A Semantically-Based Big Data Processing System Using Hadoop and Map-Reduce

Wang Wanting, Qin Zheng

A Semantically-Based Big Data Processing System Using Hadoop and Map-Reduce

Wang Wanting[1] and Qin Zheng[1,2]

[1] School of Information Management and Engineering,
Shanghai University of Finance and Economics, Shanghai, China
2014311051@live.sufe.edu.cn,
qinzheng@mail.shufe.edu.cn
[2] South University of Science and Technology of China, Shenzhen, China

Abstract. In financial industry, a wide range of financial systems generate vast amount of data in different structures, which change with compliance rules change and hard to manage due to their heterogeneity. This paper introduces a semantically-based big data processing system to integrate the data from different sources, which realizes the query and computation in semantic layer. The system provides a new data management way for the financial industry. With Semantic Web, the information can be managed, integrated, and collaborated in a more fluent way than it in traditional ETL. In order to clear the complex logical relationship among data, the system uses SPARQL to query. Through Map-Reduce, this system, based on Hadoop and Hbase can improve the processing speed for big data.

Keywords: Big data · Semantics · Data integration · Distributed computation

1 Introduction

There are some characteristics in financial big data, which bring lots of problems of data management in financial field, including cross-regional and cross-system distribution [1], multiple structured and non-standardized data formats, and rapid change in the analysis strategy of big data [2]. This paper presents a new semantically-based big data processing system, which connects data through linked data to integrate the heterogeneous data.[1]

2 Model Design

The big data in financial field is so huge-scaled and multi-structured that traditional ETL cannot integrate data efficiently. On the one hand, by using semantic analysis the data can be connected and the system can realize data sharing with RDF based on semantic data queries. On the other hand, distributed computation can provide efficient

[1] This research is supported by National Natural Science Fund of China (71302080) and Ministry of Education Research of Social Science Youth Foundation Project (13YJC630149).

M.C.C. Baranauskas et al. (Eds.): ICISO 2016, IFIP AICT 477, pp. 246–247, 2016.
DOI: 10.1007/978-3-319-42102-5

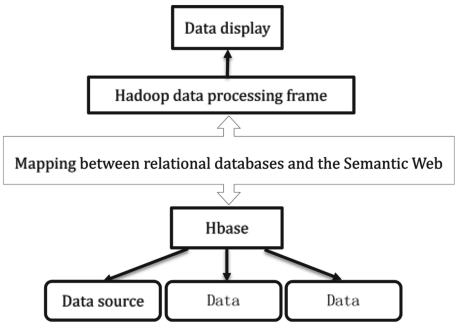

Fig. 1. The semantically-based big data processing system

big data processing [3]. An experiment of distributed semantics queries is tested in order to validate the practicability of semantic mapping of relational data. The semantically-based big data processing system is a vertical structure (Fig. 1), which is based on Hbase storage. The data collected from the data sources are uploaded to central system and permanently stored in Hbase after corresponding conversions. All the data can be connected through Linked Data, and then generate RDF data sets. With Hadoop platform, after semantic analysis and distributed computation, the results can be delivered to every application terminal in cloud.

3 Conclusion

The experiment illustrates that the system based on semantics can solve the problem of the integration of heterogeneous data to some extent. With semantics and distributed computation, the efficient process and integration of complex big data are able to be realized.

References

1. Madnick, S.E.: From VLDB to VMLDB (Very MANY Large Data Bases): dealing with large-scale semantic heterogeneity. In: Proceedings of the 21st International Conference on Very Large Data Bases, Zurich, Switzerland, pp. 11–15 (1995)
2. Rabl, T., Gómez-Villamor, S., Sadoghi, M., Muntés-Mulero, V., Jacobsen, H.A., Mankovskii, S.: Solving big data challenges for enterprise application performance management. Proc. Vldb Endowment **5**(12), 1724–1735 (2012)
3. Bizer, C., Heath, T., Idehen, K., Berners-Lee, T.: Linked Data on the Web (LDOW2008). In: Proceedings of The 17th International Conference on World Wide Web, Beijing, China, pp. 21–25 (2008)