PNPI - NRC KI

St Petersburg University

PLEKHANOV Russian University of Economics
Founded in 1907

ISP RAS

Monitoring System for the Russian Scientific Data Lake Prototype

Aleksandr Alekseev (ISP RAS), Andrey Zarochentsev (SPbSU), Andrey Kiryanov (PNPI), Tatiana Korchuganova (ISP RAS)

9th International Conference "Distributed Computing and Grid-technologies in Science and Education" (GRID 2021), 5-9 July 2021, Dubna, Russia

Introduction

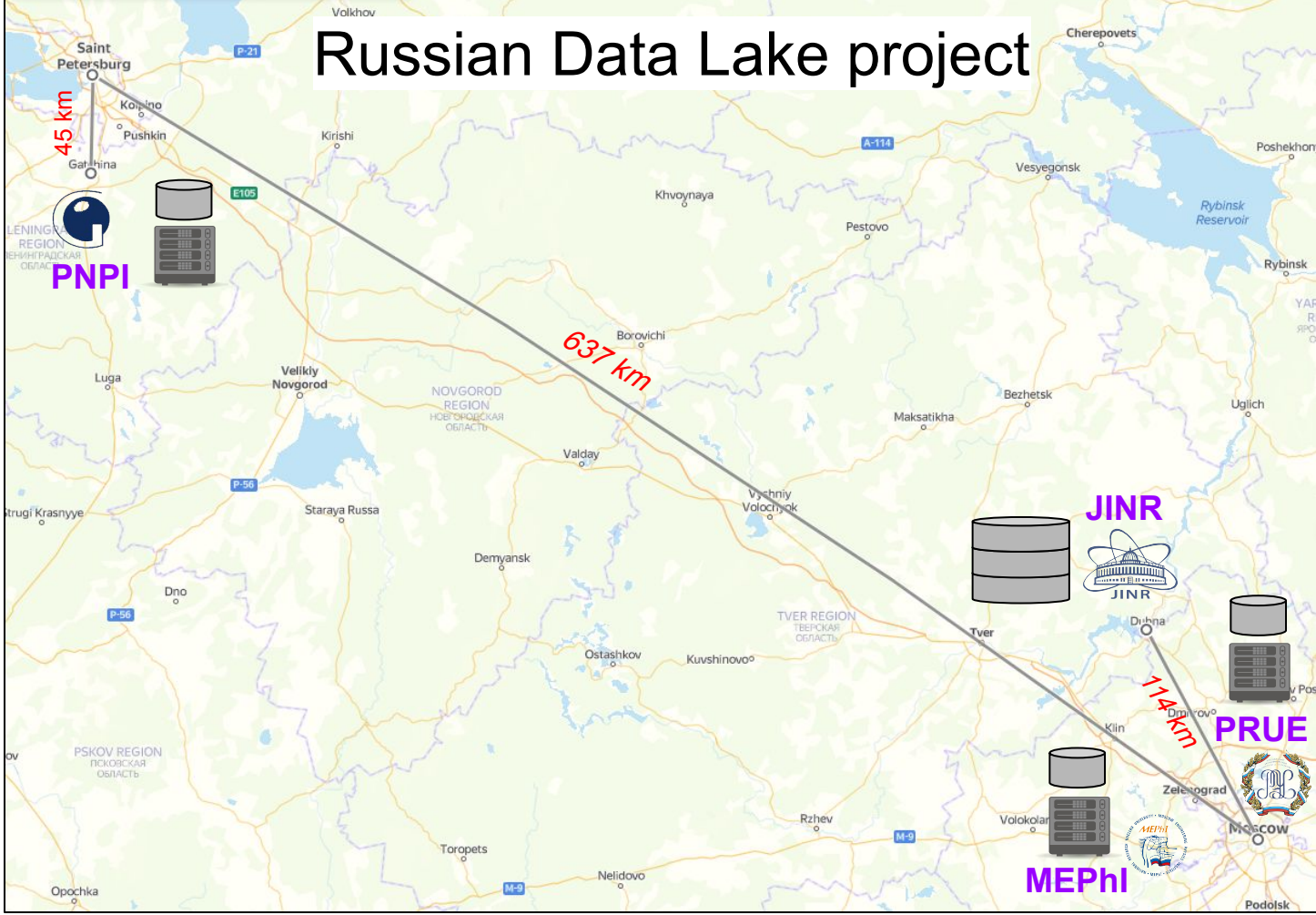
- Data Lake R&D project was launched in Russia as a continuation of the successful Federated Data Storage project and is a part of WLCG DOMA activity
- The prototype is being implemented (see talk by Andrey Kiryanov on Thursday afternoon) and it is targeted at testing of different configurations of data caching and buffering mechanisms using real ATLAS and ALICE experiments payloads
- In order to compare the efficiency of the resource usage between different configurations and control the state of the deployed infrastructure a monitoring system is needed
- ELK-stack and Django framework are used to create monitoring infrastructure of the project

