DESIGNING AND IMPLEMENTING A DATA WAREHOUSE

INSTRUCTOR: STANISLAV SELTSER, STANISLAV.SELTSER@BU.EDU

1. Course Designator/Course Number

MET CS 689 Data Warehousing

2. Course Title

Designing and implementing data warehousing

3. Textbook

(1) Required: The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, Kimball 2013
ISBN-13: 978-1118530801

(2) Required: Hadoop: The Definitive Guide by White 2012,
ISBN-13: 978-1449311520

(3) Required: Graph Databases Ian Robenson 2013
ISBN-13: 978-144935622

(4) Optional: Programming Pig by Alan Gates
ISBN-13: 978-1449302641

(5) Optional: Python for Data analysis by Wes McKinney ISBN-13: 978-1449319793

(6) Optional: Practical MDX Queries: For Microsoft SQL Server Analysis Services 2008, Art Tennick,
ISBN-13: 978-0071713368

(7) class notes and readings

4. Course Length

This is a semester long intense graduate course which surveys current state of the art in data warehousing. Depending on program needs the class can be scheduled during the Fall, Spring or Summer semester. In the Fall and Spring semesters the class meets once a week for 3 hours over a total of 15 weeks. In the Summer semester the class meets once a week for 3.5 hours over a total of 12 weeks. In all cases there is a total of 45 contact hours. The course requires significant outside of classroom, consisting of assigned readings, homework, research and project. The estimate for outside of classroom work is at least 4 times the number of contact hours or a minimum of 180.

5. Course Description

This course provides students with the engineering skills required to evaluate, implement, and scale a modern data warehouse using commercially available and open source software. We begin by surveying classical data warehousing and OLAP concepts. We then move on to overview of current state-of-the-art technologies from Massive parallel databases to Hadoop stack, Array and Graph databases. We wrap up by looking at concepts of master data management and sense-making. Students will do 10 mini-projects as weekly assignments and one final project. 4 cr

6. Prerequisites

(1) MET CS 669 or MET579 , MET CS 520

(2) Strong programming skills: Python and Java

(3) Strong knowledge of relational database theory and SQL

(4) Elementary knowledge of statistics and probability theory

7. Online tutorials

- Python - http://docs.python.org/2/tutorial/
- Python - pandas http://pandas.pydata.org/pandas-docs/stable/10min.html
- Python - pytables http://www.pytables.org/moin
- Python - scidb interface http://www.paradigm4.com/watch-the-using-scidb-from-python-webinar-form/
- Microsoft SQL Server http://technet.microsoft.com/en-us/library/bb500155(v=sql.100).aspx
- Microsoft OLAP http://technet.microsoft.com/en-us/library/ms170208(v=sql.100).aspx

- Microsoft SSIS http://technet.microsoft.com/en-us/library/ms167031(v=sql.100).aspx
- Hadoop http://www.cloudera.com/content/cloudera-content/cloudera-docs/HadoopTutorial/CDH4/Hado
  Tutorial.html
- Hadoop - Pig https://cwiki.apache.org/confluence/display/PIG/PigTutorial
- Storm - https://github.com/nathanmarz/storm/wiki/Tutorial
- Summingbird - streaming map reduce at twitter https://www.youtube.com/watch?v=Y3PETLJeP7o
- Neo4j https://github.com/jimwebber/neo4j-tutorial
- SciDB

  `http://www.scidb.org/HTMLmanual/13.6/scidb_ug/index.html`
- Tableau www.tableausoftware.com/public/download
- Netezza www.slideshare.net/bijugs/netezza-fundamentals-for-developers
- Mesos https://www.youtube.com/watch?v=37OMbAjnJn0

8. Software tools we will be using

(1) Windows VM: Erwin(Community Edition), Python (Anaconda distribution), R, Rstudio (ggplot2,reshape,data.table) , Microsoft SQL Server OLAP 2012, Microsoft SQL Server 2012, Microsoft Integration services, Excel, Tableau,Panopticon evaluation

(2) 3 Linux VMs: R, Python, Postgres, SciDB, Vertica, Cassandra, Neo4j, Hadoop Cloudera, Talend, Spark/Shark

9. Readings

– all the readings below are available for free online:
The Fourth paradigm by Jim Gray
How Vertica was the star of Obama campaign, and other revelations
MAD Skillls: New Analysis practices for big data
Mining of massive datasets Ullman
MapReduce and parallel RDBMS: Friends or Foes by Stonebraker
Scalable SQL and NoSQL data stores by Rick Cattell
Duplicate record detection by Elmagarmid
Record Linkage: Simularity measures and algorithms by Koudas
The Joy of Stats by Rosling
Tools for data enthusiasts by Hanaran
A tour through Visualization zoo by Heer and Bostock

10. Topics

10.1. **Week 1.**

(1) Use cases for data warehousing: Personal budgeting, Quantative Finance, Mobile advertisement, network-centric military strategy.

