



Prepare the Roadmap to Release 0.6.0
(*Apache Incubator*)

Core Functions

- collect DATASET MD (Schema, Profiles, Usage)
- connect datasets with DOMAIN specific ontologies
- track domain specific activities on a DATASET
- expose MD to visualization tools
- expose MD to query planners (Impala, Hive)
- expose MD via HCatalog/Kite to processing layer

The path ...



MapReduce &
“inCluster queries”

MapReduce API



“inCluster Workflows”

Oozie, Sqoop, Flume



Cluster Spanning
Data Driven Business &
Research, using linked datasets

ETOSHA

Schema REPOSITORY

- **track** the **schema lifecycle** of a DATASET and connect it to DOMAIN-Ontologies
- Each dataset has its own URI
 - all triplified facts are available as RDF graphs

EXPOSE metadata

- SPARQL interface and client side integration

EXPOSE metadata

Explore Datasets

This page collects some examples for explorative queries to our knowledge graph.

Contents [hide]

- 1 Where is trouble shooting required?
- 2 Analysis Goals
- 3 Known Dataset
- 4 Known Algorithms
- 5 Applied Algorithms
- 6 Central Nodes
- 7 Time Dependent Multilayer Networks

Where is trouble shooting required?

	• Trouble shooting required •
Apic2011Data1	
Apic2011Data2	
CIAWorldFactbook	
Hale2014Data	
USCensus	
Working Dataset for Coppock Indicator	extract the raw-time-series and create a Dashboard for each CN page for DJIA data

Analysis Goals

	• Goal is •
Apic2011Data1	
Apic2011Data2	
CIAWorldFactbook	
Hale2014Data	
USCensus	
Working Dataset for Coppock Indicator	find out, if there is a correlation between the Coppock Indicator, computed for the financial data and the Wikipedia usage data

Known Dataset

	• Info •
CN 2 en Formula One BIN=24 t7 removed dissertation DEMO 2011 merged TS 20000	
CN 4 en Influenza BIN=24 t7 removed log10 dissertation DEMO 2010 merged SO 4.Advent	
Demo	

Page Discussion

Editing Explore Datasets (section)

Warning: You are not logged in.

Your IP address will be recorded in this page's edit history.

