Scaling Up! Really Big Data

Course Description

This course attempts to provide an overview of the contemporary tool kits that are 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 the 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, covering SoftLayer and OpenStack in detail. We move on to big storage and discuss large distributed file systems. Next on the agenda is an obligatory lecture on Hadoop and MapReduce 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, OpenStack Heat, and Amazon CloudFormation. 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 up 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 (20%)
- Final project: performing an analysis on a large dataset (40%). Students will be required to organize into groups of four to five 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:

- Sicular, S. (2013, March 27). Gartner's big data definition consists of three parts, not to be confused with three "V"s. *Forbes*. Retrieved from 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. *IBM: The Big Data & Analytics Hub, Infographics & Animations*. http://www.ibmbigdatahub.com/infographic/four-vs-big-data
- Cloud computing. *Wikipedia*. http://en.wikipedia.org/wiki/Cloud_computing
- *fasp* benchmarks. *Aspera*. http://asperasoft.com/resources/benchmarks/
- Sharpe, J. (2014, April 22). Use InfoSphere streams as a sensory interface to Watson. *IBM: developerWorks*. Retrieved from http://www.ibm.com/developerworks/library/bd-streams-watson/index.html
- Silver, N., & McCann, A. (2014, May 29). How to tell someone's age when you know her name. *FiveThirtyEightLife*. Retrieved from http://fivethirtyeight.com/features/how-to-tell-someones-age-when-all-you-know-is-her-name/

Week 2: 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? *IBM Cloud*. http://www.ibm.com/cloud-computing/us/en/what-is-cloud-computing.html
- Arcangelim, A. (2008). Using Linux as a hypervisor with KVM. *Qumranet Inc*. Retrieved from http://www.linuxplanet.com/linuxplanet/reports/6503/1
- Schulz, G. (2011, December 5). Cloud, virtual and storage networking conversations part IV. *Toolbox.com*. Retrieved from http://it.toolbox.com/blogs/storage-and-io/cloud-virtual-and-storage-networking-conversations-part-iv-49637
- Hemmings, K. (2012, December 27). 3 types of cloud service models. *appcore*. Retrieved from http://blog.appcore.com/blog/bid/168247/3-Types-of-Cloud-Service-Models
- SoftLayer API: http://sldn.softlayer.com/article/SoftLayer-API-Overview
- Eadline, D. (n.d.). Moving HPC to the cloud. *Admin New Day HPC*. Retrieved from http://www.admin-magazine.com/HPC/Articles/Moving-HPC-to-the-Cloud

- Getting started with SoftLayer, an IBM Company.
    - 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: Making the cloud work for you. Open cloud: OpenStack! Introduction to OpenStack. OpenStack at work. The UE dashboard: Horizon. OpenStack components and architecture. Nova Compute: working with VMs. Glance: services for discovering, registering, and retrieving virtual machine images. Keystone: identity, token, catalog, and policy services. Cinder: block storage.

Reading:

- Bryce, J., Wilson, D., Fischer, C., & Aubuchon, G. (2013, November 8). OpenStack keynote featuring Concur, DigitalFilm Tree, Shutterstock (Video). *OpenStack Summit*. Summit conducted in Hong Kong. Retrieved from https://www.openstack.org/summit/openstack-summit-hong-kong-2013/session-videos/presentation/openstack-keynote-featuring-concur-digitalfilm-tree-shutterstock
- Sabbah, D. (2013, November 8.) IBM keynote: Managing the next era of computing with an open cloud architecture (Video). *OpenStack Summit*. Summit conducted in Hong Kong. Retrieved from https://www.openstack.org/summit/openstack-summit-hong-kong-2013/session-videos/presentation/ibm-keynote-managing-the-next-era-of-computing-with-an-open-cloud-architecture
- DevStack—an OpenStack community production: http://devstack.org/
- OpenStack documentation: https://wiki.openstack.org/wiki/Documentation/HowTo
- OpenStack architecture design guide: http://docs.openstack.org/arch-design/content/
- OpenStack command-line interface (CLI) reference: http://docs.openstack.org/cli-reference/content/

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 subblocks. The CLIs for HDFS and GPFS.

