DATA KNOWLEDGE BASE CURRENT STATUS AND OPERATION

V. Kotliar[a]

Institute for High Energy Physics named by A.A. Logunov of National Research Center "Kurchatov Institute", Nauki Square 1, Protvino, Moscow region, Russia, 142281

E-mail: [a] viktor.kotliar@ihep.ru

The Data Knowledge Base (DKB) project aims at knowledge acquisition and metadata integration. It provides fast response for a variety of complicated queries, such as summary reports and monitoring tasks (aggregation queries) and multi-system join queries. Such queries are not easy to implement in a timely manner and, obviously, are less efficient than a query to a single system with integrated and pre-processed information would be. This work describes the status of the project as well as its integration with the ATLAS Workflow Management and future perspectives.

Keywords: information integration, metadata integration, metadata, workflow pipelines

Viktor Kotliar

1. Introduction

The Data Knowledge Base (DKB) project aims at knowledge acquisition and metadata integration [1]. It started at 2016 with main purposes: integrate and link pieces of information from independent sources (pdf, indico, wiki page, etc.); reconstruct connections between research results and data samples; provide fast and flexible access to everything people might want to know about some process or object. From 2018, the main goal of the project changed to create a universal tool for multi-source queries. A python library pyDKB [2] was created to address necessaries for workflow pipelines adopted to the High Energy Physics (HEP) projects. The ATLAS [3] dataflow system is installed based on the developed software that consists of:

- ETL (Extract, Transform, Load) pipeline flow [4] based on scripts and library;
- System to run and check the flow;
- NoSQL database to store results;
- REST API to access system;
- Frontend UI for users.

This system is used in the production system at ATLAS experiment to operate with GRID computing metadata and to prepare LCH Run 3.

2. DKB environment overview

DKB project has a distributed environment over several virtual machines hosted by CERN openstack infrastructure [5]. These machines are managed by computing center virtual machine software management system which includes Puppet and Foreman profiles. The whole environment is split over production, quality assurance and development servers. CentOS7 x86_64 operating system is used as base OS for all services. Production system is shown on figure 1.

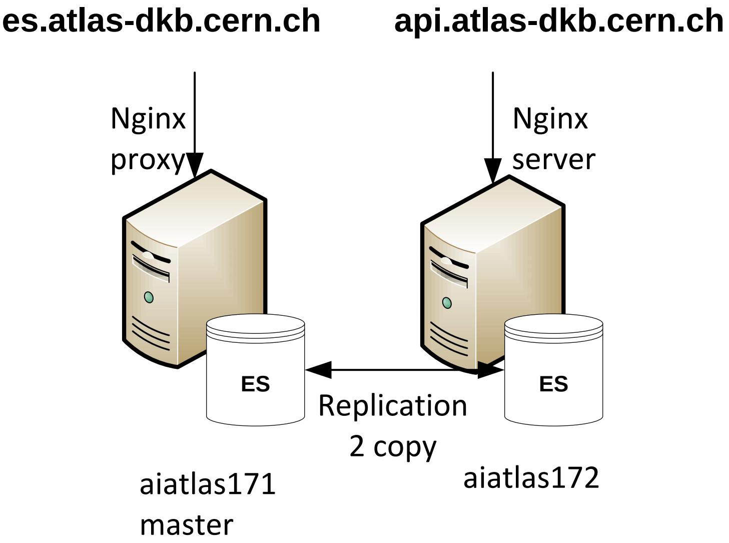

Figure 1. DKB production environment

It consists of two servers aiatlas171(master) and aiatlas172 with load balanced names assigned to them es.atlas-dkb.cern.ch and api.atlas-dkb.cern.ch accordingly. Elasticsearch [6] engine is used for data preservation and it is configured for two-copy replication mode. Such mode allows to achieve a good speed for read access and safety for data. There are two nginx servers used for system access to DKB from outside. First one works as proxy to ensure direct access levels to Elasticsearch engine for users with read-only or read-write permissions. Second one works as http server for DKB API software based on python FastCGI program. The main DKB workflow pipeline is configured to run only on the master node leaving slave node only for serving API requests.

