Persistent Storage of Data Sets in Apache Hive

Bence Bogdándy[a], **Ádám Kovács**[b], **Zsolt Tóth**[c]

[a]Eszterházy Károly University
bence.bogdandy@gmail.com, [b]Eszterházy Károly University
kovadam95@gmail.com, [c]Eszterházy Károly University
toth.zsolt@uni-eszterhazy.hu

Abstract

Building a data warehouse which stores vast amounts of data from heterogeneous data sources is a challenging task. Finding, and optimizing data steams of different origins are the first step during the creation of a data warehouse. Unfortunately, online data streams are most often formatted uniquely. Therefore, the obtained data sets must be transformed into an unified data model. The modeling and conversion of data sources serves as a key step during the unification of heterogeneous data. Storage should be persistent, and optimized for the analytical processing of data. These requirements raise technological challenges that are not common during the design of data sources. This paper presents a system that loads multiple heterogeneous data sources automatically, and store data using Hadoop, Hive and Airflow. Economical data can be often chaotic, therefore data sets should remain up to date at all times. The ETL processes of the data sources are implemented in Python and automatized with Airflow. The data is loaded in a Hive data warehouse which stores data in the distributed Hadoop File System. The outcome of this research is a data warehouse which integrates economical data which can be used for analysis of educational purposes.

Keywords: Data Warehouse, Python, Pandas, Data mining, Apache Hive, Data Processing.