INTRODUCTION TO BIG DATA ANALYTICS TOOLS

DS8003 – MGT OF BIG DATA AND TOOLS

RYERSON UNIVERSITY

Instructor: Kanchana Padmanabhan

Tell me about yourself

- Current industry you work in?
- Your interests/focus in this program?
- Programming languages?

About This Course

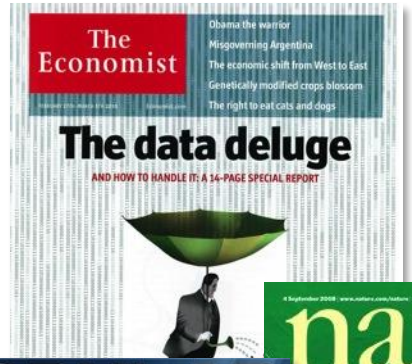
- This course focuses on practicality
  - It follows industry trends and job market trends
  - You'll be hands-on with several popular tools
  - You'll learn practical use cases and how to choose the right tools
  - You'll learn enough to be able to extend on your own after this course
- This course teaches big data tools related to analytics
  - Will focus less on Infrastructure, ETL and BI
- It mainly focuses on batch processing tools
  - May cover a little real-time streaming processing concepts depending on how well the class accept other material
- Lab exercises will be done via virtual machines

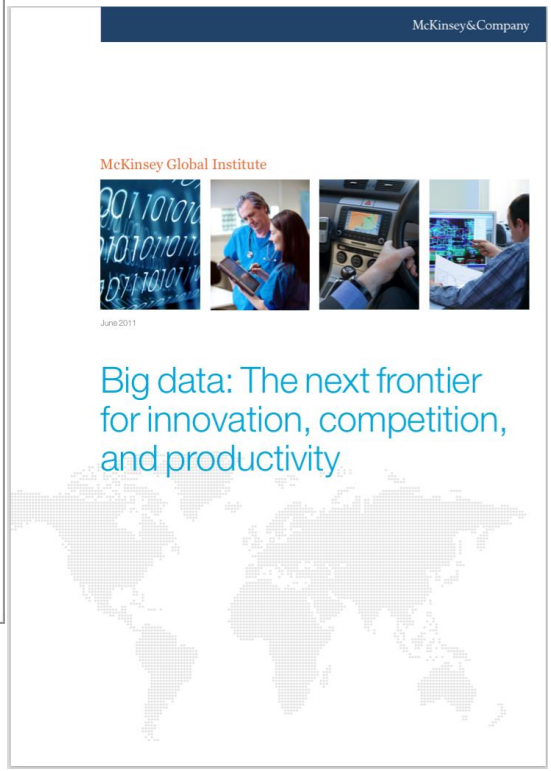
Lecture 1- Outline

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

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

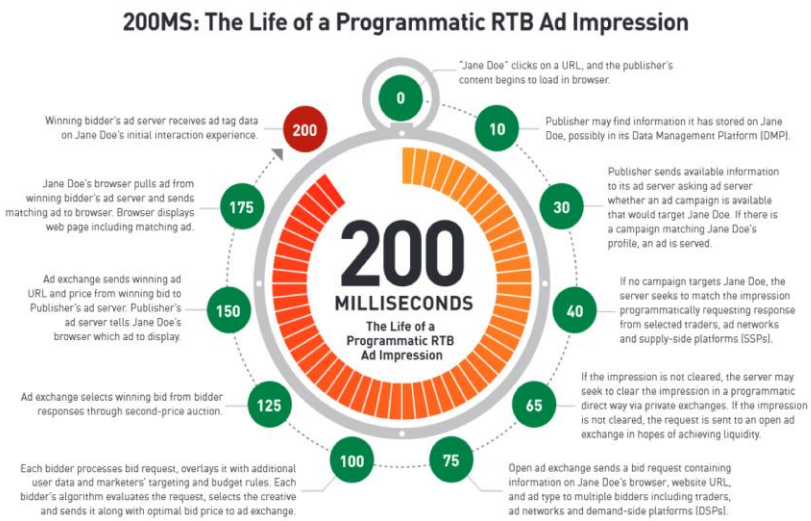
• *Internet*	○ *2.5 exabytes (2.5×10^{18}) per day – 2012* ○ *2.3 zettabytes (2.3×10^{21}) per day - 2014*
• *Facebook*	○ *500+ terabytes per day*

