Data management in heterogeneous metadata storage and access infrastructures

Marina Golosova

National Research Center "Kurchatov Institute"

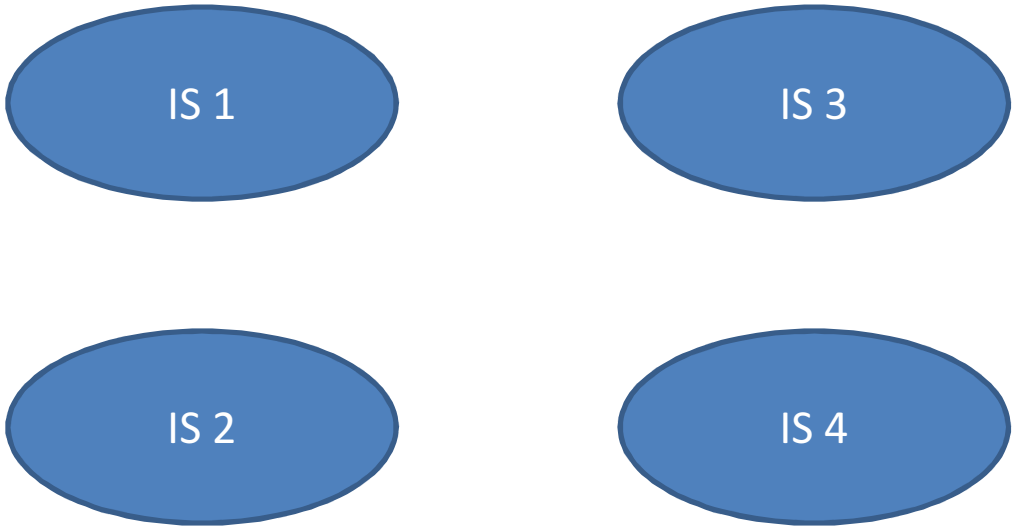
Outline

- Motivation
- Conceptual solution
- Prototype
- Future plans



