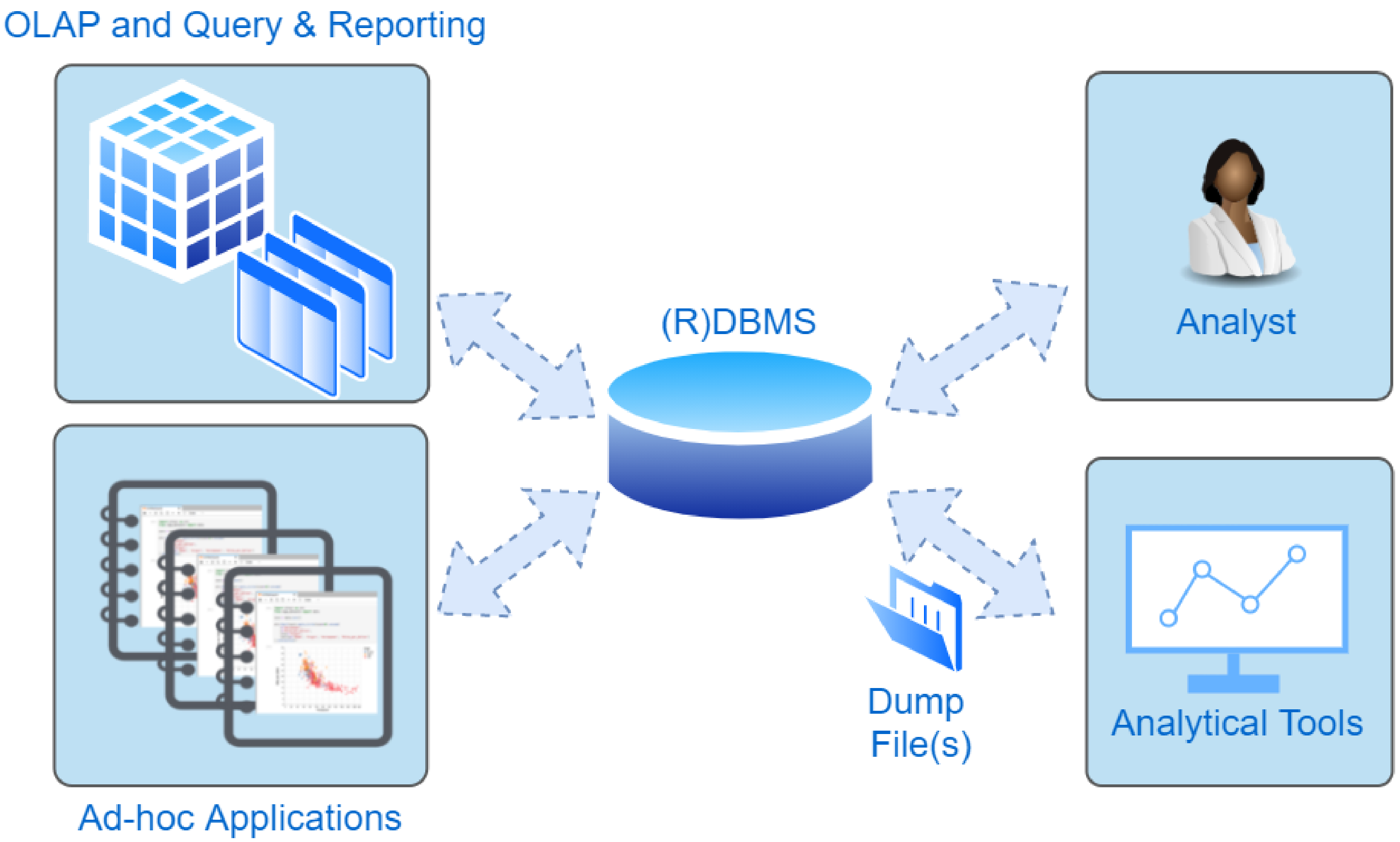
Data Management for Data Science

Oscar Romero
oromero@essi.upc.edu

DTIM Research Group
Universitat Politècnica de Catalunya

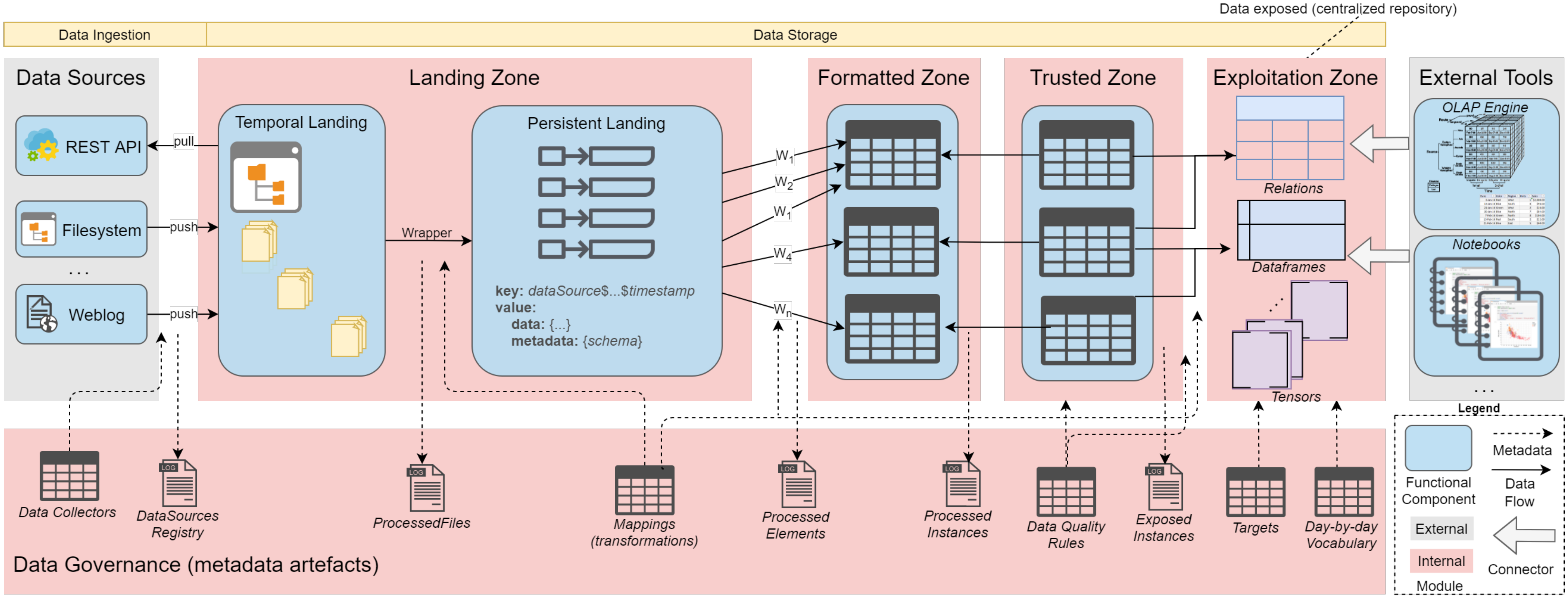
Data Analysis: As It Used to Be

- Data warehousing
  - Multidimensional modeling
  - OLAP
  - Dashboarding tools
- Query and Reporting
- Ad-hoc querying
- Ad-hoc applications
  - In-database data analysis
- Off-the-shelf analytical tools
  - Dump and load data