"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
http://highscalability.com/blog/2012/9/11/how-big-is-a-petabyte-exabyte-zettabyte-or-a-yottabyte.html

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

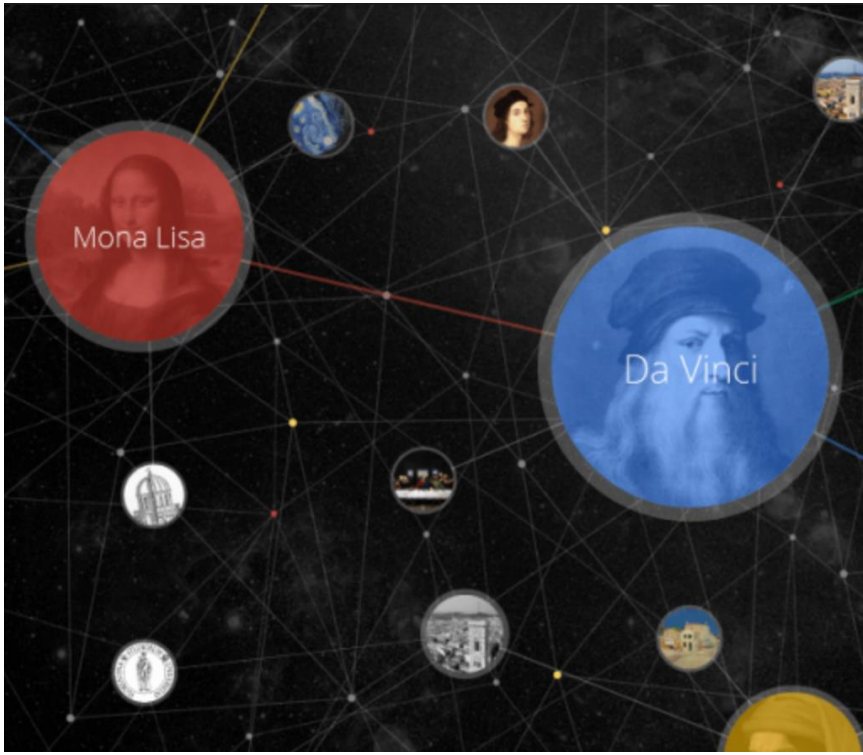


video

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



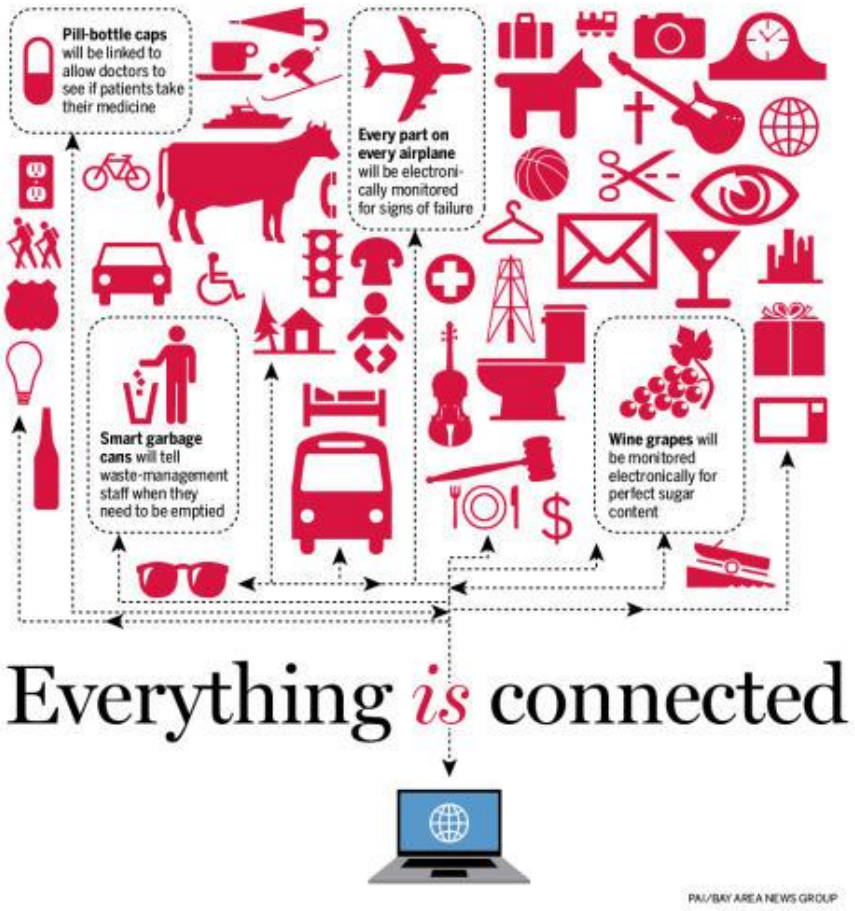
Four V's of Big Data: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Applications Driving the Need for Big Data

