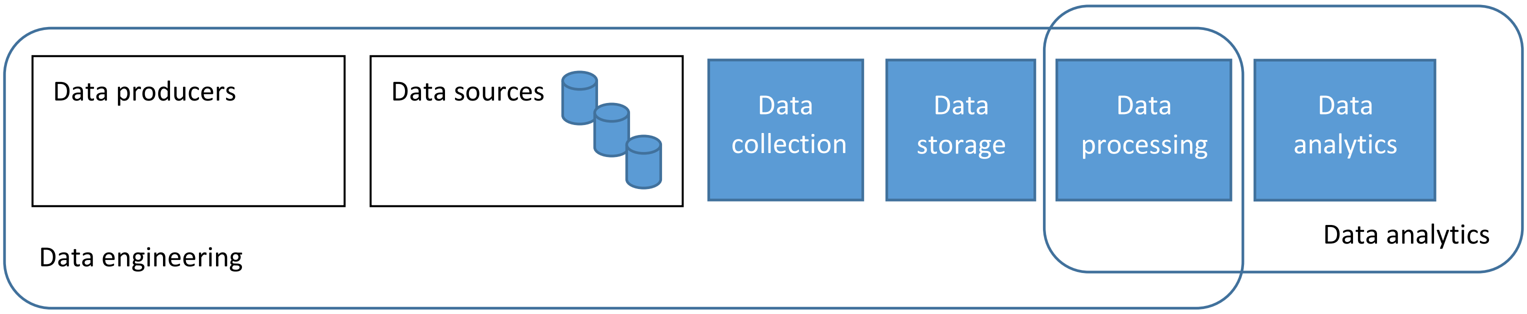
A standard architecture for big data/data science projects

Regardless the application, a data science project follows a standard pipeline:



[Ciuciu et al.]

Project description:

Within the project the students are required to implement an end-to-end application with the following requirements:

1. **Data Producers**
    a. Any sensors or data producers (e.g., websites, mobile phones, cameras, etc.) that are necessary for the use-case

2. **Data Sources**
    a. In the case of a sensor, the data source are the sensor log files containing the sensor's parameters
    b. If external data is required for the use-case (e.g., public data) the relational database of the repositories where the data is stored could represent additional data sources

    Remarks:

    - At least 2 different data sources shall be used and integrated in the project
    - At least one of the data sources should be real-time (or r.t. simulated)

3. **Data Storage**

    All data coming from the data sources should be stored in the Hadoop Distributed File System (as main storage) in json/avro/parquet format OR a similar solution (see cloud provider free solutions available). From here, the data will be consumed by the algorithms/visualization dashboards in your application.

    The real-time storage solution (e.g., Cassandra) is decided by each team.

4. **Data Processing**
    The project should consider and implement both the batch and the real-time processing layers (and show case both).

5. **The implementation of an algorithm**

   A ML (machine learning) model, statistical model which is proposed to solve the problem. The algorithms should be evaluated with proper statistical metrics in order to prove their validity.

6. **Data visualization – requirement**

   a. Visualization dashboard for real time monitoring of the sensor parameters.
   b. Distributions of sensor parameters (e.g., histograms, bar plots) - both batch and real-time visualizations.
   c. Visualization dashboard presenting the results of the algorithm in an easy to understand manner.

   Note: If the data are transmitted directly from the sensor to the visualization tool, this is **not** considered real-time processing. If you want to make the processing real-time, you will need to make the data **go through a real-time flow /pipeline** in your architecture. For this you will need to use a framework for speed processing (e.g., **Spark** or Storm) and to pass the data to that part of the pipeline for being aggregated and rendered (visualized).

Suggestion: follow the CRISP-DM process.

References:

1. I.Ciuciu, A.B. Ene, C. Lazar, An ICT Project Case Study from Education: A Technology Review for a Data Engineering Pipeline, Business Information Systems (BIS) 2019, LNBIP 353, ISBN: 978-3-030-20484-6, 2019