An Extensible Approach for Materialized Big Data Integration in Distributed Computation Environments

Vladimir Sazontev *Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University*

Sergey Stupnikov *Institute of Informatics Problems, Russian Academy of Sciences*

https://bitbucket.org/VladimirSaz/integration_system

Outline

- Motivation
- Data Integration Steps
- Targets of research
- Record Linkage Steps
- Key Points of the Approach
- Architecture of the Approach
- Prototype Implementation
- Implemented methods
- Use case
- Conclusions and Future Work

Motivation

- Modern IT world requires data integration systems to deal with the large number of heterogeneous data sources

- In the world of big data with large number of heterogenous data sources, there are number of methods that address various aspects of integration, to make the system automatic and less user-dependent

- current data integration systems are still limited by the human resource

Motivation

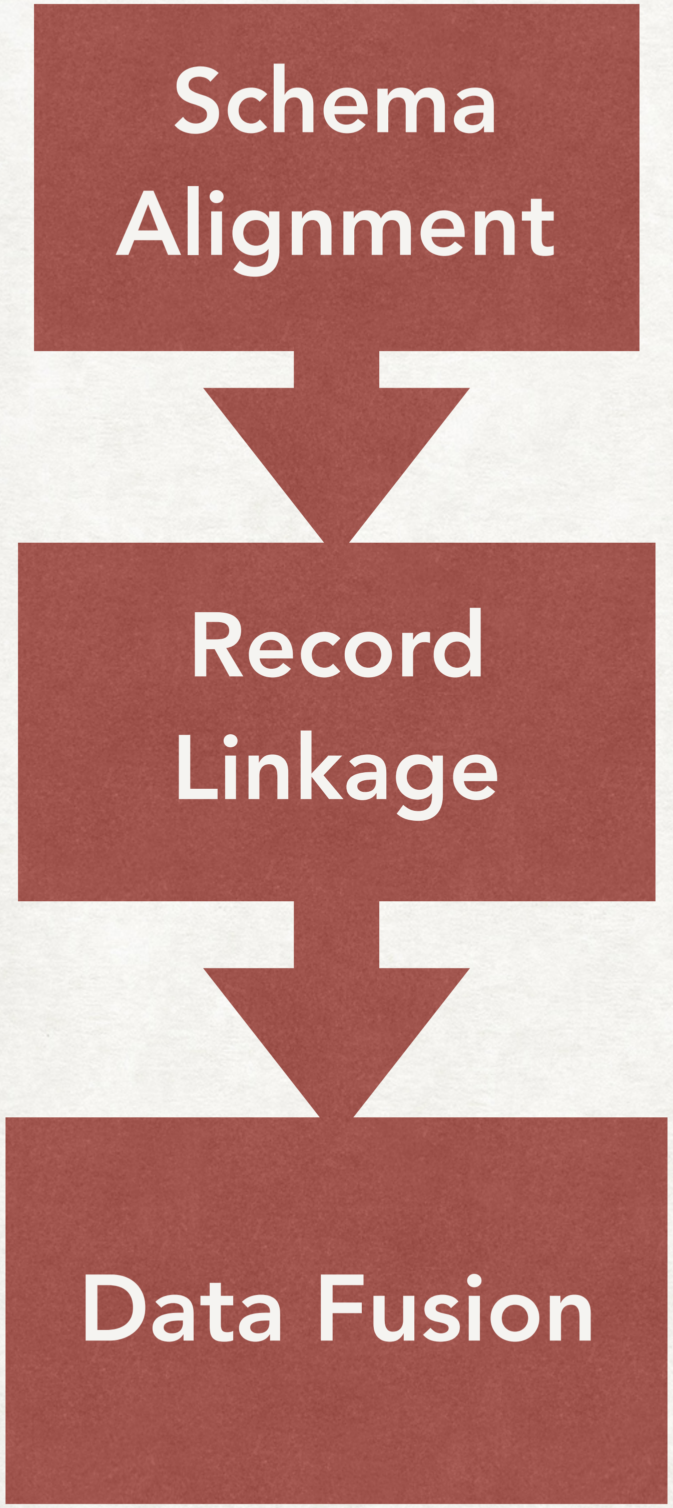
- Enterprise data integration systems:
  — in general provide end-to-end solution

