Scaling Up! Really Big Data

Course Description:

This course provides an overview of the toolkits in use for problems related to Cloud Computing and Big Data. Because the class is an advanced elective, we generally assume familiarity with the concepts and spend more time on the implementation. Every lecture is followed by a hands-on assignment, where students get to experience some of technologies covered in the lecture. By the time you complete the course, you should be able to name the Big Data problem you are facing, select proper tooling, and know enough to start applying it.

We begin the class with some life examples of how Big Data is used. There's no Big Data without Cloud today, so we dive into Cloud Computing next. We move on to Big Storage and discuss large distributed file systems. Next on the agenda is an obligatory lecture on Hadoop and Map Reduce in general, followed by a session on Apache Spark, the next generation Big Data Analytics Platform. Object Storage comes next along with a debate on how to efficiently transfer massive data sets. We learn NoSQL tooling after that, playing with Cloudant in the Cloud. High Velocity data is next on the agenda and we'll learn how to use Apache Storm and Spark D-Streams. We'll shift gears and take a session to discuss the topic of Managed Compute and Scaling, covering CloudSoft Brooklyn, Apache Mesos, OpenStack Heat, and Amazon Cloud Formation. One cannot truly understand Big Data without knowing its Big Brother – Big Compute, and that's what our next session is on. Web Search is next. We wrap the course by learning how the Big Data tools help us in the areas of Computational Genomics and Cognitive Computing – yes, the last lecture is on IBM Watson! **(3 units).**

Prerequisites:

Students must have completed W201: Research Design and Application for Data and Analysis, W203: Exploring and Analyzing Data and W205: Storing and Retrieving Data before enrolling in this course. They should be able to program in C, Python, Java and / or be able to pick up a new programming language on the fly. A degree of fluency is expected with the basics of operating systems (e.g. Linux) and the Internet Technologies.

Course Evaluation:

- Homework (40%)
- Participation and Group Assignments (20%)
- Final Project: performing an analysis on a large dataset (40%). The students will be required to organize into groups of 4-5 and prepare a final presentation (slides) + video (10 min)

List of Topics by Week:

Week 1: General Course Overview. What is Big Data? Big Data at work. The four Vs of Big Data. Data Distribution. Compute Distribution. Data Transfer. Scaling up and down. What is Cloud? Introduction to the class project options and structure.

Reading:

- The three Vs of Big Data: http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/
- The four Vs of Big Data: http://www.ibmbigdatahub.com/infographic/four-vs-big-data
- Wikipedia on Cloud Computing: http://en.wikipedia.org/wiki/Cloud_computing
- Aspera Benchmarks: http://asperasoft.com/resources/benchmarks/
- Streaming and Watson: http://www.ibm.com/developerworks/library/bd-streams-watson/index.html
- How to tell someone's age when you know her name: http://fivethirtyeight.com/features/how-to-tell-someones-age-when-all-you-know-is-her-name/

Week2: Cloud Computing 101. Defining the Cloud. How Clouds are used. Hypervisors in a nutshell. Types of Clouds. Service Types. The API. Using the API. High Performance Computing in the Cloud. Setting up your account in the Cloud and accessing your environments. The project incubator.

Reading:

- What is Cloud Computing?: http://www.ibm.com/cloud-computing/us/en/what-is-cloud-computing.html
- Linux as a Hypervisor: http://www.linuxplanet.com/linuxplanet/reports/6503/1
- Types of clouds: http://it.toolbox.com/blogs/storage-and-io/cloud-virtual-and-storage-networking-conversations-part-iv-49637
- Cloud Service Types: http://blog.appcore.com/blog/bid/168247/3-Types-of-Cloud-Service-Models
- SoftLayer API: http://sldn.softlayer.com/article/SoftLayer-API-Overview
- Key Cloud Frameworks and technologies: Apache Mesos, Docker, Kubernetes, Salt, Ansible, OpenStack
- Getting Started with SoftLayer:
    - http://knowledgelayer.softlayer.com/gettingstarted/meet-softlayer
    - http://knowledgelayer.softlayer.com/gettingstarted/how-to
    - http://knowledgelayer.softlayer.com/gettingstarted/how-to/set-up-your-account

Week 3: Internet of Things and how it is different from the Cloud. Data of the Internet of Things [that usually does not make it to the cloud]. Edge devices and gateways, sensors, extracting and transmitting to the cloud. Analytics and clustering at the edge.

