INTRODUCTION TO BIG DATA ANALYTICS TOOLS

CKME 134 – BIG DATA ANALYTICS TOOLS

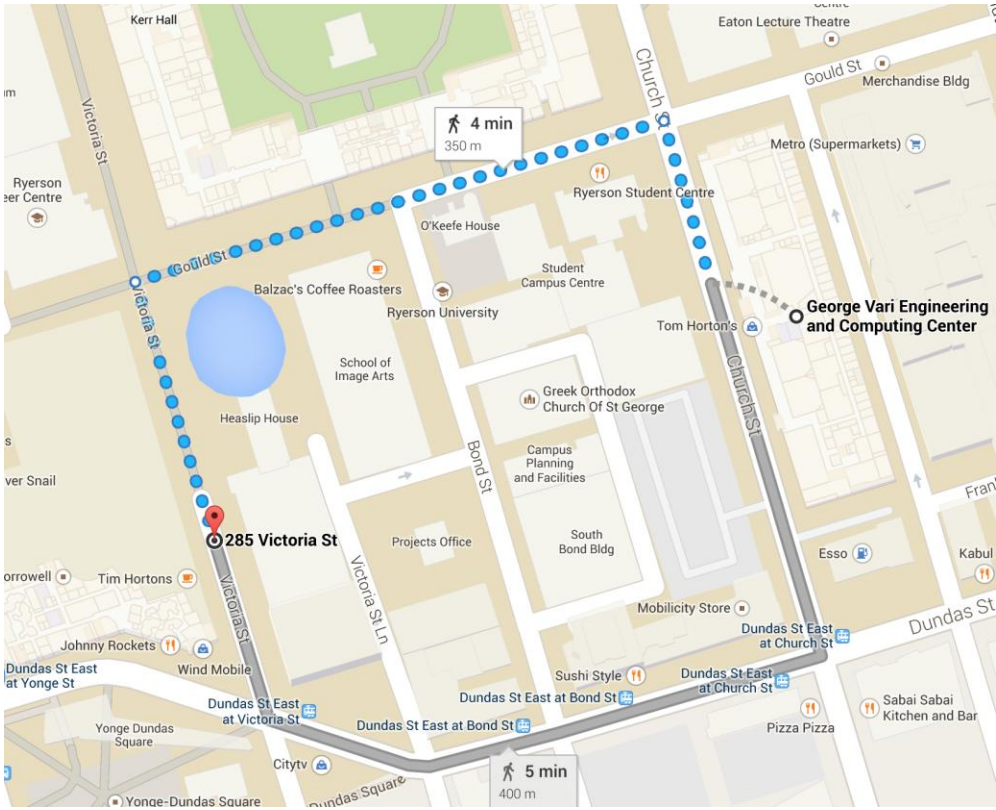
RYERSON UNIVERSITY

SPRING 2015

Instructor: Shaohua Zhang

General Course Information

- Instructor
  - Shaohua Zhang
  - Ryerson shaohua.zhang@ryerson.ca
  - Personal shaohua.zhang@live.com
- GA
  - Behjat Soltanifar
  - behjat.soltanifar@ryerson.ca
- Lectures
  - 6:30~8:30
  - ENGLG06
- Lab
  - 8:30~9:30
  - 285 Victoria St (403/404)
    - Take the elevator to 4FL



About This Course

- This course focuses on practicality
  - It follows industry trends and job market trends
  - You'll be hands-on with several popular tools
  - You'll learn practical use cases and how to choose the right tools
  - You'll learn enough to be able to extend on your own after this course
- This course teaches big data tools related to analytics
  - Will focus less on Infrastructure, ETL and BI
- It mainly focuses on batch processing tools
  - Will introduce streaming processing
- Lab work will be done via virtual machines

Course Outline (*subject to change*)

1. Intro to Big Data
2. Distributed Computing and MapReduce
3. Hadoop Ecosystem
4. Intro to Hive
5. Pig
6. Advanced Pig
7. Hadoop Performance Optimization
8. Big Data Use Cases: Location Intelligence and Marketing Analytics
9. Big Data Use Cases: Recommendation Engine and Computational Advertising
10. Hadoop In Action: Building Data Pipelines
11. Beyond Hadoop: Spark
12. Beyond Hadoop: Real-Time Analytics

Course Project/Assignments

- Analysis of Geo-tagged Tweets
  - Twitter data analysis
  - Simple NLP
  - Location clustering
  - User segmentation
  - Census demographic

Exams

- No mid-term (probably)
- Final Exam
  - Hands-on lab test
  - Apply everything you've learned to complete a small real-life task
    - Labs and assignments are important

Survey

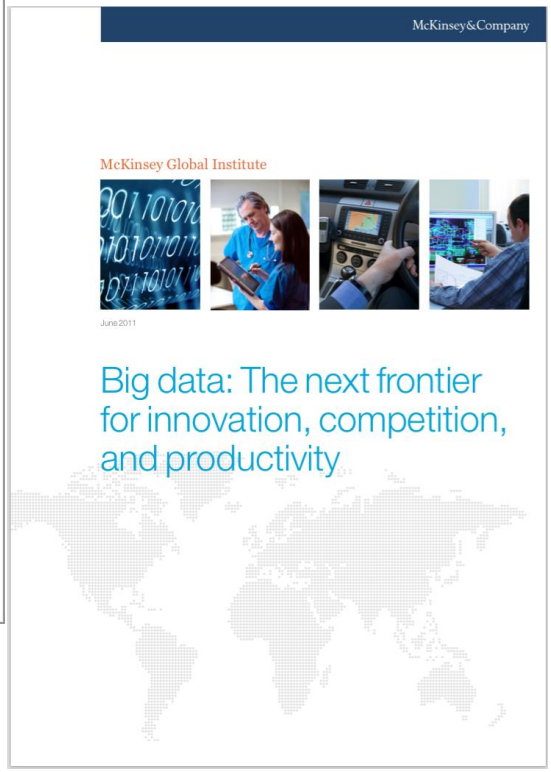
- Self-introduction
  - How many of you are already data analysts?
  - What background do you come from?
  - What kind of programming languages you know?
  - Size of data you worked with?
- Why do you take this course?
  - Curious about big data; want to gain knowledge
  - Starting career
  - Switching career

Lecture 1- Outline

1. Big Data Introduction
2. Big Data Use Cases
3. Data Analytics Tooling
4. Big Data Job Market
5. Big Data Challenges

1. Intro to Big Data

Big Data Is Hot!



Data Scientist:

The Sexiest Job of the 21st Century

Big Data – Why Now?

- Data at scale (Volume)
  - Since when was 1TB not big data any more :-(
- Speed (Velocity)
  - Near Realtime response is key to the modern web/mobile experience
- Data in many forms (Variety)
  - Structured
  - Unstructured
    - Location
    - Text
    - Image
    - Video
  - Semi-structured
    - Graph

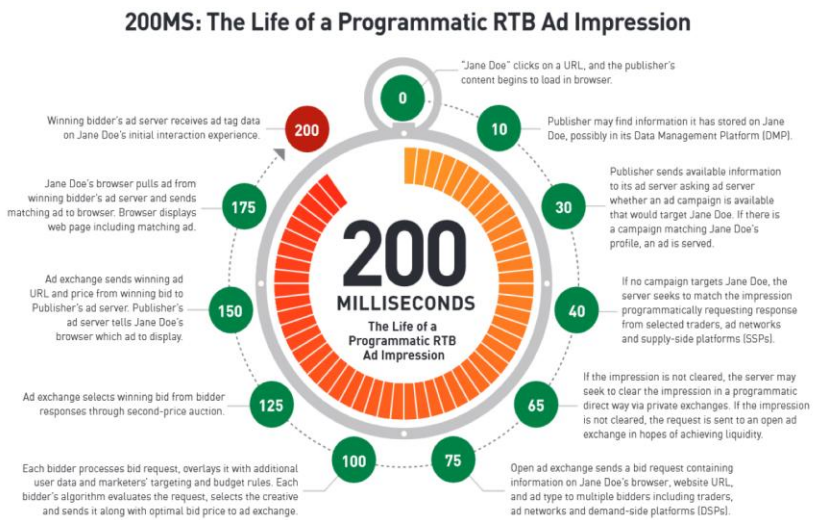
• *Internet*	○ *2.5 exabytes ($2.5x10^{18}$) per day – 2012* ○ *2.3 zettabytes ($2.3x10^{21}$) per day - 2014*
• *Facebook*	○ *500+ terabytes per day* ○ *100+ petabytes in a single Hadoop cluster*

"More data cross the internet every second than were stored in the entire internet just 20 years ago" - Big Data: The Management Review (HBR)

Four V's of Big Data: http://www.ibmbigdatahub.com/infographic/four-vs-big-data
Big Data (wiki): http://en.wikipedia.org/wiki/Big_data

Big Data – Why Now?