- Data-driven Applications
  - Location-based services
  - Social media apps
  - Image/voice recognition
  - Advertising

Applications Driving the Need for Big Data

- Quantified Self
- Internet of Things
- Data-driven Economies
  - Monetization needs - data the new oil

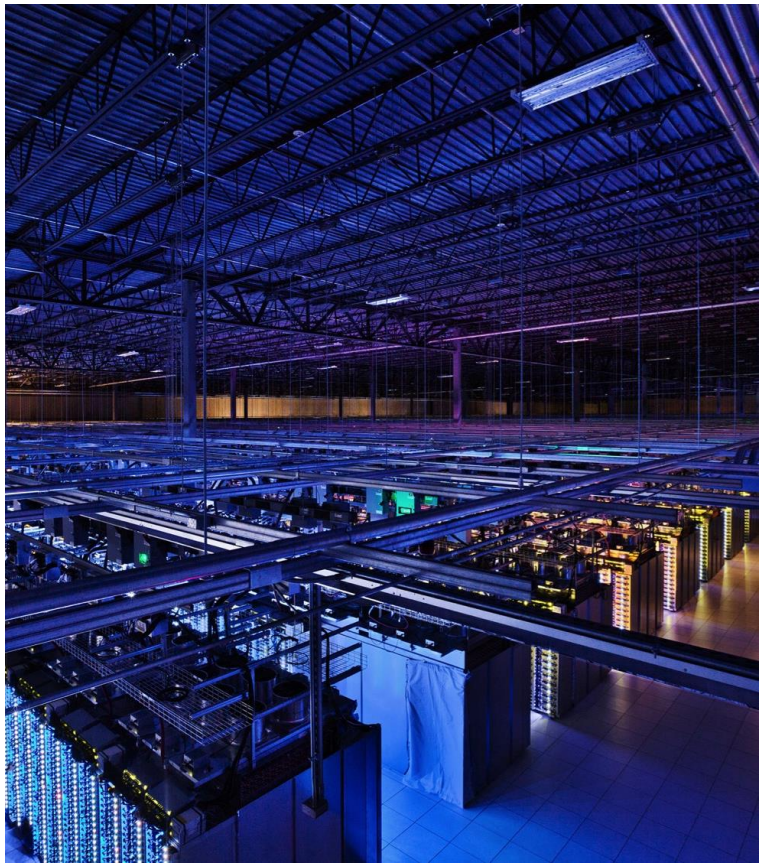


Communities Driving the Need for Big Data

- Big Data vendors
  - Cloudera
  - Hortonworks (public)
  - MapR
  - DataStax
  - *Databricks*
- Traditional Vendors
  - Oracle
  - SAS
  - IBM
  - Revolution Analytics (now part of Microsoft)
- Open Source Communities
  - RHadoop
  - RapidMiner
  - NoSQL

