WHAT IS BIG DATA to you?
**SO WHAT IS A PETABYTE ANYWAY?**

Source – www.mozy.com

**WHAT IS A PETABYTE?**

To understand a petabyte we must first understand a gigabyte.

- 1 petabyte = 7 minutes of HD-TV video
- 2 petabytes = 20 yards of books on a shelf
- 4.7 petabytes = size of a standard DVD-R
- There are a million gigabytes in a petabyte

"Let me repeat that: we create as much information in two days now as we did from the dawn of man through 2003." (That’s something like 5 Exabytes of Data). - Eric Schmidt – Google 8/10

**A PETABYTE IS A LOT OF DATA**

- 1 petabyte = 20 million four-drawer filing cabinets filled with text
- 1 petabyte = 13.3 years of HD-TV video
- 1.5 petabytes = size of the 10 billion photos on Facebook
- 15+ petabytes = internet user’s data backed up on Mozy.com
- 20 petabytes = the amount of data processed by Google per day
- 20 petabytes = total hard drive space manufactured in 1995
- 50 petabytes = The entire written works of mankind, from the beginning of recorded history, in all languages

**A ZETABYTE IS ONE MILLION PETabytes!**

Facebook:
More that 750 Million Users.
Average user creates 90 Pieces of content each month.
More than 30B pieces of content shared each month.
What Brings You Here?
A QUICK INTRODUCTION OF ME

http://infosense.cs.georgetown.edu/
WHAT BRINGS ME HERE

• To have a wonderful semester with you :-)

• As an educator, I want to teach a class that is timely and useful, helping you in the job market

• For myself, to be honest, I am not a big fan of huge data volume. However, big data not just means bigger volume, it also means higher data variety and faster data change rate (velocity)

• I am a fan of complexity. ;-)
WHAT BRINGS ME HERE

BIG DATA RELATES TO EVERYONE
COURSE PURPOSE

• able to code and design large scale data analytics tools

• master spark programming

• understand how web search engine works

• focus on text analytics, but the techniques we will learn are generic

• have fun!
LET’S START TO THINK
IF YOU ARE THE GOD OF DATA

• What are the typical uses of your data?
  • understand trends and patterns
  • prediction
  • search
HOW GOOGLE WORKS
IF YOU ARE THE GOD OF DATA

• What will be the challenges/problems when your data is big?

• What is your solution?
  • divide-and-conquer
  • parallelization
  • compression

“Really? — my people always say multiply and conquer.”
MAP REDUCE

Map Work Shuffle Reduce
**MAPREDUCE IS A LITTLE BIT OUTDATED**

- It is great at one-pass computation

- but not efficient enough for multiple-pass algorithms
  - things that require repeatedly hashing or other operations

- states go to file systems
  - a lot of I/Os

- slow
SPARK

• Key idea:
  • Load things in the memory
  • Resilient Distributed Datasets (RDDs)
  • Clean APIs in Java, Scala, Python, R
  • not for c++
  • We will learn Scala
COURSE PLAN

• Key topics
  • Spark
  • Web search engine

• September - Spark Essentials

• October - Text Processing and Basic Search Engine

• November - PageRank and Web search

• December - Other Apps (Recommender Systems, Dynamic Search, Social Search)
“A homework is worth a thousand lectures.”

–GRACE HUI YANG
ASSIGNMENTS

• Build Google’s pagerank algorithm over Wikipedia
• A big project broken into small pieces
• (nearly) weekly
• 10 + 1 of them
• (almost) all due on some Wednesday 11:59PM
Syllabus

1. A B
   C
2. A
   B
   C
3. A B
HIGHLIGHT OF TODAY

• Install Spark
• First Spark Program
SPARK INSTALLATION

• Note: Please do not install/run Spark using:
  • Homebrew on MacOSX
  • Cygwin on Windows
STEP 1 - INSTALL JAVA JDK 6/7 ON MACOSX OR WINDOWS

• oracle.com/technetwork/java/javase/downloads/jdk7-downloads-1880260.html

• follow the license agreement instructions

• then click the download for your OS

• need JDK instead of JRE (for Maven, etc.)
STEP 2: GET SPARK

- We will use Spark 1.1.0
- 1. copy from the USB sticks
- 2. connect into the newly created directory
- or you could download from spark.apache.org/downloads.html
STEP 3: RUN SPARK SHELL

• we’ll run Spark’s interactive shell…

• within the “spark” directory, run:

  • ./bin/spark-shell

• then from the “scala>” prompt,

• let’s create some data…

• val data = 1 to 10000
STEP 4: CREATE AN RDD

- create an RDD based on that data…
- val distData = sc.parallelize(data)
- then use a filter to select values less than 20…
- distData.filter(_ < 20).collect()
ASSIGNMENT 1

• Use a filter to select values less than your age

• Submit the screen captures of the results of above programs
SUMMARY

- big data
- spark
- search engine
- syllabus
- installation of spark
- assignment 1
- due next Wednesday