Please don’t use this information as a benchmark. Sky is the limit for the ideas and implementation of the project.

Projects should propose solutions that are scalable, i.e., "big data" solutions. You can use "small data" to develop, test, and demo your solutions, but your solutions should scale to larger datasets. This implies that you need to use a big data framework, e.g., Spark to implement your solution.

1) Propose your own project that involves some useful data analysis (e.g., related to your work or thesis topic)
2) Develop a commute time prediction model for Deerfoot trail data (data and Spark code fragments will be provided)
3) Variation of 2 - Develop a classifier that can predict whether a given day will be "good" or "bad" in terms of rush hour commute
4) Analysis of stack overflow data - https://www.ics.uci.edu/~duboisc/stackoverflow/ - build a classifier for each user that can predict whether that user will be able to answer a given question or not
5) Predict trip times of NYC taxis (https://www.kaggle.com/c/nyc-taxi-trip-duration/data)
6) Predicting the response times of various Web transactions at various load levels from log file data
7) Variation of 6) predicting whether the response time of a transaction will likely exceed a certain threshold at a certain load
8) Variation of 6) - Find groups of transactions that have similar response time behaviour, i.e., similar response times under similar loads.
9) Find publicly available Web server access logs - from these logs try to identify users with similar navigational behaviour, e.g., in terms of URLs visited and/or in terms of whether they are robots or real users.