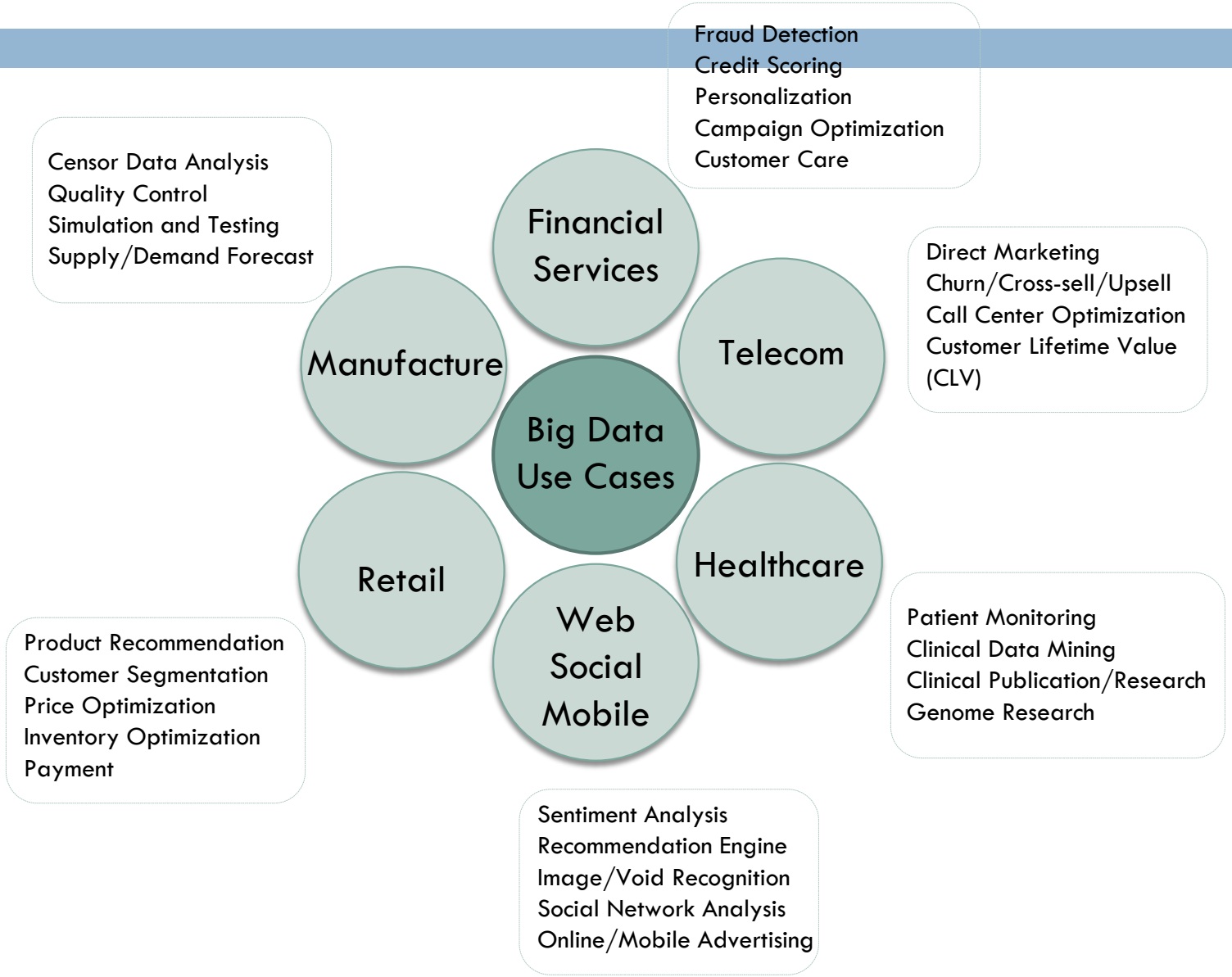
Big Data Made Possible

- 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)



