

# LAB 1 INSTRUCTIONS

CKME 134 – BIG DATA ANALYTICS TOOLS

RYERSON UNIVERSITY

SPRING 2015

Instructor: Shaohua Zhang

# Assignments

---

- Assignment #1 is optional and is not graded
- Assignment #2 due date extended till March 9

# In-Class Survey

- How many have completed both Hive labs?
- How many students find the material covered too difficult?
- How many students find the pace of lectures fast?
- How many find it too easy?
- How many still have troubles with Hadoop installation?

# Outline

---

- Hadoop Installation
- Regular Expression
- GIS with Hadoop
- Github
- Lab

# Hadoop Installation

DEMO

- Options
  - ▣ HDP Sandbox (Virtualbox)
  - ▣ Vanilla Hadoop on Linux (Ubuntu, CentOS)
  - ▣ HDP on Linux Using Vagrant and Ambari
  - ▣ Vanilla Hadoop on AWS
  - ▣ **Multi-node Cluster Setup AWS**
- *If you're interested in getting Hadoop up and running on AWS, please go home and create an amazon account. I will teach you how to do that in the next lab*

# Regular Expression

- Java Regex Tutorial

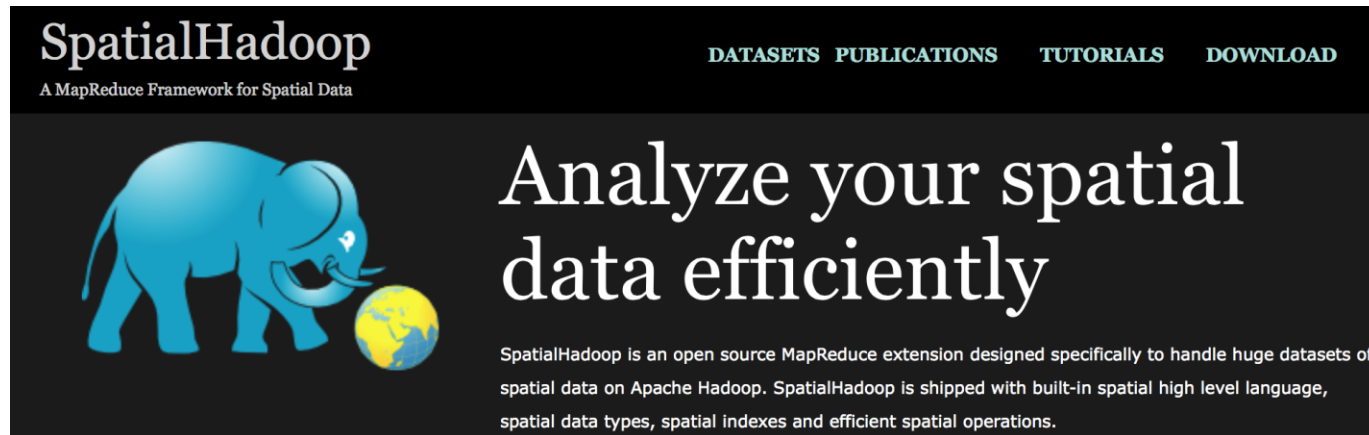
- ▣ <http://www.vogella.com/articles/JavaRegularExpressions/article.html>

- Test your regex with Regex Pal

- ▣ <http://regexpal.com/>

# GIS with Hadoop


- SpatialHadoop ([Link](#)) → VM



SpatialHadoop

A MapReduce Framework for Spatial Data

DATASETS PUBLICATIONS TUTORIALS DOWNLOAD



## Analyze your spatial data efficiently

SpatialHadoop is an open source MapReduce extension designed specifically to handle huge datasets of spatial data on Apache Hadoop. SpatialHadoop is shipped with built-in spatial high level language, spatial data types, spatial indexes and efficient spatial operations.

- GIS Tools for Hadoop by Esri ([Link](#))



## GIS Tools for Hadoop

Big Data Spatial Analytics for the Hadoop Framework



View project on  
GitHub

# GitHub

- Use GitHub for Version Control and start building your Data Science repositories
- Example
  - ▣ <https://github.com/ipython/ipython/wiki/A-gallery-of-interesting-IPython-Notebooks>



[Explore](#) [Gist](#) [Blog](#) [Help](#)



# Today's Lab

- Continue to work on all previous labs
- Lab preparations for future labs
  - ▣ Download SpatialHadoop Virtual Machine Image
    - <http://spatialhadoop.cs.umn.edu/SpatialHadoop-vm-2.2.ova>
- (Optional) Set up single-node Hadoop in Ubuntu Virtual Machine
  - ▣ Download Ubuntu LTS version to your computer ~1 G
  - ▣ Import Ubuntu Linux in Virtualbox
  - ▣ Install Hadoop from Ubuntu
  - ▣ Tutorial on BlackBoard
- (Optional) Register AWS account
- (Optional) Register and Download Github
  - ▣ Download Git – refer to the git book ([link](#))
  - ▣ Register a Github account