# CIS 450/550: Database and Information Systems
# Fall 2016

## Course Project: Summer Olympics Dataset

### PROJECT OVERVIEW

The goal of this project is to create an application of your own choosing over an Olympics dataset. The dataset in itself is not rich enough for developing interesting applications, so you will need to find supplemental data sources to complement it. As part of developing this project, you will exercise: schema design, cloud hosting, data extraction, entity resolution, SQL queries, noSQL queries, and performance considerations. The project is designed to offer flexibility in choosing features to implement, and to think about how to make these features efficient.

You are allowed to develop any application/website/game which uses the Olympics dataset and other complementary data sources, so be creative! It is your choice as to what you want to emphasize: countries, athletes, sports, medals, or all of them. The intent of the project is to help you understand the importance of good database design, to introduce you to techniques for extracting information from partly structured sources, and to learn database and application hosting on Amazon Web Services platform.

You should build your application using a relational technology and augment your basic design with non-relational (noSQL) technologies.

We are also open to your developing an application over a different dataset of your choosing, but you must clear the idea with us first before moving ahead.  You must also follow the project requirements listed on page 3 of this handout.

## DATASETS

The Summer Olympics dataset (SummerOlympicMedallists1896-2008.xlsx) can be downloaded from the Guardian in CSV format https://www.theguardian.com/sport/datablog/2012/jun/25/olympic-medal-winner-list-data. The dataset contains information about all medalists at the summer Olympic games by year, sport, discipline, gender and event between 1896 and 2008. The spreadsheet also includes other analysis over the main datasets, such as the breakdown of medals by gender, country, etc. Depending on your application idea, you will need to extract complementary datasets to make your application more informative (and interesting). You will need to use content extractors (e.g., the Jackson JSON parser, the Tika reader for many file formats) that read the contents of the raw data items, interpret the file format, and extract the required information. You will likely want some of the following libraries to extract data from (partly) structured files, such as html (Wikipedia, World Factbook, DBpedia), CSV, XML, JSON, and text files.

- https://tika.apache.org/ (reads Word docx, PDF, … text and headers)
- https://github.com/FasterXML/jackson (reads JSON)
  https://docs.oracle.com/javase/tutorial/jaxp/ (reads XML)
- https://commons.apache.org/proper/commons-csv/ (reads comma separated)
- https://jsoup.org/ (reads html)

Here are some examples of complementary data sources:

- Wikidata, https://www.wikidata.org/wiki/Wikidata:Main_Page
- DBpedia, http://wiki.dbpedia.org/
- Wikipedia, https://en.wikipedia.org/wiki/Main_Page
- World Bank Open Data, http://data.worldbank.org/
- World Factbook, https://www.cia.gov/library/publications/the-world-factbook/
- Greatest Sports Nation, http://www.greatestsportingnation.com/

However, there are many other sources of interesting data that you may choose instead.

## PROJECT REQUIREMENTS

Your project should have the following features:

- Information drawn from at least two other datasets other than the Summer Olympics dataset
- Diversity of data types: the Summer Olympic dataset can be downloaded in CSV format, so at least one of the other datasets should be something different, e.g. HTML, XML or Word.
- Data cleaning: frequently, online data is incorrect and incomplete. As you import the data, you should run simple scripts to check for errors.
- Entity resolution: Since you are importing data from multiple sources, you will need to "connect" the data. For example, if you are focusing on Athletes in your application you may need to use athlete name (and possibly country information) to relate the imported data to the Summer Olympic dataset. Note that there may be misspellings of names (data cleaning issues!).
- Normalized relational schema with more than four relations.
- More than one interaction page with the database.
- Complex SQL queries, i.e. with multiple joins, subqueries, aggregation etc. These can be in the application queries, or part of the schema design (e.g. triggers).
- Consideration of performance, including indexing. (See experimental validation requirement in Milestone 4.)

## EXTRA CREDIT

- Trigger Bing Search, see http://datamarket.azure.com/dataset/bing/search, to return additional information.
- Import login and user information from Facebook, Google, or Microsoft authentication services.
- View-based access control (e.g. if personal data is part of the application and there are privacy concerns)
- Adding streaming data (e.g. from Twitter feeds)
- Anything else you think is intuitive and adds interest to the application.

# Project Milestones

## Milestone 1 (October 10th)

### Form a team (size 3 or 4), develop initial idea, and set up infrastructure.

The initial step is to select your teammates and do the following:

1. Develop an initial idea, and determine the technologies you wish to standardize on as a group. Amazon Web Services should be used for hosting your database and deploying your application.
2. Provide 6-10 questions (in English) that someone might want to ask about the domain of your intended application. (You will be permitted to revise these questions later if needed.)
3. Setup Subversion/Git to share source code and starter data files. See http://www.seas.upenn.edu/cets/answers/subversion.html for details, and be sure that whoever sets it up grants access to everyone in the group. You should also add your assigned TA and Professor Davidson to it so we can see what you are doing.
4. One group member should then upload a PDF document via Canvas, stating who is in the group, what your initial idea is, and what technologies you plan to use. The document should also include a timeline for the different milestones of your project, and a preliminary division of responsibilities.

Based on this description, we will assign each group a TA who will follow your progress throughout the remainder of the semester. You should consult them early and often, and get their input as you refine your ideas of the features to be implemented in your project (Milestone 2).

Each group member should apply for AWS Educate, which grants $100 in usage credits. You need to use your ***.upenn.edu*** email address to register. The approving process may take about one week, so **apply early**!

