LABORATOIRE DE RECHERCHE EN INFORMATIQUE

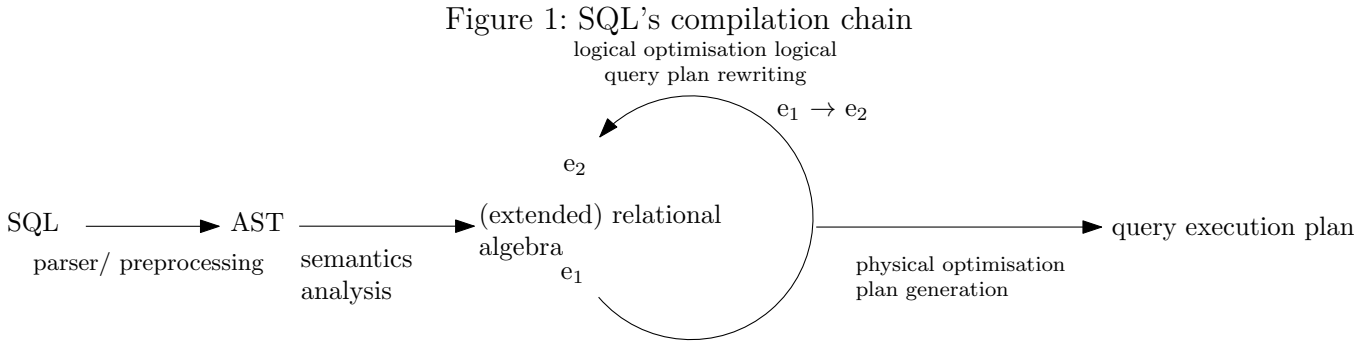
Deep specification and verification of SQL compilation chain

Junior Conference on Data Science and Engineering — September 2017

Data systems and languages, proofs

Until now there have been little usages of formal methods towards SQL and DBMS, yet, the compilation chain of SQL queries is complex.

Figure 1: SQL's compilation chain



Certified parsing, from SQL to SQLCoq

```
SELECT s.sname FROM sailors s WHERE EXISTS (SELECT * FROM reserves r WHERE
↪  r.bid = 103 AND s.sid = r.sid);
```

- SQL parser in menhir/OCamllex
- certified using the CompCert parser generator
- automatic translation into SQLCoq (in OCaml)

Involving DBMS, from SQL to unified execution plan

- PostgreSQL and Oracle query execution plan parser
- both translated to our own unified plan format

From unified plan to Extended Algebra

```
<Plan>
  <Node-Type>Hash Join</Node-Type>
  <Join-Type>Inner</Join-Type>
  <Hash-Cond>(s.sid = r.sid)</Hash-Cond>
  <Plans>
    <Plan>
      <Node-Type>Seq Scan</Node-Type>
      <Parent-Relationship>Outer</Parent-Relation
      ↪  ship>
      <Relation-Name>sailors</Relation-Name>
      <Alias>s</Alias>
    </Plan>
    ...
```

- Extended Algebra (ExtAlg) is similar to Relationnal Algebra but more powerful
- translation from an unified plan to an ExtAlg expression
- automatic annotation of ExtAlg expression with algorithms
- ability to execute our Coq version of those algorithms

```
Omega {
  _ Fine
  _ r.dday  r.dday, r.bid  r.bid, r.sid  r.sid,
  ↪  s.age  s.age, s.rating  s.rating, s.sname
  ↪  s.sname, s.sid  s.sid
  _ x_0.s.sid = x_0.r.sid
  _ {
      Omega {
        _ Fine
        _ sid  s.sid, sname  s.sname, rating
        ↪  s.rating, age  s.age
        _
        _ {
          Seq Scan on sailors
        }
      }
      X (DependentJoin) (Hash Join)
      ...
```

Equivalence of two ExtAlg queries

- translation from SQLCoq to ExtAlg
- two ExtAlg expressions, one from the DBMS, one from the SQL query/SQLCoq
- proof that two ExtAlg expression have the same semantic
- strong guarantees about the fact that the DBMS kept the semantic of the SQL query
- algorithms used by the DBMS have not been proved yet
- Coq version of some algorithms