Schedule (subject to change)

All readings should be completed by the following week.
All assignments are due on the date listed, prior to the start of class, at 7pm.

Date	Topic / Guest / Readings	Assignments due
2015-09-01	Introductions; Jupyter and command line basics; VM setup. **Guest**: Shmuel Ben-Gad, Gelman Library Readings Required: Software Carpentry Lesson: The Unix Shell, http://software-carpentry.org/lessons.html Required: JHU Reproducible Research on Coursera, week one videos, https://www.coursera.org/course/repdata/ (about one hour) Recommended: Data Science at the Command Line, chapters 1-5	None
2015-09-08	The command line; input, output, and pipelines; csvkit; data types. Book review project. Readings Required: Software Carpentry Lesson: Using Databases and SQL, http://software-carpentry.org/lessons.html Required: Wickham, "Tidy Data." http://vita.had.co.nz/papers/tidy-data.pdf Recommended: Data Science at the Command Line, chapters 6-8	#1
2015-09-15	Command line filters; parallel processing; introduction to R/dplyr. Readings Required: Database System Concepts, chapters 1-3 (slides at http://codex.cs.yale.edu/avi/db-book/; text recommended) Optional: Learning SQL, chapters 1-4	#2
2015-09-22	**No class**	None
2015-09-29	RDBMS: schema, keys, basic SQL operations, aggregate functions, subqueries	#3, and book reviews start

	Readings Required: Database System Concepts, chapters 4, 5, 7, 8 (slides at http://codex.cs.yale.edu/avi/db-book/; text recommended) Optional: Learning SQL, chapters 5, 6, 7, 9, 10	
2015-10-06	RDBMS: joins, integrity, schema design and E-R models, normal forms; using MySQL from Python and R. Group project. Readings Required: Database System Concepts, chapters 11-13 (slides at http://codex.cs.yale.edu/avi/db-book/; text recommended) Optional: Learning SQL, chapters 12, 13, 14	#4
2015-10-13	**No class**	
2015-10-20	RDBMS: transactions, functions, triggers, indexes, query processing and optimization Readings Required: Star Schema, chapters 1-5	#5
2015-10-27	Warehouses: facts and dimensions, architectures, schemas Readings Required: Star Schema, chapters 6-8	#6
2015-11-03	Warehouses: dimension design Readings Required: Star Schema, chapters 11-13	#7
2015-11-10	Warehouses: fact table design Readings Required: Star Schema, chapters 14-18	#8
2015-11-17	**Midterm exam** Warehouses: performance, tools, documentation Readings Required: Dean and Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters." http://research.google.com/archive/mapreduce.html Required: Drake, "Command-line tools can be 235x faster than your Hadoop cluster." http://aadrake.com/command-line-tools-can-be-235x-faster-t	#9

	han-your-hadoop-cluster.html Required: Redis project. http://redis.io/ and Try Redis http://try.redis.io/ Optional: Chang et al. "Bigtable: A Distributed Storage System for Structured Data." http://research.google.com/archive/bigtable.html Optional: DeCandia et al. "Dynamo: Amazon's Highly Available Key-value Store", http://www.read.seas.harvard.edu/~kohler/class/cs239-w08/decandia07dynamo.pdf	
2015-11-24	noSQL and beyond: map/reduce, Hadoop, Redis, Dynamo Readings Required: CAP theorem. https://en.wikipedia.org/wiki/CAP_theorem Required: Apache Spark. https://spark.apache.org/ Required: Apache Storm. http://storm.apache.org/ Required: Apache Drill. https://drill.apache.org/ Required: AWS Redshift. https://aws.amazon.com/redshift/ Required: AWS Kinesis. https://aws.amazon.com/kinesis/	#10, book reviews end
2015-12-01	Spark and PySpark	#11
2015-12-08 (?)	Group projects	Group projects