Big Data Use Cases

Big Data Use Cases



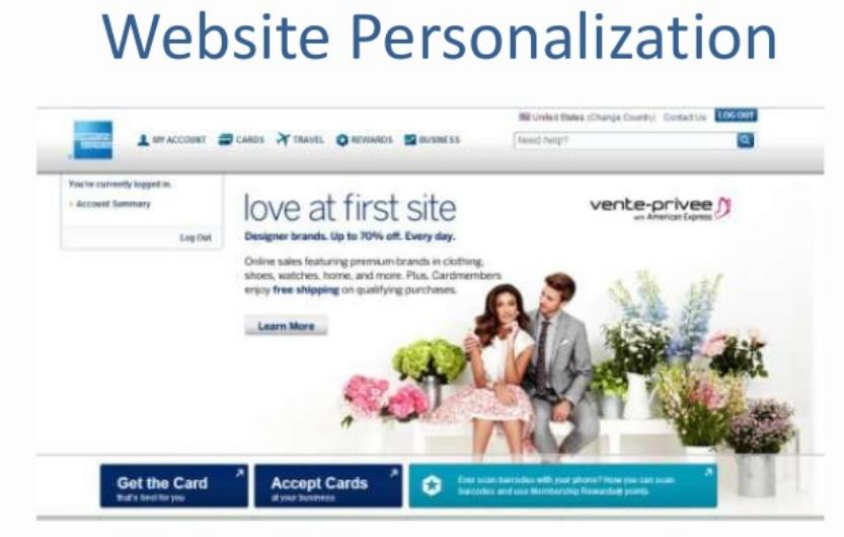
Big Data Use Case – Search & Media

- 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

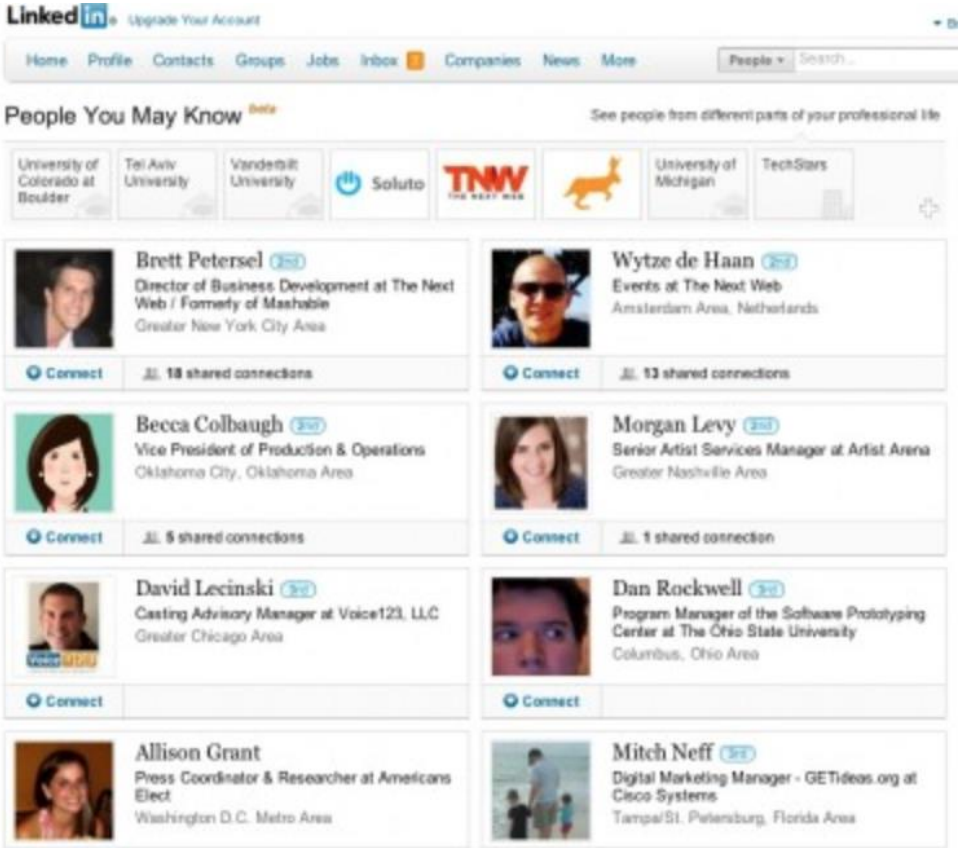
- Hadoop is the new backbone of American Express
  - Recommender systems
  - Graph algorithms
  - Machine learning for Fraud and Marketing
  - Data products
  - Experiments