- Data at scale (Volume)
  - Since when was 1TB not big data any more :-(
- Speed (Velocity)
  - Near Realtime response is key to the modern web/mobile experience
- Data in many forms (Variety)
  - Structured
  - Unstructured
    - Location
    - Text
    - Image
    - Video
  - Semi-structured
    - Graph

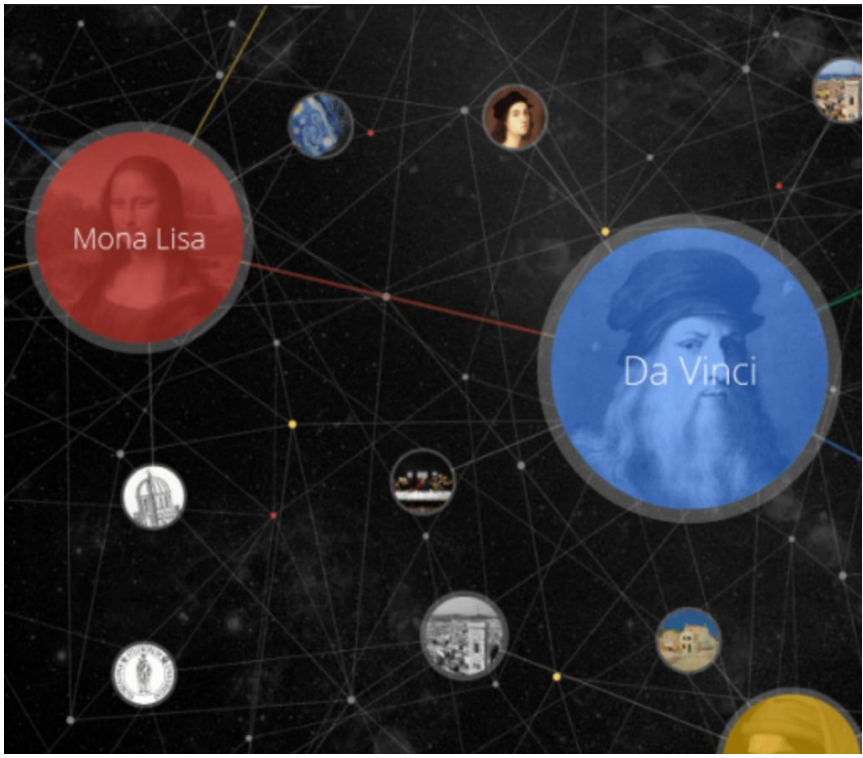


video

Four V's of Big Data: http://www.ibmbigdatahub.com/infographic/four-vs-big-data
200MS: The Life of a Programmatic RTB Ad Impression: http://bit.ly/1iPqAlt

Big Data – Why Now?

- Data at scale (Volume)
  - Since when was 1TB not big data any more :-(
- Speed (Velocity)
  - Near Realtime response is key to the modern web/mobile experience
- Data in many forms (Variety)
  - Structured
  - Unstructured
    - Location
    - Text
    - Image
    - Video
  - Semi-structured
    - Graph



Four V's of Big Data: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Applications Driving the Need for Big Data

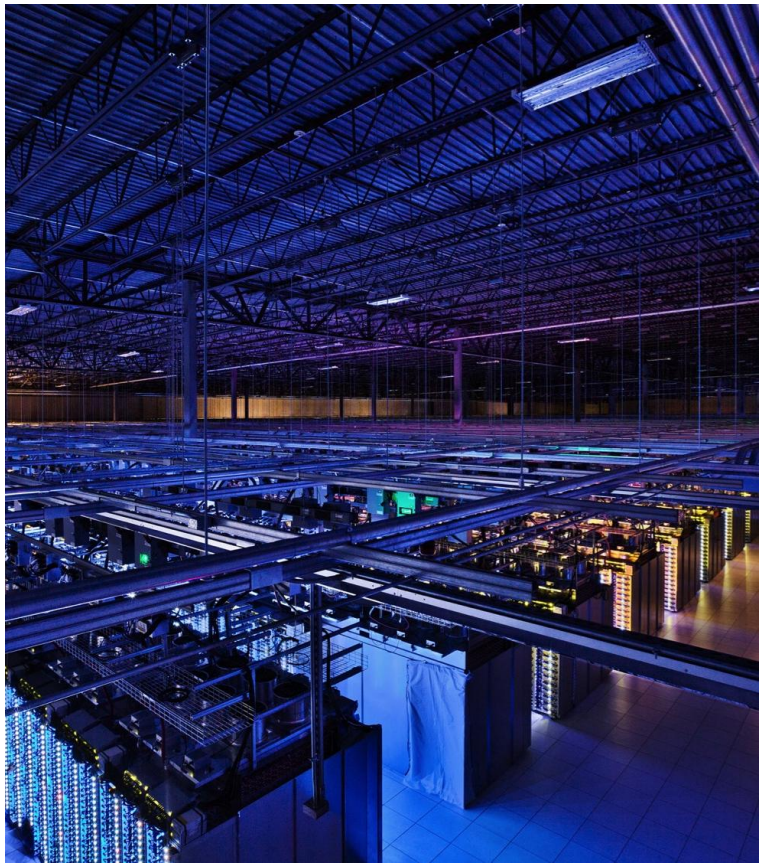
- Data-driven Applications
  - Location-based services
  - Social media apps
  - Image/voice recognition
  - Personal Assistant
  - Advertising
  - Internet of Things
- Internet Economies
  - Monetization needs - data is the new oil

Communities Driving the Need for Big Data

- Hadoop vendors
  - Cloudera
  - Hortonworks (public)
  - MapR
- Traditional Vendors
  - Oracle
  - SAS
  - IBM
  - Revolution Analytics
- Open Source Communities
  - RHadoop
  - RapidMiner
  - NoSQL

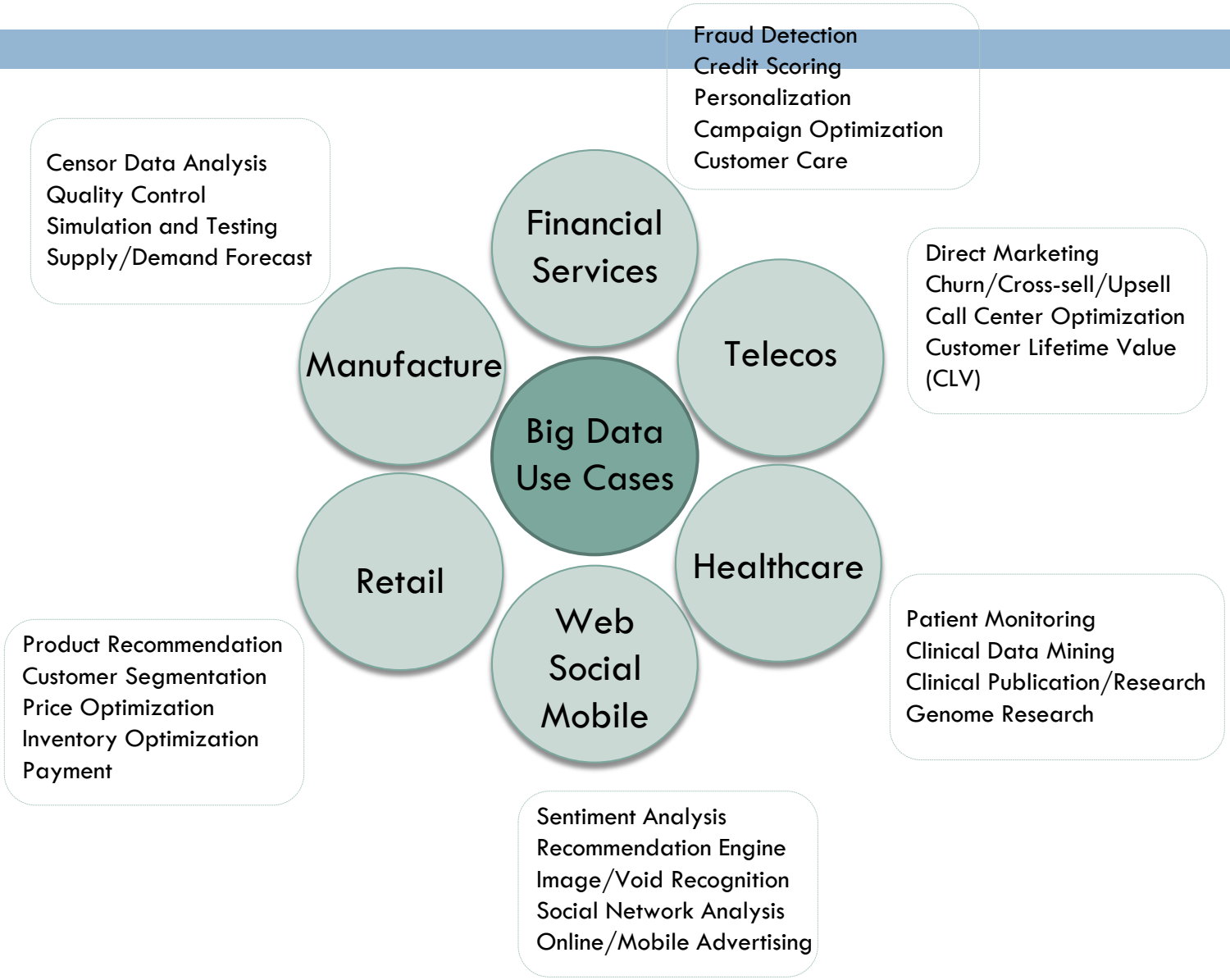
Big Data Made Possible

- Hardware
  - Big cluster of commodity machines at lower cost
    - Faster processor
    - Cheaper memory
    - Bigger hard drive space
    - Faster network bandwidth
- Software
  - Algorithms to allow parallel computing (map-reduce)



