The objectives of the term project are,

1. Performing large-scale analysis
2. Using technologies typically used in modern data centers
3. Interpreting your results

To accomplish above objectives, you should select an analysis run over one of the datasets listed in the public EBS catalog (http://aws.amazon.com/datasets/). Example datasets and analyses include the following:

E1) “Google Books Ngrams”: Contrasting British and American English expressions by performing a bi-gram likelihood analysis
E2) “Google Books Ngrams”: Contrasting language evolution in 10-20 year intervals using TF-IDF Cosine similarity
E3) “Million Song Dataset”: Finding similar songs based on metadata using Euclidian distances
E4) “Wikipedia XML data”: Keyword search using an inverted index
E6) “Transportation Databases”: Correlation analysis between features available within the dataset
E7) E6 can be extended by combining “Transportation Databases” and other datasets (e.g. economics dataset or census datasets)
E8) “Labor Statistics Databases”: Correlation analysis among features (you may consider time as a feature as well)
E9) “Common Crawl Corpus”: Running the PageRank algorithm over the available crawl datasets

Your topic is not limited to the above examples. You may come up with something on your own as well. The requirements for your analysis are:

1. You should use complete (or reasonably large portion of) dataset from what’s available on AWS publically.
2. You should run your analysis on EC2 and/or the EMR (Elastic Map Reduce) environment
3. Select an additional service in AWS besides (1) and (2), and “improve” the quality of your system
4. Your interpretation of results should be included in the final report

Your proposal should include: (1) the targeted dataset, (2) the goal of your analysis, (3) your analysis methodology and (4) the selection of services. Please send your proposal to cs480@cs.colostate.edu by March 6, 2015 5:00 pm. Please send only one email per team, and cc to all of the team members.