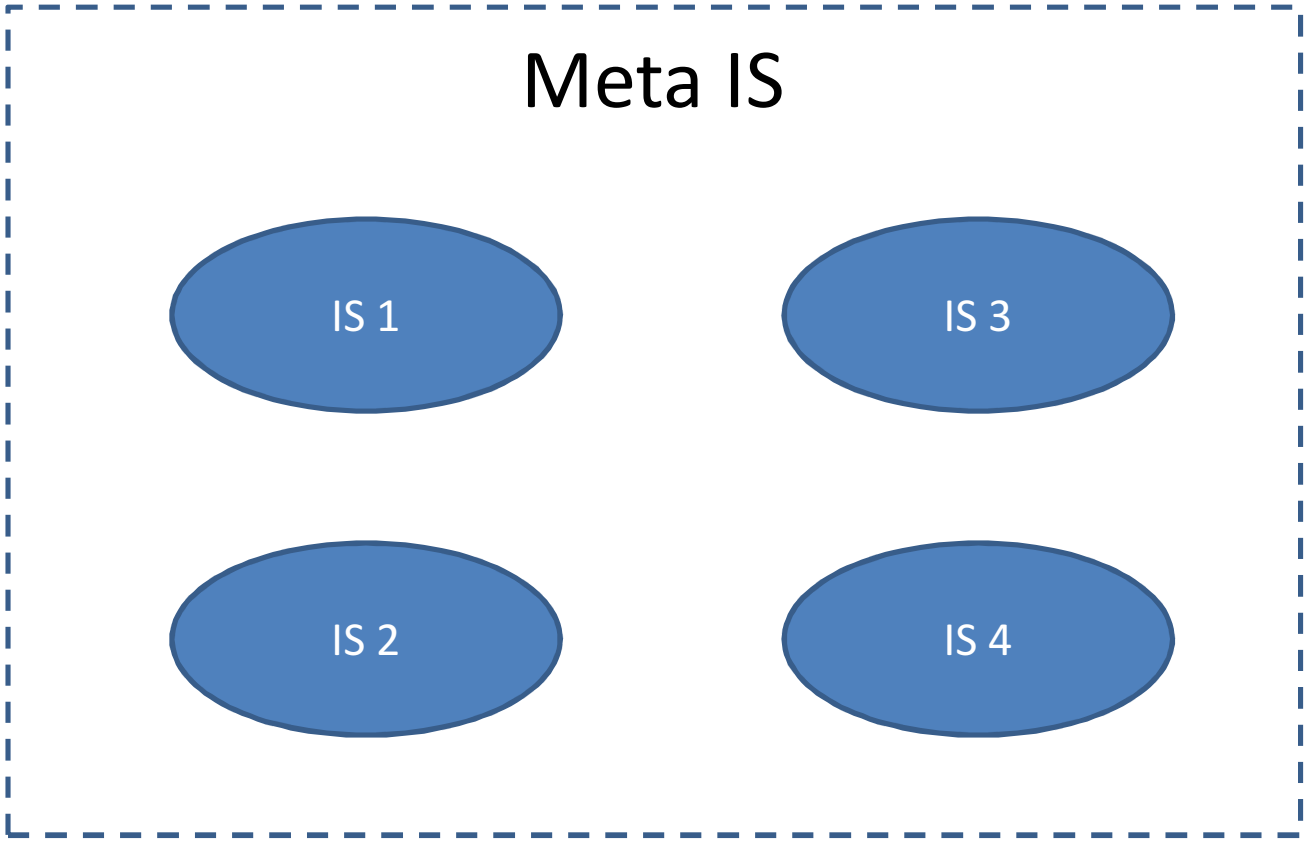



Motivation



IS – Information System

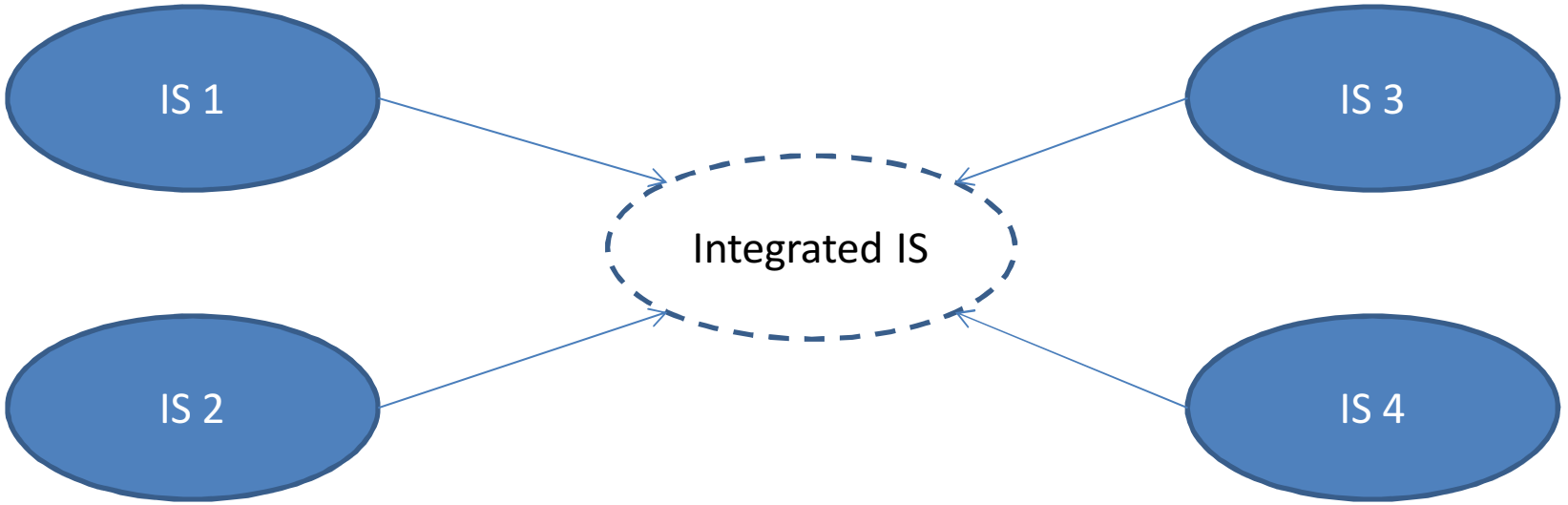
Motivation



IS – Information System

Motivation

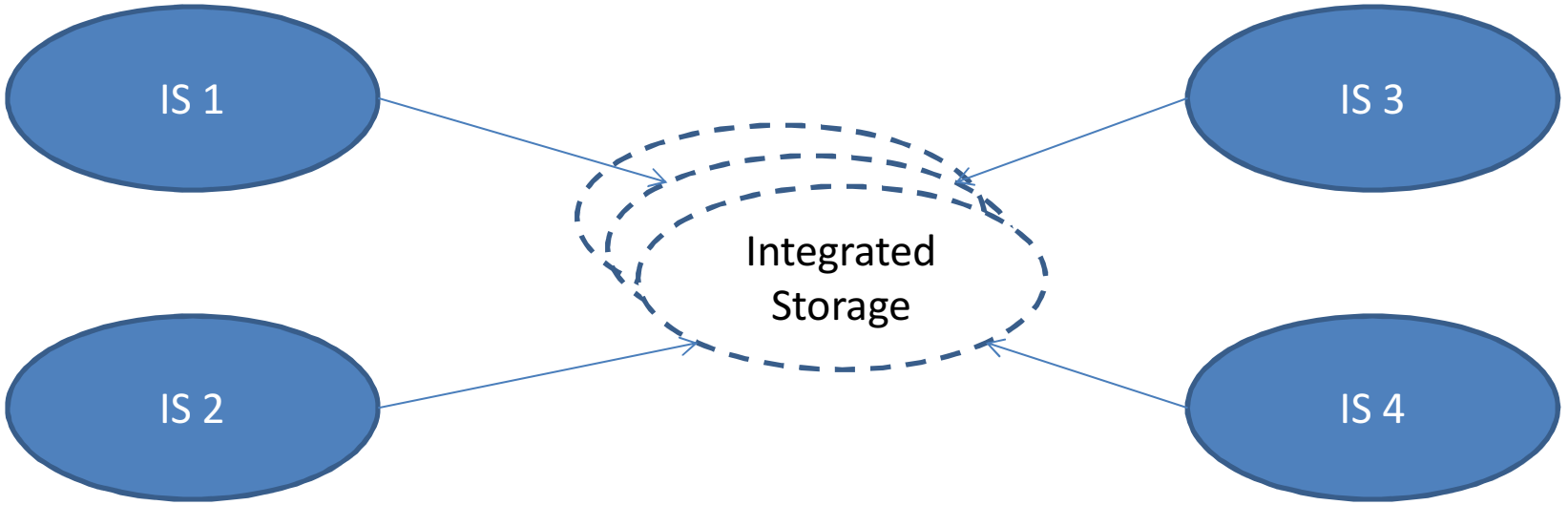
Meta IS
(integrated, unified, …)



IS – Information System

