**ISIT 420 Course Project**

Pick two of the public data files.

I listed a few sources at the end of this document, but also search for others. (If you use the stock market one we already used in class, make sure you do something with it “very different” than we did in class.)

Download the data, import it into PIG, and “massage” the data into a specific subset useful for your chosen analysis. Do things such as eliminating columns so you go from 8 variables to 4, eliminate rows so you go from 20000 rows for the US down to 498 rows for just the state of WA. Aggregating and computing totals is a good thing to do.

At some point you will need to do some kind of inner join between the data sets you imported. So figure out what will be the primary key from one dataset and what will be the corresponding foreign key in the other data set..

Export the data out from PIG and import into SQL as at least twotables, but they should be normalized, so it will likely be 3, 4 or even more tables.

For example, assume I started with the city data from some online data source that has csv data:

locId, country, region, city, postalCode, latitude, longitude, metroCode, areaCode,

Using PIG, I would get rid of all entries that are not in the USA, then I would drop the locId, country and metro code columns. I would also change “region” to “State”, and I would get rid of any rows that did not have valid data.

So as I left PIG, I would have

State City PostalCode Latitude Longitude AreaCode

I would then normalize this into 3 SQL tables,

ZIPtable

ZIPcodeID ZipCode CityID Latitude Longitude

CityTable

CityID CityName AreaCode StateID

StateTable

StateID StateName

To extend upon my example above, maybe I would import data that included the average temp by month for major US cities. I would use either zip code or CityName to link the 2 data sets with primary/foreign keys) I could then do some interesting queries using the data from both my imported data sets.

(Given one original pig data source file, how do you map that into 2 or more tables with relationships? I will do a little PPT explanation of that in class.)

Then using ASP API with a JavaScript client for writing an application

Create an ADO.NET Entity Data Model, using EF Designer From Database approach

Using that “object context” (data model), along with LINQ to Entities, provide a web user interface that lets the user analyze the data by providing at least 3 interesting queries.

Projects will be due Wednesday night, 11:59 June 17 , and you will be required to make a 15-minute presentation to the class Friday June 19 (class starts at 1:30, not 5:00!) showing your process, code, and results.

To complete your assignment for a grade; I will need:

1. links to your 2 data sources
2. Your PIG scripts
3. The output of your PIG runs, as a zipped file.
4. A SQL database diagram showing the tables and relationships you created from the import of the PIG-data
5. A zipped copy of your ASP API / JavaScript project
6. A PPT that you use to present your project to the class. This should include screen shots of your PIG process and outputs. You will have to explain the Pig code steps. When you present, you should run your ASP program live for the class, but also please include screen shots in the PPT in case something bad happens.)

Just a few examples of places you can download data sets from:

* 20 Big Data Repositories You Should Check Out
  + <http://www.datasciencecentral.com/profiles/blogs/20-free-big-data-sources-everyone-should-check-out>
* NYT’s The Movie Reviews API (this one can be fun, but I remember a student had a problem with this one and could not get the data downloaded. If you choose this one, make sure you can download the data before settling on that decision.
  + <http://developer.nytimes.com/docs/movie_reviews_api/>
* The online book has a long list at the end
  + <http://hadoopilluminated.com/hadoop_illuminated/index.html>
* Govt data (this one has many many good ones)
  + <https://www.data.gov/>