EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Optimised Transformation Algorithm For Hadoop Data Loading in Web ETL Framework

Gaurav Gupta[1]*, Neelesh Kumar[2] and Indu Chhabra[3]

[1] Research Planning & Project Management, CSIR-Indian Institute of Petroleum, Dehradun, India, gaurav.gupta@iip.res.in
[2] BioMedical Instrumentation, CSIR-Central Scientific Instruments Organisation, Chandigarh, India, neel5278@gmail.com
[3] Department of Computer Science & Applicaions, Panjab University, Chandigarh, India, chhabra_i@rediffmail.com

Abstract

Web ETL unlike conventional ETL framework requires considerable improvements in all the three layers i.e. Extraction, Transformation and Loading due to the inherent nature of web input data. Websites are huge and are unique source of information, out of such huge information available on the websites, finding and analysing the required and relevant data is critical as the data may be foul consisting of redundant data or misspelled. Determining integrated record that stands for identical real world entities in abundant ways is the major problem to be analysed for any database. Hence, Web ETL transformation layer functionality of data transformation becomes mandatory in determining the pertinent information to be examined. Since the data on the web is "very voluminous" hence loading only clean data in data warehouse is necessary for fast processing to achieve accurate result. The present research focuses on data transformation in web ETL framework and proposes a modified technique to employ token wise sentence sorting to remove redundant records from the patent database along with Levenshtein distance used for string matching. Afterwards the cleaned data is transformed and loaded from this staging area to hadoop environment. The integration of proposed transformation technique with hadoop system delimits the constraint of data processing, storage and retrieval of large data structure from conventional data warehouse system.

Keywords: Redundant Data; Data Transformation; Data Loading; Levenshtein Distance Matching; Hadoop

Received on 11 May 2019, accepted on 01 October 2019, published on 02 October 2019

doi: 10.4108/_____

*Corresponding author. Email:gaurav.gupta@iip.res.in

1. Introduction

Web ETL Framework requires web documents mainly web-pages as input for extraction of data. This existing voluminous data is highly unstructured and heterogeneous due to high expectations and needs of user thus requires pre-treatment through multiple transformation techniques. After discussing the major limitations of the existing scenario, this research has contributed into a modified transformation technique for the data being loaded in the distributed nodes under Hadoop file system. Web documents are enormous source of knowledge and information. These contain information which is used by public, private and government sectors to observe the progress going on their respective domains. Data which is acquired and processed for mining is identified and cleaned before loading it to the data warehouse. Massive amount of data is generated every minute of the day and hence there is copious data present in the world. Data in the databases of the firms and social media is usually in petabyte and zetabyte and obtaining useful data is the real task because without accurate data correct result extraction is a major issue. Hence the need of data transformation arises. Many independent sources of data are merged into a gigantic vault known as data warehouse for querying and analysis purpose [1]. The data present in the staging area can be heterogeneous. Since our focus is on web data which is not based on conceptual schema and therefore explicit use of semantic model is required for web based data

Computational Intelligence, Communications, and Business Analytics pp 267-277

[20] Shaker H., Abdeltawab M. and Bastawissy H. (2011), A proposed model for data warehouse ETL processes *Journal of King Saud University – Computer and Information Sciences*, 23, 91–104

[21] Guang Sun, YingJie Song, ZiQin Gong, Xiya Zhou, Xinyi Zhou, YiLin Bi, Survey on streaming data computing system, Conference: the ACM Turing Celebration Conference - China, May 2019 DOI: 10.1145/3321408.3326687

[22] Michael Frampton, ETL with Hadoop, In book: Big Data Made Easy, December 2015 DOI: 10.1007/978-1-4842-0094-0_10