Russian Data Lake project

~800 km of cables from PNPI to JINR

~8 ms latency

10 Gbps



Russian Data Lake monitoring architecture

perfSONAR

BigPanDA monitoring

PanDA

Xrootd logs

Filebeat

dCache

Billing

ARC CE

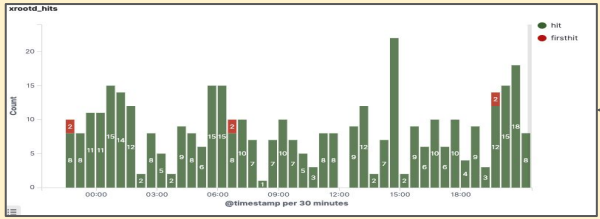
Accounting

JDBC

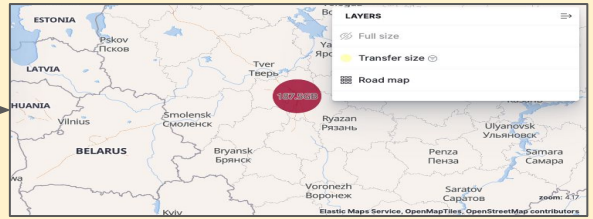
Logstash

ElasticSearch cluster

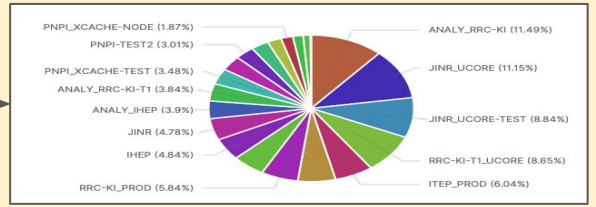
Kibana



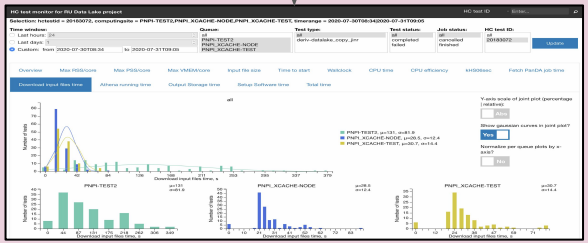
Xrootd monitoring



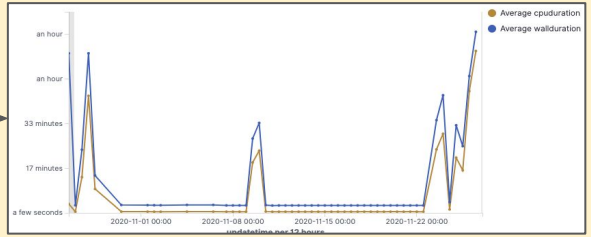
Billing monitoring



Job monitoring



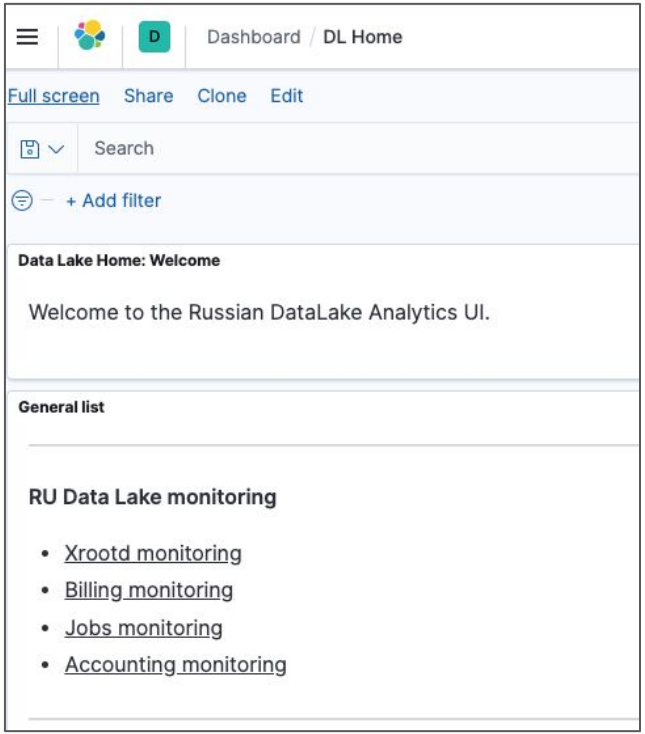
Custom monitoring and analytics



Accounting monitoring

ELK stack and datasources

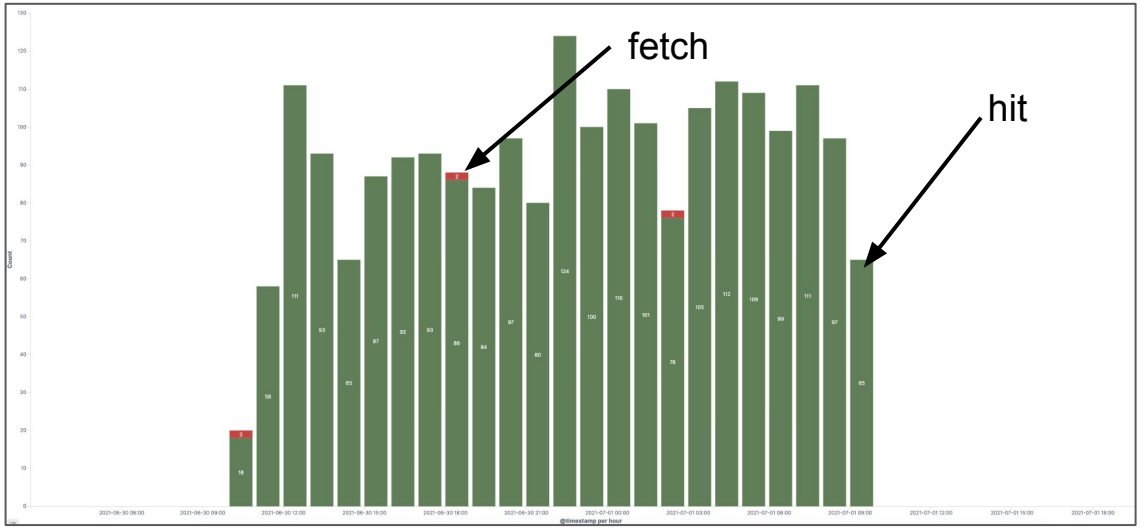
- **ELK stack**
  - **Filebeat** is a special open-source software to collect messages from logs files
  - **Logstash** is a special open-source software for collecting, filtering and normalizing data
  - **ElasticSearch** is a distributed open-source software for storing and searching information
  - **Kibana** is an open-source data visualization plugin for Elasticsearch. It is used for visualization of data from Elasticsearch cluster
- **Datasource**
  - Xrootd logs
  - Billing database
  - Accounting database
  - BIgPanDA API



Xrootd monitoring. Overview

- It allows to monitor the state of the Xcache storage and file accesses in the cache
- Fully based on information from Xcache logs
- Filebeat is installed and configured on all nodes with Xcache to send log messages to Logstash
- Logstash processes these messages using special filters and stores processed information to ElasticSearch cluster
- Data is stored in a dedicated ElasticSearch index
- Xrootd monitoring consist of 9 plots (including map visualization)

