Project
Implementation of a (Big) Data Management Backbone

P2 Description

Big Data Management – FIB – UPC

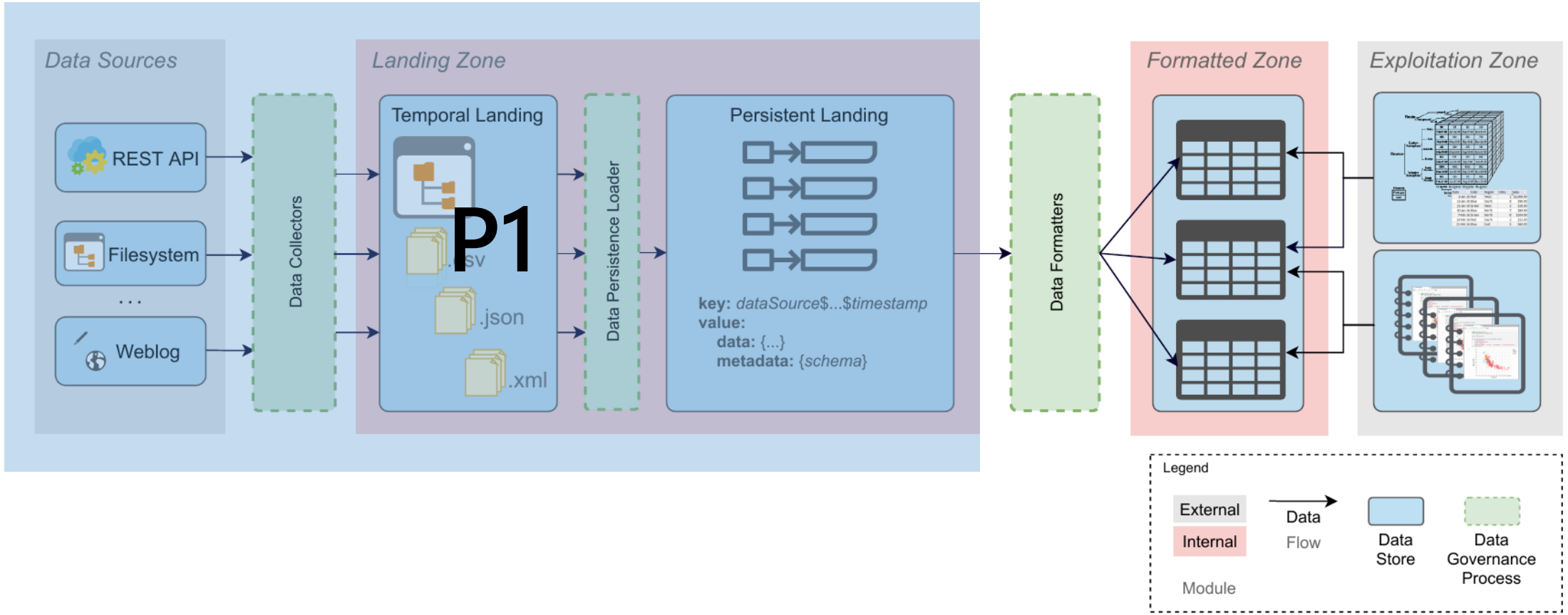
Project's objective

- Descriptive and predictive analysis of data related to Barcelona's housing and the relationship with its economy
- Examples of descriptive analysis KPIs
  - Average number of new listings per day
  - Correlation of rent price and family income per neighborhood
- Examples of predictive analysis KPIs
  - Estimate the rental price for a new apartment
  - Evaluate the deviation of a predicted price with respect to the real average price in a neighborhood