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

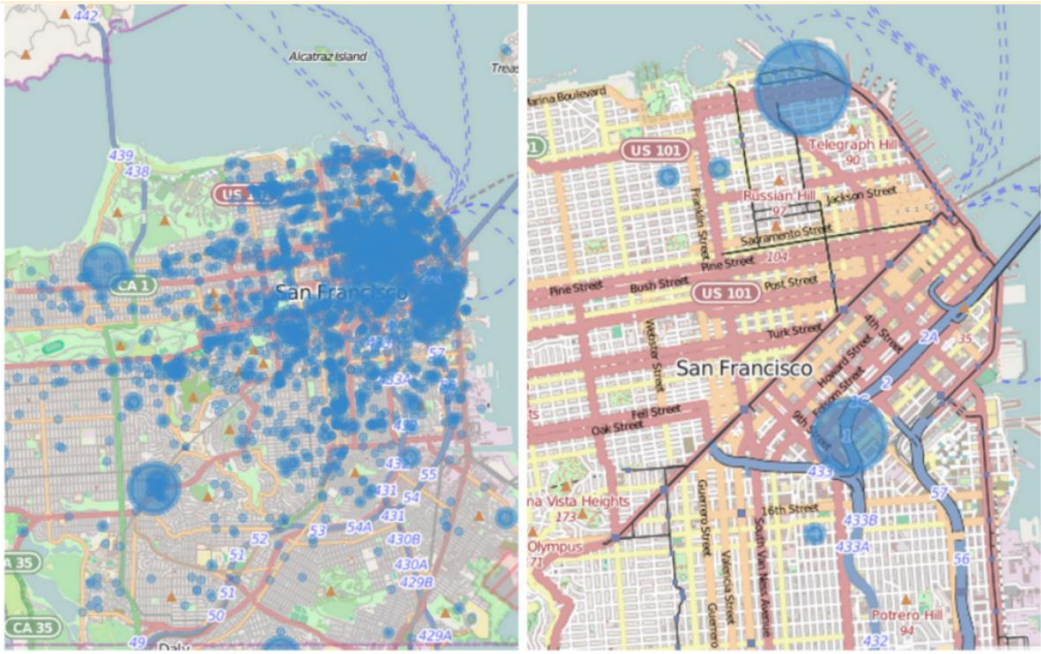
Big Data Use Case – Social Media

- Facebook (designed Hive, Giraph)
  - New Feeds
  - Friend recommendation
  - Ads
  - Graph Search
  - Video Search
  - Artificial Intelligence
- 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



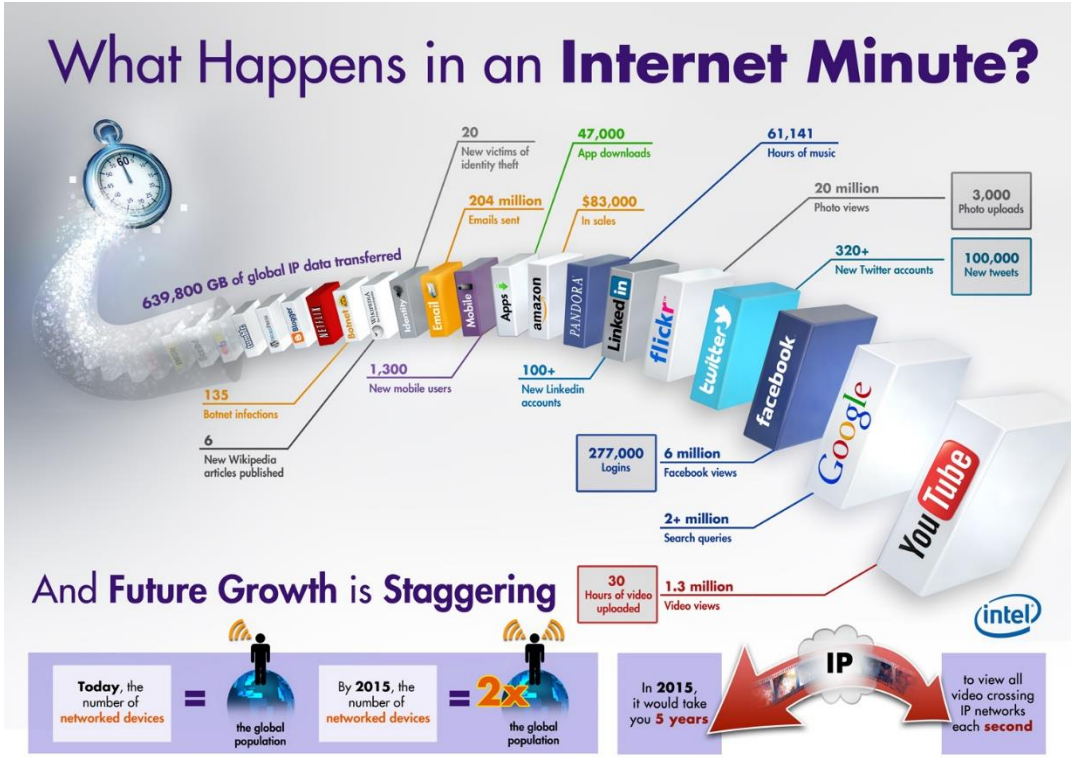
Big Data Use Case – E-Commerce

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

Big Data & Analytics Tooling

Big Data – Exciting Future

Great! Cool! Promising! Exciting!