Motivation

Meta IS
(integrated, unified, …)



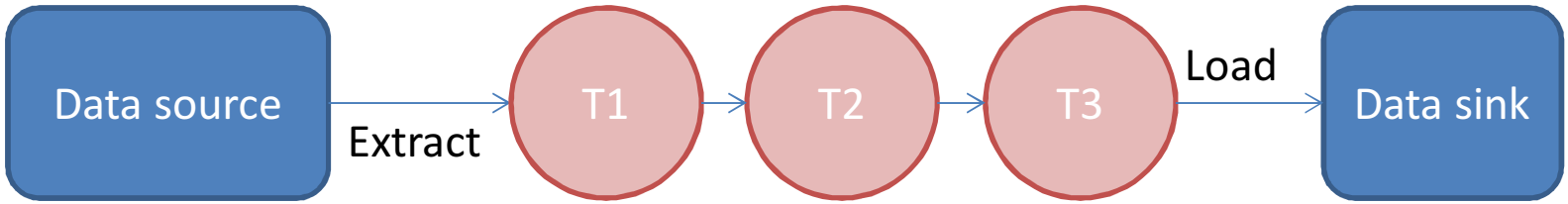
IS – Information System

Common ETL Issues

- Transformation chaining:
  - data delivery
  - T-chain (topology) management
- Flexibility
- Scalability
- Process supervising

