PADS: Processing Arbitrary Data Sources
by
Kathleen Fisher
AT&T

ABSTRACT

Vast amounts of useful data are stored and processed in ad hoc formats. Traditional databases and XML systems provide rich infrastructure for processing well-behaved data, but are of little help when dealing with ad hoc data. Examples that we face at AT&T include call detail data, web server logs, netflows capturing internet traffic, log files characterizing IP backbone resource utilization, and wire formats for legacy telecommunication billing systems. Such data may simply require processing before it can be loaded into a data management system, or it may be too large or too transient to make such loading cost effective. Typically, data consumers have no control over the format of ad hoc data. Therefore, they must invest significant effort in understanding such a data source and writing a custom parser, a process that is both tedious and error-prone. Often, the hard-won understanding of the data ends up embedded in parsing code, making both sharing the understanding and maintaining the parser difficult. Typically, such parsers are incomplete, failing to specify how to handle situations where the data does not conform to the expected format.

PADS is a declarative data description language that allows data analysts to describe both the physical layout of ad hoc data sources and semantic properties of that data. From such descriptions, the PADS compiler generates libraries and tools for manipulating the data, including parsing routines, statistical profiling tools, translation programs to produce well-behaved formats such as XML or those required for loading relational databases, and tools for running XQueries over raw PADS data sources. The descriptions are concise enough to serve as “living” documentation while flexible enough to describe most of the ASCII, binary, and Cobol formats that we have seen in practice. The generated parsing library provides for robust, application-specific error handling.

This is joint work with Mary Fernandez, Bob Gruber, Yitzhak Mandelbaum, and David Walker.

Thursday, November 09 at 4:15 p.m.
Rose Hills Theatre, Smith Campus Center
Pomona College

Refreshments at 4:00pm