Review Report

Evaluation of query execution times for complex XSD in Hive and Spark: A case study for performance management files in mobile networks (#60399)

Authors: Diana Martinez-Mosquera, Rosa Navarrete, and Sergio Luj´an-Mora

Overview of Complete Work

Authors have proposed complex XML schemas processing with Big data tools, Hive and Spark. They have worked on two real datasets provided from performance management of two mobile network vendors (Files named cs-60399-vendorA and cs-60399-vendorB). The three methods- a catalog, deserialization, and positional explode are applied for parsing XML files. Finally, queries are applied on Hive internal and external tables and Spark data frames for 1000 rows and query execution times are evaluated (shown in File: Experiments.xlsx). Based on chosen case study, a comparison is drawn which showed that Spark is a better tool for handling queries on complete data frame for complex XML schema definitions (XSD) applications and Hive internal and external tables are better when querying individual values or attributes derived from XML files.

Major comments:

- The paper is well-organized and contributes to novel research work which falls in Computer Science Research domain of the journal.
- In terms of experimental technique, this paper provides original algorithms designed to solve the efficient query processing for XML files with complex schemas.
- The statistical analysis in this paper is suitable and justify the proposed work.
- Some of the fundamental papers in the field are not cited, among these are

  Dmitry Vasilenko, "An Empirical Study on XML Schema Idiosyncrasies in Big Data Processing", in International Journal on Computer Science and Engineering, October 2015.

  Dmitry Vasilenko, Mahesh Kurapati,." Efficient Processing of XML Documents in Hadoop Map Reduce, IJCSE, 2014, Vol.6, No.9,

  p.329–333.
- I would like to know which type of queries have been selected to evaluate the proposed algorithm.

Minor comments:

- Few sentences need to be reframed/reexamined:

  46-48

  193-194

  273-274

  280-281

  307-308

  - At line number 444-445- I request if you can elaborate or reference why it is needed to create the raw table at first?

I appreciate authors for their research contribution. Paper can be accepted with minor revisions.

Reviewer Name: Aarti Chugh