ETL – (E)xtract, (T)ransform, (L)oad

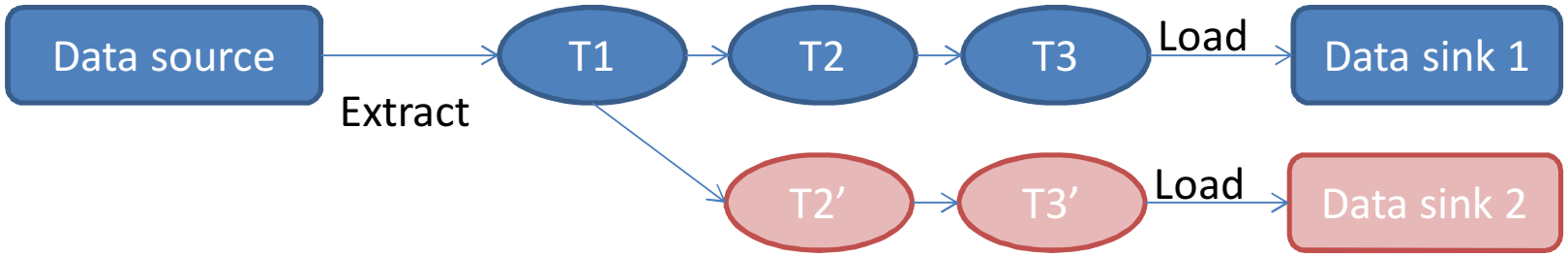
Common ETL Issues



- Transformation chaining:
  - **data delivery**
  - T-chain (topology) management
- Flexibility
- Scalability
- Process supervising

ETL – (E)xtract, (T)ransform, (L)oad

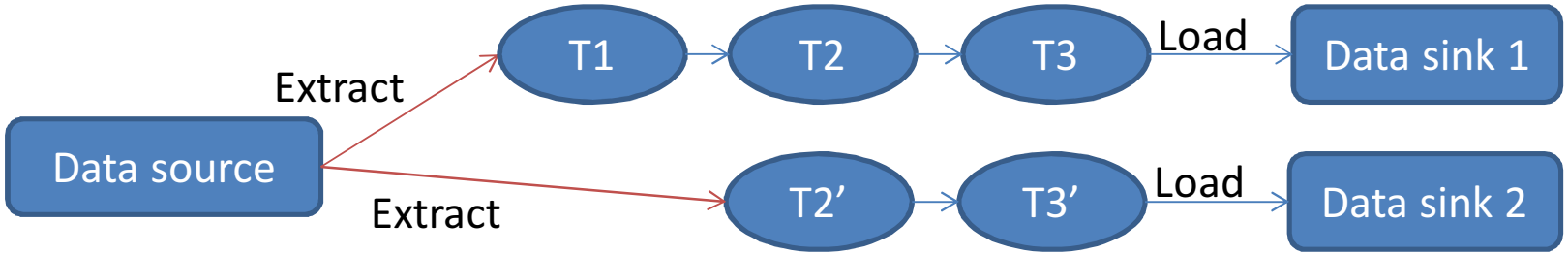
Common ETL Issues



- Transformation chaining:
  - data delivery
  - **T-chain (topology) management**
- Flexibility
- Scalability
- Process supervising