With a total group amount of about $300, you should have enough to complete the project. However, if you exceed this amount through carelessness you will be responsible for overages. By this, we mean that you should turn off instances whenever they are not being used and NEVER publicly share your id and password or put them somewhere that they can be compromised. We had an incident last year where hackers used the AWS Keys and deployed several EC2 instances and a bill of more than 1000$ was generated.

## Milestone 2 (October 31st)

### Project outline and schema design.

In this phase, you will explain your project idea in more detail, and discuss how it uses the Olympics dataset. Your project idea should contain the following:

- Motivation for the idea
- Features that will definitely be implemented in the application
- Features that might implemented in the application, given enough time
- Technology and tools to be used
- Description of the complimentary sources you intend to use for data, and how you intend to ingest the data into your database
- Member responsibility for project components

It is important to establish early on specific project component responsibilities – each group member should have aspects of the project that they "own" and are responsible for. "Own" does not necessarily mean they will be doing all of the coding / development, but rather that they are responsible for making sure the feature is complete.

You should also design a relational schema for your application, as well as a description of the noSQL component, if you choose to add one. Your schema should be based on the application rather than a straightforward copy of the dataset provided. The relational schema should be represented as an ER diagram, as well as through (normalized) SQL DDL.

For the web technology, we prefer you to use Node.js – however, if your team feels this is too difficult, then you may use something simpler (e.g. PHP). It is most important for you to be able to complete the project.

For this milestone you should submit a file with the information above, along with the relational schema and noSQL description.

## Milestone 3 (November 14th)

### Populate the database.

Now that you have a baseline schema to work on, the next part is to populate the database. You must extract from 1) the Olympics dataset that we provided, and 2) the complementary resources that you decided to use. You should clean and format data from your main and complementary sources, as needed, and perform entity resolution.

For each of your 6-10 questions (from Milestone 1) provide its translation to an SQL query. (If you needed to change any of your original questions, provide the new questions in English and explain why you needed to change or replace them.)

You should use the AWS Getting Started handout to create your own MySQL (cheaper!) or Oracle database on Amazon RDS. For the milestone you should submit a text file with a full JDBC/SQLPLUS connect string, including guest user ID and password and database schema name, to us via Canvas. (From this we should be able to dump your SQL tables.)

## Milestone 4 (November 23rd)

### Demo basic functionality

In this phase, you should have a running application with some basic features. Submit the source code and a brief document of the list of features through Canvas; you should also set up a time to demo what you did to your overseeing TA to get their feedback.

## Final Milestone (December 13th-14th)

### Project Demonstrations

Your final demo should contain all the basic and/or advanced features mentioned in your report along with any extra credit implemented. You should also give the instructors an **updated copy of your project description** (Milestone 2) prior to starting the demo.

# FINAL DELIVERABLES (DECEMBER 16TH)

## A. EXPERIMENTAL VALIDATION AND REPORT

A modern software infrastructure project isn't done until you understand how it performs, and where the bottlenecks are. Instrument your application to collect timings on various aspects. You should at least be able to determine what the latency in handling each request is, and extra credit will be awarded if you can also see what happens under multiple concurrent requests.

Your final report should include a write-up of:

1. Introduction and project goals
2. Basic architecture (not a dump of the classes)
3. Key features of the project.
4. Technical challenges and how they were overcome
5. Description of your complementary data and how you extracted it
6. Performance evaluation
7. Potential future extensions

## B. CODE

The entire project code along with the final report should be zipped and submitted on Canvas.

## SAMPLE IDEAS TO TRIGGER YOUR CREATIVITY

- **2020 Olympics Nations Performance Predication App**

  Create a web application, which could predict the performance of nations in 2020 Olympics using data from the Olympics and complementary datasets. You could use the Olympics dataset to perform an analysis over past games to predict how would a given country perform (i.e., the number of gold, silver and bronze medals it will receive) at the 2020 Olympics. You could use data from other sources, such as the World Bank or the CIA World Factbook, to get a country's GNP, population, etc. Then use such data as input to your prediction algorithm. In addition, you could use visual tools to depict the background analysis.

- **Olympics Athletes App**

  Create a web app that provides information about Olympic athletes. You could use the main Olympics dataset to provide some basic information about athletes, including their country, events they participated in; medals they won, etc. For famous Olympic athletes, you could also use other data sources, such as Wikipedia, to provide other information, such as photos, personal life, non-Olympic competitions, honors and awards, world records and more. Some of such information (e.g., photos) can be stored in a noSQL system, such as MongoDB.

- **Greatest Olympic Nations App**

  Since Judo was first included in the Olympics in 1964, Japan has been dominating this sport. Kenya runners claimed the top spots in the men's 3,000-meter since 1984. It would be interesting to know why some nations are superior in some Olympic sports. For that, you could develop an app that uses the Olympics dataset to find the best nation for each Olympic sport. Then, your app should crawl other resources to find any unique features, such as geographic location, weather, government policies, food culture, sports origin, people demographics, and historical events, that could explain the reasons behind a country's supremacy in an Olympic sport. You app could be more sophisticated by being able to decide which features to look for based on the Olympic sport. It could also predict future dominating nations for a given sport based on the unique features similarity.

## Plagiarism Policy

You can refer to web or any other resource for ideas, but you are STRICTLY NOT ALLOWED to use other people's code directly. In case you would like to use some code or snippets, please consult your mentoring TA before you do so. Please make sure that you cite the original author/source if you are approved to use it.
If you are caught under Plagiarism, academic measures will be taken as directed by:
http://gethelp.library.upenn.edu/PORT/documentation/plagiarism_policy.html