  — lack automation in the field of entity resolution and data fusion

- Research data integration systems provide:
  — wide variety of automation

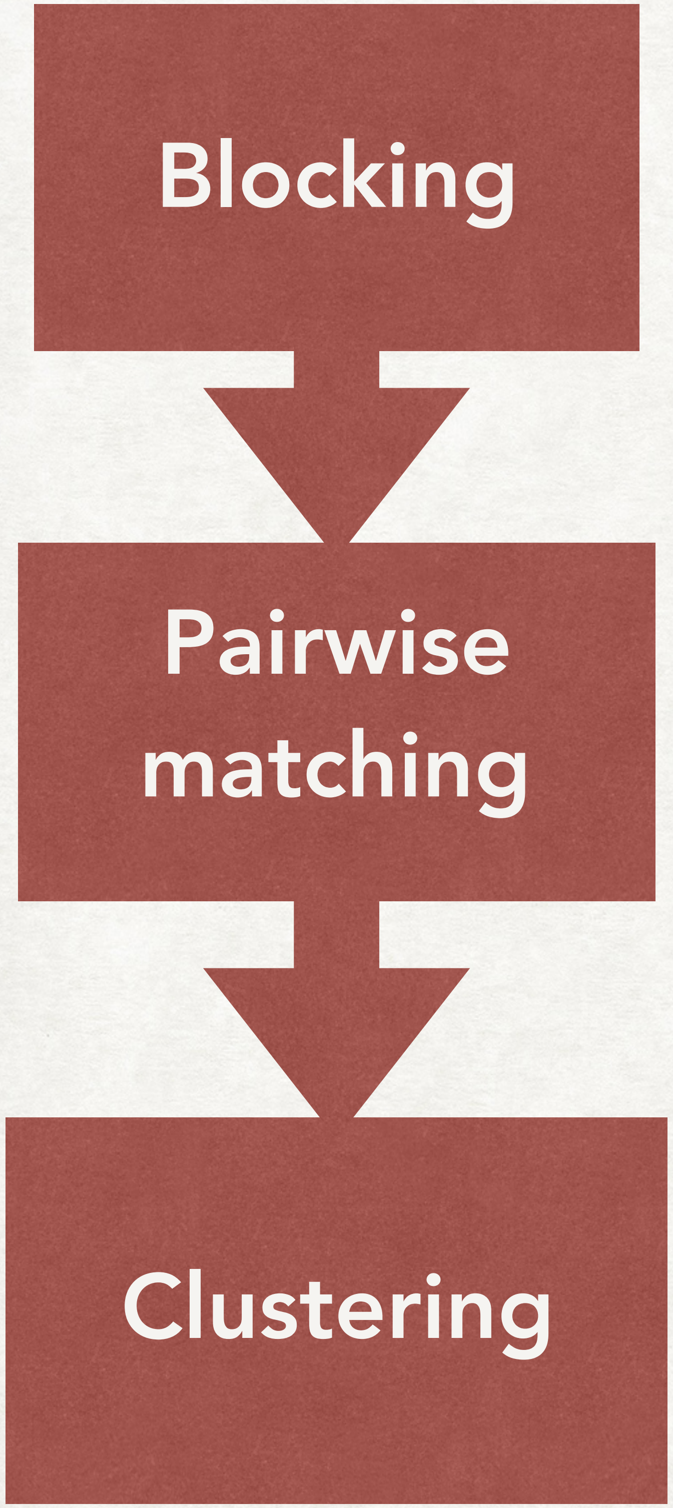
  — not able to support external implementation of other methods or applicable only for specific problem

Data Integration Steps



- The *schema alignment* step is performed to merge different data source schemas into one unified data schema
- The *record linkage* step is performed over extracted data with unified schema. The main goal is to identify those records that describe the same real-world entity
- On *data fusion* step different representations (records) of the same entity are found and combined into a single representation while inconsistencies in the data are resolved

Record Linkage Steps



Blocking

Pairwise matching

Clustering

- The *blocking* step is performed to significantly lower the number of comparisons. Instead of performing pairwise comparison of all records, we do this for subsets of records
- The *pairwise matching* step is performed to calculate the similarity of pair of records
- The clustering step is performed to identify those records that describe the same real-world entity

Targets of Research

Previous work of the authors [1] presented an architecture for big data integration in distributed environment

This work concentrates on:

- extensible approach for development of the system to perform materialized integration
- prototype of the data integration system including advanced methods for big data integration. The prototype is applied in e-commerce domain.

