Toward Data Certification

Véronique Benzaken, Evelyne Contejean
Equipe Proval - LRI - INRIA Saclay

March 5, 2012

Context

Current data management applications and systems involve huge and increasing data volumes. Data can be numeric, e.g. output of measure instruments, textual, e.g. corpora studied by social scientists which may consist of news archives over several years, structured as is the case of astronomy or physics data, or highly unstructured as is the case of medical patient files. Data, in all forms, is increasingly large in volume, as a result of computers capturing more and more of the companies activity, the work scientists used to do based on paper, and also as a result of better and more powerful automatic data gathering tools, e.g. space telescopes, focused crawlers, archived experimental data (mandatory in some types of government-funded research programs) and so on. The availability and reliability of such large data volumes is a gold mine for companies, scientists and simply citizens.It is then of crucial importance to protect data integrity and reliability by means of robust methods and tools.

Program verification and certification have been intensively studied in the last decades yielding very impressive results and highly reliable software (e.g., the Compcert project). However, and surprisingly, while the amount of data stored and managed by data engines has drastically increased, little attention has been devoted to ensure that such (complex) systems are indeed reliable.

The aim of this thesis is to certify data intensive systems such as relational database systems, XQuery and semi structured engines using formal tools such as SMT provers (e.g., Alt-ergo) embedded in the Why(3) platform as well as interactive theorem provers (e.g., Coq).

Possible research plan

In a first step of this thesis will consists in equiping data systems with a mean to automatically and statically (at compile time) prove that updates will preserve invariants if such is the case. To do so we shall rely on the Why3 platform. Why3 takes as input a program equipped with its specification, and computes a logical formula the validity of which is equivalent to the fact the input program satisfies the given specification. A second step of the PhD thesis will consist in adapting the previous approach to different data models with a particular emphasis to XML and JSON semi-structured formats and NoSQL languages. In a third, more ambitious and prospective part of the work,

we expect the PhD student to model other databases functionalities with the Coq proof assistant

Prerequisite

The candidate should have a very strong background in theoretical computer science with emphasis in logic as well as a concrete practical experience of functional programming in OCaml-like programming languages. A good knowledge of Coq will be a plus.