Reviewer's report

Title: A User-friendly Tool to Transform Large Scale Administrative Data into Wide Table Format using a MapReduce Program with a Pig Latin based Script

Version: 1 Date: 19 October 2012

Reviewer: Bram Adams

Reviewer's report:

The authors propose user-defined functions for the Pig environment in order to enable users to easily convert medical data into wide table format. The functions also improve Pig’s ability to deal with dates.

In favour:
* useful tool
* detailed performance analysis

Against:
* functionality and implementation of user-defined functions not clearly explained, so hard to know what the tool really does

I ordered my review comments based on the questions asked to reviewers. The comments for questions 2, 3 and 6 are Major Compulsory Revisions, the detailed comments at the bottom are Minor Essential Revisions. Everything else are Discretionary Revisions.

1. Does the software address a novel task? Alternatively, if there is already software available that performs this task, does the software outperform it in terms of speed, reliability, efficiency, or breadth of application?

The authors motivate their tool well. There are existing tools available to achieve the same goal, for example SQL databases and queries, but the paper provides numbers to show how the proposed tool scales. Unfortunately, no comparison is made between the Pig solution and an SQL-based one. Similarly, other environments like the R programming language with the reshape package (http://had.co.nz/reshape/) might be interesting as well to compare to.

2. Is it easy to use?

Yes and no. The authors claim that it is easy to use and they show examples to prove this, however the naming of the user-defined functions is rather vague and hard to understand. Furthermore, the description of each function is very unclear. Even with the examples, I was unable to say what each function really does, even for the major ones like GroupFilterFormat and PickupSequenceValues. For
example, what are “key/item codes”, “unit bases”, “re-categorize groups”, “unique function”, …?

3. Does it satisfactorily address the task or application the authors intend?

Performance-wise yes, as shown in the detailed performance study. User-friendliness-wise, I cannot really tell, as mentioned for question 2.

4. Is the software freely available for non-commercial use (note that this is a condition of publication)? And is the availability of the software and any restrictions on use clearly stated in the manuscript?

Yes, it’s open source, which is stressed by the manuscript.

5. Does the manuscript clearly describe the problem the software is designed to address?

Yes, the motivation and context were very clear.

6. Does the manuscript clearly describe how the software is implemented?

No, as mentioned for question 2, it is (a) not clear what each function really does, and (b) no explanation at all is given about the implementation. Towards the end, a limitation is mentioned regarding having to run reduce twice, but it is not clear at all how the implementation of each function links to map and reduce. Of course, Pig automatically translates into map and reduce, but even the Pig implementation is not explained.

Since there is quite some Java code to wade through, the paper should explain the goal of each function and their implementation much better. For example, for each function, it would be handy to have a separate visualization like Fig. 2 and 3, but on a much smaller example. Fig. 2 and 3 are interesting, but rather complicated, since no explanation is given of what really happens per step and a lot happens on the data at once.

7. Does the manuscript clearly describe how the software performs and its advantages / limitations over existing applications?

Yes, a detailed performance analysis is given. No comparison is made in the study to other techniques, but the limitations of those techniques are explained textually.

8. Does the manuscript state the software’s operating requirements?

Yes.

9. Are the discussion and conclusions of the manuscript well balanced and adequately supported by the data?
Yes, except for the points made in 6.

10. Do the title and abstract of the manuscript accurately convey what has been found?
Yes.

11. Is the writing acceptable?
The writing is not perfect, but acceptable.

Some more detailed points:
* Abstract: not clear what “wide format” means here
* Refered Technologies: a picture of the MapReduce architecture would have been nice
* Results & Discussion: in fact, adding extra nodes yields diminished returns after a while
* Results & Discussion: is Durok official Apache software?
* Results & Discussion: what is “ordinal appreciation software”?

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Needs some language corrections before being published

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**
I declare that I have no competing interests