[1] Sazontev V.V.: Methods for Big Data Integration in Distributed Computation Environments, Conference on Data Analytics and Management in Data Intensive Domains, 2018

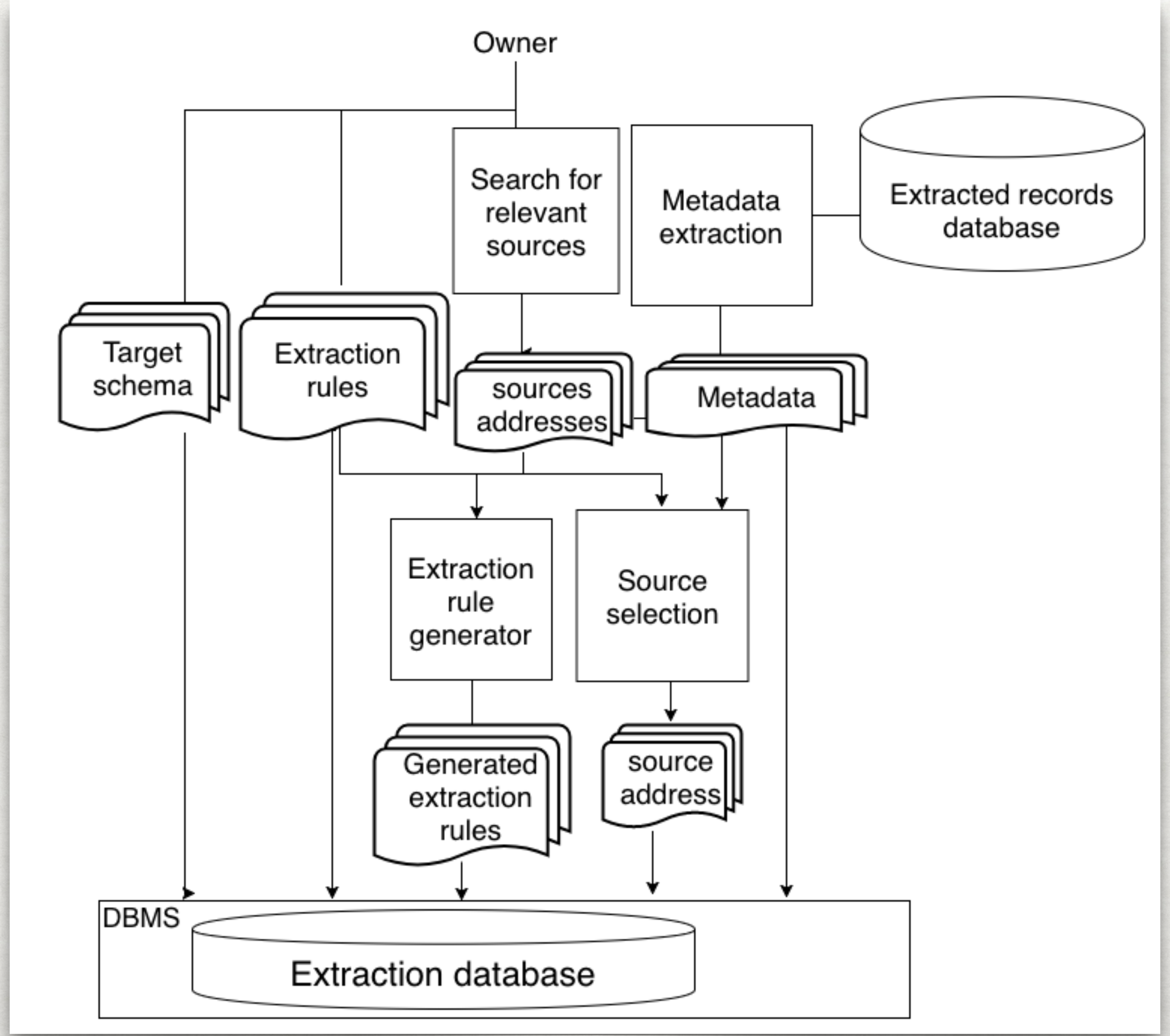
Key Points of the Approach

The aim of the approach is to provide end-to-end data integration extensible solution and overcome the limitations of known existing research and enterprise systems