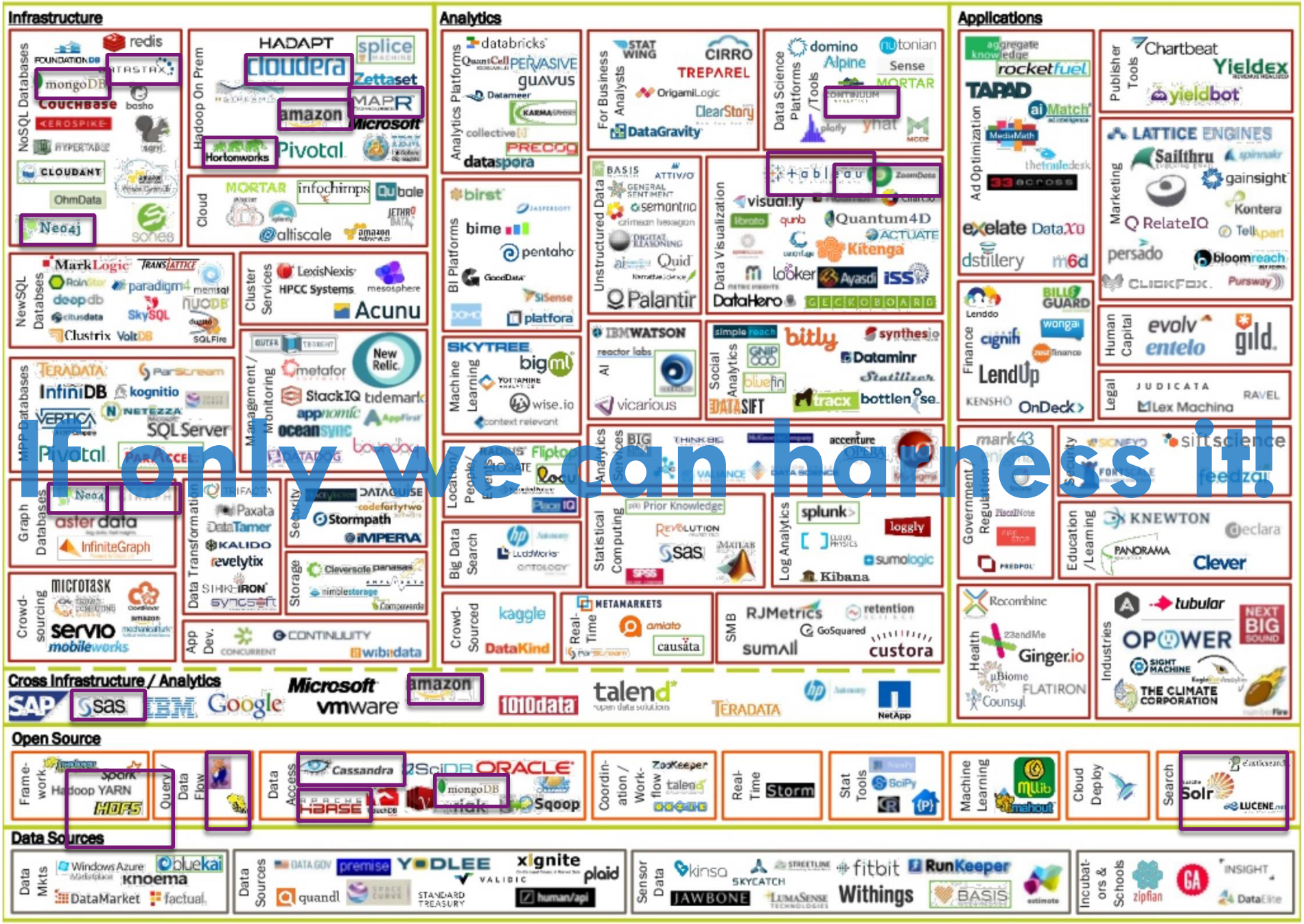


Big data: The next frontier for innovation, competition, and productivity: http://bit.ly/1pCOqom

http://archive.tiecon.org/content/big-data-landscape-%C3%A2%E2%82%AC%E2%80%9C-why-should-you-care

