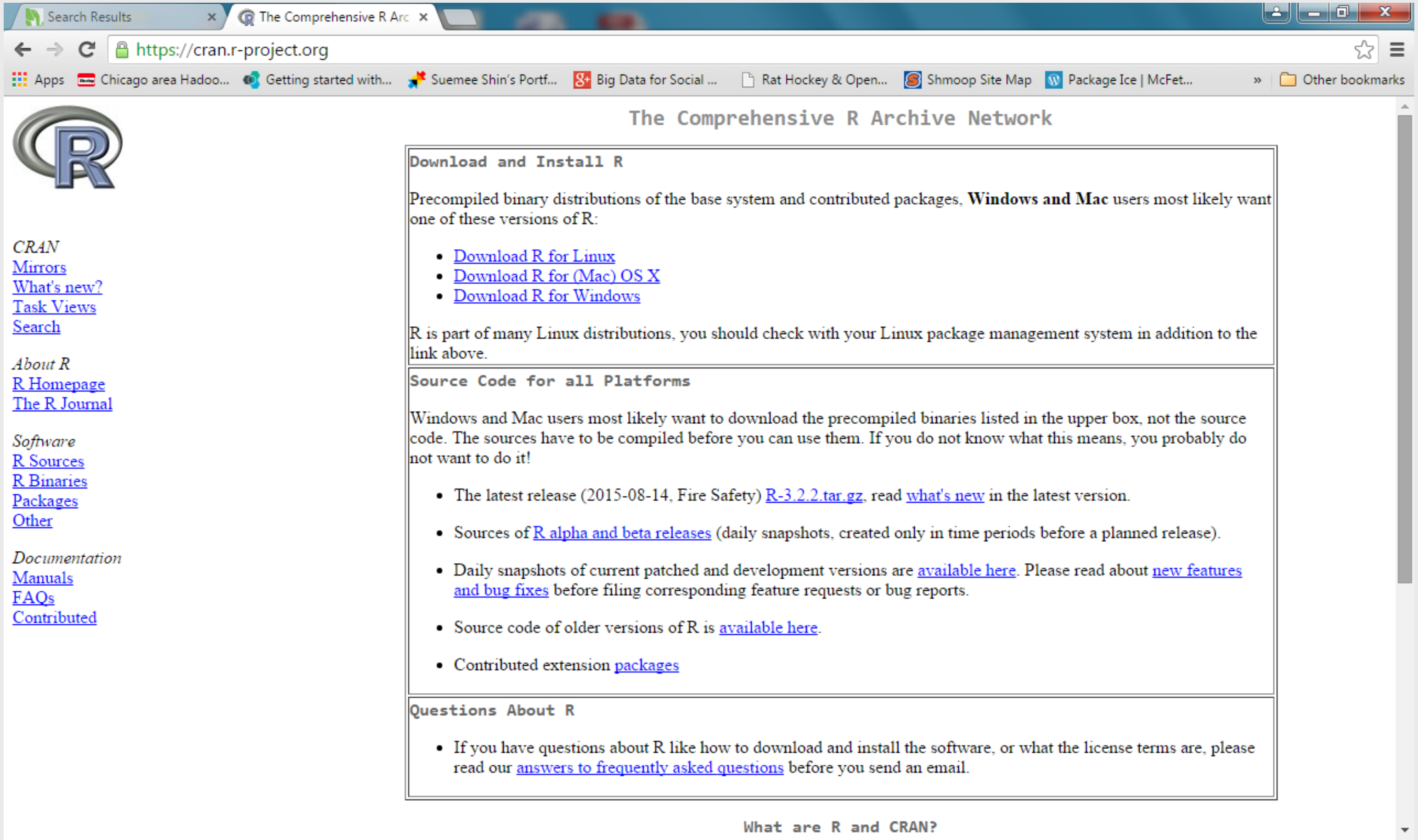
If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- **R version 3.2.2 (Fire Safety)** has been released on 2015-08-14.
- **The R Journal Volume 7/1** is available.
- **R version 3.1.3 (Smooth Sidewalk)** has been released on 2015-03-09.
- **useR! 2015**, will take place at the University of Aalborg, Denmark, June 30 - July 3, 2015.
- **useR! 2014**, took place at the University of California, Los Angeles, USA June 30 - July 3, 2014.