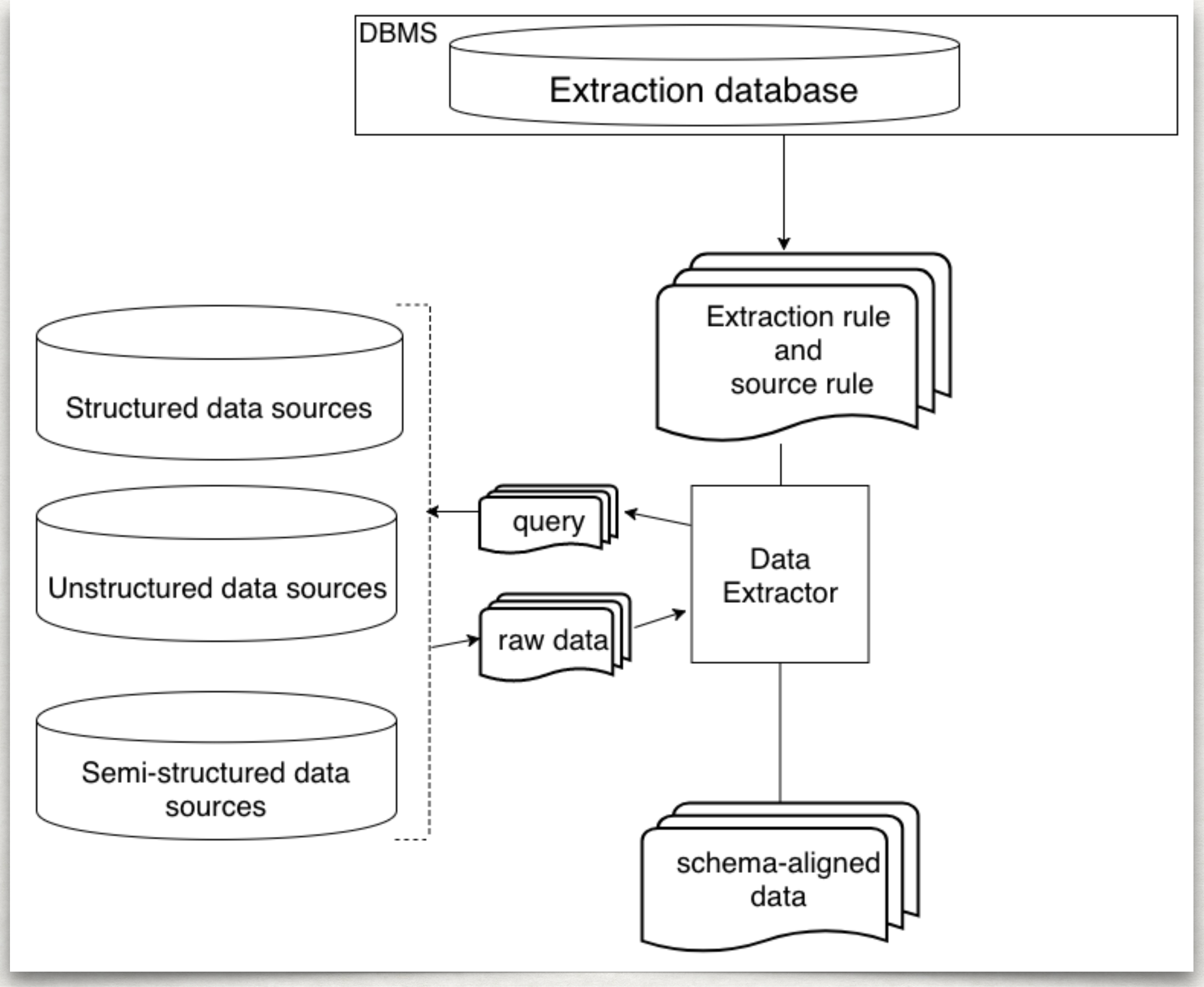
The system should be:

- extensible
    - the system defines separate steps of data integration
    - consistent interfaces and data manipulation of steps
- with advanced methods for big data integration
- operate in a distributed computational environment and scale effectively
- perform materialized data integration (similar to ETL in conventional data warehouses)

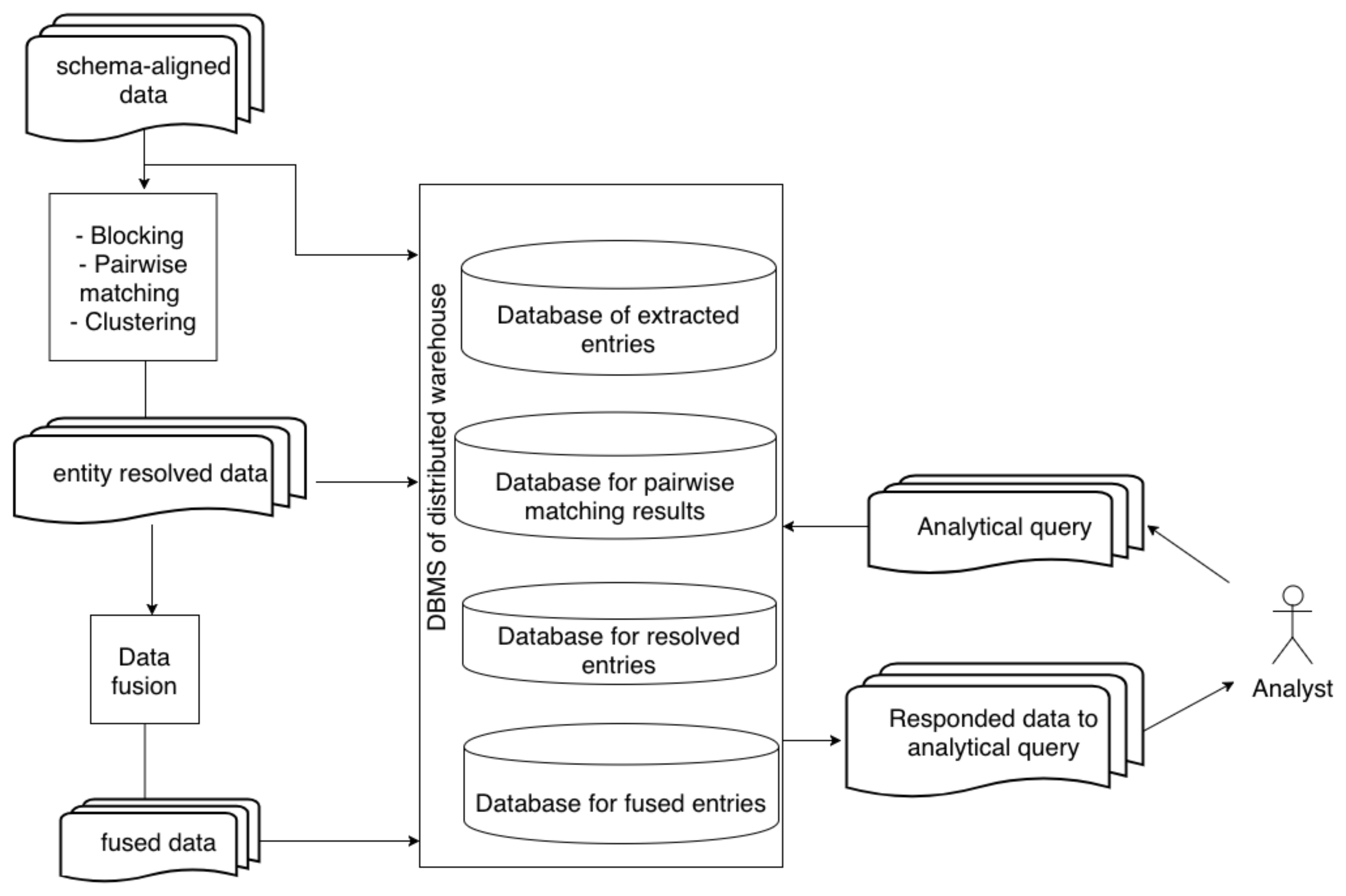
Architecture of the Approach



Architecture of the Approach



Architecture of the Approach


schema-aligned data
- Blocking
- Pairwise matching
- Clustering
entity resolved data
Data fusion
fused data
DBMS of distributed warehouse
Database of extracted entries
Database for pairwise matching results
Database for resolved entries
Database for fused entries
Analytical query
Responded data to analytical query
Analyst

Prototype Implementation

The prototype of the approach is implemented in the distributed computational environment Hadoop.

- Programming language - Python
- Computational model - Spark
- Warehouse - Hive

Implemented methods (I)

