Abstract—The data warehouse is a very famous solution for analyzing business data from heterogeneous sources. But, unfortunately, the data that a data warehouse can analyze is structured data, whereas nowadays, with the growing popularity of social media, and as it becomes more accessible for people and organizations to use web applications to create data, we are currently experiencing a flood of unstructured data. Therefore we need an approach that can "structure" the unstructured data into structured data so that it can be processed by the data warehouse and produce facts and insights. To do this, we propose a schema extraction approach that will produce a schema from unstructured data. The existing work uses Hadoop for extracting the schema, and we notice some limitations of this approach. Therefore we propose using Google Cloud Platform to do the schema extraction. Based on the experimentation result, our approach successfully produces a schema for unstructured data that a data warehouse can further process. We also prove that our approach helps the database developer to understand the unstructured data better.

Keywords: unstructured data, schema extraction, data warehouse