2. Big Data Use Cases

Big Data Use Cases



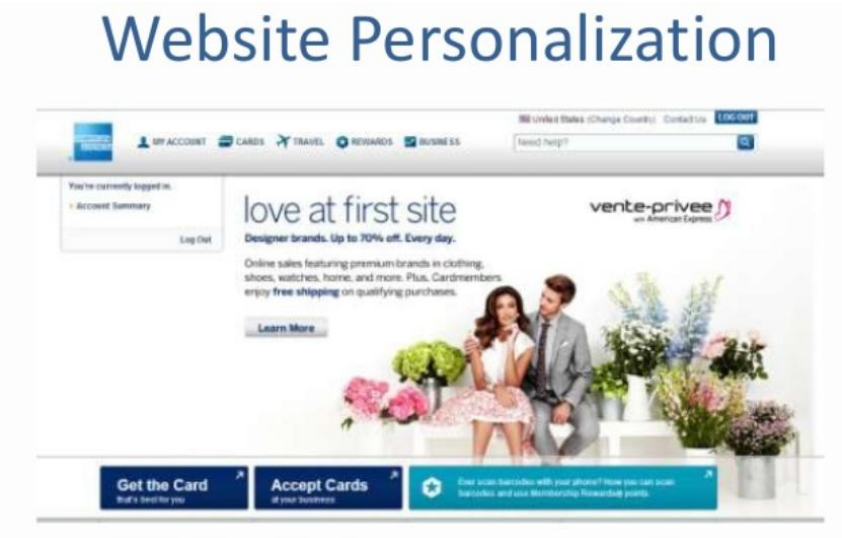
Big Data Use Case – Search & Media

- Google
  - Original MapReduce paper 2004
  - Search & Advertising
  - Image Recognition
  - Google Voice
  - Etc.
- Yahoo!
  - Hadoop 2005 (Doug Cutting)
  - Page personalization
  - Flickr Image Recognition
  - Advertising

Google MapReduce Paper 2004: http://bit.ly/1ADYtjC
Google, Stanford build hybrid neural networks that can explain photos: http://bit.ly/11owuZ2

Big Data Use Case - Banking

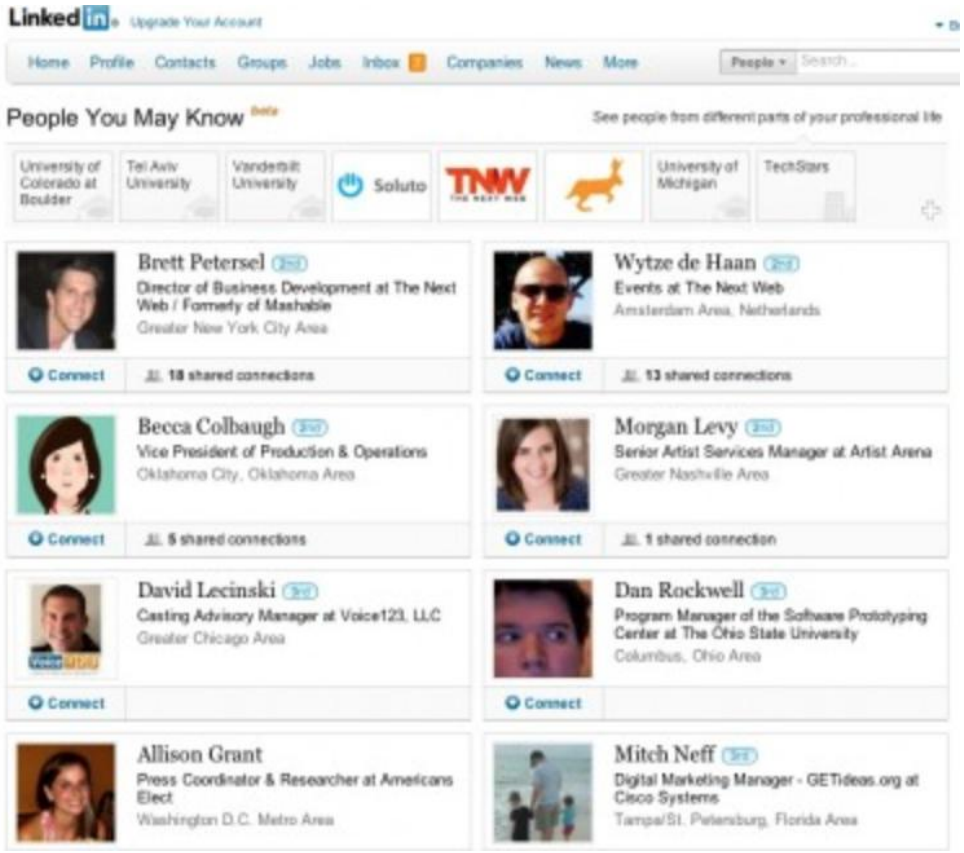
- Hadoop is the new backbone of American Express
  - Recommender systems
  - Graph algorithms
  - Machine learning for Fraud and Marketing
  - Data products
  - Experiments



Why Hadoop Is The New Backbone Of American Express : http://www.enterprisetech.com/2014/10/17/hadoop-new-backbone-american-express

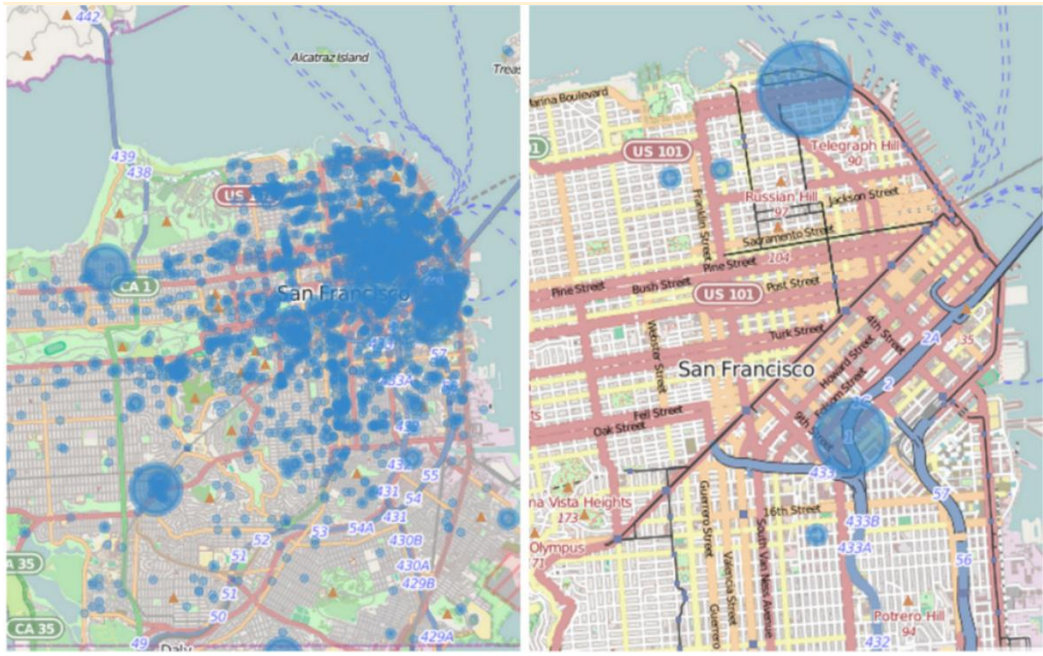
Big Data Use Case – Social Media

- Facebook (designed Hive, Giraph)
  - New Feeds
  - Friend recommendation
  - Ads
  - Graph Search
- Twitter
  - Follower recommendation
  - Tweet search
  - Timeline
- LinkedIn
  - PYMK
  - Job recommendation



Big Data Use Case – LBS

- Foursquare
  - Location Recommendation
  - Local Search
  - Location-based Social Network
  - Location-based Interest
- Yelp
  - Sentiment Analysis
  - POI Recommendation
  - Text Classification
  - Personalized Star Rating
- Uber
  - Trip Prediction
  - Location-based User Segmentation
  - Location-based Demographic Prediction