Week 11 Topic: Introducing CRAN

http://cran.r-project.org/



CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2015-08-14, Fire Safety) R-3.2.2.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

What are R and CRAN?

Week 11 Topic: R and CRAN

What are R and CRAN?

- R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc. Please consult the R project homepage for further information.
- CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. Please use the CRAN mirror nearest to you to minimize network load.

Submitting to CRAN

- To "submit" a package to CRAN, check that your submission meets the CRAN Repository Policy and then use the web form.
- If this fails, upload to ftp://CRAN.R-project.org/incoming/ and send an email to CRAN@R-project.org following the policy. Please do not attach submissions to emails, because this will clutter up the mailboxes of half a dozen people.
- Note that we generally do not accept submissions of precompiled binaries due to security reasons. All binary distribution listed above are compiled by selected maintainers, who are in charge for all binaries of their platform, respectively.

Week 11 Topic: Getting started with R

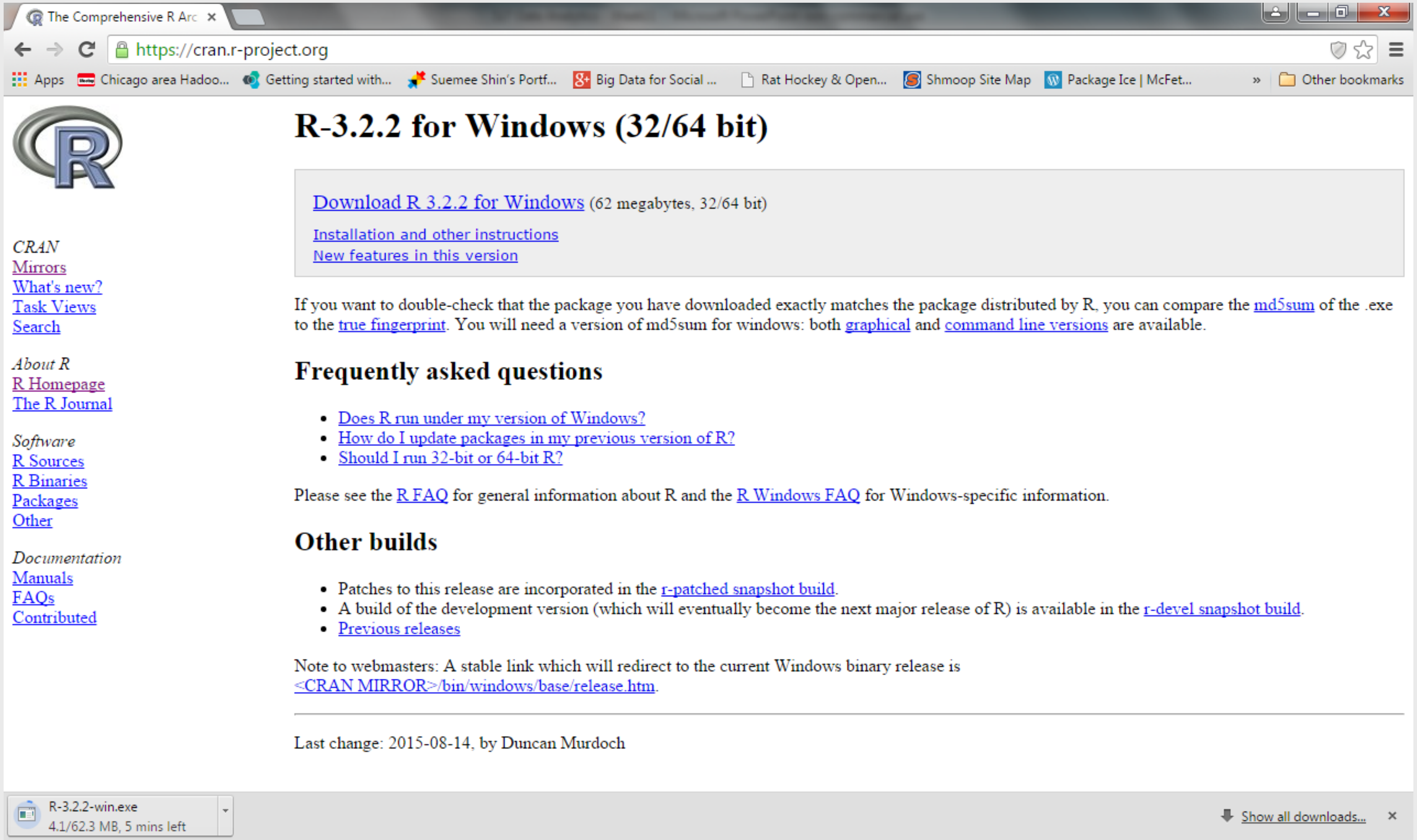
- R is mainly used when the data analysis task requires standalone computing or analysis on individual servers. It's great for exploratory work, and it's handy for almost any type of data analysis because of the huge number of packages and readily usable tests that often provide you with the necessary tools to get up and running quickly. R can even be part of a big data solution.
- When getting started with R, a good first step is to install the amazing RSTUDIO IDE (http://www.rstudio.com/products/rstudio/) . Once this is done, we recommend you to have a look at the following popular packages:
- dplyr, plyr and data.table to easily manipulate packages,
- stringr to manipulate strings,
- zoo to work with regular and irregular time series,
- ggvis, lattice, and ggplot2 to visualize data, and
- caret for machine learning