The project sources are available on github [1] and development to production workflow goes through github pull requests [fig. 2]

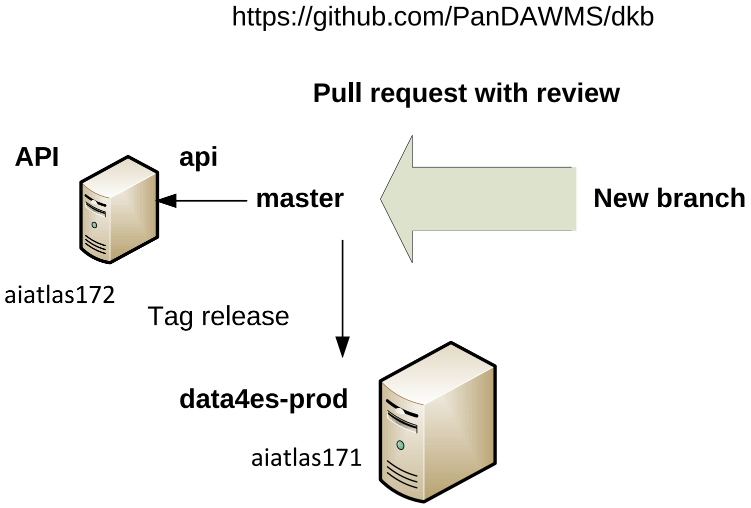

Figure 2. DKB development workflow

After a new functionality or bug fix are added and tested in a new git branch, a new pull request is created for merging changes to the master branch. All pull requests go through careful review from another person in the project and only after that merges into master. Master branch automatically applies to the API server (based on puppet profile) and manually gets tags and applies as data4es-prod branch on the production workflow server. For the moment DKB provides API with version 0.3.3 and DKB production workflow runs version 0.2-0.

Current environment stores near 15GB data for ATLAS production tasks and 50GB of data for ATLAS analysis tasks. Every hour it loads and stores metadata information about around 1500 tasks and 5000 datasets from ATLAS experiment.

3. Metadata integration

At present DKB serves for ATLAS collaboration Production System [7] as metadata integration service for the metadata at the level of Task and Dataset objects [fig. 3].

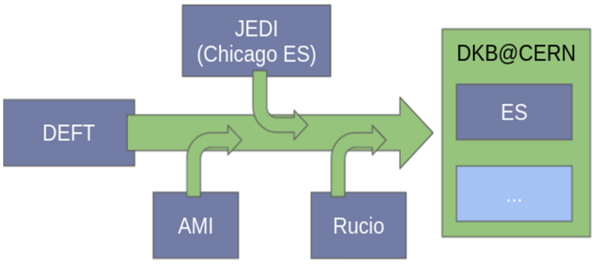

Figure 3. DKB ATLAS metadata integration

Information updates are based on “task timestamp” from ProdSys database. Main information comes from DEFT (Database Engine for Tasks) and is extended with additional metadata from other systems like:

- AMI – ATLAS Metadata Interface;

- JEDI - Job Execution and Definition Interface;
- Rucio - scientific data management system.

At the end as soon as the new integrated metadata stored in the single Elasticsearh it simplifies search queries for the whole systems and such queries integrated into the ProdSys user interface through web access [6].

From implementation point of view, this workflow pipeline presents ETL process which is shown on figure 4. It is implemented through one Linux bash script calling different stages (workflow parts). These stages could use any software inside but to simplify communications between them and simplify building of such stages DKB python library is used.

