Abstract

Increased data generation has led to an increase in the availability of rich information online. However, complications occur in the form of heterogeneity in the data storage. In order to have complete information, all the data sources must be utilized. Hence a data integration mechanism is required. However, integrating heterogeneous data leads to conflicting data in the system. This paper presents a fast and efficient mechanism to identify and resolve conflicts on huge streaming data using Spark. A wrapper based query formulation module constructs queries depending on the underlying data sources. The retrieved data is converted to a structured format and similarity between the data is identified, followed by distributed conflict identification and resolution. Experiments were conducted on streaming data. Effective conflict detections and a speed up from ~589 seconds to 10 seconds was achieved.

References

Coding and Transmission (pp. 491-520). Springer Berlin Heidelberg.


Index Terms
Computer Science
Information Sciences

Keywords
Conflict identification; conflict resolution; Spark; Streaming Data; Wrappers