Master Project 2021: Data Lake

Alexander Martin, Marcel Thiel, Robin Kuller

Hochschule Niederrhein, Krefeld, Germany

ABSTRACT

"Brief Abstract"
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Phasellus iaculis purus eget neque semper, quis dictum est volutpat. Phasellus tempus lacinia velit vitae scelerisque. Nulla lacinia enim eu dui ullamcorper, vitae faucibus libero ultricies. Donec tristique nulla quis massa facilisis, ut posuere arcu ullamcorper. Nulla malesuada velit tortor, a aliquet est consectetur eget. Curabitur tincidunt euismod feugiat. Aliquam velit turpis, posuere nec risus non, sollicitudin accumsan velit. Nullam vitae mi nisi. Duis sed cursus magna.

INTRODUCTION

"This should include a short background of the topic to set the context and state the main aims and objectives of your piece of work. What differentiates your work from your competition? Why is your work novel in the field?"
Copied from Quix: Data lakes have been proposed as a solution to deal with the heterogeneity of big data, as they should provide a storage system for any kind of raw data. Metadata is of particular importance in such a system to have information about the structure and semantics of the data. The group has been working on the enhancement of the data lake system Constance. It is based on a modular architecture with components for schema matching, schema mapping, query rewriting, and wrapping of data sources. This system is applied in other research projects, e.g., mi-Mappa, HUMIT, and charMAnt, as basic platform for data

Own: Developed our own data lake based on the schemata defined in the shown figure;

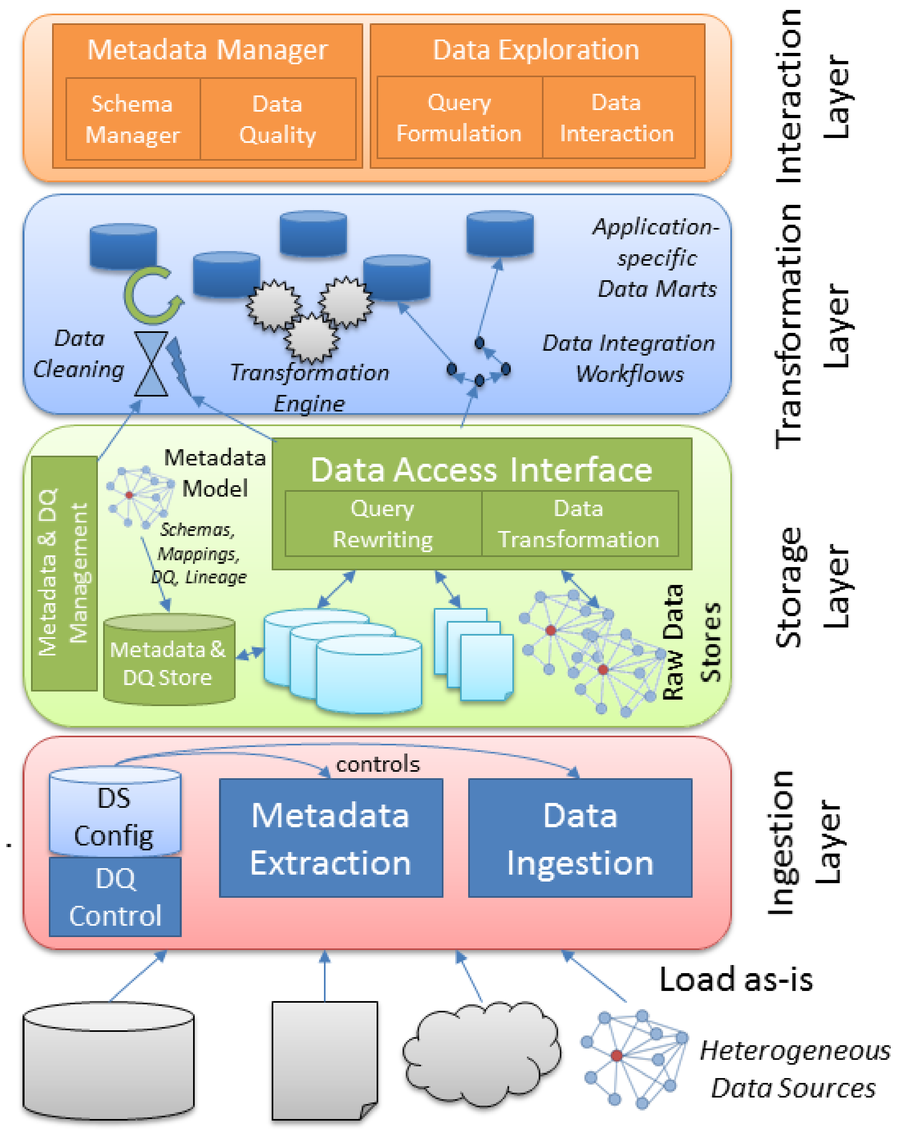

Figure 1: Data lake architecture[3]

METHODS

"The methods section (poster space permitting) should include basic parameters including target sample, setting, duration of study, inclusion/exclusion criteria, statistical techniques, key interventions assessed and primary outcome measures."
used flask, angular, react, spark and all different kinds of dbs; for easy deployment and scaling we used docker.

RESULTS

"The results section should include data analysis and stratification and should only include the results which answer the stated hypothesis. Moreover, essential to the results section is the inclusion of pertinent and key graphs, graphics, images and tables. These need to be large enough for the audience members to see and be as attractive and clutter-free as possible."

- Ingestion Layer
  - Own: Ingestion of structured (sql) and semi-structured (json, xml) data is possible; for each ingested dataset a metadata entry is generated (maybe show the schema);
- Storage Layer
  - Own: The datasets can be stored in their source format and storage-system; For files the default target storage-system is HDFS, for sql it is postgresql and for document-orientated data it is mongodb; However, if the user wants he can select a different target storage-system at his own risk
- Transformation Layer
  -
- Interaction Layer
  -

CONCLUSION

"The conclusions must derive directly from the results section and answer solely what has been proposed at the start of the paper. Obvious confounders and limitations should also be acknowledged. Key improvements as well as potential for project expansion should also be considered."

- Achievements
  - additional components can be included easily
- Confounders / Limitations
  - using Bytestream for uploading files
- Improvements & Project Expansion
  - Machine-Learning integrated in Apache Spark
  - Data Cleaning
  - Data Quality Check at Ingestion Layer
  - Sharded MongoDB Cluster

REFERENCES

"Only cite key references integral to your study, as references are wordy and space consuming. Use a smaller font to the main body text to reduce this."

[1] L. R. Dice. "Measures of the Amount of Ecologic Association Between Species". In: *Ecology* 26.3 (1945), pp. 297–302.

[2] A. F. Frangi et al. "Multiscale vessel enhancement filtering". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 1998, pp. 130–137.

[3] C. Quix. "Data Lake". In: *Ecology* 26.3 (1945), pp. 297–302.

[4] M. Wibowo, S. Sulaiman, and S. M. Shamsuddin. "Machine Learning in Data Lake for Combining Data Silos". In: *Data Mining and Big Data*. Ed. by Y. Tan, H. Takagi, and Y. Shi. Cham: Springer International Publishing, 2017, pp. 294–306.