Week 11 Topic: R Pros and Cons

- ***Pro: A picture says more than a thousands words:*** Visualized data can often be understood more efficiently and effectively than the raw numbers alone. R and visualization are a perfect match. Some must-see visualization packages are ggplot2, ggvis, googleVis and rCharts.
- ***Pro: R ecosystem:*** R has a rich ecosystem of cutting-edge packages and active community. Packages are available at CRAN, BioConductor and Github. You can search through all R packages at Rdocumentation (http://www.rdocumentation.org/).
- ***Pro: R lingua franca of data science:*** R is developed by statisticians for statisticians. They can communicate ideas and concepts through R code and packages, you don't necessarily need a computer science background to get started. Furthermore, it is increasingly adopted outside of academia.
- ***Pro/Con: R is slow:*** R was developed to make the life of statisticians easier, not the life of your computer. Although R can be experienced as slow due to poorly written code, there are multiple packages to improve R's performance: pqR, renjin and FastR, Riposte and many more.
- ***Con: R has a steep learning curve:*** R's learning curve is non-trivial, especially if you come from a GUI for your statistical analysis. Even finding packages can be time consuming if you're not familiar with it.

Week 11 Topic: Installing R

http://cran.r-project.org/



CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

R-3.2.2 for Windows (32/64 bit)

Download R 3.2.2 for Windows (62 megabytes, 32/64 bit)

Installation and other instructions
New features in this version

If you want to double-check that the package you have downloaded exactly matches the package distributed by R, you can compare the md5sum of the .exe to the true fingerprint. You will need a version of md5sum for windows: both graphical and command line versions are available.

Frequently asked questions

- Does R run under my version of Windows?
- How do I update packages in my previous version of R?
- Should I run 32-bit or 64-bit R?

Please see the R FAQ for general information about R and the R Windows FAQ for Windows-specific information.