- Pasley DSL[5] language is used to describe rules of extraction (implemented with Paslepy on Python). The extraction is performed by web-crawler developed using Scrapy
- The blocking functions applied on blocking step to decrease the quadratic complexity of pairwise matching step.
- Pairwise matching, to compute similarity of pairs of records and identify records of same entity
- Agreement/Disagreement Decay temporal metrics, to identify the evolution of same entity over the period of time

Implemented methods (II)

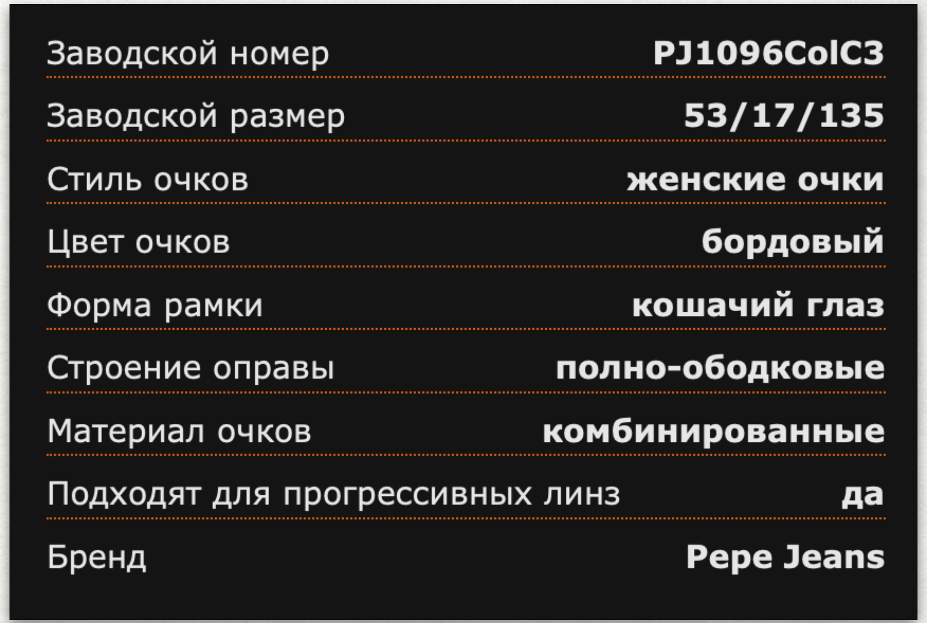
- The *PairRange* method evens out the load of clusters
- *ConnectedComponent* as clustering algorithm - marks each cluster of records as distinct entity
- Incremental Record Linkage algorithm clusters records (where each cluster corresponds to unique entity). In case of stream of small incremental updates, it is inefficient to start whole entity resolution from scratch, this algorithm stores the results of previous integration process and update clusters of records according