Reading:

- Amazon IoT Portal: https://aws.amazon.com/iot/
- IBM Internet of Things Portal: http://www.ibm.com/internet-of-things/
- Amazon Lambda: https://aws.amazon.com/lambda/
- http://www.intel.com/content/www/us/en/internet-of-things/overview.html
- Pi Foundation: https://www.raspberrypi.org/
- Software Defined Radio blog: http://www.rtl-sdr.com/

Week 4: Storing Big Data and limitations of centralized storage. Hadoop HDFS: name nodes and data nodes, block placement strategies, data replication and pipelining, recovery from failure and rebalancing. A POSIX-compliant alternative: IBM General Parallel File System (GPFS). The goals of GPFS. The Features. Architecture Highlights. Disk layout and data files. Replication: blocks and sub-blocks. The CLIs for HDFS and GPFS.

Reading:

- HDFS Design: http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html
- Hadoop API: http://hadoop.apache.org/core/docs/current/api/
- GPFS: A Shared disk file system for large computing clusters: https://www.usenix.org/legacy/events/fast02/full_papers/schmuck/schmuck.pdf
- GPFS Web Link: http://www.ibm.com/systems/platformcomputing/products/gpfs/

Week 5: Map Reduce & Apache Hadoop in detail. A Map Reduce Example. Parallelizing Map Reduce. Projecting onto hardware. Data affinity. Input Splits. Data Partitioning. The Shuffle. The pre-reduce sort and merge. The optional post-map. The overall flow. Apache Hadoop. Tuning Hadoop. V1 vs v2 vs Yarn. Hadoop Satellite Projects. Hands-on with Hadoop Streaming. Hands-on with Pig.

Reading:

- Apache Hadoop: http://hadoop.apache.org/
- Apache Pig: http://pig.apache.org/
- IBM Plaftorm Computing Blog: https://www-304.ibm.com/connections/blogs/platformcomputing/?lang=en_us
- Cloudera raises $900M: http://www.zdnet.com/cloudera-raises-900-million-plots-expansion-7000027879/

- Tom White: Hadoop, the definitive guide: http://www.amazon.com/Hadoop-Definitive-Guide-Tom-White/dp/1449311520

Week 6: Apache Spark. Limitations of MapReduce. What is Spark? Batch vs Stream processing. Scala, Java, Python Examples. Resilient Distributed Datasets (RDDs). Parallel Operations. Shared Variables. Text Search Example. ALS Example. GraphX.

Reading:

- Spark Overview: http://spark.apache.org/docs/latest/
- Batch versus Stream Processing: http://datatactics.blogspot.com/2013/02/batch-versus-streaming-differentiating.html
- Code examples: http://spark.apache.org/docs/latest/programming-guide.html
- RDD's: http://spark.apache.org/docs/latest/programming-guide.html#resilient-distributed-datasets-rdds
- Mlib: http://spark.apache.org/docs/latest/mllib-statistics.html
- GraphX: http://spark.apache.org/docs/latest/graphx-programming-guide.html

Further Reading:

- Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, Learning Spark: http://shop.oreilly.com/product/0636920028512.do

Week 7: Storing even larger volumes of data & data transfers. Scalability. Key Constraints to scaling a data store. Vertically and horizontally scaled solutions. The I/O bottleneck. S3. Swift. Ceph. Moving Data in and out of the Cloud. Aspera and the FASP Protocol.

