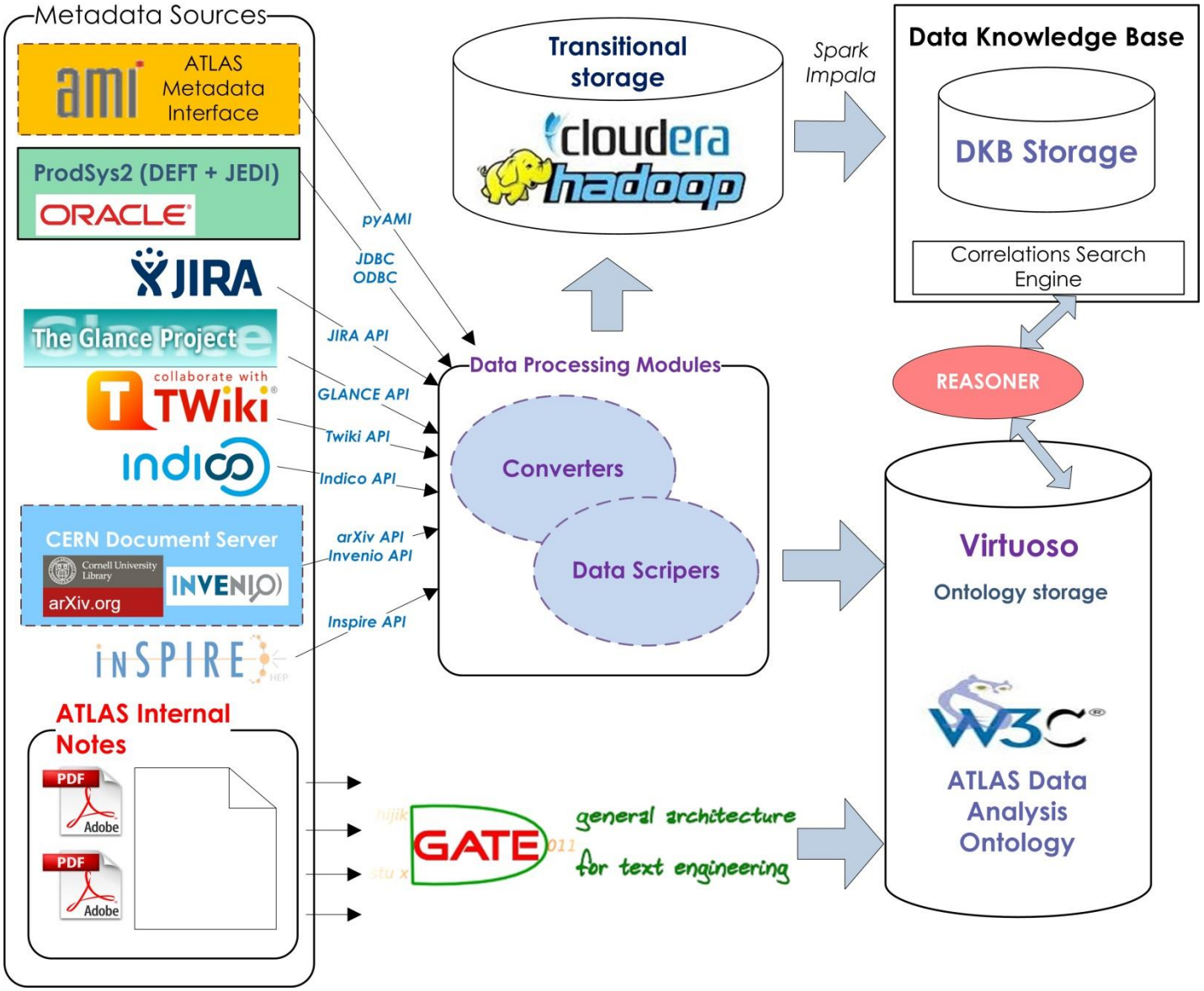
The basic ideas of the DKB development

- To represent all stages of data processing and analysis by the ATLAS Collaboration in unified information space we need to formalize the data analysis life-cycle. So, now we are working on the development of the Data Analysis OWL Ontology for ATLAS
- To "feed" this ontology with the information we will use the following methods:
  - Data acquisition from *ATLAS metadata sources*: AMI, GLANCE, Rucio, JIRA, Indico, CDR, CERN Twiki
    - *Data findings: energy, luminosity, Run Number, generators, papers, conf Notes, managers, JIRA-tickets*
  - Mining full texts of *ATLAS Internal Notes* (PDF documents)
    - *Data findings: dataset names, and other data required by ontology and absent in other metadata sources*
    - METHODS:
      - Preparing PDFs for the analysis, parse PDF to TXT
      - Linguistic analysis & Machine Learning algorithms for the automatic text markup
  - Mining *ProdSys2*
    - *Data findings: aggregated detailed information about tasks, input/output datasets*
- DKB storage backend: Hadoop clusters in NRC KI and TPU
  - Currently ProdSys1 & ProdSys2 have been exported to NRC KI Cluster in AVRO files
  - ProdSys data aggregation will be tested with Impala (NRC KI) & Spark (TPU)
- Ontology storage Virtuoso in TPU (in progress…)
- Test the GATE (*https://gate.ac.uk/*) - a full-lifecycle open source solution for text processing – for the Internal Notes processing.

DKB Architecture Prototype



Anomaly detection

- First step: simple “cold” task completion time predictions.
- Second step: machine learning-based models for both “cold” and “hot” predictions.

Task duration distribution

Project	Type	Provenance	TTC (days)
data15_13TeV	merge	GP	50.8
data15_13TeV	recon	GP	2.2
data15_13TeV	recon	AP	8.9
data15_13TeV	merge	AP	4.9
data15_5TeV	merge	GP	10.1
data15_5TeV	merge	AP	2.2
data15_5TeV	recon	AP	2