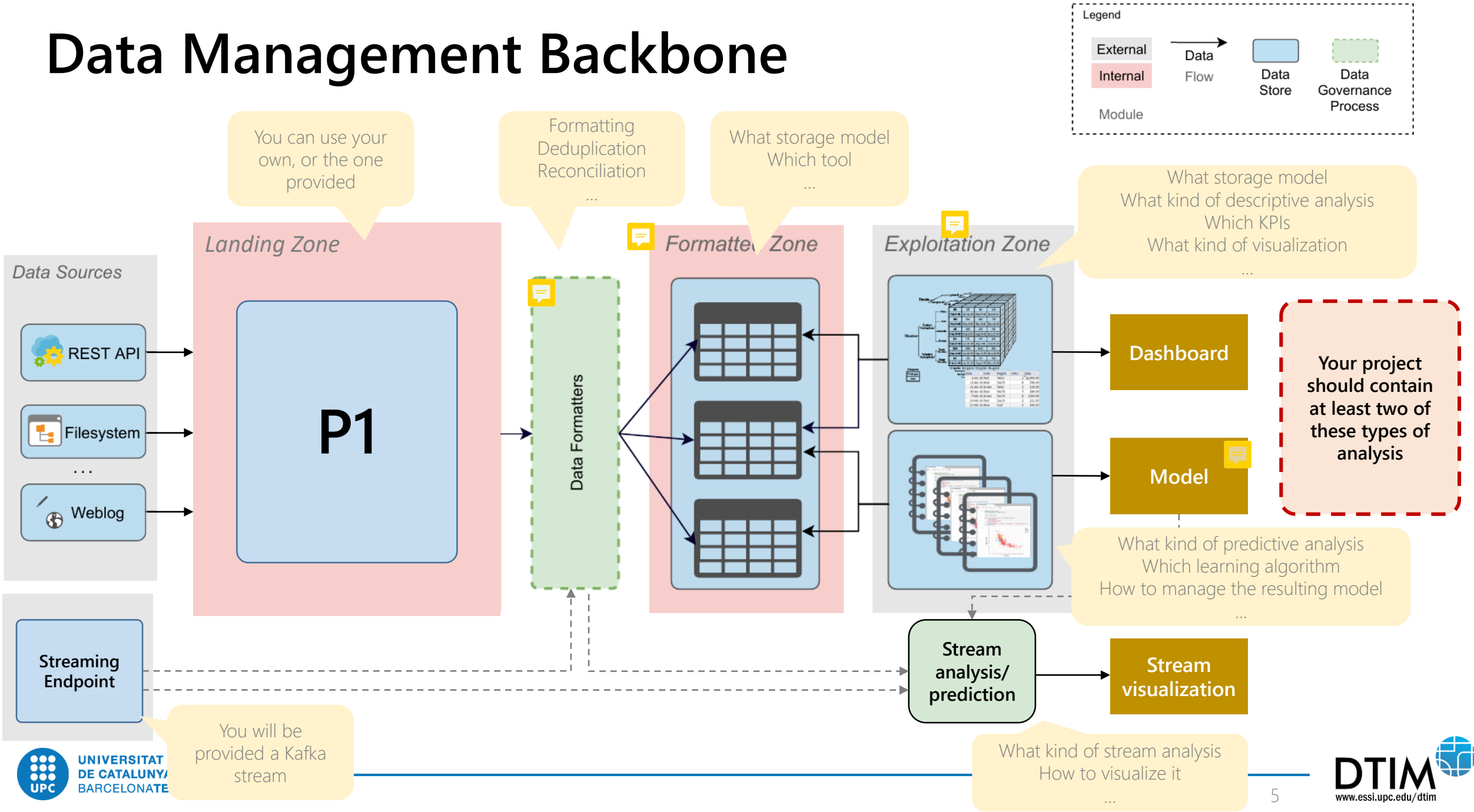
Two parts

- P1 – Data design (Landing Zone)
  - Conceptualization and Data Lake design
  - Technologies: Apache Hadoop (+ file formats), Apache HBase, MongoDB
- P2 – Descriptive and predictive analysis (Formatted and Exploitation Zones)
  - Data integration and reconciliation
    - Technologies: Apache Spark, a visualization tool (e.g., Tableau)
    - **Bonus point**: Apache Spark Core (RDDs)
  - Distributed machine learning and real-time data prediction
    - Technologies: Apache Spark (MLlib, Streaming), Apache Kafka, a visualization tool for streams (e.g., Kibana)

Data Management Backbone



Data Management Backbone



P2 objectives

- Integrate the three provided datasets in the Formatted Zone
  - Handle duplicates, reconciliate data, clean, etc.
- Implement the calculation of three KPIs
  - Store them in the Exploitation Zone
- Prepare the input data and train an ML model
  - Store it in disk
- Ingest a data stream
  - Perform predictions applying the model on the data stream elements
  - Describe the data stream using approximate stream analysis algorithms
- Graphically display the results of the analysis

Analytical needs

- Examples of descriptive analysis KPIs
  - Average number of new listings per day
  - Correlation of sale price and family income per neighborhood
  - Top-seller neighborhoods in a time window
- Examples of predictive analysis KPIs
  - Predict the rental price for a new apartment; or
  - Predict the family income index of a neighborhood based on its sale price

Datasets

- Mandatory datasets
  - Barcelona rentals
    - idealista
  - Territorial distribution of income
    - Open Data Barcelona
  - Lookup tables
- Extra dataset
  - You can check out OpenData BCN portal or other Open Data portals
  - You might need to implement your own reconciliation process
    - See LearnSQL for a guideline on how to run a reconciliation process with OpenRefine

