SUMMARY:
Christopher Olston, Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, and Andrew Tomkins.
Pig latin: A not-so-foreign language for data processing.
In Proc. *ACM SIGMOD Int'l Conference on Management of Data*, pages 1099-1110, 2008.

nested operations for situations where we might have

DATE: 8 February 2010

This paper describes *Pig*, which is an open-source data processing environment developed at Yahoo and intended for analysis of extremely large data sets (e.g. the web). Pig has two major components: a *programming language* and a *debugging environment*. The programming language is named *Pig Latin* while the debugging environment is referred to as *Pig Pen*.

Pig Latin is a new language created with the desire of having the capabilities of both high-level declarative query languages (e.g. SQL) and the low-level procedural programming languages (e.g. MapReduce). The data model that is used by Pig Latin consists of four distinct data types: *atom*, *tuple*, *bag* and *map*. An *atom* "contains simple atomic values such as a string or an number". A *tuple* is a "sequence of fields which can be of any data type". A *bag* is a "collection of data items" in the form of key-data which can be looked up by key values. Generally speaking, each Pig Latin program has three major stages. Firstly, the input data files need to be read by the program. This step is done by the `LOAD` command. This command returns a *handle* that provides access to the bags of input data. Secondly, the read input data needs to be processed. For this purpose, a set of commands have been implemented, like `FOREACH` which provides the capability of processing all the data tuples in a specific way. This set of commands operate in way that there is no correlation between the processing of different tuples of the input. Hence, they can provide an efficient parallelism that is essential for processing large data sets. In addition, most of these commands allow nested operations for situations where we might have nested bags within tuples. Thirdly, the output of the program can be written to a file by using the `STORE` command. Pig Latin programs are then compiled to MapReduce jobs which run on Hadoop.

Pig Pen is a debugging environment developed for Pig Latin. Since the conventional run-debug-run process is very costly and inefficient when dealing with large data sets, Pig Pen has been developed to allow incremental debugging of Pig Latin programs. It enables users to write a small portion of the whole program, examine the output on a small sample data set and then append to the code incrementally.

SUMMARIZED BY: Hani Khoshdel Nikkhoo