ETL – (E)xtract, (T)ransform, (L)oad

Common ETL Issues

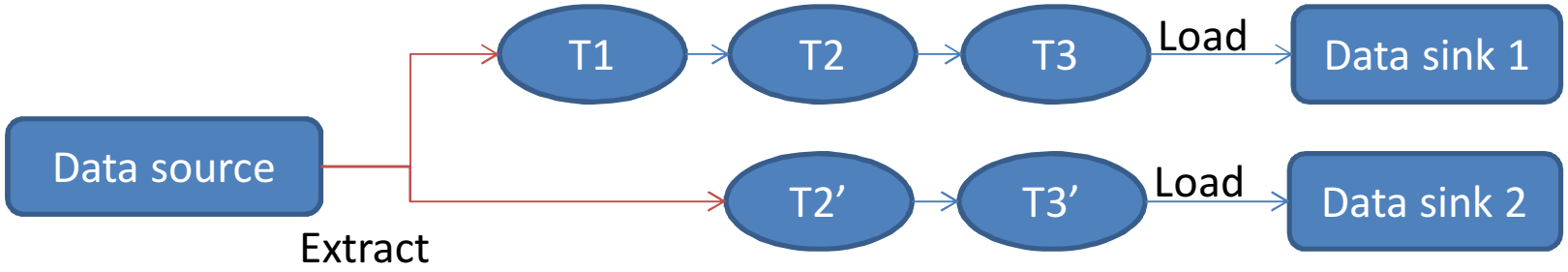


- Transformation chaining:
  - data delivery
  - T-chain (topology) management

- **Flexibility**
- Scalability
- Process supervising

ETL – (E)xtract, (T)ransform, (L)oad

Common ETL Issues

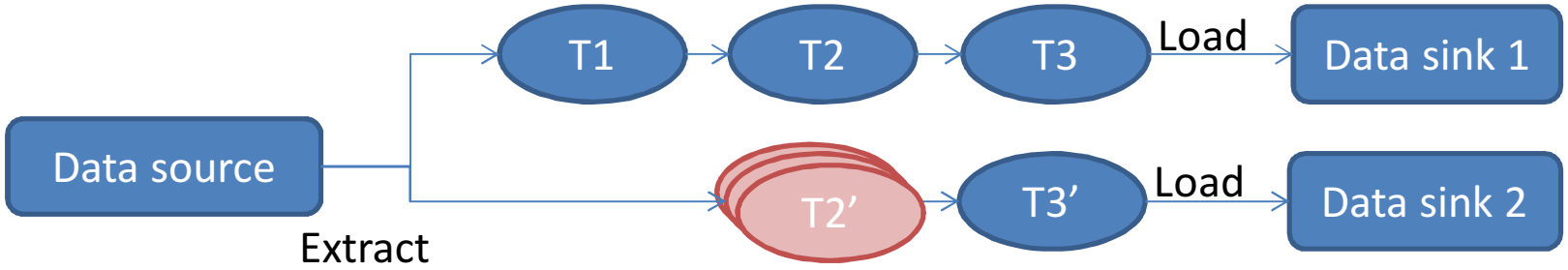


- Transformation chaining:
  - data delivery
  - T-chain (topology) management

- **Flexibility**
- Scalability
- Process supervising

ETL – (E)xtract, (T)ransform, (L)oad

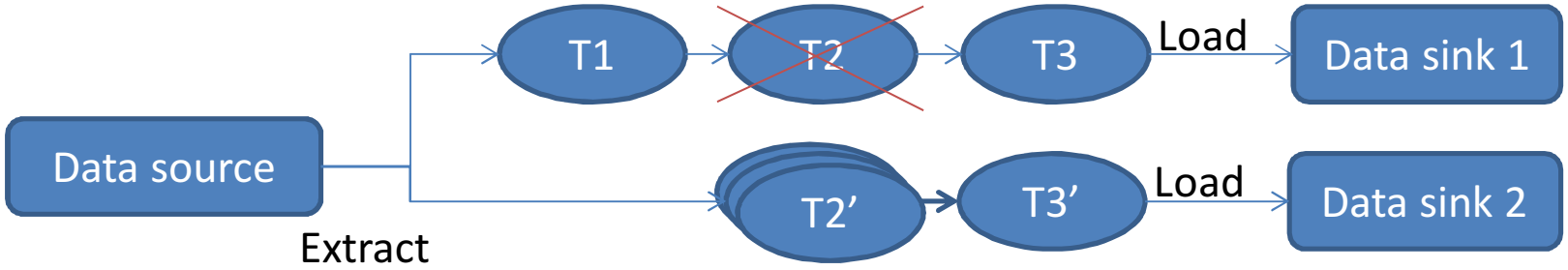
Common ETL Issues



- Transformation chaining:
  - data delivery
  - T-chain (topology) management
- Flexibility
- **Scalability**
- Process supervising