A (given) solution for P1

- Barcelona rentals
  - Each JSON file has been converted to a Parquet file
- Territorial distribution of income
  - CSV converted to one MongoDB collection
    - Use mongoimport to import the file into a collection
- Lookup tables
  - Four MongoDB collections
    - rent_lookup_district
    - rent_lookup_neighborhood
    - income_lookup_district
    - income_lookup_neighborhood
  - Prefixed attributes

    - di → district
    - n → name
    - ne → neighborhood
    - re → reconciled

Distributed machine learning

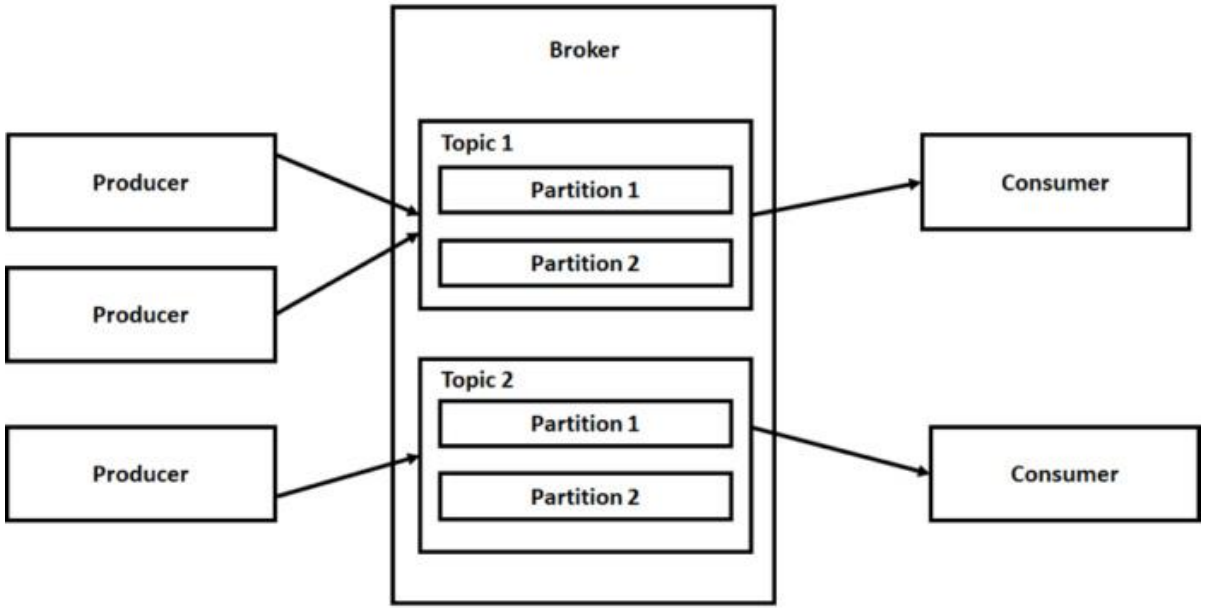
- Create two datasets – perform the necessary transformations, cleaning
  - Training
  - Validation
- Use the training dataset to create a classifier using Spark MLlib (RDD-based)
  - https://spark.apache.org/docs/latest/mllib-guide.html
- You are free to choose the kind of model
  - The objective of the course is not to optimize this part
- Validate the model
  - Compute recall and accuracy
- Store the model
- Ingest and process a data stream to perform predictions using the stored model

Technologies

- Apache Spark
  - Integration and reconciliation using lookup tables
  - Calculate KPIs and store them in views
  - Your pipeline must be optimal from the perspective of…
    - Minimizes the number of wide dependencies
    - Caches results when required
    - Exploits parallelism
    - …
- Apache Kafka
  - Endpoint for stream ingestion
- Apache Spark MLlib
  - Classifier and evaluation
- Apache Spark Streaming
- Visualization tool
  - Choose the one you prefer
  - Provide online access or a video of the resulting solution

Kafka endpoint

- A message queue for raw data streams that are pushed from the data sources



- Available at *sandshrew.fib.upc.edu:9092*, topic *bdm_p2*
  - Check out the example code for integrating Spark Streaming + Kafka

Delivery

- Document (max 5 pages)
  - Describe the pipelines to integrate and to calculate/store of KPIs
    - Sketch the pipelines at a higher abstraction level. If you use RDDs you can use the notation seen in the lectures to describe the Spark job
    - Elaborate on your assumptions. Refer to any specificity of your solution that should help the lecturer to understand the decisions you made in your code that, otherwise, might look like controversial
  - Describe the extra dataset and new KPIs
  - Describe and justify the data model used in the Formatted and Exploitation Zones
- Code
- Extra material
  - Online access to visualization tool, videos, etc.

Closing