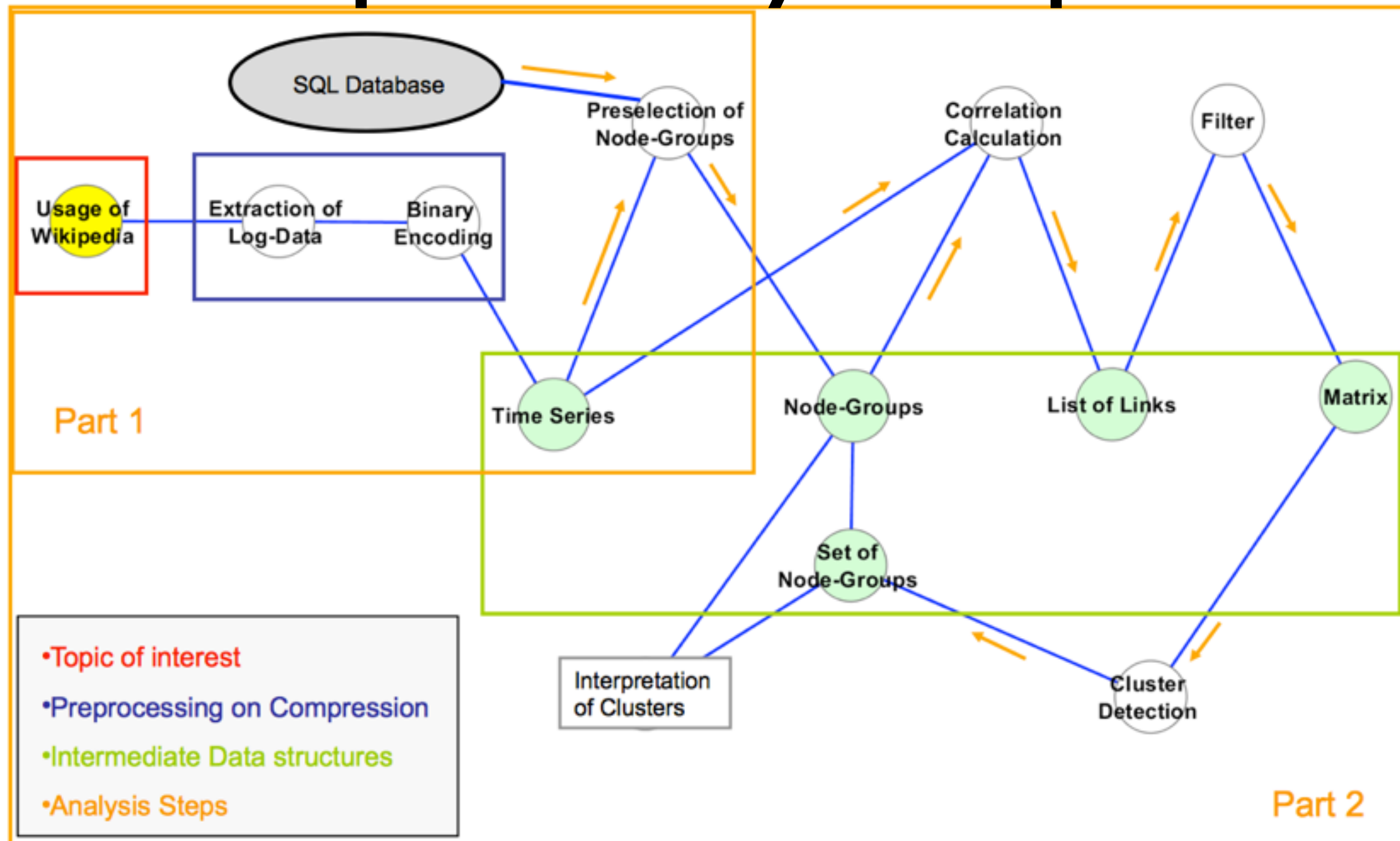


```
==Where is trouble shooting required?==
{{#ask: [[Category:DOAD]]
| ?trouble shooting required
}}
```

EXPOSE metadata

- visualize dependencies and correlations between datasets, projects, processes, resources and objectives
 - >>> **Visual and Quantitative**
RISK ANALYSIS is based on
network metrics from
complex systems research

Analysis Scenario ... Flow & Dependency Graph



Social Media Analysis - more than a buzzword! - Based on daily access rates to Wikipedia pages (or even groups in a given semantic context) one can study complex systems like financial markets or technology evolution and emerging markets, like the market around the Hadoop Ecosystem.

A fresh Web-UI

Business Perspective



Big-Data-Asset Manager

Business Perspectives

Strategic goals and short term operational goals must be in line. But, both can be orthogonal to each other, at least if short time-ranges are considered and temporary views are used for orientation.

Define goals and baselines, explore recent research results and technical white papers to find objectives, and metrics which allow a comparison with your own process metadata.

Open Module

Risk-Assessment

Deviation from the optimal path - if such one exists - and differences between the current state and the target contribute to the overall risk of projects. Define the objectives, analyze the current state and find out the influencing aspects.

Beside technical risks we manage ethical and privacy issues. You have to start with defining privacy roles and ethics statements.

Open Module

Technical Perspectives

Data availability, data quality, cluster operations, and the status of tool-development projects are the operational dimensions, with high impact on you success in a data driven world. But because many different tools are used in different contexts, we have to provide a homogenous view to the management level.


Open Module

Top Metrics



- explore the path, node by node, walk on it ...
- merge multiple layers
- interconnect thing from multiple categories ...

A Dataset Profile (in the backend)



84.175.42.6

Talk for this IP address

Create account

Log in

Page

Discussion

Read

Edit with form

Edit

View history

Search

Main page

WikipediaExplore (WE)

collect local neighborhood graph

extract time series

analyze time series

DATASET CATALOG (GDC)

register a new dataset

explore datasets

Toolbox

What links here

Related changes

Special pages

Printable version

Permanent link

Page information

Browse properties

Infos

Community portal

Current events

Recent changes

Help

HadoopMarketStudiesDS

This data set contains a set of local neighborhood networks around selected pages (CN) from Wikipedia. The Representation Index **REP** was calculated for the page data set. We also collected click count time series for all nodes in those neighborhood networks and calculated the Timresolved Relevance Index **TRRI**.