ETL – (E)xtract, (T)ransform, (L)oad

Common ETL Issues

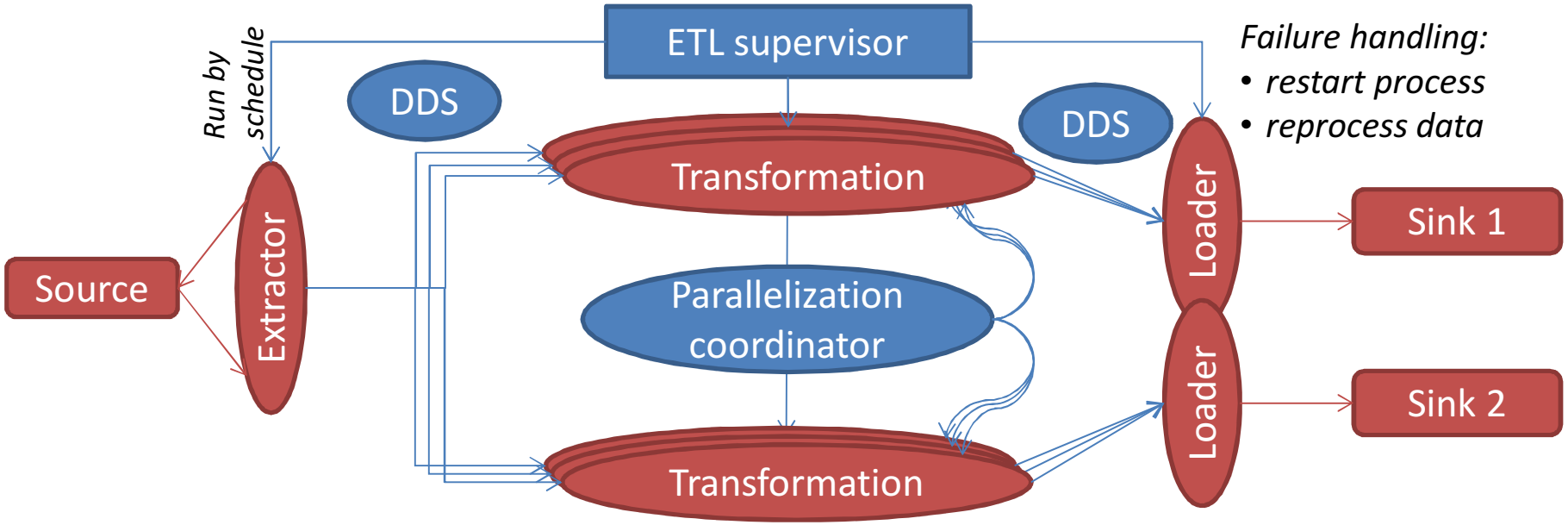


- Transformation chaining:
  - data delivery
  - T-chain (topology) management
- Flexibility
- Scalability
- **Process supervising**

ETL – (E)xtract, (T)ransform, (L)oad

Dataflow Framework Concept

- General framework, taking care of the common issues
- Custom (subject area specific) ETL modules