Reading:

- HDFS architecture: http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html
  Apache Hadoop Main 2.5.2 API: http://hadoop.apache.org/core/docs/current/api/

- Schmuck, F., & Haskin, R. (2002, January). GPFS: A shared-disk file system for large computing clusters. In *Proceedings of the FAST 2002 Conference on File and Storage Technologies*. Retrieved from https://www.usenix.org/legacy/events/fast02/full_papers/schmuck/schmuck.pdf
- Elastic storage (GPFS web link): http://www.ibm.com/systems/platformcomputing/products/gpfs/

Week 5: MapReduce and Apache Hadoop in detail. A MapReduce example. Parallelizing MapReduce. Projecting onto hardware. Data affinity. Input splits. Data partitioning. The shuffle. The prereduce sort and merge. The optional postmap. The overall flow. Apache Hadoop. Tuning Hadoop. Version 1 vs. version 2 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 Platform Computing blog: https://www-304.ibm.com/connections/blogs/platformcomputing/?lang=en_us
- Dignan, L. (2014, March 31). Cloudera raises $900 million, plots expansion. *ZDNet*. Retrieved from http://www.zdnet.com/cloudera-raises-900-million-plots-expansion-7000027879/
- White, T. (2012). *Hadoop, the definitive guide.* Yahoo Press. Available at 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/
- Hornbeck, R. L. (2013, February 18). Batch versus streaming: Differentiating between tactical and strategic big data analytics. *L-3 Data Tactics Big Data Insights*. Retrieved from http://datatactics.blogspot.com/2013/02/batch-versus-streaming-differentiating.html
- Spark programming guide (code examples): http://spark.apache.org/docs/latest/programming-guide.html
- Resilient distributed datasets (RDDs): http://spark.apache.org/docs/latest/programming-guide.html#resilient-distributed-datasets-rdds
- MLlib—Basic statistics: http://spark.apache.org/docs/latest/mllib-statistics.html
- GraphX programming guide: http://spark.apache.org/docs/latest/graphx-programming-guide.html

Further reading:

- Karau, H., Konwinski, A., Wendell, P., & Zaharia, M. (2014). *Learning Spark*. O'Reilly Media. Available at http://shop.oreilly.com/product/0636920028512.do

Week 7: Storing even larger volumes of data and 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:

- Sliwa, C. (n.d.). Troubleshooting and identifying data storage performance bottlenecks. *TechTarget*. Retrieved from http://searchstorage.techtarget.com/report/Troubleshooting-and-identifying-data-storage-performance-bottlenecks
- Shafer, J. (2010). I/O virtualization bottlenecks in cloud computing today. In *Proceedings of the 2nd Conference on I/O Virtualization*. Available at http://dl.acm.org/citation.cfm?id=1863186&preflayout=flat#source
  Vajgel, P. (2009, April 30). Needle in a haystack: Efficient storage of billions of photos. *Facebook*. Retrieved from https://www.facebook.com/note.php?note_id=76191543919
  Gustafson's law. *Wikipedia.* http://en.wikipedia.org/wiki/Gustafson%27s_law
- Boudjnah, C. (2013, February 12). Ceph and Swift: Why we are not fighting. *eNovance*. Retrieved from http://techs.enovance.com/6427/ceph-and-swift-why-we-are-not-fighting
- FASP. *Apera.* http://asperasoft.com/technology/transport/fasp/

Week 8: Databases 2.0—the NoSQL movement. Taxonomy review. Consistent hashing introduction. CAP theorem and quorum algebras. Cloudant/CouchDB API introduction. Cloudant account creation. Cloudant MapReduce introduction. The obligatory word count. Graph databases.

Required:

- Metz, C. (2012, August 8). If Xerox PARC invented the PC, Google invented the Internet. *Wired*.
- DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., . . . Vogels, W. (n.d.). Dynamo: Amazon's highly available key-value store.
- NoSQL. *Wikipedia.* http://en.wikipedia.org/wiki/NoSQL
- Sign-up for free Cloudant.com account.
- Cloudant For Developers interactive tutorial
  - Reading and writing, primary index, secondary indexes, search indexes
- Cloudant API Reference (skim)

Optional:

- Stoica, I., Morris, R., Karger, D., Kaashoek, M. F., & Balakrishnan, H. (2001, August). Chord: A scalable peer-to-peer lookup service for Internet applications. *SIGCOMM'01*, San Diego.
- Gheawat, S., Gobioff, H., & Leung, S-T. (2003, October). The Google File System. *SOSP'03*, New York.
- Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified data processing on large clusters. *OSDI*.
- Chang, F., Dean, J., Ghemaat, S., Hsieh, W., Wallach, D. A., . . . Gruber, R. E. (2006, November). Bigtable: A distributed storage system for structured data. *OSDI'06*, Seattle.
- 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:

- Zaharia, M., Das, T., Li, H., Shenker, S., & Stoica, I. (n.d.). Discretized streams: An efficient and fault-tolerant model for stream processing on large clusters. Retrieved from http://www.cs.berkeley.edu/~matei/papers/2012/hotcloud_spark_streaming.pdf
- Stream computing in the cloud (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
- Amazon Kinesis (Amazon Web Services Blog post about launch): Launching Kinesis
- Amazon Kinesis: http://aws.amazon.com/kinesis/
- Amazon Kinesis Service API Reference: http://awsdocs.s3.amazonaws.com/kinesis/latest/kinesis-api.pdf
- Goetz, P. T. (2014, August 11). Apache storm vs. Spark streaming (Video). Retrieved from http://www.slideshare.net/ptgoetz/apache-storm-vs-spark-streaming?qid=dfbb7c09-6f87-40ca-a69d-d76837efd236
- Goetz, P. T. (2014, April 7). Hadoop Summit Europe 2014: Apache Storm architecture (Video). Retrieved from http://www.slideshare.net/ptgoetz/storm-hadoop-summit2014
- Zaharia, M., Das, T., Li, H., Shenker, S., & Stoica, I. (n.d.). Discretized streams: An efficient and fault-tolerant model for stream processing on large clusters. Retrieved from http://tinyurl.com/dstreams
- Spark streaming programming guide: http://spark.incubator.apache.org/docs/latest/streaming-programming-guide.html
- ETE 2012—Nathan Marz on Storm (Video). Retrieved from https://www.youtube.com/watch?v=bdps8tE0gYo
- Apache Spark tutorial: https://storm.incubator.apache.org/documentation/Tutorial.html

- AMPLab Camp on Spark streaming: http://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-centric, 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.

Reading:

- AWS CloudFormation: http://aws.amazon.com/cloudformation/
- 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 3.0 getting started: 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:

- van de Geijn, R. A., & Watts, J. (n.d.). SUMMA: Scalable universal matrix multiplication algorithm. Retrieved from http://www.cs.utexas.edu/ftp/techreports/tr95-13.pdf
- Message Passing Interface (MPI and 2D Cartesian communicators): https://computing.llnl.gov/tutorials/mpi/
- Introduction to parallel computing (HPC tutorial): https://computing.llnl.gov/tutorials/parallel_comp/
- MPI tutorial: A comprehensive MPI tutorial resource: 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 example.

Reading:

- Brin, S, & Page, L. (n.d.). The anatomy of a large-scale hypertextual web search engine. Retrieved from http://infolab.stanford.edu/~backrub/google.html
- Floyer, D. (2014, July 5). The growth and management of unstructured data. Wikibon.org. Retrieved from http://wikibon.org/wiki/v/The_Growth_and_Management_of_Unstructured_Data
- Google: 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
  Atreya, A. (n.d.). Nutch and Lucene framework (Presentation). Retrieved from http://tinyurl.com/k79hofu
- Baeza-Yates, R., & Cambazoglu, B. B. (2014). Scalability and efficiency challenges in large-scale web search engines. Yahoo Labs: Tutorial at SIGIR 2014, Gold Coast, Australia. Retrieved from 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. MapReduce/Platform Symphony. Speed vs. cost of large jobs and making the outputs available to users. An example: aligning a chromosome.

Reading:

- Computational genomics. *Wikipedia*. http://en.wikipedia.org/wiki/Computational_genomics
- Human population genomics. *IBM Research.* http://researcher.watson.ibm.com/researcher/view_group.php?id=2303
- Langmead, B., Schatz, M. C., Lin, J., Pop, M., & Salzberg, S. L. (2009). Searching for SNPs with cloud computing. *Genome Biology, 10*, R134. doi: 10.1186/gb-2009-10-11-r134. Retrieved from: http://genomebiology.com/2009/10/11/r134
- Wall, D. P., Kudtarkar, P., Fusaro, V. A., Pivovarov, R., Patil, P., & Tonellato, P. J. (2010, May). Cloud computing for comparative genomics. *BMC Bioinformatics*, 11, 259. doi:10.1186/1471-2105-11-259. Retrieved from: http://www.biomedcentral.com/1471-2105/11/259

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.

Reading and viewing:

- Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., . . . Welty, C. (2010, Fall). Building Watson: An overview of the DeepQA project. *AI Magazine*. Retrieved from http://www.youtube.com/watch?v=DywO4zksfXw
- Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., . . . Welty, C. (2010, Fall). Building Watson: An overview of the DeepQA project. *AI Magazine*. Retrieved from http://www.aaai.org/Magazine/Watson/watson.php
- Shenoi, M. (2014, September 18). IBM Watson analytics—Powerful analytics for everyone (Blog). Retrieved from http://mayashenoi.com/category/softlayer/
- WatsonPaths. *IBM Research.* http://www.research.ibm.com/cognitive-computing/watson/watsonpaths.shtml#fbid=OMIfqbmoDZr
- IBM Watson. (2014, August 27). *Introducing IBM Watson Discovery Advisor*. (Video). Retrieved from https://www.youtube.com/watch?v=qry_zGZFjOc