Other builds

- Patches to this release are incorporated in the r-patched snapshot build.
- A build of the development version (which will eventually become the next major release of R) is available in the r-devel snapshot build.
- Previous releases

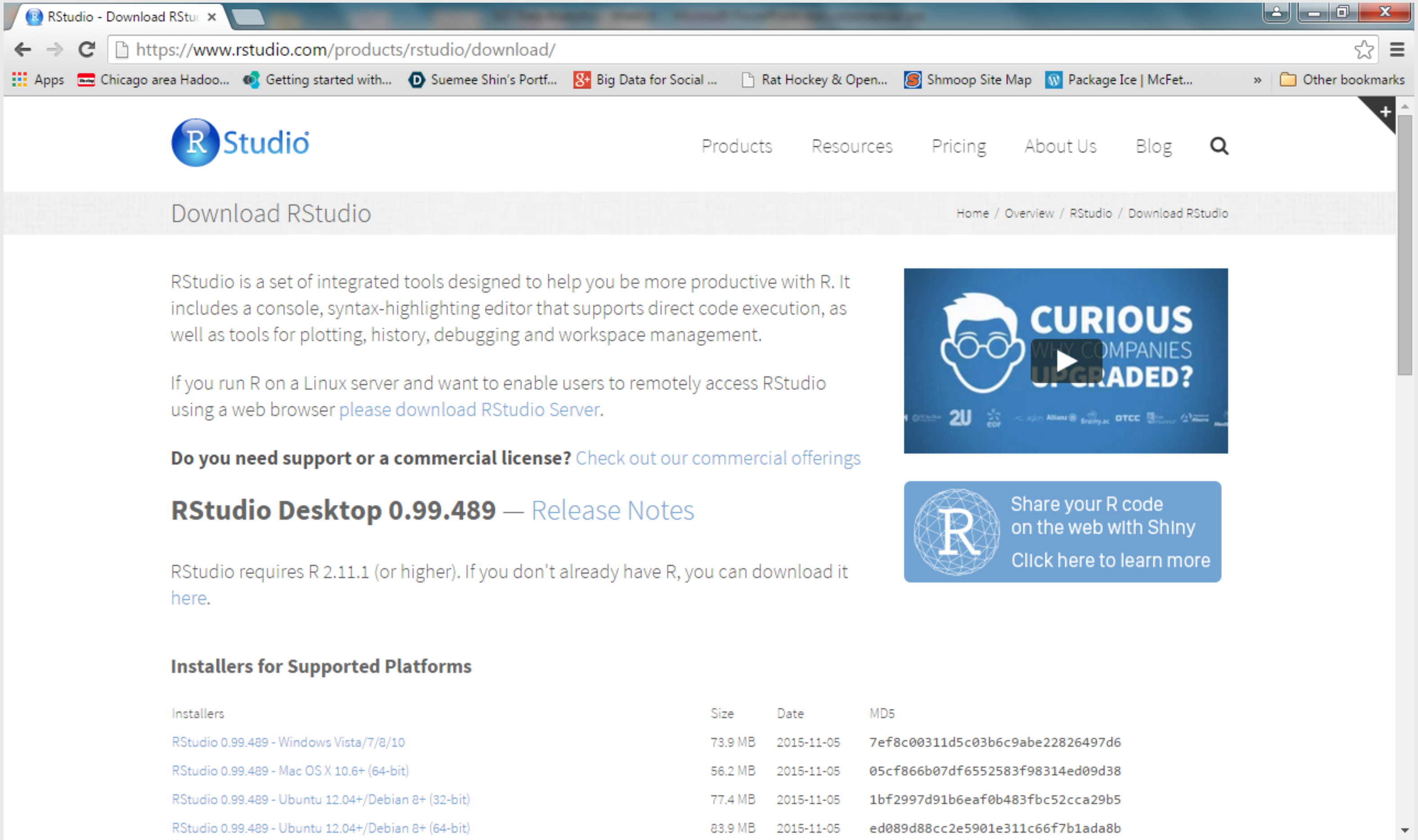
Note to webmasters: A stable link which will redirect to the current Windows binary release is
<CRAN MIRROR>/bin/windows/base/release.htm.