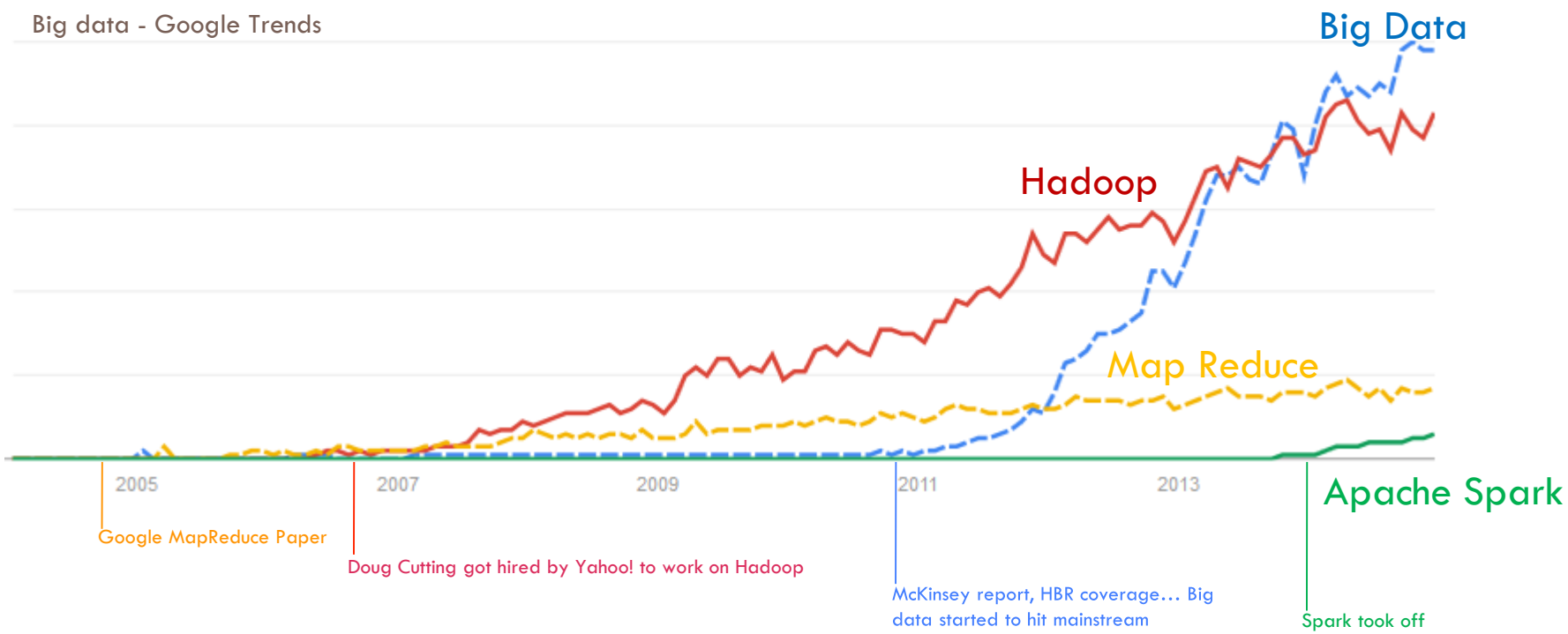
BIG DATA LANDSCAPE, VERSION 3.0

Exited: Acquisition or IPO



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	Databricks	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
  - Data scientists/engineers preference
  - Scalability
  - 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

Considerations	SAS	R	Python	Java/Hadoop	Pig/Hive	Spark
Scalability	●	○	●	●	●	●●
Ease of data manipulation	●●	●	●	○	●●	●
Algorithms/Libraries	●	●●●	●●	●	●	●
Operations/Production Readiness	●●	●	●	●●	●●	●●
Cost (low)	○	●	●	●	●	●●
Support & Service	●●	●	●	●	●	●
Business/Data analyst	●●	●	●	○	●●	○
Statistician	●	●●	●	●	●	●
Data engineer	●	●	●	●●	●	●●
Data scientist	●	●	●	●	●	●●

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*	*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,	Spark

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: http://research.google.com/pubs/pub41344.html
Memsql Pinterest Use Case: http://blog.memsql.com/pinterest-apache-spark-use-case/

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
  - HBASE/MongoDB
- 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
  - Toronto Data Science Group (Meetup.com)
  - 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