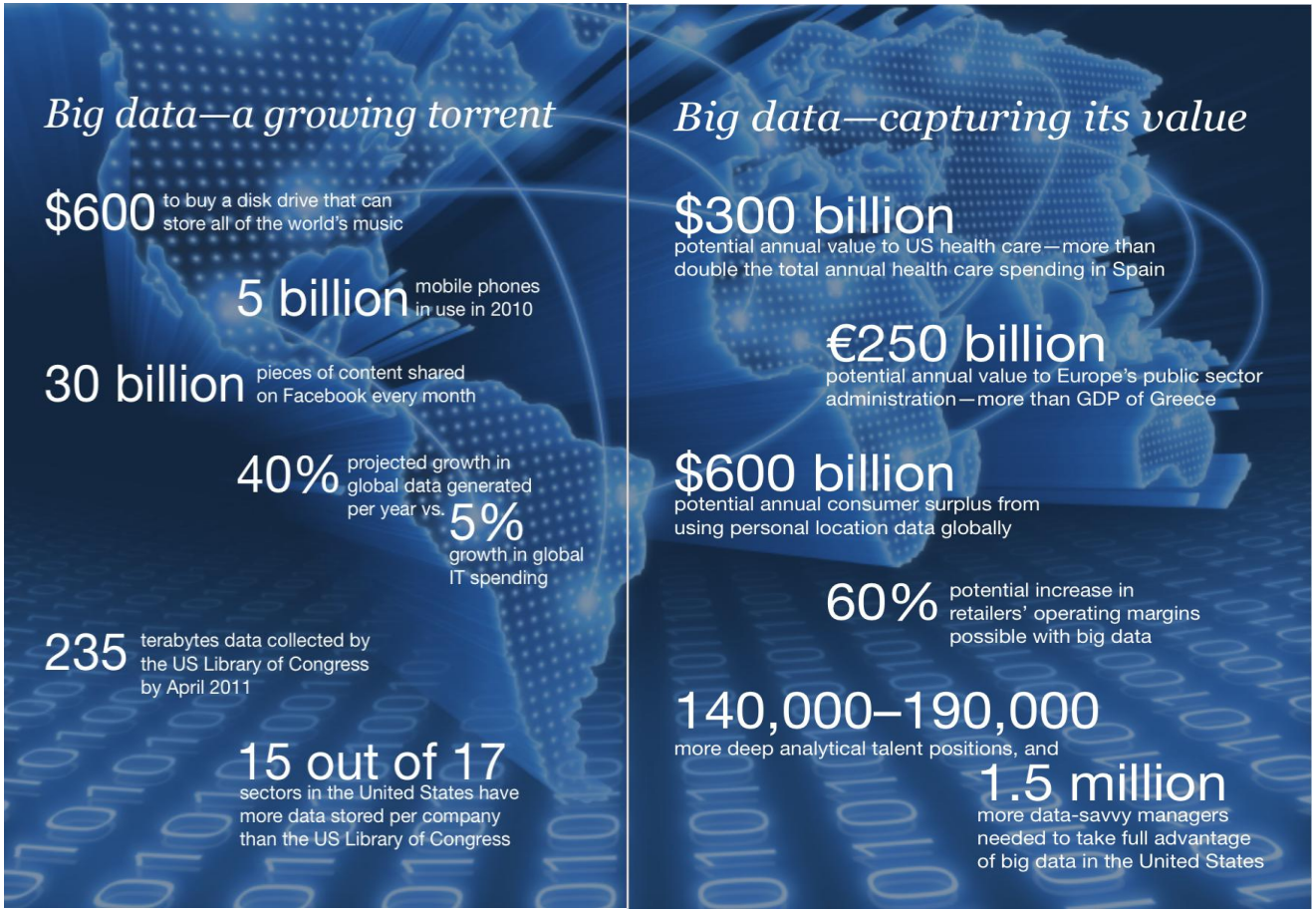
Big Data Use Case – E-Commerce

- Amazon
  - Product Recommendation
    - People who bought this also …
  - Fire Phone Image Recognition
- Ebay
  - Product Tagging
  - User Taste/Interest Graph
  - Fraud Detection
  - Personalization

