Partial Data Materialization Techniques for Virtual Data Integration
Dmitriy Pochitaev
dmitriy.pochitaev@put.poznan.pl
Supervisors:
Robert Wrembel - Poznan University of Technology; Oscar Romero, Petar Jovanovic - Universitat Politècnica de Catalunya

1. Background
Organizations and data scientists are facing a challenge of integrating data from a great number of disparate data sources.

Data Integration
- Physical Integration
  - Materialized
    - "+" Query performance – fast
    - "-" Up-to-date data not available
  - Virtual Integration
    - Not materialized
    - "+" Up-to-date data available
    - "-" Query performance - slow
  - Hybrid Data Integration
    - Partial Materialization
      - "+" Query performance – faster then virtual
      - "-" Up-to-date data available

2. Challenges
- Which query results and intermediate query results to materialize?
- Where to materialize data, on disk and/or in RAM?
- How to manage materialized data: which data to refresh incrementally, which data to refresh fully, and when to mark materialized data invalid or outdated?
- Data prefetching - what data to prefetch and when?

3. Objectives
Development of partial data materialization techniques for:
- Proactive materialization of relevant data by use of prediction techniques. Based on system usage statistics prediction techniques should be developed to prefetch and materialize data in advance.
- Improve proactive materialization by considering additional factors such as data sources properties, e.g., capacity, performance, data format, data volatility.

4. Architecture

5. Materialized Data Refreshment
- Which query results and intermediate query results to materialize?
- Where to materialize data, on disk and/or in RAM?
- How to manage materialized data: which data to refresh incrementally, which data to refresh fully, and when to mark materialized data invalid or outdated?
- Data prefetching - what data to prefetch and when?

6. Materialized Data Stages

6. Publications
Submitted:

In Progress:
D.Pochitaev, P. Jovanovic, O. Romero, R. Wrembel. Hybrid Data Integration: the Case of Web Data Sources