Reading:

- Carol Sliwa: Troubleshooting and Identifying Data Storage Performance Bottlenecks: http://searchstorage.techtarget.com/report/Troubleshooting-and-identifying-data-storage-performance-bottlenecks
- Jeffrey Shafer: I/O Virtualization Bottlenecks in Cloud Computing today: http://dl.acm.org/citation.cfm?id=1863186&preflayout=flat#source

- Peter Vajgel: Needle in a haystack: efficient storage of billions of photos: https://www.facebook.com/note.php?note_id=76191543919
- Gustafson's Law: http://en.wikipedia.org/wiki/Gustafson%27s_law
- Ceph Versus Swift: http://techs.enovance.com/6427/ceph-and-swift-why-we-are-not-fighting
- Apera FASP: http://asperasoft.com/technology/transport/fasp/

Week 8: Databases 2.0 – The NoSQL Movement. Taxonomy Review. Consistent Hashing Intro. CAP Theorem and Quorum Algebras. Cloudant/CouchDB API Intro. Cloudant Account Creation. Cloudant MapReduce Intro. The Obligatory Word Count. Graph Databases

Required:

- Cade Metz: If Xerox PARC Invented the PC, Google Invented the Internet
- G. DeCandia et al: Dynamo: Amazon's Highly Available Key-value Store
- NoSQL Overview: http://en.wikipedia.org/wiki/NoSQL
- Sign-up for cloudant.com account.
- Complete Cloudant For Developers interactive tutorial
    - Reading and Writing, Primary Index, Secondary Indexes, Search Indexes
- Skim Cloudant API Reference

Optional:

- Ion Stoica and R. T. Morris: Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications
- Google File System
- Google MapReduce
- Google BigTable
- Cloudant Online Training Presentations

Week 9: Dealing with High Velocity Data. Why is Streaming different? The real world examples. Continuous Queries. Active Databases. Pubsub. Complex Event processing systems. The data view. The operator view. SPAs and SPLs. The leading streaming processing systems: Amazon Kinesis, Apache Storm, IBM SPL, Spark D-Streams

Reading:

- Spark D-Streams: http://www.cs.berkeley.edu/~matei/papers/2012/hotcloud_spark_streaming.pdf
- IBM InfoSphere Streams: https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=sw-infomgt&S_PKG=ov24587&S_TACT=109HF53W&S_CMP=is_bdebook8
- Launching Kinesis: Amazon Web Services Blog post
- AWS Kinesis: http://aws.amazon.com/kinesis/
- Kinesis API: http://awsdocs.s3.amazonaws.com/kinesis/latest/kinesis-api.pdf
- Storm vs Spark: http://www.slideshare.net/ptgoetz/apache-storm-vs-spark-streaming?qid=dfbb7c09-6f87-40ca-a69d-d76837efd236
- Introduction to Apache Storm: http://www.slideshare.net/ptgoetz/storm-hadoop-summit2014
- D-Streams: http://tinyurl.com/dstreams
- Apache Spark Programming Guide: http://spark.incubator.apache.org/docs/latest/streaming-programming-guide.html
- Nathan Marz on Apache Storm: https://www.youtube.com/watch?v=bdps8tE0gYo
- Apache Spark Tutorial: https://storm.incubator.apache.org/documentation/Tutorial.html
- AMP Camp on Spark Streaming: Shttp://ampcamp.berkeley.edu/wp-content/uploads/2013/07/Spark-Streaming-AMPCamp-3.pptx

Week 10: Scaling up and slimming down: how to reconfigure your clusters dynamically in response to load. What is Managed Compute? Three generations of tooling: Infrastructure-centrilc, Application-centric, and Platform-Centric. Examples of Infrastucture level tools: OpenStack Heat . AWS CloudFormation. An example of application level platform: CloudSoft Brooklyn. Deploying a three tier topology with either type of tooling. Apache Mesos.

Reading:

- Apache Mesos: http://mesos.apache.org/
- Mesosphere DCOS: https://mesosphere.com/product/
- AWS CloudFormation : http://www.cs.utexas.edu/ftp/techreports/tr95-13.pdf
- AWS OpsWorks: http://aws.amazon.com/opsworks/
- AWS Elastic BeanStalk: http://aws.amazon.com/elasticbeanstalk/
- IBM BlueMix: http://ibm.biz/HackBluemix
- OpenStack Heat: https://wiki.openstack.org/wiki/Heat
- CloudSoft Brooklyn Walkthrough: http://brooklyncentral.github.io/start/walkthrough/index.html
- GigaSpaces Cloudify QuickStart: http://getcloudify.org/guide/3.0/quickstart.html

Week 11: Big Compute. An overview and a bit of history. HPC vs HTC/Big Data. Matrix-matrix multiply vs matrix elementwise product. Typical HPC problems. Architecture of supercomputers. Interconnect topologies: fat tree, torus. FLOPs, Top500. Scaling: strong, weak, Amdahl's law. Programming for HPC systems. Landscape overview: MPI, OpenMP, PGAS (UPC, Global Arrays), active messaging (Charm++, X10). SUMMA. OpenML.