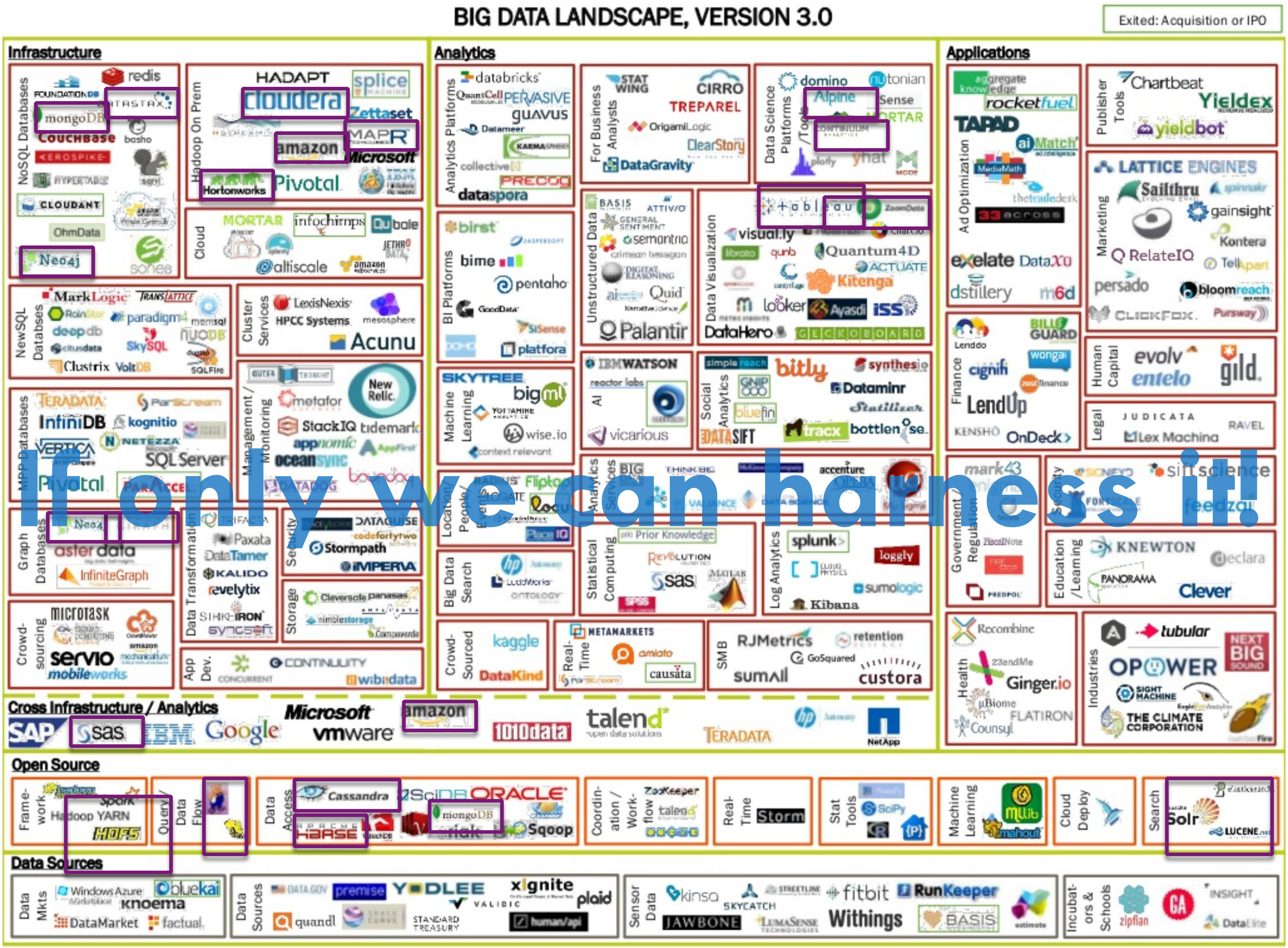
3. Big Data & Analytics Tooling

Big Data – Exciting Future

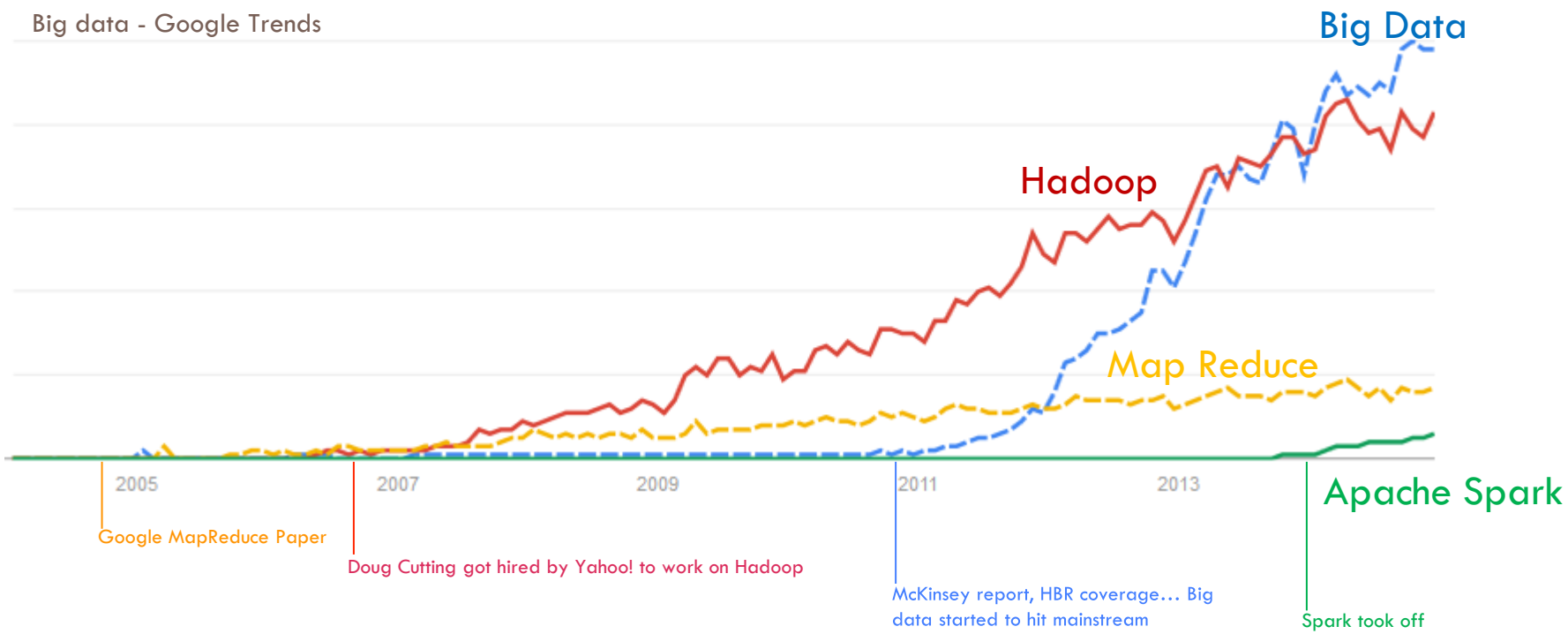
Great! Cool! Promising! Exciting!



Big data: The next frontier for innovation, competition, and productivity: http://bit.ly/1pCOqom


BIG DATA LANDSCAPE, VERSION 3.0
Exited: Acquisition or IPO
Infrastructure
Analytics
Applications
NoSQL Databases
Hadoop On Prem
Cloud
NewSQL Databases
Cluster Services
MPP Databases
Management / Monitoring
Graph Databases
Data Transformation
Security
Crowd-sourcing
Storage
App Dev.
Analytics Platforms
For Business Analysts
Data Science Platforms / Tools
BI Platforms
Unstructured Data
Data Visualization
Machine Learning
AI
Social Analytics
Location / People / Events
Statistical Computing
Log Analytics
Big Data Search
Crowd-Sourced
Real-Time
SMB
Ad Optimization
Publisher Tools
Marketing
Finance
Human Capital
Legal
Government / Regulation
Security
Education / Learning
Health
Industries
Cross Infrastructure / Analytics
Open Source
Framework
Query
Data Flow
Data Access
Coordination / Workflow
Real-Time
Stat Tools
Machine Learning
Cloud Deploy
Search
Data Sources
Data Mkts
Data Sources
Sensor Data
Incubators & Schools
If only we can harness it!

Big Data History



Google MapReduce Paper: http://research.google.com/archive/mapreduce.html
Big data: The next frontier for innovation, competition, and productivity: http://bit.ly/1pCOqom

Big Data Landscape - Simplified

	Open Source	Commercial	Comments
Big Data Platform	Hadoop (MR, Pig, Hive etc.)	Cloudera, Hortonworks, MapR	Hadoop is going mainstream
	Spark	DataStax	Spark is HOT! considered as next-generation big data platform
		AWS	Elastic MapReduce (EMR) and EC2 from AWS is most popular among startups.
Machine Learning & Statistical Learning	Mahout		Mahout was one of the earliest ML libraries for MapReduce. It is being revamped to take advantage of Spark currently
	MLlib (Spark)		MLlib is Spark's machline learning library. It's written in Scala and also provides Python and Java API
	H2O		H2O is the latest buzzing big data machine learning tool, backed by 0xdata. It works with a Hadoop cluster but also works on Standalone cluster. It has an amazing lineup of algorithms and even supports Deep Learning. The GUI-based predictive analytics suites works like a charm
		SAS	SAS integration with Hadoop will be very powerful. Imaging writing your data steps that runs procedures on hadoop
		Revolution Analytics	Commercial version of open source R. Enterprise-class big data analytics capability
		Alpine	World's first code-free in-cluster web analytics platform to analyze big data and hadoop

Big Data Landscape - Simplified

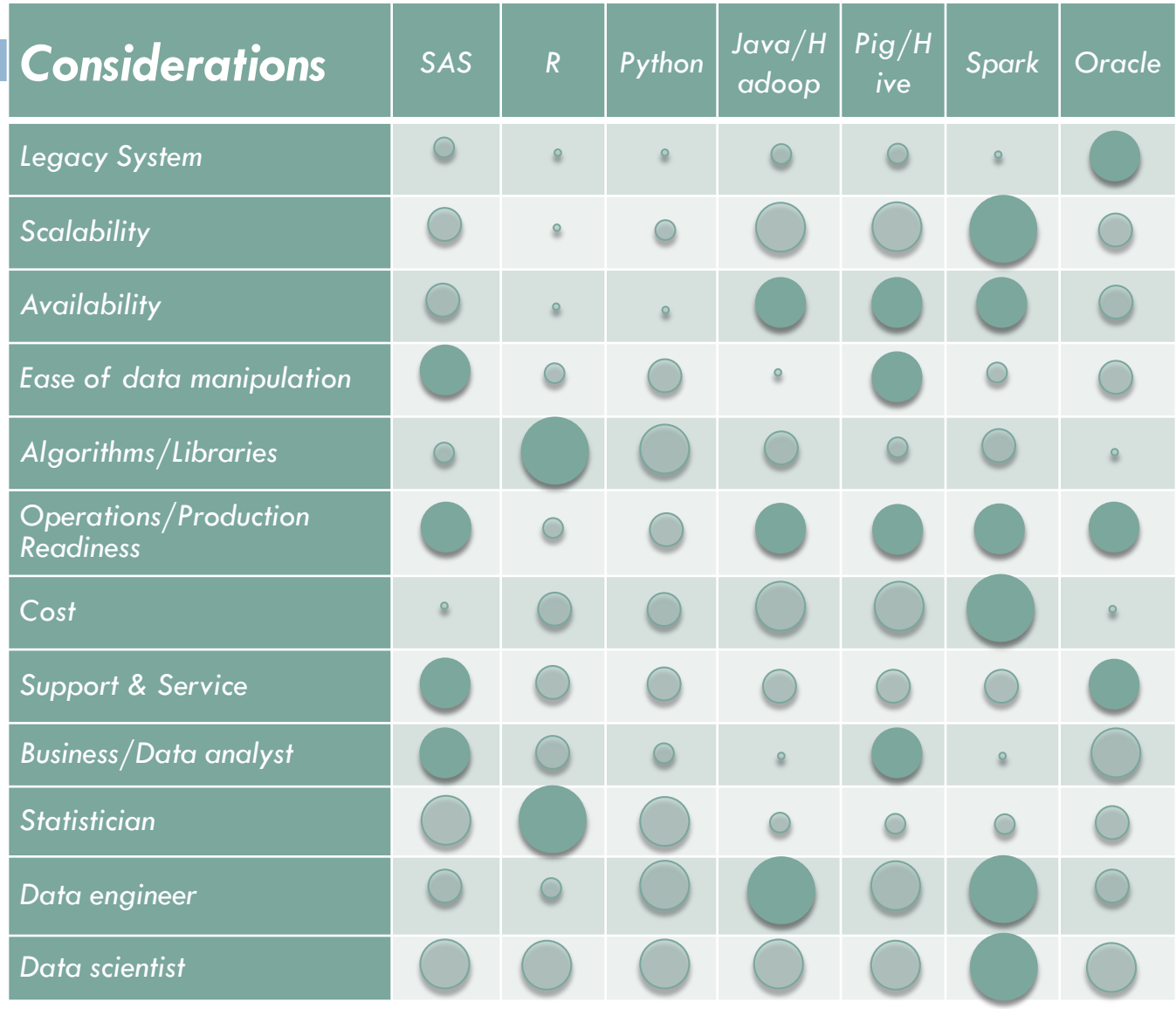
	Open Source	Commercial	Comments
Graph Processing	Giraph		Graph processing framework on top of Hadoop. Used extensively at Facebook for large-scale graph algorithms
	GraphLab		Developed at CMU by Dr. Carlos and his team. Superior graph processing performance. Building the tools to make data scientists' lives easier. Great as a standalone graph processing and machine learning tool but won't fit well into the existing hadoop cluster
	GraphX (Spark)		Graph processing on Spark platform
Search	Solr		Open source search server based on Lucene Java library
	Elastic Search		Open source search and analytics engine
Stream Processing	Storm		Real-time stream processing framework developed at Twitter. Most popular streaming processing tool
	Spark Streaming		Streaming processing on Spark. Less mature than Storm at the moment but growing rapidly
Visualization	d3.js		Fantastic javascript library for visualization
		Tableu, Qlikview, Zoomdata	Popular visualization tools widely adopted
	Kibana		Log and time series data visualization tool from Elasticsearch