This dataset is used in our research project to develop a tool set around the time resolved relevance index.

Data is available in the following clusters:

BDATVM4.5.0v1


DEV1

Query: Applied Algorithms

n.a. at this time

Static Network

File:Hadoop-market-static-network.tiff



Category: DataSet

Facts about "HadoopMarketStudiesDS"

Is available in cluster

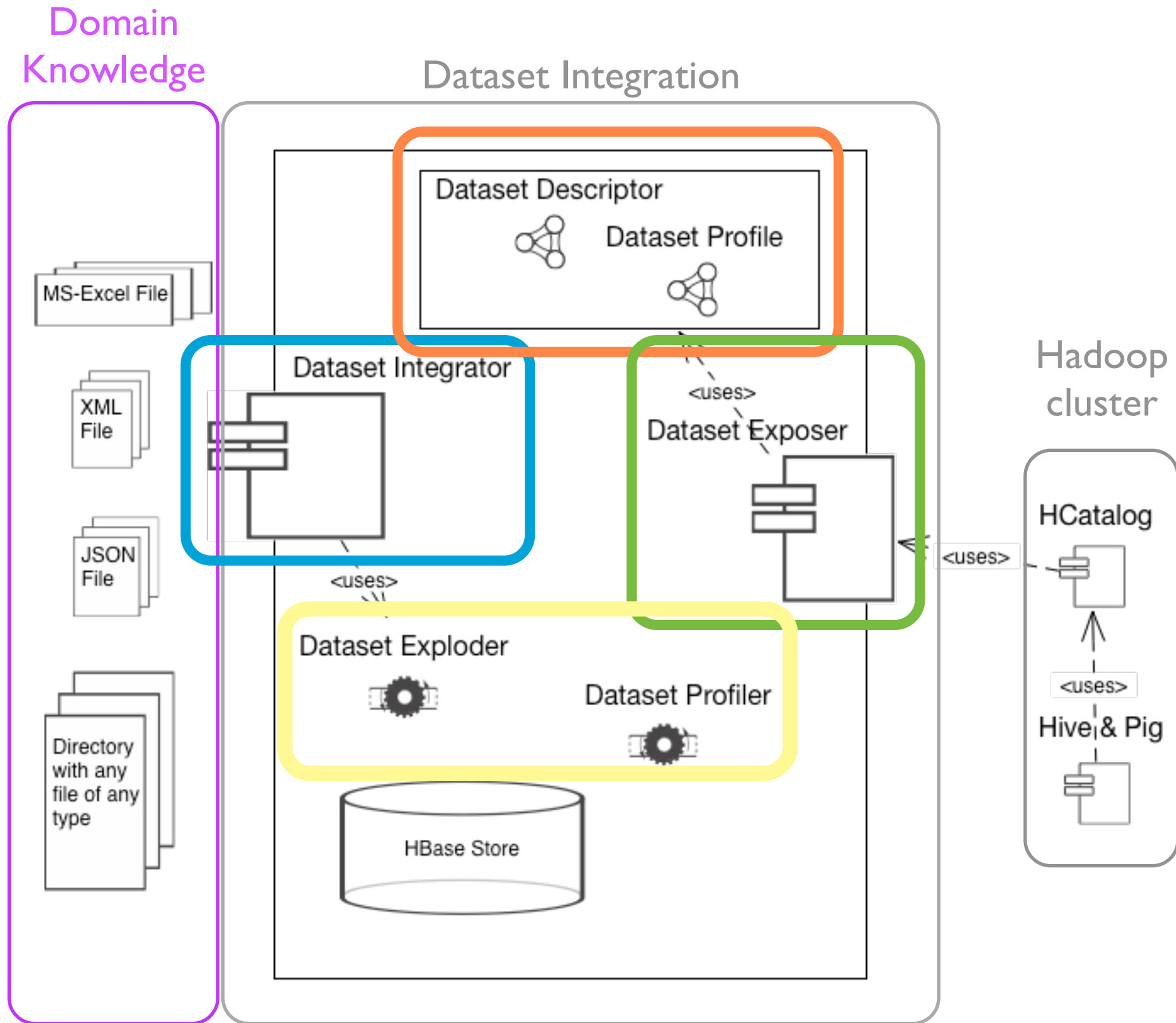
BDATVM4.5.0v1

 and

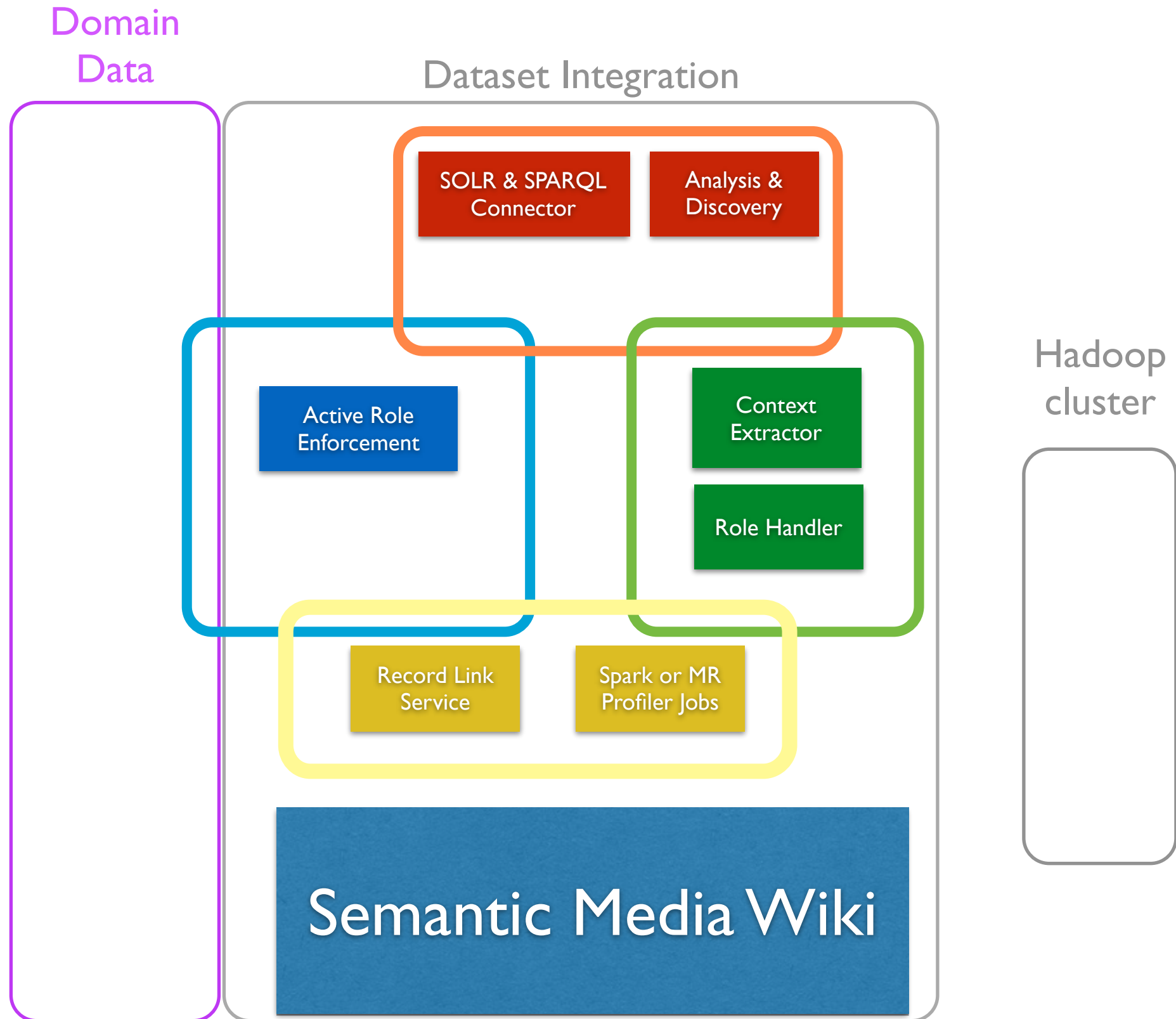
DEV1

RDF feed

Cluster Spanning Dataset Management



Cluster Spanning Dataset Management



Cluster Spanning Dataset Management