(2) Metadata: Knowledge representation: Relational, Graph, Key-Value pair. Concept of metadata.

(3) Metadata: Extracting Knowledge from data (relational sources, unstructured data (ETL, data scraping) using metadata.

(4) Data quality monitoring. Bad data detection. Proxy rules on bad data.

(5) ETL engines: Microsoft SSIS, Talend. Pentaho Kettle/GeoKettle. Informatica PowerCenter. Hadoop as ETL engine for staging and transforming.

10.2. **Week 2.**

(1) Metadata: Temporality and bitemporality, Historic and Current views of data. Provenance. Change data capture, incremental data calculation. Concept of data fusion from multiple sources - geo sensoring, numeric tensors, text, audio, video, imagery, social

(2) Logical Data models overview: Kimball, Anchor, Data Vault. Dimensional modelling. Ontology. Protege(OWL)

(3) Logical data models: Dimensions, Measures, Grain, Facts, Many-to-Many modeling, Hierarchies, Attributes and Attribute space, Change Tracking. Attribute vs Dimension. Hierarchies changing over time.

10.3. **Week 3.**

(1) Logical data models: Derived Measures, Scope calculations, Aggregations across time and space. Relative time calculations. Notion of Lattice of aggregations. Partial materialization. Curse of dimensionality. Hyperroll.

(2) Compute engines: OLAP vs Hadoop. in-memory data grids.In-memory compute grids - Grid Gain.

10.4. **Week 4.**

(1) OLAP operations: slicing, dicing, pivot and unpivot, drill down, drill-across, write-back

(2) OLAP measures: additive, semi-additive, ratios

(3) OLAP aggregation functions: sum, max, weighted average, count

10.5. **Week 5.**

1. OLAP: proxy on storage vs proxy on retrieval. Importance of 3-valued boolean logic.
2. Physical data models for DW: Relational, Array, Graph, Key-Value pair, Cube
3. Physical data models: Languages for data warehousing: SQL, MDX, GraphQL, AQL, SparQL

10.6. **Week 6.**

1. Relational databases: Postgres, MySQL, VoltDB. Sharding as cheap way to scale.
2. Bitmapping as way to navigate vast multidi-dimensional spaces - FastBit, Druid
3. OLAP engines - Microsoft OLAP, Pentaho Mondrian OLAP, Microstrategy, IBM Cognos, Hyperion.

10.7. **Week 7.**

1. MPP Relational databases - Netezza, Vertica, Teradata, EMC Greenplum, Sybase IQ, Oracle Exadata. Column stores vs row stores.
2. Array databases: SciDB
3. GraphDatabases - Graphlab, Neo4j, Apache Giraph, Titan; Lightweight graph libraries: PyGraphviz and applications in GraphOLAP.
4. File-based databases - HDF5

10.8. **Week 8.**

1. NoSql databases - MongoDB, HBase, Casssandra, Spark/Shark, Sqrrl
2. Cloud databases Amazon Redshift, EC2, S3. CloudFront
3. Real-time streaming databases aka CEP: Streambase, Storm. Event-drven data warehousing discovery based on real-time and historical pattern matching.
4. Hybrids of real-time and streaming databases - Summingbird from Twitter.

10.9. **Week 9.**

1. Text search engines - Solr. Query splitting engines with hybrid storage - Hadapt and Aster data. HAWQ and Pivotal HD.
2. Dissemination of information: publish/subscribe, mobile views. Message-centric (JMS, Red Hat AMQP) vs data centric (OMG DDS) pub/sub.
3. Cluster management systems: Mesos

10.10. **Week 10.**

(1) Sense making techniques and disambiguation of information. Entity/Identity resolution. Record linkage Detection of ambiguity. Sense-making based on sparse matrix representation (Saffron, Grok Solutions comparison). Context-based probabilistic mapping as disambiguation of information technique.

10.11. **Week 11.**

(1) Data visualization with Tableau and R(shiny, GoogleVis). Streaming visualization with Panopticon. Data visualization using D3.js and iPython notebook/matplotlib.

(2) Data warehousing over Hadoop. Platfora , Karmasphere, DataMeer

10.12. **Week 12.**

(1) Benchmarking and evaluation methodology for data warehouse. Yahoo cloud serving benchmark. Intel-Hadoop HiBench. Pavlo - Comparison of Approaches to Large-Scale Analysis of Big Data Benchmark - CS Berkeley AmpLab.

10.13. **Week 13.**

(1) Final Project is due