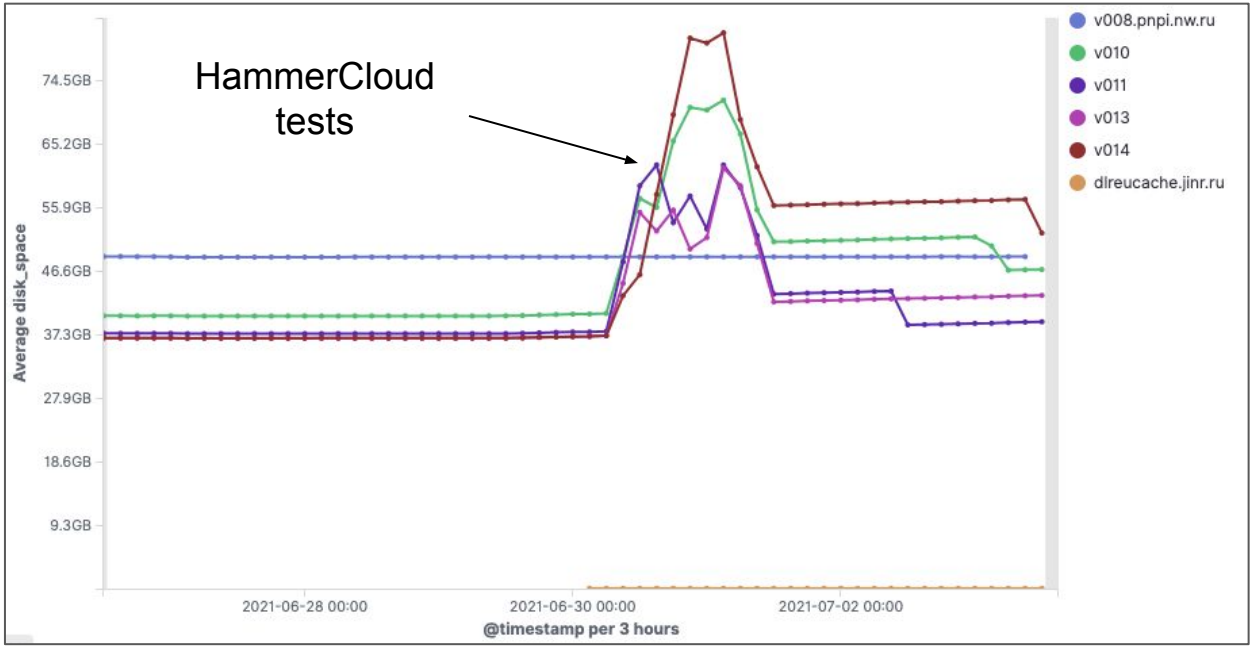
Xrootd monitoring. Accessing XCache files



- Bar plot provides information about number of hits (fetch - first access to the file after copying to the cache) to files in the Xcache distributed by time
- Table contains information about files in Xcache: number of accesses to files, path to file and file name, type of access, hostname, size of files

file.keyword: Descending	hits.keyword: Descending	agent.hostname.keyword: Descending	Count	Average datasize	Sum of datasize
/pnfs/jinr.ru/data/atlas/atlasdatadisk/rucio/data18_13TeV/8e/e4/data18_13TeV.00349263.physics_Main.merge.AOD.f937_m1972._lb0165._0002.1	firsthit	v014	1	4.2GB	4.2GB
/pnfs/jinr.ru/data/atlas/atlasdatadisk/rucio/data18_13TeV/a0/6b/data18_13TeV.00349263.physics_Main.merge.AOD.f937_m1972._lb0161._0003.1	firsthit	v010	1	4.4GB	4.4GB
/pnfs/jinr.ru/data/atlas/atlasdatadisk/rucio/data18_13TeV/8e/e4/data18_13TeV.00349263.physics_Main.merge.AOD.f937_m1972._lb0165._0002.1	firsthit	v011	2	4.2GB	8.4GB
/pnfs/jinr.ru/data/atlas/atlasdatadisk/rucio/data18_13TeV/a0/6b/data18_13TeV.00349263.physics_Main.merge.AOD.f937_m1972._lb0161._0003.1	firsthit	v013	2	4.4GB	8.9GB
/pnfs/jinr.ru/data/atlas/atlasdatadisk/rucio/data18_13TeV/8e/e4/data18_13TeV.00349263.physics_Main.merge.AOD.f937_m1972._lb0165._0002.1	hit	v014	397	4.2GB	1.6TB
/pnfs/jinr.ru/data/atlas/atlasdatadisk/rucio/data18_13TeV/a0/6b/data18_13TeV.00349263.physics_Main.merge.AOD.f937_m1972._lb0161._0003.1	hit	v010	389	4.4GB	1.7TB
/pnfs/jinr.ru/data/atlas/atlasdatadisk/rucio/data18_13TeV/8e/e4/data18_13TeV.00349263.physics_Main.merge.AOD.f937_m1972._lb0165._0002.1	hit	v011	692	4.2GB	2.9TB
/pnfs/jinr.ru/data/atlas/atlasdatadisk/rucio/data18_13TeV/a0/6b/data18_13TeV.00349263.physics_Main.merge.AOD.f937_m1972._lb0161._0003.1	hit	v013	695	4.4GB	3TB