Last change: 2015-08-14, by Duncan Murdoch

Week 11 Topic: (Optional) R Training & References

- Revolution Analytics offers good video training on R: http://www.revolutionanalytics.com/academyr-training-education
- An Introduction to R (In Blackboard)
- Big Data (In Blackboard)

Week 11 Topic: Installing RStudio



Download RStudio

Home / Overview / RStudio / Download RStudio

RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management.

If you run R on a Linux server and want to enable users to remotely access RStudio using a web browser please download RStudio Server.

Do you need support or a commercial license? Check out our commercial offerings

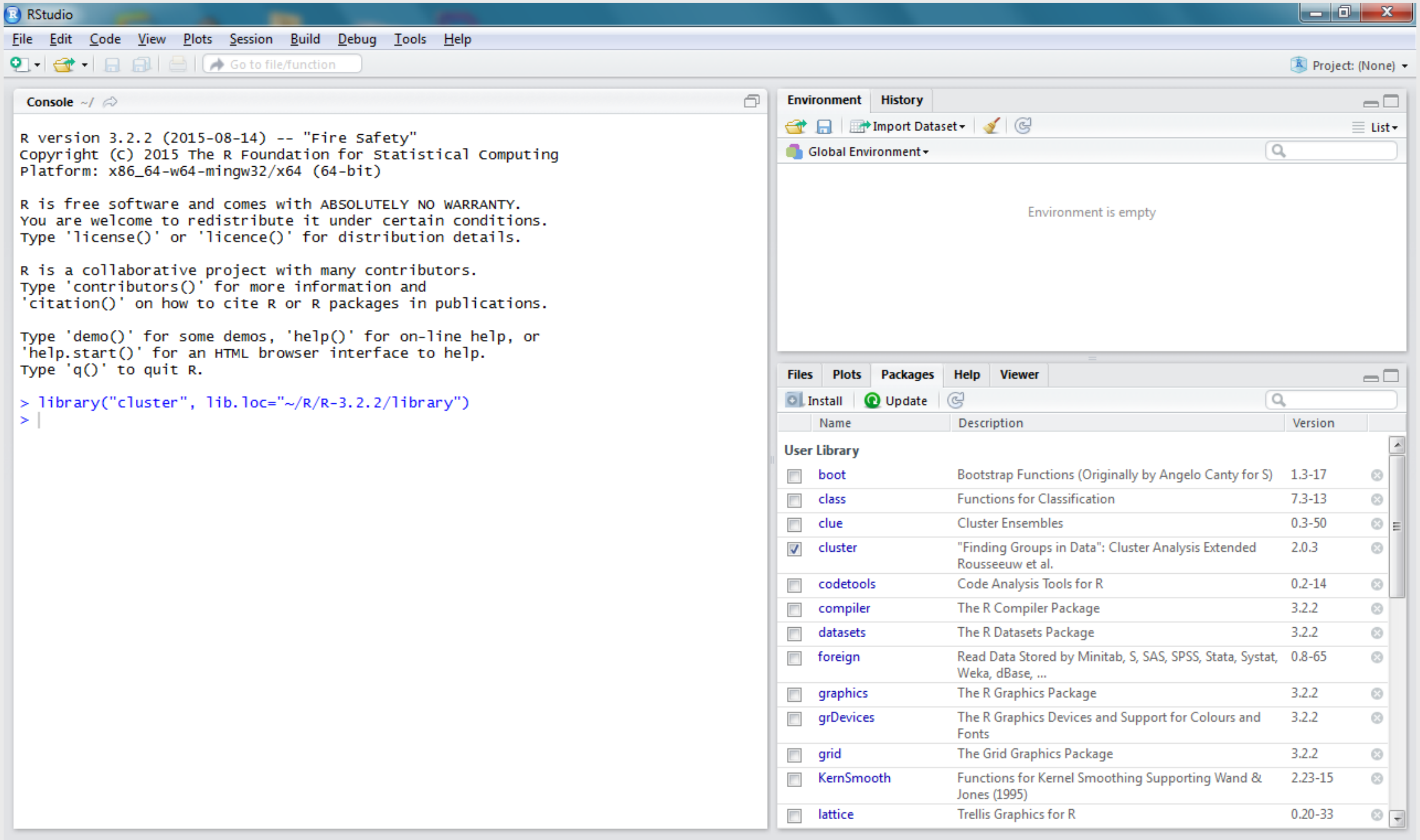
RStudio Desktop 0.99.489 — Release Notes