ATLAS DKB Prototype

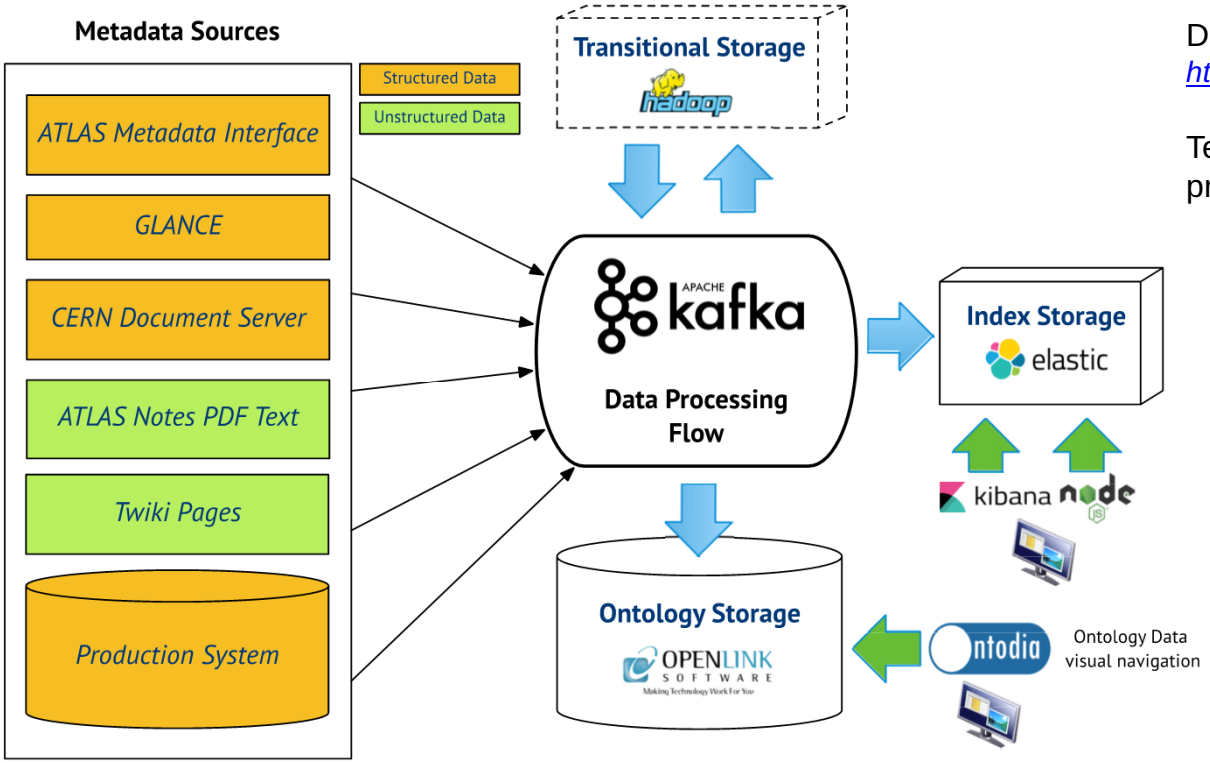
We had:

- a bunch of executables written in Bash, Python, PHP, …
- the set of data sources (ATLAS/CERN ISs: GLANCE, CDS, ProdSys, …) and sinks (Integrated storages: Hadoop, Virtuoso, …)

We added:

- stream processing approach: simple protocol (specified EOMessage and EOProcessing markers)
- Apache Kafka as a temporary storage for messages between executables
- Apache Kafka Streams as executable topology manager (data delivery, scheduling, …)
- Apache Kafka extensions:
  - to use any executable as a topology nodes
  - to configure topology (not to hardcode it)
- Python library to simplify executables development/adaptation (keeping in mind that things must be kept as simple as possible to be implemented without any specialized library)

DKB Architecture Prototype



DKB program code on GitHub:
https://github.com/PanDAWMS/dkb

Technology evaluation and system prototype architecture:

- ✓ RDF storage: **Virtuoso**
- ✓ Transitional storage: **Hadoop**
- ✓ Metadata streaming&processing: **Apache Kafka**
- ✓ Knowledge Base navigation (Web GUI): **Ontodia**
- ✓ Index storage: **ElasticSearch**
- ✓ Index search web-interfaces: **Kibana, NodeJS**
- ✓ Documents metadata mining: **PdfMiner utility and implemented PDF Analyzer tool**