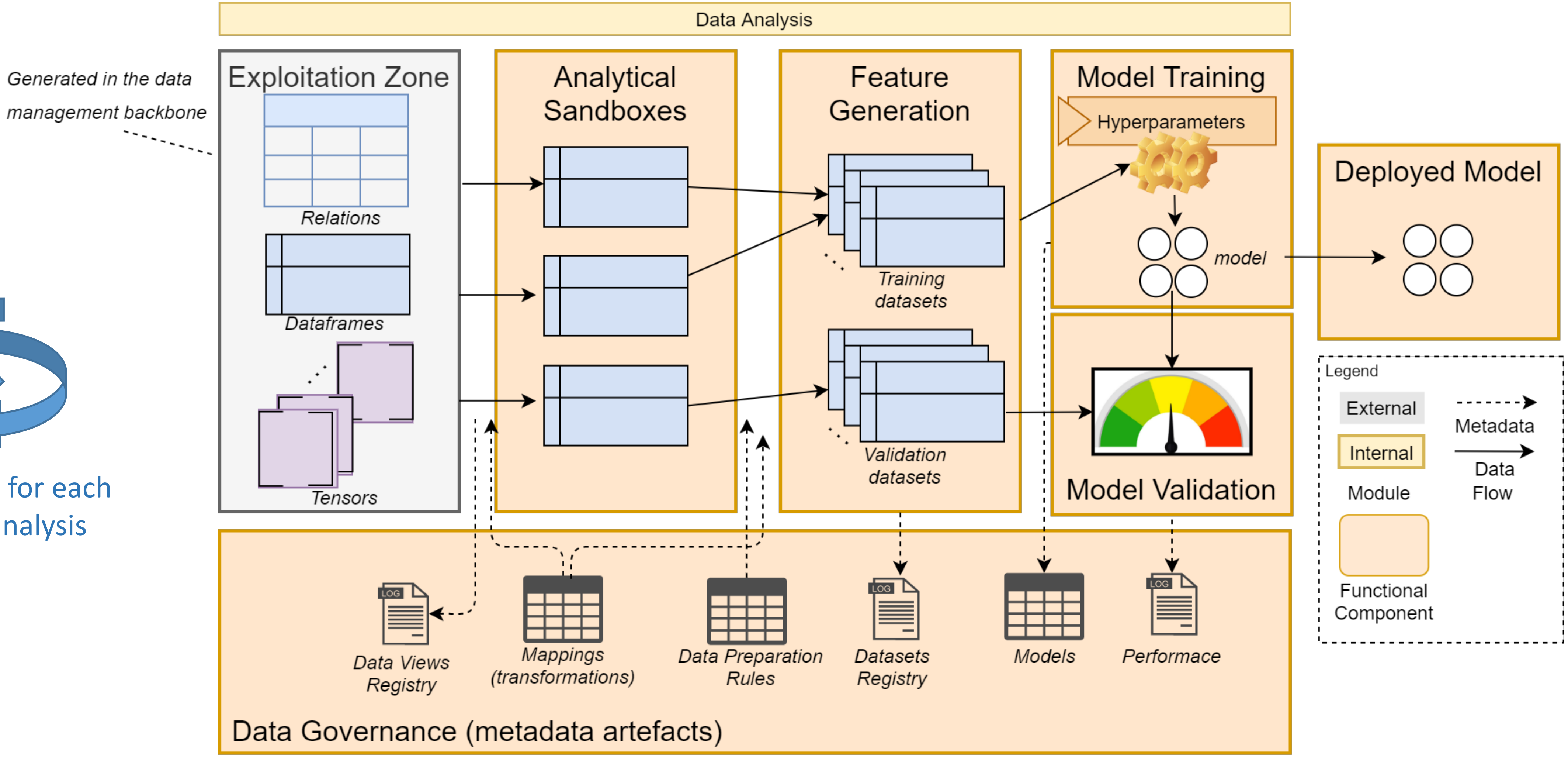


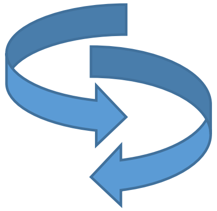
(Relational) databases as main drivers to manage data: ingest, store, model, process and query

Data Science: The Data Management Backbone



As-Is Today: The Data Analysis Backbone





Repeats for each data analysis

The Whole Ecosystem Maps to a DBMS

Open Challenges:

- There must be a common governance of the whole ecosystem
  - Traceability / Lineage
  - Explainability
  - Collaborative analysis
  - Etc.
- Deal with the whole ecosystem as a single DBMS to welcome data engineering good practices
  - Single source of truth for data analysis (*exploitation zone*)
    - Code / data sharing and reusage
  - Operationalize and automate processes (DataOps, MLOps…)
    - Global optimizations throughout the different stages