Schema Profiling of Massive Nested Key-Value Data and its Application to Effective Machine Learning

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

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 [11], 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. At the same time, several machine learning framework [10, 5] support nested-value data formats for submitting training data.

In the literature, there has been some attempts for profiling relational data as witnessed by a recent survey [6]. 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 [9] and [8]. 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 [9]. However, there is no clear understanding of the semantics of their classification approach since no formal documentation is available.

The first goal of this PhD project 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 [7] 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 means to exploit informative schemas for optimizing the data preparation phase of machine learning pipelines. This phase is acknowledged to raise a big challenge since extracting relavant features and transforming them in a way that is

suitable for the target algorithm requires a good understanding of the underlying data. Without such an understanding, it is impossible to write complete extraction programs that account for all possible issues that can arise in the data like an incompatibility in the type or in the structure of data.

The third goal is to use of informative schemas for data exploration purposes. The idea is to guide users while formulating their queries for expressing meaningful feature extraction programs but also to inject some constraints expressed in the schema into the inference process itself.

Pre-requisites and expected results

The current project lies in the intersection of three majors domains: data management, machine learning 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 one of: Java, Scala or Python. A good proficiency of database internals and systems in the Hadoop echo-system and in the Tensor Flow framework 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 [11] 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] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[6] 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.

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

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

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

[10] Akshay Naresh Modi, Chiu Yuen Koo, Chuan Yu Foo, Clemens Mewald, Denis M. Baylor, Eric Breck, Heng-Tze Cheng, Jarek Wilkiewicz, Levent Koc, Lukasz Lew, Martin A. Zinkevich, Martin Wicke, Mustafa Ispir, Neoklis Polyzotis, Noah Fiedel, Salem Elie Haykal, Steven Whang, Sudip Roy, Sukriti Ramesh, Vihan Jain, Xin Zhang, and Zakaria Haque. Tfx: A tensorflow-based production-scale machine learning platform. In *KDD 2017*, 2017.

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