Data access

Jens Lechtenbörger

Winter Term 2023/2024

This text is meant to provide pointers for data access in general and in the context of your projects in particular. In the past, lots of teams set up a relational database (usually free/libre and open source (FLOS) servers such as PostgreSQL or MariaDB/MySQL) to store their integrated data, which you may want to do as well. Here are some pointers for setting up PostgreSQL.

To access data in the first place, you need to locate sources on your own. Usual options include downloads of data in various file formats (often CSV), API access (usually REST), or Web scraping (to "scrape" Web data means to download and parse the HTML contents of Web pages to extract relevant pieces of data).

Whenever you work with data and aim to extract value or knowledge, you may first want to get to know that data. Indeed, Chapter 2 in the data mining book [HKP11] is called "Getting to Know Your Data" (with suggestions to inspect types of attributes; their statistical properties such as mean, media, mode, quartiles, variance; their visualization; their similarity). Then, you likely need to preprocess data (Chapter 3 in that book), which involves data cleaning (to improve data quality) and data integration.

To integrate data in the context of projects, two major options were chosen by student teams in the past: First, some used an extract-transform-load (ETL) tool to clean and integrate data, namely the FLOS community edition of Pentaho Data Integration (PDI). Second, others used a programmatic approach, for which the Python tools recommended by BigGorilla (a scientific endeavour to collect reusable FLOS components for data integration), provide a good starting point; note that we also use BigGorilla code in other sessions, namely FlexMatcher (for schema matching) and similarity joins (for duplicate detection and data fusion).

Maybe familiarize yourself with PDI: Click "Download" at SourceForge; 1.8 GB zip archive; requires Java 8, 64-bit version; if extraction under Windows fails with "path too long", try 7-zip or, better yet, switch to GNU/Linux. Note that sub-directory "samples" contains lots of examples (in particular, data and transformations). In any case, these tutorials offer a good overview (Chapter 1 - Chapter 6 take ca. 25 min total).

Lots of teams use Scrapy, a popular FLOS Python tool (also recommended as part of BigGorilla), to scrape Web data. Check out the first 5 tutorials here (ca. 27 min total) to learn how easy Web scraping can be. If you should scrape data, regular expressions are a useful tool to extract relevant pieces of data.

Bibliography

[HKP11] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. 3rd ed. Morgan Kaufmann Publishers, 2011.

License Information