Xrootd monitoring. Xcache disk usage



host.name.keyword: Descending	AVG used disk space	MIN used disk space	MAX used disk space
v008.pnpi.nw.ru	48.7GB	48.7GB	48.8GB
v014	47.2GB	36.8GB	86GB
v010	46.7GB	40GB	80.8GB
v011	41.6GB	37.4GB	69.1GB
v013	41.2GB	36.8GB	68GB
dlreucache.jinr.ru	32.3MB	32.3MB	32.3MB

agent.hostname.keyword: Descending	Sum of removed_data_files	Sum of bytes_to_remove_files	Sum of bytes_to_remove	Sum of bytes_to_remove_c
v011	2	0B	749.2GB	749.2GB
v013	2	0B	605.8GB	605.8GB
v010	1	0B	17.3TB	17.3TB
v014	1	0B	2.6TB	2.6TB
dlreucache.jinr.ru	0	0B	0B	0B
v008.pnpi.nw.ru	0	0B	0B	0B

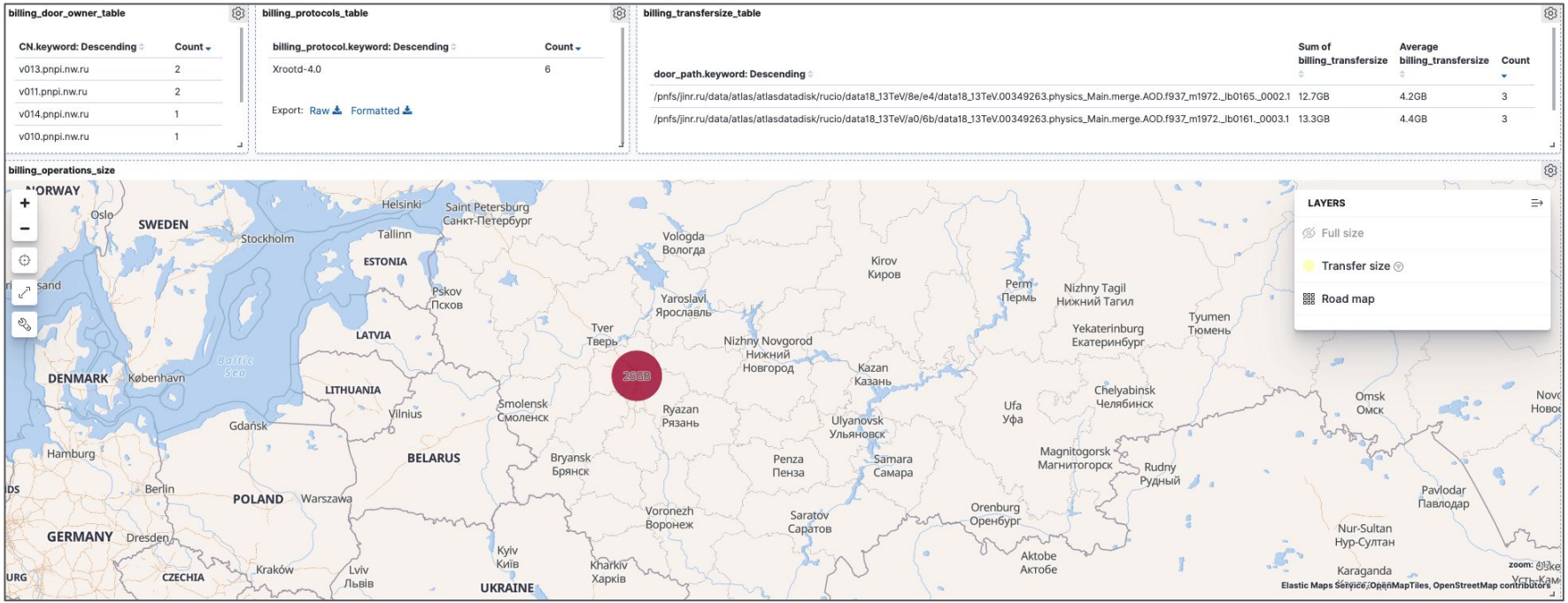
Billing monitoring. Overview

- It allows to get information about operations and requests from *dCache* database
- dCache uses PostgreSQL as storage
- Logstash using JDBC filter extracts and combines data from *billinginfo* (operations) and *doorinfo* (requests) tables in dCache database
- Information from tables is enriched with meta information using the Logstash filter and Geoip2 base
- Meta information includes: geographic location of the initiators of requests for operations: geographic coordinates, city, country, etc
- Combined information is stored in a dedicated ElasticSearch index which contains fields from both tables
- Billing monitoring consists of 11 visualizations (including map visualization)