Big Data Analytics Tooling

- Choosing the right tools – considerations
  - Legacy systems
  - Data scientists/engineers preference
  - Scalability
  - Availability
  - Data manipulation capability
  - Algorithms/libraries supported
  - Operations (use in production)
  - Cost
  - Industry/vertical standards
    - Security
    - Support and service
  - University programs

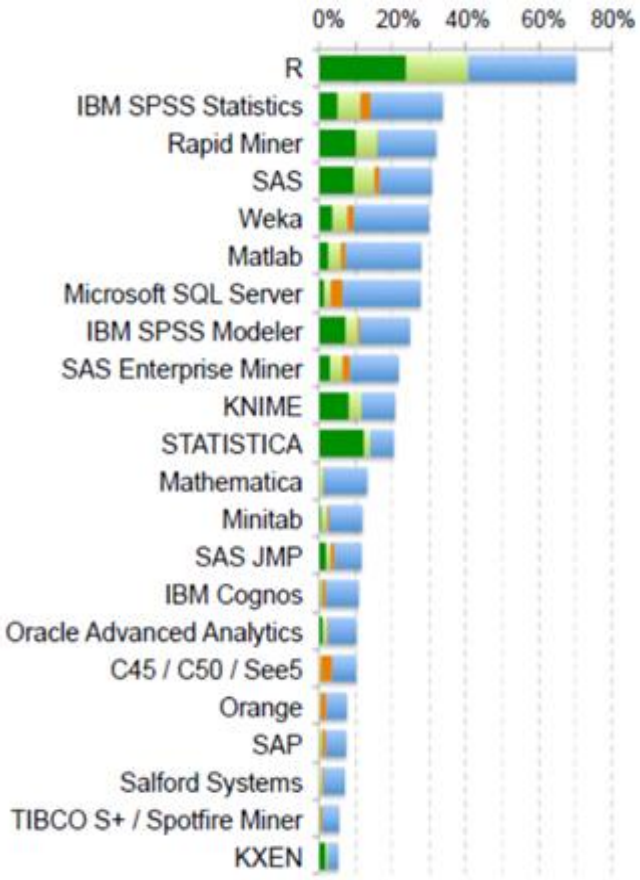
Big Data Analytics Tooling

- Case Study
  - Data size
    - 100TB
  - Formats
    - Structured
    - Unstructured
  - Tasks
    - ETL
    - Data analysis
    - Machine Learning



Analytics Tools Comparison

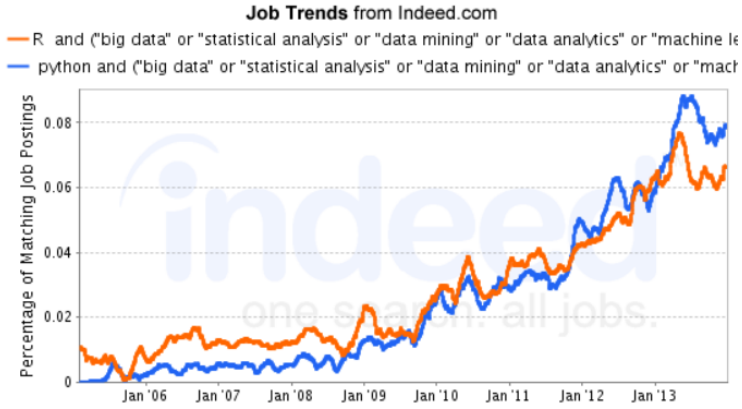
- Ranked by usage
- Overall trend is insightful
- But don't get too serious about the results
  - Should probably break down by industry and business functions
- Statistically significant?
  - Depends on how you interpret it!
- R gets all the love
- SAS is expensive, but still popular in big companies

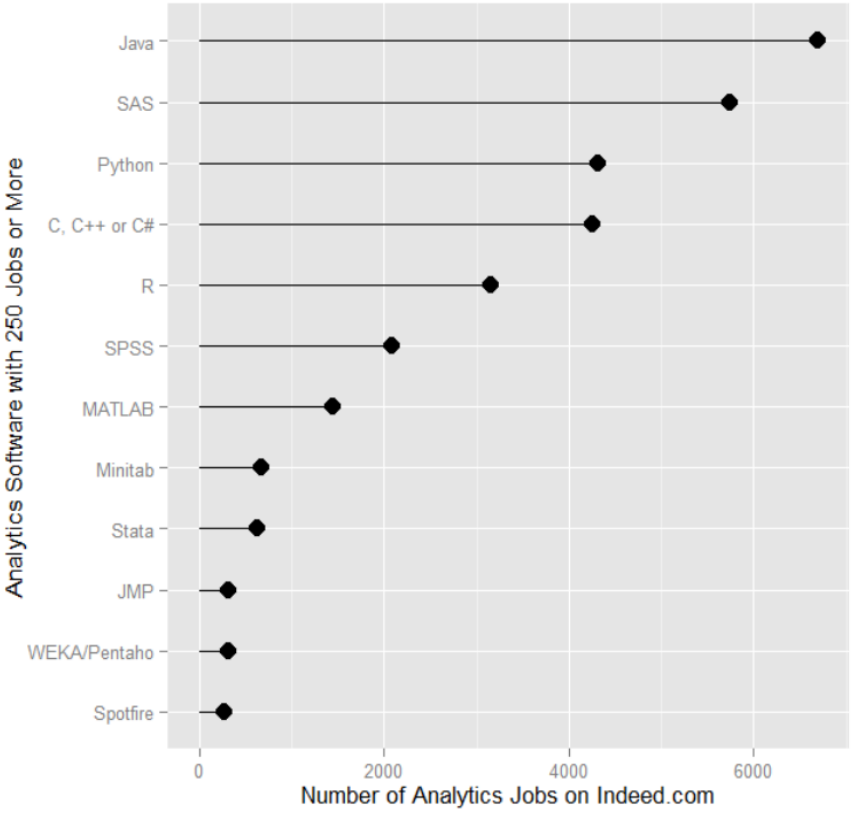


Rexer Analytics 2013 Data Miner Survey Highlights: http://bit.ly/1sgb6lE

Analytics Tools – Popular Tools

- SAS vs. R vs. Python





The Popularity of Data Analysis Software 2012 http://r4stats.com/articles/popularity/

Indeed.com Job Trends: SAS vs R http://www.indeed.com/jobtrends/R+statistics,+SAS+statistics.html

Choosing The Right Tools - Previously

	SAS	R	Python
Prototyping	*SAS Base, SAS EG*	*R (requires sampling)*	*Python (requires sampling)*
Data Manipulation	*SAS, Oracle SQL*	*R, Oracle*	*Python*
Modeling	*Enterprise Miner, SAS Base, SAS EG*	*R*	*Scikit-learn*
Scoring	*Enterprise Miner, SAS Base, PMML*	*R, PMML*	*Python*