Reading:

- SUMMA in detail: http://www.cs.utexas.edu/ftp/techreports/tr95-13.pdf
- MPI and 2D Cartesian communicators: https://computing.llnl.gov/tutorials/mpi/
- Basic tutorial on parallel computing (HPC): https://computing.llnl.gov/tutorials/parallel_comp/
- Basic MPI tutorial: http://mpitutorial.com

Week 12: Web Search and Related Problems. Definition and Examples. Unstructured Data and Scope of Data. Data Discovery. Crawling Techniques. Data analysis Searchable Indices. Lucene and Nutch. Metadata. Context Discovery and Inference. High Volume Queries and Search Performance. Scaling out. Making results available. Defining dictionaries examples.

Reading:

- Sergey Brin and Lawrence Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine: http://infolab.stanford.edu/~backrub/google.html
- David Floyer: The Growth and Management of Unstructured Data: http://wikibon.org/wiki/v/The_Growth_and_Management_of_Unstructured_Data
- Google on Crawling & Indexing: http://www.google.com/intl/en/insidesearch/howsearchworks/crawling-indexing.html
- Solr Tutorial: http://lucene.apache.org/solr/4_10_0/tutorial.html
- Arjun Atreya V; Nutch and Lucene Framework Presentation; http://tinyurl.com/k79hofu
- Ricardo Baeza-Yates , B. Barla Cambazoglu – Yahoo Labs; Tutorial on: Scalability and Efficiency Challenges in Large-Scale Web Search Engines; http://tinyurl.com/pmnlzdw

Week 13: Computational Genomics. Definition and Examples. Genetic sequencing and data sourcing. Data formats. Data collection and transmission. Analysis, tooling, crunching the data. Data preparation. Platform Process Manager. Map Reduce / Platform Symphony. Speed vs. Cost of large jobs and making the outputs available to users. An example: aligning a chromosome.

Reading:

- Wikipedia definition of Computational Genomics: http://en.wikipedia.org/wiki/Computational_genomics
- Human Population Genomics: http://researcher.watson.ibm.com/researcher/view_group.php?id=2303

- Ben Langmead; Searching for SNPs with Cloud Computing: http://genomebiology.com/2009/10/11/r134
- Dennis P Wall; Cloud Computing for Comparative Genomics: http://www.biomedcentral.com/1471-2105/11/259
- Adam for Scala and Spark: http://bdgenomics.org/
- The Adam paper: https://www.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-207.html
- The Avocado paper: http://www.cs.berkeley.edu/~kubitron/courses/cs262a-F13/projects/reports/project8_report.pdf
- One thousand genomes: http://www.1000genomes.org/

Week 14: IBM Watson. What is language and why it is hard for computers to understand it? Overview of Watson Deep NLP Process. Knowledge Corpus. Question Parsing and Context Derivation. Hypothesis generation. Comparative Reasoning. Scoring, Confidence Level assignments, candidate selection. SyNapse. The future of cognitive computing. Systems and Architecture of Watson. Course Wrap up, summary, and Presentation of Final Projects

Reading and Viewing:

- Watson services: http://www.ibm.com/cloud-computing/bluemix/solutions/watson/
- IBM Watson: The Science behind an Answer: http://www.youtube.com/watch?v=DywO4zksfXw
- *David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty* ; **The AI Behind Watson**; Published in AI Magazine Fall, 2010: http://www.aaai.org/Magazine/Watson/watson.php
- Watson on the cloud: http://mayashenoi.com/category/softlayer/
- Watson Paths: http://www.research.ibm.com/cognitive-computing/watson/watsonpaths.shtml#fbid=OMIfqbmoDZr
- IBM Watson Discovery Advisor: https://www.youtube.com/watch?v=qry_zGZFjOc