Use case

Domain: e-commerce, eyewear stores

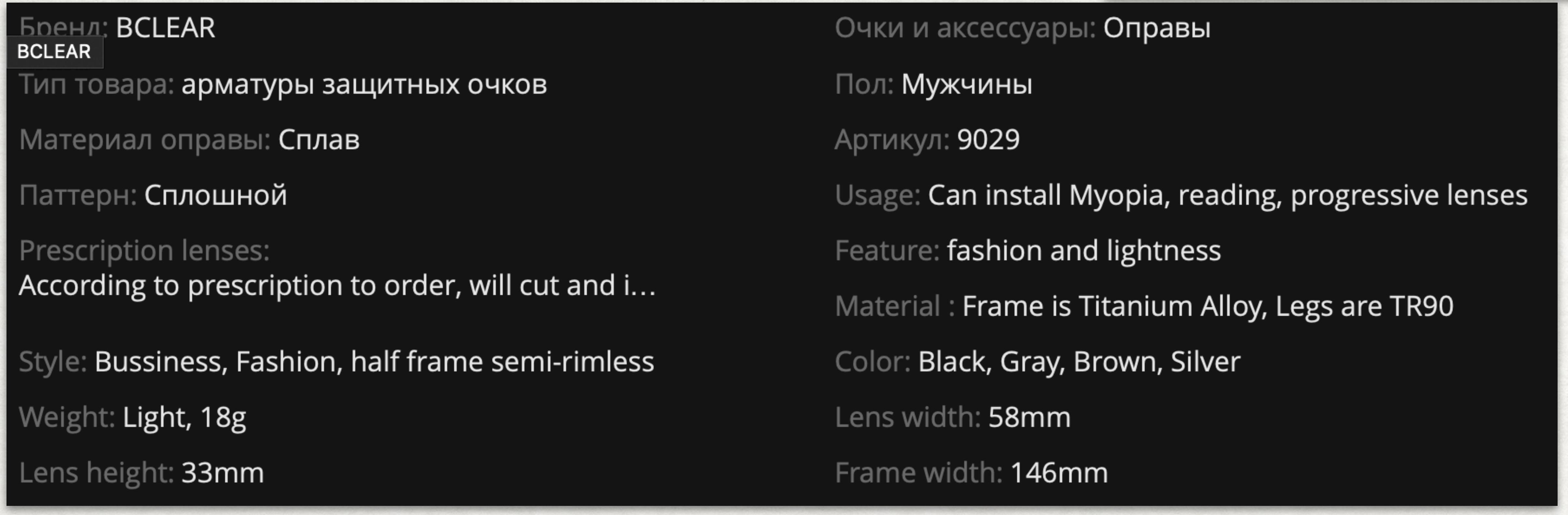
This domain has wide variety of items, frequent updates and large number of heterogenous data sources

Opticbox.ru:



Заводской номер	PJ1096ColC3
Заводской размер	53/17/135
Стиль очков	женские очки
Цвет очков	бордовый
Форма рамки	кошачий глаз
Строение оправы	полно-ободковые
Материал очков	комбинированные
Подходят для прогрессивных линз	да
Бренд	Pepe Jeans

Aliexpress.com:



Бренд: BCLEAR

BCLEAR

Тип товара: арматуры защитных очков

Материал оправы: Сплав

Паттерн: Сплошной

Prescription lenses:
According to prescription to order, will cut and i…

Style: Bussiness, Fashion, half frame semi-rimless

Weight: Light, 18g

Lens height: 33mm

Очки и аксессуары: Оправы

Пол: Мужчины

Артикул: 9029

Usage: Can install Myopia, reading, progressive lenses

Feature: fashion and lightness

Material : Frame is Titanium Alloy, Legs are TR90

Color: Black, Gray, Brown, Silver

Lens width: 58mm

Frame width: 146mm

Conclusions and Future Work

The main contributions of this work are:

- description of key points of an extensible approach for development of a materialized big data integration system
- a prototype of data integration system implementing the approach
- application of the prototype in e-commerce domain

As a future work it is planned to implement greater level of automation in extraction from the Web. The automatic extraction approach CERES can be used.

The current work lacks methods for schema alignment. Probabilistic Schema Alignment method is an example of such method that could be implemented.

It is planned to implement more advanced data fusion methods, such as AccuCopy method, since this work offers only simplest solution

It is planned to evaluate the system and methods over various synthetic data sources to check the scalability of the system