Big Data Challenge for high school students

Erik Spence

SciNet HPC Consortium

5 November 2014

Big Data Challenge, introduction

The Canadian Young Scientist Journal is leading a Big Data Challenge for high school students.

- Students from high schools across the country are invited to participate.
- Prize for the top team: $1000.
- Assistance and sponsorship provided by SAS, SciNet, and iEcarus.
- The goal is to expose students to the techniques and approaches used in Big Data analytics.

Big Data Challenge, overview

Students will be supplied with a real data set of grocery store transactions. These data include

- the number of visits to the stores by the customers, per week.
- categorized shopping lists of the customers.
- amount of money spent by the customers.
- information about the shoppers themselves
  - date of birth,
  - gender,
  - address (postal code).
- information about the 5 stores being visited.

The shopping data itself consists of over 50,000 transactions.

SciNet compute • calcul CANADA

Big Data Challenge, overview

Using the provided data, students are asked to build a model to analyze the data, and address one or more of the following questions:

- Predict next weeks shopping list for each consumer.
- Analyse changes in consumer behaviour.
- Make recommendations regarding grocery-store layout.
- Attempt to classify the consumers by household size, income, demographics.
- Classify consumers by life style, political preferences.
- Any other indicators the students find interesting - be creative!

Students are free to use any supplementary open/freely available data sets they can find in order to support their model and analysis.

SciNet compute • calcul CANADA

Tools

Students are free to use any modeling/computational tools available to them for their analysis. Some common tools available include:

- R: by far the most-common open source tool for Big Data analytics. Very common in the Big Data community.
- Python: another open-source tool. Not as commonly used as R.
- SAS: a commercial code with extensive functionality.

Other languages and commercial codes are available. Students have complete freedom in this regard.

SciNet compute • calcul CANADA

SAS resources

SAS has provided the following resources, which students are welcome to use if they so desire.

- University Edition of SAS: http://www.sas.com/en_us/software/university-edition.html
- SAS On-Demand For Academics: a free cloud version of 5 of SAS' products including Enterprise Miner.
- SAS will be providing a course for the competition http://www.sas.com/govedu/edu/programs/od_academics.html:
- Free E-Learning from U of T. See this PDF; the Access code is G70072789. There are twelve different courses for students learn how to use SAS.
- http://www.lexjansen.com, a great website for students to search and learn about different analytical techniques.

Challenge notes

Further notes about the Challenge:

- Students should form teams of 2-4 participants.
- Students are encouraged to have a mentor.
- SciNet analysts are available for consultation throughout the competition. They can be contacted at bigdatachallenge@scinet.utoronto.ca.
- General inquiries and submissions can be directed to bigdata@cysjournal.ca.

Students are encouraged to request assistance if they need it. The dataset itself, as well as pointers to resources, can be found at http://wiki.scinethpc.ca/wiki/index.php/BigDataChallenge2014.

SciNet compute • calcul CANADA

Challenge timeline

The sequence of events for the Challenge:

- November 15, 2014:
  - Students should provide a list of participants, school affiliations, contact information and mentor information.
  - Submit the $100 participation fee.
  - Submit a 1-2 page abstract, describing the team's motivation and goal for the challenge.
- January 12, 2015:
  - Submit the analysis report, describing hypotheses, methodology, results and discussion.
  - Submit the computer codes used to analyse the data, for evaluation and reproducibility.

Challenge timeline, continued

The sequence of events for the Challenge, continued:

- end of January, 2015:
  - Judges will announce the short list of 5 finalist teams.
  - These teams will be invited to present their work at the University of Toronto, or online if travel is not possible.
- February 13, 2015:
  - The 5 finalist teams will present their projects to a panel of experts at the University of Toronto.
  - Judges will select the top 3 winning teams.

Top prize: $1000. Second and third prizes will be determined by the number of participanting teams.

SciNet compute • calcul CANADA