Projects

Alberto Belussi

anno accademico 2020 - ’21

Project assignment

- Each student has to choose one system in the list and one dataset
- The project consists in:
  1. Download the current version of the system
  2. Install it on your computer
  3. Test it by creating simple data structure, loading some data and executing a basic query.
  4. Analyse the dataset content and produce a conceptual data model in UML of its content
  5. Design a physical data model for the dataset starting from the conceptual data model
  6. Load the dataset in the system
  7. Write a query or a portion of code to execute the computation requested

Project presentation

- Each student will present her/his project in an oral test.
- The presentation consists of:
  - A set of slides presenting the characteristics of the analysed system:
    - Constructs for data modelling
    - Query language
    - Distributed architecture
    - Consistency model
  - A set of slides presenting the characteristics of the dataset:
    - Conceptual data model
    - Physical data model
  - A demo of the implemented queries or computation.

List of systems

- HBase
- Couchbase
- MongoDB
- InfluxDB
- Cassandra
- Vertica or C-Store
- Amazon DynamoDB

List of datasets

- **SITAVR**: in **GeoJSON** from **GeoServer** => two files: one with the Information Sources (IS) the other one with the Archaeological Partitions (PA). The two datasets need to be integrated first.
- **VeronaCARD**: in **CSV**, data from 2015 until today. Attributes: ID VeronaCARD, Date and Time fo the badge swipe, POI, type of card (72, 48 or 24 hours). Additional data about the POIs to be integrated.
- **Events for tourists**: in **GeoJSON** from **GeoServer** => Attributes: name and description in multi-language with timestamp and geolocation.
- **Auditel**: in **CSV.** It contains the visualizations of TV programs by a group of users. Attributes: id utente, id program, starting date and time, ending date and time, etc...). Additional data about programms and users to be integrated.
- **Twitter**: in **CSV.** It contains a set of Tweets about COVID. Attributes: id, user_id, timestamp is_retweet, retweeted_status_id, is_quoted quoted_status_id, lang, text.