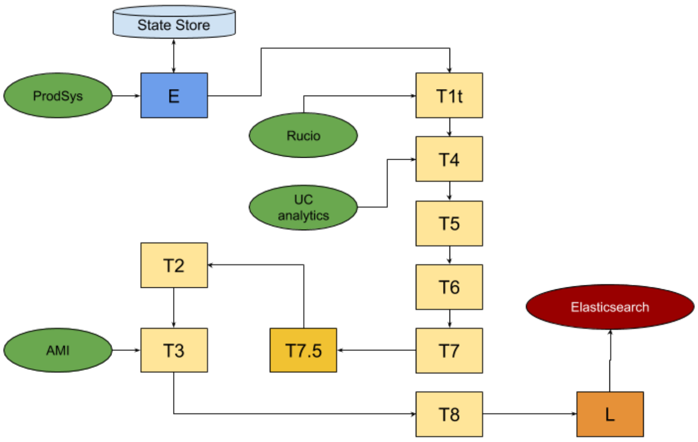

Figure 4. DKB ATLAS ETL workflow pipeline

4. Resent changes and plans

Several changes have been made to the DKB project recently, mainly aimed at improving system performance and upgrading it to use a new version of the Elasticsearch engine. To improve the performance of user operations, a new metadata indexing model is implemented for ATLAS integrated metadata. It takes into account the specifics of the already addressed use-cases, and the most noticeable change is that the output datasets properties are now stored together with the Task object, in the form of nested documents (instead of parent/child documents). It is made to simplify queries to the Elasticsearch index, used in the most problematic requests from the addressed use-cases. Some investigations are made on internal communication protocol for DKB stages to use batch processing instead of serial one which is in place in production. The nearest plan for DKB is to fully migrate to the CERN production Elasticsearch infrastructure and split data storage from the project to special dedicated outside service [fig. 5].

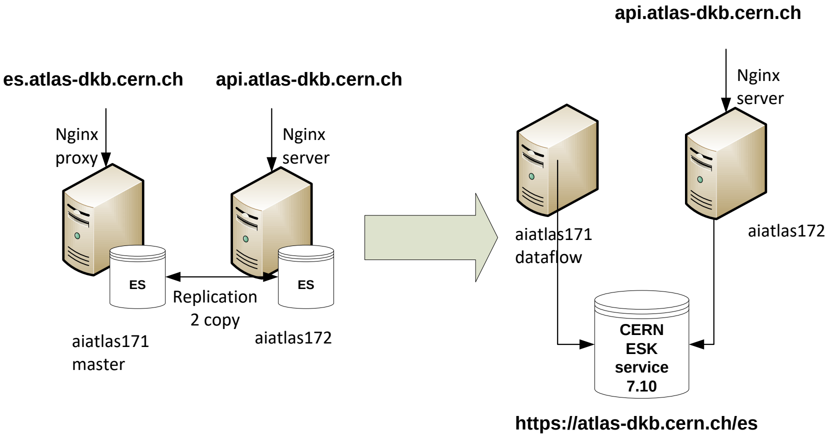

Figure 5. DKB migration to CERN Elasticserach

5. Conclusion

The Data Knowledge Base project is successfully integrated with the Production System of the ATLAS experiment and it allows execution of complex analytical requests, requiring information from different information systems and from different levels of abstraction in a timely manner. The developed library and resent changes allows implementation of multiple different scenarios for metadata integrations, providing flexible tool for building metadata workflow pipelines. Stable run in production and good availability and accessibility allowed to use DKB metadata integration service in processing ATLAS metadata for tasks and datasets to prepare LHC Run 3.

6. Acknowledgement

The DKB ATLAS metadata integration services is supported by CERN IT and by CERN ATLAS IT support team. The UI for DKB added by ATLAS ProdSys development team by M. Borodin.

DKB project is supported by NRC "Kurchatov Institute".

Special thanks to M. Golosova and V. Aulov for the project development.

References

[1] Grigoryeva M., Golosova M., Klimentov A., Wenaus T. Data Knowledge Base for HENP Scientific Collaborations.// Journal of Physics: Conference Series, vol. 1085, issue 3, 2018

[2] The Data Knowledge Base for HENP experiments [DKB]. Available at: https://github.com/PanDAWMS/dkb (accessed 22.09.2021)

[3] ATLAS Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider [ATLAS]. Available at: https://nordberg.web.cern.ch/nordberg/PAPERS/JINST08.pdf (accessed 22.09.2021)

[4] Extract, transform, load procedure in computing [ETL]. Available at: https://en.wikipedia.org/wiki/Extract,_transform,_load (accessed 22.09.2021)

[5] CERN OpenStack Private Cloud Guide [CERN OpenStack]. Available at: https://clouddocs.web.cern.ch/ (accessed 22.09.2021)

[6] The Elastic Stack [ESK]. Available at: https://www.elastic.co/elastic-stack/ (accessed 22.09.2021)

[7] The ATLAS collaboration Production System [ProdSys]. Available at: https://prodtask.cern.ch/dkb/ (accessed 22.09.2021)