Schema Inference for Massive JSON Datasets

Mohamed-Amine Baazizi, Houssem Ben Lahmar, Dario Colazzo, Giorgio Ghelli, Carlo Sartiani

(1) Université Pierre et Marie Curie, France (2) University of Stuttgart, Germany (3) Université Paris-Dauphine, France
(4) Università di Pisa, Italy (5) Università della Basilicata, Italy

Java Script Object Notation (JSON)
- Flexible format for data exchange, NoSQL systems
- Large datasets, complex structure
- Incomplete data description, schema-less

Problem

- No abstraction to formulate meaningful queries
- Difficulty to reason about the structure properties
- Miss schema-based optimization techniques

Solution

Efficient inference of succinct descriptive schemas from large JSON datasets

Scenario

Map phase: infer a schema for each input document
generalize values to basic types, compact array content
Reduce phase: merge schemas
collapse identical types, capture irregularities

Characteristics

- Captures record nesting, arrays content
- Reflects optional values, structural variation
- Large scale schema inference
- Produces succinct schemas

Experiments

Implemented in Scala by extending the Json4s parser
Run on Spark 1.6.1, 6 nodes with 2*10 cores, 64GB RAM
Tested on Github, Twitter, Wikipedia and NYTimes datasets

Future directions

- study the precision-succinctness tradeoff
- enrich the schemas with cardinalities
- investigate the impact on query optimization

Contact: Mohamed-Amine BAAZIZI
baazizi@lip6.fr