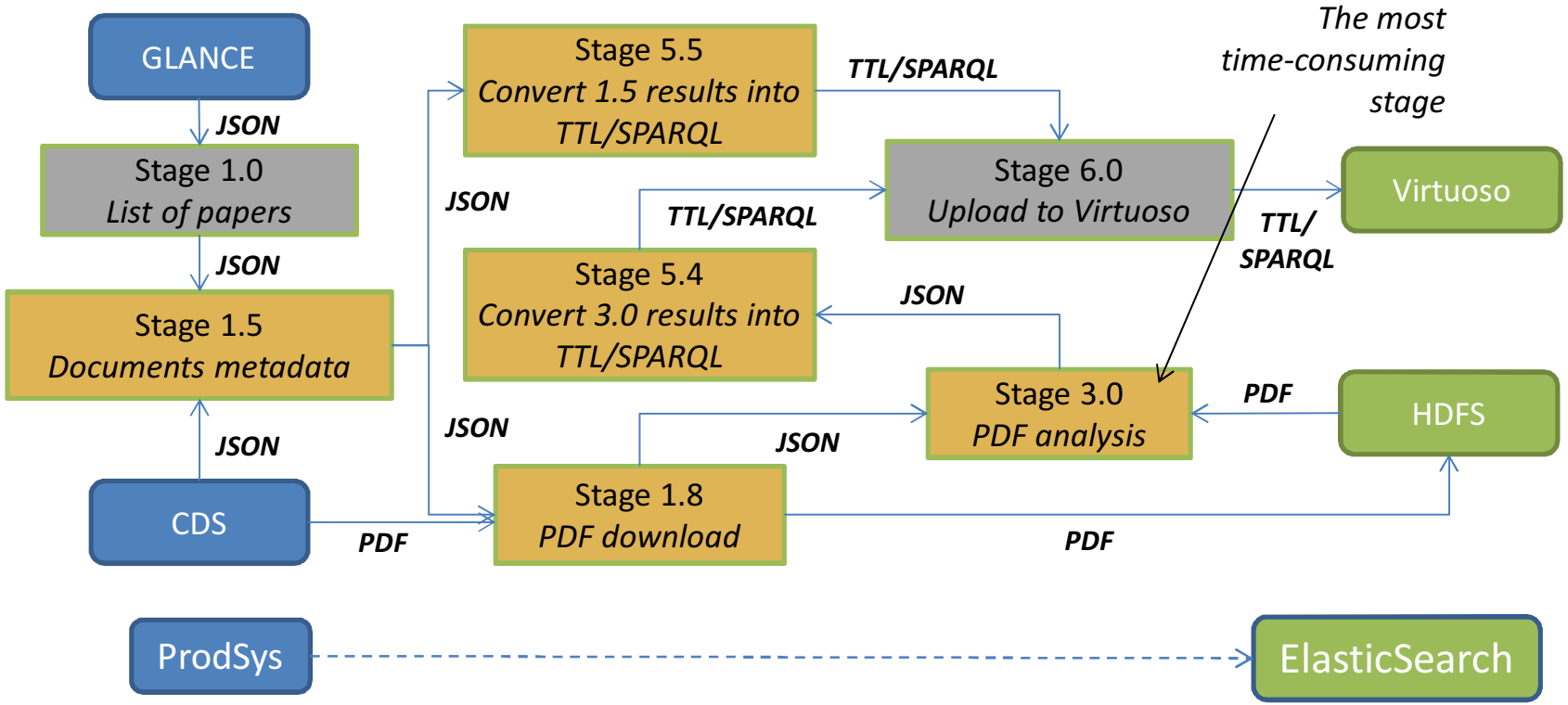
Basic programming language - python

Maria Grigorieva

ATLAS DKB Dataflow Prototype

The ETL process consists of two chains:

- GLANCE -> 1.0 -> 1.5 -> 5.5 -> 6.0 -> Virtuoso
- GLANCE -> 1.0 -> 1.5 -> 1.8 -> 3.0 -> 5.4 -> 6.0 -> Virtuoso



Summary

- Fully automated dataflow for DKB is implemented for the instance in KI

- The dataflow is configurable, meaning that the chains and stages can be added or removed via config files, without manual scheduling and chaining

Future Plans

Near-term:

- Puppetized Dataflow installation
- Extended protocol to make the ETL topology configuration more flexible (improved data routing within the topology)
- Improved management tools to minimize manual operations

Long-term:

- Visualized monitoring
- GUI for management tools
- Authorized access (allowing external user (IS) to add their own data sources to the DKB / get DKB feedback)

Acknowledgment

- The work was supported by Russian Foundation for Basic Research (RFBR) under contract No. 16-37-00246.

Thanks

Many thanks to these wonderful people:

- Maria Grigorieva, *for being our team leader and an extraordinary hard worker*
- Alexei Klimentov, *for guidance*
- Eugene Ryabinkin, *for treasurable discussions*
- Torre Wenaus, *for encouraging us to the new deeds*
- Anastasia Kaida, *for joining the pack*
- *many people from* ATLAS collaboration, *for being open-minded and cooperative*

Thank you