Lecture 02 - Data Collection From Web
Before getting into Lecture 02, let us look at the roadmap for next few weeks.

- Last week (1st week), we discussed Cloud Computing paradigm.
- This week's Lab (1st week), you will practice how to provision machines in the Cloud and make them talk to each other.
- In this week's lecture (2nd week), we will take a deeper look at some techniques to collect data from the web, running Multithread computation, and discuss how we could automate provisioning machines in the Cloud.
- Next week's Lab (2nd week), you will practice how to automate provisioning machines in the Cloud to make them talk to each other and practice how to collect data from a web source using Multithread.
Before getting into Lecture 02, let us look at the roadmap for next few weeks.

- In next week lecture (3rd week), we will take a look at ways of storing and representing data using XML, JSON, and MySQL relational databases.
- In 3rd week lab, you will practice how to write programs to store and represent data in XML, JSON, and Mysql databases and how to write procedures to process the data on those storage engines.
- In 4th week lecture, we will take a look at ways of automating data storage in distributed data storage platforms such as Dropbox, and Google Drive.
- In 4th week lab, you will practice how to use the API provided by Dropbox and Google Drive to store your data in different distributed storage platforms.
- Let us plan the course organization as we move forward.
So Why is it important to collect Data?

- To perform distributed computation, we need data.
- No big data no distributed system
- Where is big data?
Web Scraping

- Web Scraping is data scraping used for extracting data from websites.
- Automate data scraping procedures using a bot.
- Data copied from the web and stored in a separate repository for later retrieval or analysis.
Basic HTML

- Look at the PDF in course web page
Jsoup Library

- Jsoup is a Java library for working with real-world HTML. It provides a very convenient API for extracting and manipulating data from the Web.
- Reading: Look at the PDF in course web page
Let us look at some code to collect web data
Class activity: to discuss on identifying data sources for web scraping
Questions