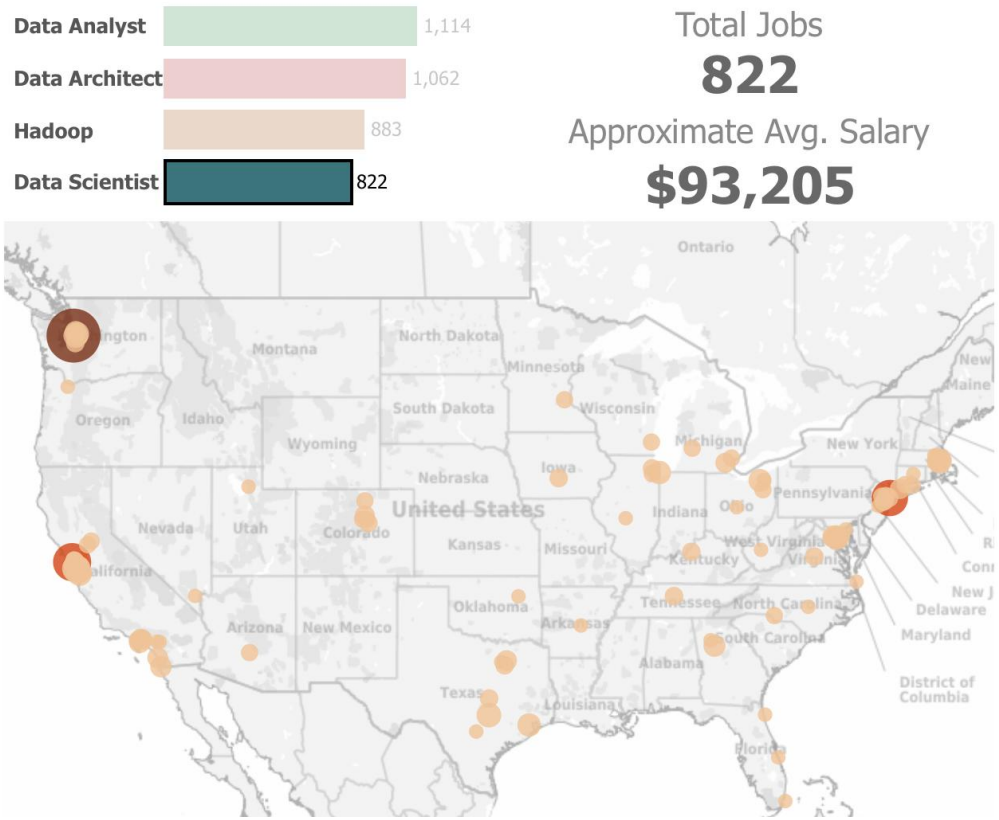
Choosing The Right "Big Data" Tools - Today

	Java	SAS	R	Python	Spark
Prototyping	*Weka, Java*	*SAS Base, SAS EG*	*R*	*Python*	*Spark/R*
Data Manipulation	*Hadoop, Pig/Hive*	*SAS Connector for Hadoop*	*RHadoop*	*Hadoop Streaming Pig/Hive*	*Spark*
Modeling	*Weka, Mahout*	*Enterprise Miner, SAS Hadoop*	*RHadoop*	*Hadoop Streaming*	*Mllib, GraphX*
Scoring	*Hadoop, Mahout*	*Enterprise Miner, SAS Base, Hadoop PMML*	*RHadoop*	*Hadoop Streaming, Pig*	*Spark*

4. Big Data Job Market

Job Market

- Where are big data jobs?
  - North America
    - Sillicon Valley, Seattle, NYC, Toronto
  - India/China



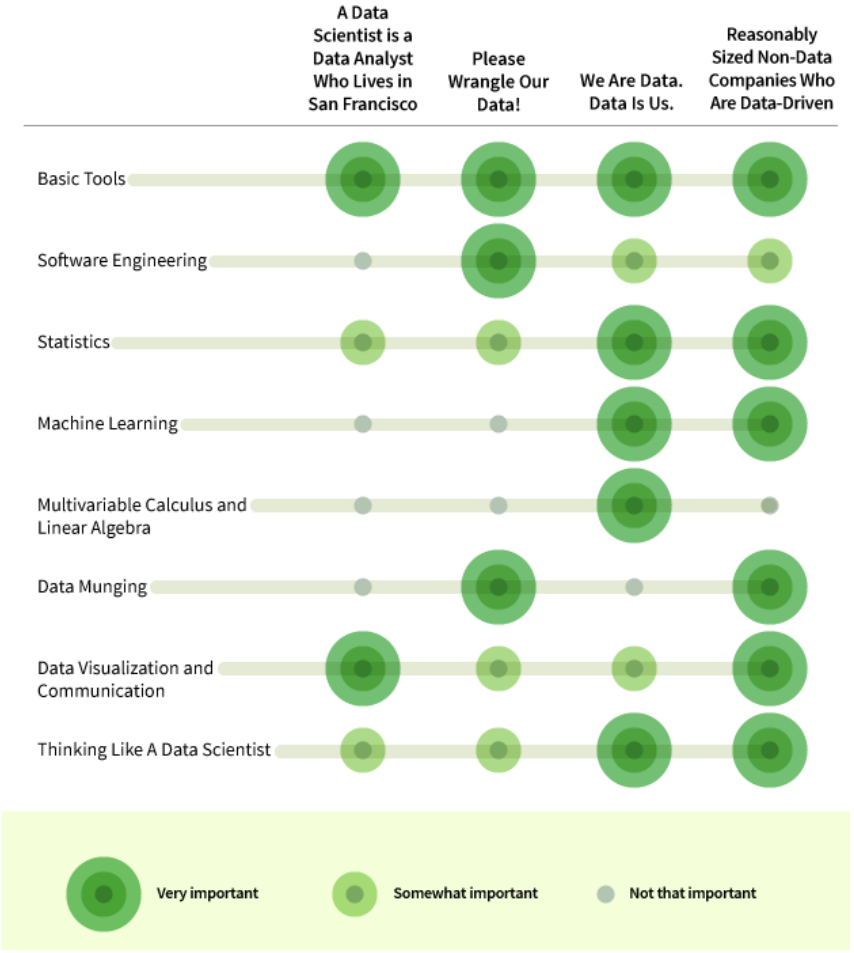
Big data jobs around the nation: http://www.tableausoftware.com/public/gallery/big-data-jobs
Big Data Salary: http://goo.gl/4et998
Oreilly Media Data Science Salary Survey: http://www.oreilly.com/data/free/files/stratasurvey.pdf
KDNuggets 2014 Analytics/Data Science Salary Poll: http://goo.gl/VhO9IW

Big Data/Analytics Jobs (Toronto)

- Banks
  - RBC, TD, CIBC, Scotiabank, AMEX, CapitalOne, ING Direct
- Telcommunications
  - Rogers, Telus, Bell, etc.
- Technology
  - BlackBerry, Huawei, CGI
- Manufacture/Services
  - GM, Canada Post, Workopolis
- Insurance
  - SunLife, Manulife
- Web/Mobile/Startup
  - Google, Mozilla
- Digital Media/Agencies
  - Globe and Mail, Kobo
- Consulting
  - Accenture, IBM, Deloitte, SAS
- Retail/e-commerce
  - Amazon, HR, Hudson Bay, Sears, Shoppers, Canadian Tire, Sobeys
- Pharmaceutical/Healthcare
  - Hospitals, Clinical Research Companies etc.

Data Organizations

- Data Analyst Who Lives in San Francisco
- Please Wrangle Our Data!
- We Are Data. Data Is Us
- Reasonably Sized Non-Data Companies Who Are Data-Driven



Reference:

Big Data Skill Sets (Analytics)

- Data munging
  - SAS, Oracle, Hadoop Pig/Hive
- Database
  - SQL
  - Hive
- Programming
  - Java, Scala, Python
- Machine learning, predictive modeling at scale
  - SAS Enterprise Miner
  - Mahout, Spark Mllib, H_2O
  - R, Python
- Algorithms
  - Regression, Trees, Clustering, SVM, KNN, etc.
- Data Engineering
  - linux, shell, cron, automation
- Reporting
  - Pivot tables
  - Dashboards
- Visualization/presentation
  - d3.js, Tableu, Qlikview, Excel

Job Functions/Titles

- Teams
  - Platform/Infrastructure
  - Data Science
  - Business Intelligence
  - Analytics (Marketing, Sales, Product etc.)
  - Startup - Jack of all trades
- Titles
  - Data Engineer (Database Developer, Data System Developer, etc.)
  - Data Scientist/Statistician
  - Big Data Researcher
  - Machine Learning Researcher/Scientist
  - Platform Engineer
  - Platform/Infrastructure Architect
  - Infrastructure Specialist
  - Data Analyst
  - Software Engineer
  - Database Administrator
  - ETL Developer/Specialist
  - Web developer
  - Data Ninja
  - Data Evangelist

How to build a hadoop data scienice team: http://hortonworks.com/blog/hortonworks-hadoop-data-science/

Salary

- Great compensation!
- But range is wide…

Big data jobs around the nation: http://www.tableausoftware.com/public/gallery/big-data-jobs
Big Data Salary: http://goo.gl/4et998
Oreilly Media Data Science Salary Survey: http://www.oreilly.com/data/free/files/stratasurvey.pdf
KDNuggets 2014 Analytics/Data Science Salary Poll: http://goo.gl/VhO9IW
Big Data Salaries: An Inside Look Data Scientist Salaries, Data Analyst Salaries, DBA Salaries, etc: https://datajobs.com/big-data-salary

4. Big Data Challenges

Getting Over The Big Data Hype

- "Big Data" is NOT about "big"
  - we've done it for many many years (costly)
  - isn't it expected anyway with the growth of the Internet
  - it is a mentality
- You don't need big data sometimes
- Having big data and Hadoop cluster doesn't solve your problems… it may create new problems if you can't harness it
  - you need the right tools, right talent, right management support and team structure
- Just a different tool or platform
  - How you do analytics haven't fundamentally changed
- Bigger doesn't mean better
  - Big data vs small data

