Schema Profiling for Exploring and Querying Massive Nested Key-Value Data

Dario Colazzo (LAMSADE) and Mohamed-Amine Baazizi (LIP6, Sorbonne Université)

11/2/2019

Description

Nested Key-value data like JSON are very popular as they allow for overcoming the rigidity of relational databases by adopting flexible, schema-less models. This flexibility is a desirable property, especially when data is produced by uncontrolled sources, but it also complicates the processing and the analysis of data due to their variable structure. Major NoSQL systems like MongoDB [9], Couchbase [3], Apache Drill [1] and Spark [4] already adopt some schema extraction mechanism to reveal the structure of the data when it is loaded. However, the extracted schemas are purely structural and do not allow for expressing richer semantic constraints such as correlations or dependencies.

In the literature, there has been some attempts for profiling relational data as witnessed by a recent survey [5]. In the context of JSON, data profiling is in its infancy and the only few approaches require to flatten the data before applying standard classification or clustering techniques devised for relational data [8] and [7]. Moreover, scalability is not addressed although JSON datasets are expected to be large and running classification or clustering algorithms may be prohibitive. Recently, Couchbase introduced a schema extraction module for classifying JSON documents based on their structure [3] using a kind of decision tree like in [8]. However, there is no clear understanding of the semantics of their classification approach since no formal documentation is available.

The first goal the study to be curried on in this PhD thesis is to devise and study techniques for extracting constraints in a distributed fashion over large JSON datasets. A possible direction is to investigate the use of the distributed schema inference approach developed in [6] which allows for extracting statistical information about the structure of JSON datasets, by extending it in several directions, just to mention some of them : counting enumeration, constraints and statistics on simple values contained in records and arrays, tuple types and set operators like difference.

The second goal is to study the notion of *informativeness* or *precision* of a schema. This is important issue which naturally arrises in many problems related to data summarization

and which may have important applications in this project where schemas are expected to be used a means for exploring massive and complex datasets.

The third goal is to study the limitations of existing approaches for storing and querying massive JSON data in order to identify opportunities for improving the underlying techniques by exploiting schema information. The plan to start with a simple class of queries mainly used for data exploration purposes then to deal with more complex workflows expressed a wider class of queries. The plan is also to tackle these issues for both styles of processing: batch and streaming while devising a specific solution for each style.

Pre-requisites and expected results

The current project lies in the intersection of three majors domains: data management, data mining and type theory. Good proficiency in one these domains is sufficient but in general the candidate is expected to have good modeling and programming skills. The language of choice is usually Java or Scala. A good proficiency of database internals and systems in the Hadoop echo-system is desirable. The expected outcome of the thesis consists of both formal material and system development. Our goal is to apply the solutions of the problems described above in mainstream frameworks for shared-nothing parallelism and distribution like Apache Spark [4] or Apache Flink [2] but also for more specific systems like MongoDB [9] and Couchbase [3], when applicable. This entails that a study of recent approaches for optimizing JSON representation and storage in such frameworks to be carried on.

Contact information: `dario.colazzo@dauphine.fr,mohamed-amine.baazizi@lip6.fr`

References

[1] Apache Drill. `http://drill.apache.org`.

[2] Apache Flink. `https://flink.apache.org`.

[3] Couchbase auto-schema discovery. `https://blog.couchbase.com/auto-schema-discovery/`.

[4] Spark Dataframe. `https://spark.apache.org/docs/latest/sql-programming-guide.html`.

[5] Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. Profiling relational data: a survey. *The VLDB Journal—The International Journal on Very Large Data Bases*, 24(4):557–581, 2015.

[6] Mohamed-Amine Baazizi, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani. Counting types for massive JSON datasets. In *DBPL '17*, 2017.

[7] Michael DiScala and Daniel J. Abadi. Automatic generation of normalized relational schemas from nested key-value data. In *SIGMOD '16*, pages 295–310, 2016.

[8] Enrico Gallinucci, Matteo Golfarelli, and Stefano Rizzi. Schema profiling of document-oriented databases. *Inf. Syst.*, 75:13–25, 2018.

[9] Peter Schmidt. mongodb-schema, 2017. `https://github.com/mongodb-js/mongodb-schema`.