Assignment 3 - Your Project Database¶

Due Tuesday, October 31, 2017

Submission: Put the data and Jupyter notebook files in a folder. Make sure all links to data are relative to the folder so the TAs can run the notebooks.

You MUST use the data set that you will use in your project.

In this assignment, you will create a database for analysis for your project. The assignment can be done together for group projects.

Part A - Create a database (30 points)¶

- Create a database for your project. If a relational database it must be in at least third normal form. Your database can be SQL, NoSQL or Hadoop Distributed File System (HDFS). Show your schema.

Part B - Queries (30 points)¶

- Generate a list of five common queries for your data.
- Implement the five common queries.
- Time the implementation of the five common queries and give an estimate of how they scale.

Part C - Pipeline/automation (20 points)¶

- Design at least one pipeline or sequence of analysis.
- Implement at least one pipeline or sequence of analysis. This can by done with Celery, Dask, Luigi, Airflow or through shell scripting.

Part D - Back-up your data (20 points)¶

- Create a scheme to back-up your data.
- Implement your scheme to back-up your data.

Last update October 3, 2017