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: 4 October 2012

Reviewer: Ronald Taylor

Reviewer's report:

Summary:
The authors fill in a gap in column field management in Pig, and provide open source software that can serve as a starting point for further enhancements. Not an earthshaking paper, but valuable to researchers and data managers using Pig and Hadoop for large scale parallelized data management and analytics. Since there are a number of such people now – and with said number rapidly growing – the paper and the software it describes are worth presenting to the community in an established journal.

Also, the article is well-written. The software (with a function breakdown) and its purpose are clearly described. The reason for software creation is nicely placed in context in the Background section. Readers are walked through the data transformations in decent figures, and the software testing results are well-presented in tables and graphs.

The testing shows that the Java software is working as one would expect with Hadoop-based software in regard to changes in data size and number of nodes employed.

I went to the authors’ web site and could easily download the Java *.jar file for the software and read the Quick Start instructions that are available in English. Also, the sample program script used for benchmarking is given in Appendix 2, and can be used to get users started with the new Pig Latin functions.

Major Compulsory Revisions
NONE

Minor Essential Revisions
NONE

Discretionary Revisions
A few minor syntax suggestions:
1) Change
The usefulness of Hadoop, however, is still limited to many epidemiologists, health service researchers, and health policy analysts who are familiar with the declarative style of existing statistical software and SQL commands, because
MapReduce programming is too rigid and difficult for these end users to write procedural code.

The usefulness of Hadoop, however, is still limited. Many epidemiologists, health service researchers, and health policy analysts who are familiar with the declarative style of existing statistical software and SQL commands do not use Hadoop because MapReduce programming is too rigid and difficult for these end users to write procedural code.

2) Change

Besides, these researchers often need trial-and-error ad-hoc analysis for data description and planning optimal analytic strategy, for which a more user-friendly framework that allows iterative, easy, and quick transformation of ultra-large scale administrative data into an analytic dataset, is in great demand.

Also, such researchers often perform trial-and-error ad-hoc analysis for data description and planning optimal analytic strategy. For such purposes, we desire a more user-friendly framework that allows iterative, easy, and quick transformation of ultra-large scale administrative data into an analytic dataset.

3) Change

Pig’s infrastructure layer consists of a compiler that turns Pig Latin programs into sequences of MapReduce programs.

Pig’s infrastructure layer consists of a compiler on the user’s client machine that turns Pig Latin programs into sequences of MapReduce programs that run in parallel on the nodes of a Hadoop cluster.

4) Change

As shown in the graph, there is a clear linear dose-relationship between the processing time and data size.

As shown in the graph, there is a clear linear relationship between the processing time and data size.

5) Change (remove the comma)

activities in the performance test, took one hour

activities in the performance test took one hour

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

**Quality of written English:** Acceptable

**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.