RStudio requires R 2.11.1 (or higher). If you don't already have R, you can download it here.

Installers for Supported Platforms

Installers	Size	Date	MD5
RStudio 0.99.489 - Windows Vista/7/8/10	73.9 MB	2015-11-05	7ef8c00311d5c03b6c9abe22826497d6
RStudio 0.99.489 - Mac OS X 10.6+ (64-bit)	56.2 MB	2015-11-05	05cf866b07df6552583f98314ed09d38
RStudio 0.99.489 - Ubuntu 12.04+/Debian 8+ (32-bit)	77.4 MB	2015-11-05	1bf2997d91b6eaf0b483fbc52cca29b5
RStudio 0.99.489 - Ubuntu 12.04+/Debian 8+ (64-bit)	83.9 MB	2015-11-05	ed089d88cc2e5901e311c66f7b1ada8b

Week 11 Topic: Using R Packages

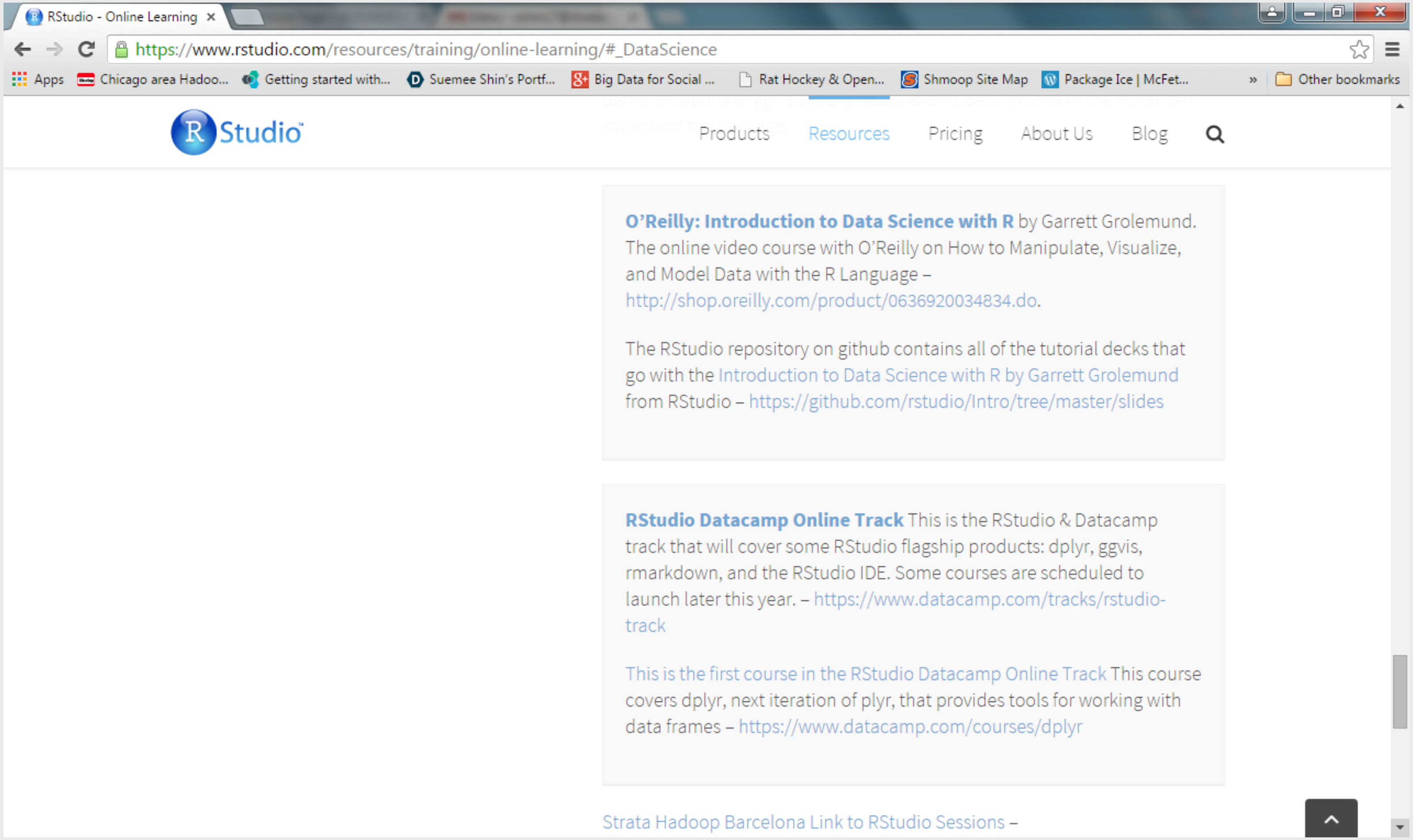


```
R version 3.2.2 (2015-08-14) -- "Fire Safety"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library("cluster", lib.loc="~/R/R-3.2.2/library")
> |
```