Big Data – The Challenges

- Reality is that Hadoop is still hard to use (usability for business analysts)
  - Requires low-level Map Reduce programming to achieve sophisticated task
  - Mostly command line, GUI is not user friendly (improving)
- SQL-on-hadoop not delivering the promise yet
  - The SQL vs. NoSQL war
  - NewSQL (Google's F1 paper)
- Rapid growth causes confusions
  - Emerging stack such as Spark
  - Uncertainty
  - Vendors and confusions

Google F1 Paper:

Big Data – The Challenges

- Most companies are still in very early stage, leveraging hadoop for data storage and ETL, not really taking the full advantage of the stacks
- Building data pipelines is hard
  - pipelines are the glues
  - different platforms/tools cause frictions
  - Hadoop stack works
  - Spark is the challenger
- Talent gap
  - High quality data scientists/engineers hard to find
  - Unicorns are rare

Lecture 1 - Summary

- Data science/analytics is a competitive market, you need to master a set of new tools to stay competitive
- Data size is growing exponentially. You need to choose the right tools for your analytics needs
- Tools you'll learn in this course
  - Hadoop – basic MapReduce concept
  - Pig/Hive large-scale data processing
  - Building automated data pipelines
  - Apache Spark Introduction
- Use cases you'll learn in this course
  - Location analytics
  - Marketing analytics
  - Recommendation engine
  - Computational advertising
  - Real-time analytics

Big Data Analytics Resources

- Blogs/Talks
  - Datasciencecentral
  - DataTau
  - DataScience Weekly
  - Meetups (HakkaLab) – youtube/slideshare
  - SF Machine Learning
  - Engineering blogs - FB, Yahoo, Twitter, 4SQ etc.

Big Data Analytics Resources

- Conference/Workshop
  - Oreilly Strata/Hadoop World
  - Hadoop Summit
  - Cassandra Summit
  - H2O World
  - Solr Revolution
  - GraphLab Conference
  - Qcon
  - MLconf
  - PAW – Predictive Analytics World
  - SAS Conference

Lab & Assignments

- Lab Computer
  - Username: datastudent
  - Password: datastudent
- Lab 1
  - Virtualbox
  - Hadoop setup
- **Assignment 1**
  - Pick a big data job title that fits your interests
  - Find 3 companies (U.S. or Canada) that offer the job position you identified
  - Read the job postings and summarize the skillsets required by the position
    - Send to Instructor's email: shaohua.zhang@ryerson.ca by putting "[Assignment1]" in the title

Lab 1 – Environment Setup

Lab 1 – Install VirtualBox

- Installation of VirtualBox
  - https://www.virtualbox.org/wiki/Downloads

Our Lab Director Shannon has already downloaded and installed VirtualBox for us ☺

Lab 1 – Install Putty

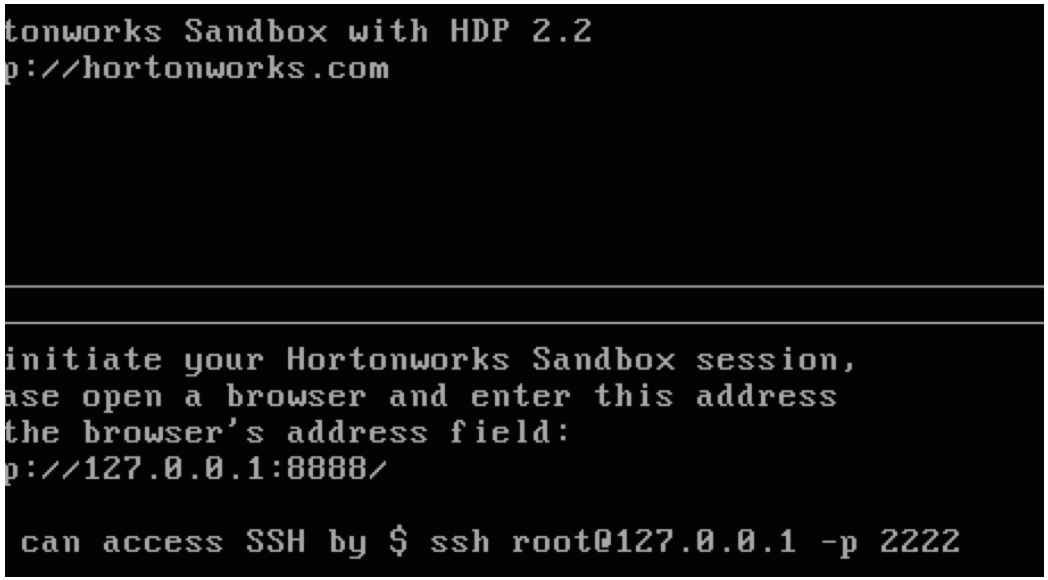
- Windows users need to install Putty in order to use SSH to access hadoop
  - http://www.chiark.greenend.org.uk/~sgtatham/putty/

Our Lab Director Shannon has already downloaded and installed Putty for us ☺

Lab 1 – Install Hortonworks HDP Sandbox

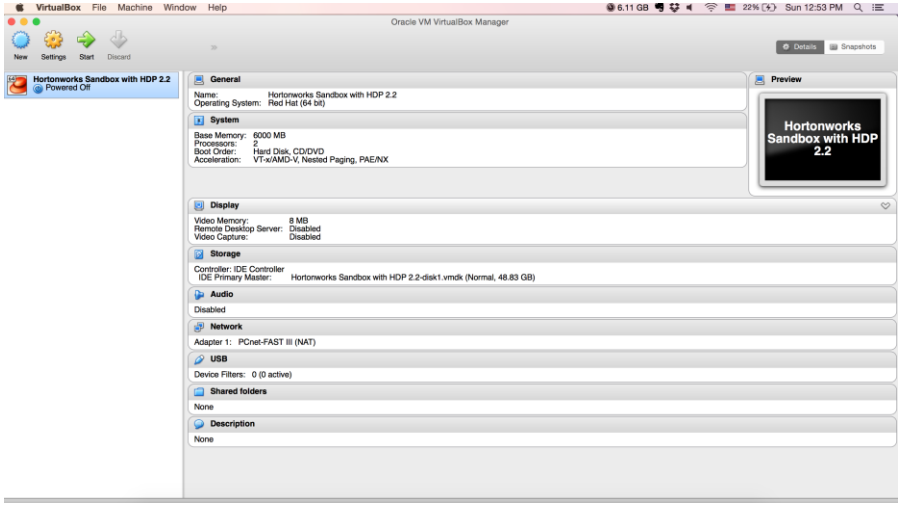
- Download HDP hadoop vm image: http://hortonworks.com/products/hortonworks-sandbox/
- Tutorial for Windows: http://hortonworks.com/wp-content/uploads/unversioned/pdfs/InstallingHortonworksSandbox2onWindowsusingVB.pdf

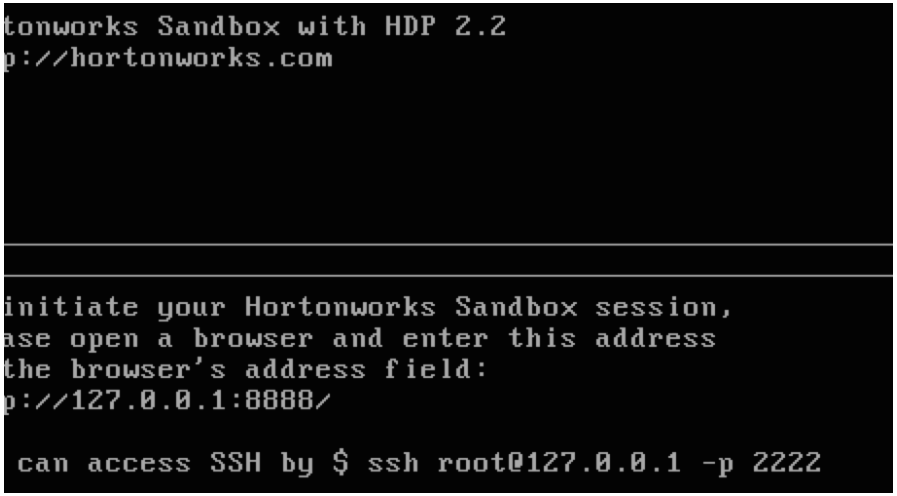
- The HDP VM Sandbox has been downloaded to C drive on windows.
- After installation and importing the instance, you start the vm and will see the screen on the right



Lab 1 – Log On and Test Hive

1. Open VirtualBox
2. Start the HDP 2.2 VM
3. Open Putty and connect to 127.0.0.1 port 2222
   - Username: root
   - Password: hadoop
4. Test Hive
   - $ hive





Lab 1 – Access Hadoop via Browser Using Hue

- Access Hue via Browser
  - http://127.0.0.1:8000