DESIGNING AND IMPLEMENTING A DATA WAREHOUSE

INSTRUCTOR: STANISLAV SELTSER, SSELTSER@BU.EDU

1. Course Designator/Course Number

MET CS 689 Data Warehousing

2. Course Title

Scalable infrastructure for analytics

3. Course objectives

(1) Undersand analytics types and mapping to infrastructure

(2) Student is able to pick and choose appropriate physical and logical DW methodology: Relational DW, Distributed Relational DW, OLAP, or unstructured storage(Hadoop). Student is able to explain cost-performance tradeoffs and justify the choice of technology.

(3) Student is able to define baseline benchmarks and implement them and compare the numbers across platforms being considered.

(4) Student will demonstrate familiarity with principal languages of DW: Python, SQL, MDX, Apache Pig, Apache Spark

(5) Student will be able to perform data transformations using ETL tool of choice - Python , SQL

4. Textbook

(1) Required: OLAP Solutions: Building Multidimensional Information Systems, Eric Thompsen
2002, ISBN-13: 978-0471400301

(2) Required: Python for Data analysis by Wes McKinney ISBN-13: 978-1449319793

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

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

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

(6) Optional: Learning Spark: Lightining fast data analysis, 2015
ISBN-13: 978-1449311520

5. Course Length

This is a semester long intense graduate course which surveys current state of the art in data warehousing. The class meets once a week for 3 hours over a total of 15 weeks. The course requires significant effort 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 160 hours

6. 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 of HDFS + Accumulo + Spark. We wrap up by looking at concepts of master data management, sense-making, data visualization Students will do 6 2 week long assignments, weekly quizzes and one final project. The final project will have weekly milestones 4 cr

7. Course grading

We will do 6 assignment projects and one final project. The final project is worth 50% of the grade, quizzes are 10% assignments are 40% total the grade is based on relative scale.

8. Prerequisites

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

(2) Recommended elementary knowledge of Python programming language or any other programming language.

(3) Basic knowledge of SQL

(4) Basic knowledge of relational data modeling

(5) Elementary knowledge of Windows and Linux OS basic commands

9. Online tutorials

- Python `https://docs.python.org/2/tutorial/`
- Python Pandas library `http://pandas.pydata.org/pandas-docs/stable/tutorials.html`
- Vertica `https://my.vertica.com/docs/5.1.6/HTML/index.htm#8871.htm`
- Analytical functions in Vertica `https://my.vertica.com/docs/5.1.6/HTML/index.htm#10955.htm`
- Microsoft OLAP `http://technet.microsoft.com/en-us/library/ms170208(v=sql.100).aspx`
- Hadoop `http://www.cloudera.com/content/cloudera-content/cloudera-docs/HadoopTutorial/CDH4/Hadoop-Tutorial.html`

10. Software tools we will be using

(1) Windows VM: Microsoft SQL Server OLAP 2012, Microsoft SQL Server 2012, Microsoft Integration services, Excel, Tableau public

(2) Linux VMs: Vertica, Hadoop

11. Readings

(1) Thomsen Chapters 4 – 7

(2) Kimball Chapters 1 – 2, 18 – 21

12. Topics

12.1. **Week 1-2 - Introduction to Data warehousing and OLAP.**

(1) Python as prototyping language for data acquision

(2) Use cases for data warehousing: Fusion of heterogeneous data sources.

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

(4) Metadata: self-describing knowledge representations - JSON

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

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

(7) ETL engines and data munging: Microsoft SSIS and Python

(8) Analytic functions in SQL: lag/lead, row_number

(9) Technology used in course: Vertica, Microsoft OLAP and Hadoop

(10) Virtual Machines: VMWare and Oracle Virtual Box

12.2. **Week 3-4 Data modeling for Relational DW.**

(1) Dimensional modelling

(2) Metadata: Temporality and bitemporality, Historic and Current views of data.

(3) Provenance. Change data capture, incremental data calculation.

(4) Concept of data fusion from multiple sources

(5) Logical Data models overview: Kimball Dimensional modelling. Ontology.

(6) Logical data models: Dimensions, Measures, Grain, Facts, Many-to-Many modeling,

(7) Hierarchies, Attributes and Attribute space, Change Tracking. Attribute vs Dimension.

(8) Hierarchies changing over time. Snowflakes

12.3. **Week 5-6 Data modeling for OLAP.**

(1) Logical data models: Derived Measures, Scope calculations,

(2) Aggregations across time and space. Relative time calculations.

(3) Notion of Lattice of aggregations.

(4) Partial materialization. Curse of dimensionality.

12.4. **Week 7-8 Data modeling for OLAP - Advanced topics.**

(1) OLAP: default context
(2) OLAP: operations: slicing, dicing, pivot and unpivot, drill down, drill-across, write-back
(3) OLAP: measures: additive, semi-additive, ratios
(4) OLAP: aggregation functions: sum, max, weighted average, count,first, last, rank,top N
(5) OLAP server side calculations vs client side.
(6) OLAP: storing data on multiple levels of hierarchy
(7) OLAP: proxy on storage vs proxy on retrieval. Importance of 3-valued boolean logic.
(8) OLAP for graphs
(9) OLAP for images
(10) Physical data models for DW: ROLAP vs MOLAP
(11) Physical data models: Languages for data warehousing: SQL, MDX

12.5. **Week 9-10 Vertica and Hadoop with OLAP on top.**

(1) Vertica integration with Hadoop
(2) Concept of invisible data load and in-situ data processing
(3) How OLAP integrated with Vertica
(4) How MDX is compiled into SQL
(5) Hadoop stack: HDFS
(6) Hadoop stack: Accumulo
(7) Hadoop stack: Spark
(8) Hadoop stack: Parquet
(9) Hadoop stack: Hive and Pig
(10) Hadoop stack: Spark python
(11) Hadoop stack: Spark SQL

12.6. **Week 11-12 Decision making using OLAP.**

(1) Pivot Tables
(2) Dissemination of information: publish/subscribe.
(3) Data visualization with Tableau and R(ggplot, ggmap).
(4) Data visualization using D3.js and iPython notebook/matplotlib.
(5) Big data platforms for datawarehousing and data mining
(6) Heterogenious platforms

12.7. **Week 13.**

(1) Sense making techniques and disambiguation of information.
(2) Entity/Identity resolution.
(3) Record linkage
(4) Detection of ambiguity.
(5) Benchmarking and evaluation methodology for data warehouse.