User Library

	Name	Description	Version
□	boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-17
□	class	Functions for Classification	7.3-13
□	clue	Cluster Ensembles	0.3-50
☑	cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.	2.0.3
□	codetools	Code Analysis Tools for R	0.2-14
□	compiler	The R Compiler Package	3.2.2
□	datasets	The R Datasets Package	3.2.2
□	foreign	Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ...	0.8-65
□	graphics	The R Graphics Package	3.2.2
□	grDevices	The R Graphics Devices and Support for Colours and Fonts	3.2.2
□	grid	The Grid Graphics Package	3.2.2
□	KernSmooth	Functions for Kernel Smoothing Supporting Wand & Jones (1995)	2.23-15
□	lattice	Trellis Graphics for R	0.20-33

Week 11 Topic: RStudio Resources



O'Reilly: Introduction to Data Science with R by Garrett Grolemund. The online video course with O'Reilly on How to Manipulate, Visualize, and Model Data with the R Language – http://shop.oreilly.com/product/0636920034834.do.

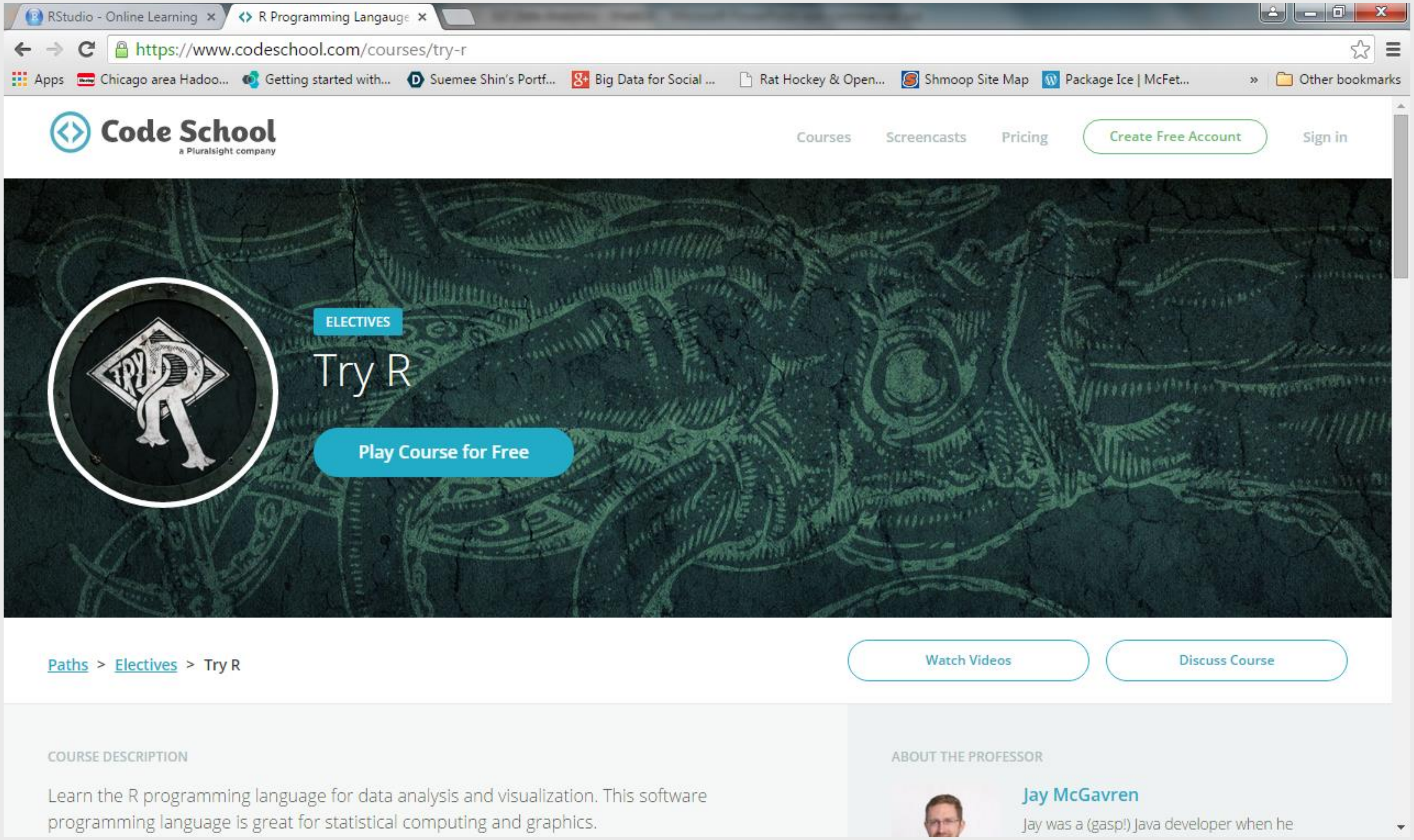
The RStudio repository on github contains all of the tutorial decks that go with the Introduction to Data Science with R by Garrett Grolemund from RStudio – https://github.com/rstudio/Intro/tree/master/slides

RStudio Datacamp Online Track This is the RStudio & Datacamp track that will cover some RStudio flagship products: dplyr, ggvis, rmarkdown, and the RStudio IDE. Some courses are scheduled to launch later this year. – https://www.datacamp.com/tracks/rstudio-track

This is the first course in the RStudio Datacamp Online Track This course covers dplyr, next iteration of plyr, that provides tools for working with data frames – https://www.datacamp.com/courses/dplyr

Strata Hadoop Barcelona Link to RStudio Sessions –

Week 11 Topic: TryR



COURSE DESCRIPTION

Learn the R programming language for data analysis and visualization. This software programming language is great for statistical computing and graphics.

ABOUT THE PROFESSOR

Jay McGavren

Jay was a (gasp!) Java developer when he