Billing monitoring. File transfer information

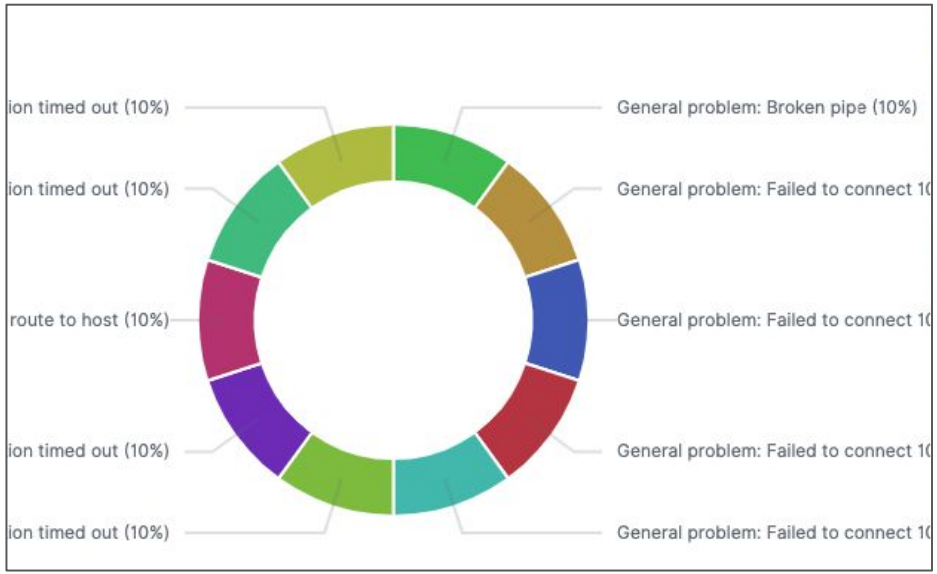
CN.keyword: is one of v010.pnpi.nw.ru, v008.pnpi.nw.ru, v011.pnpi.nw.ru, v013.pnpi.nw.ru, v014.pnpi.nw.ru ×

← Default filter

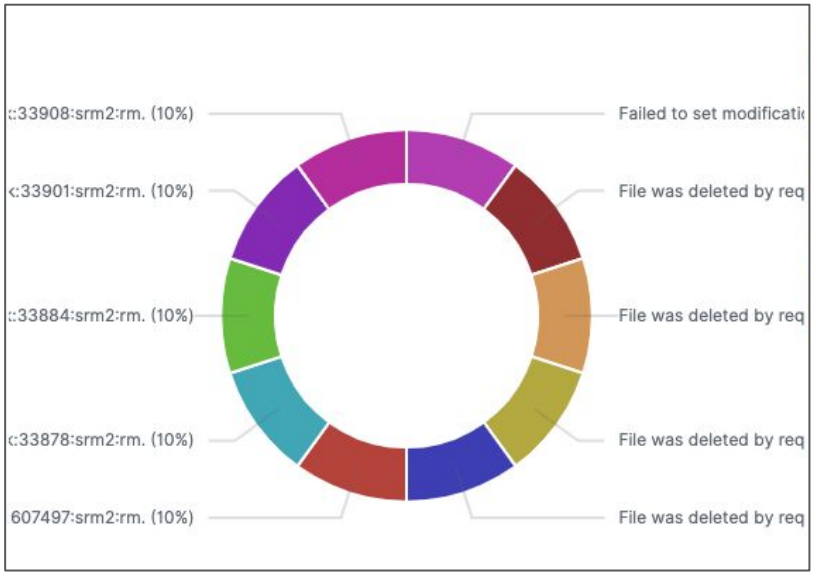


Billing monitoring. Errors information

Billing (operations) errors



Door (requests) errors



Jobs monitoring. Overview

- It provides information about test jobs that are launched using HammerCloud system
- Collecting data using BigPanDA API (the monitoring system for PanDA WMS in ATLAS)
- Logstash every 10 minutes makes HTTP request to this API and gets information about jobs in RU cloud
- The information about the test jobs is stored in the ElasticSearch index
- Kibana dashboard consists of 30 visualizations which allow evaluating the efficiency of jobs execution on computingsites

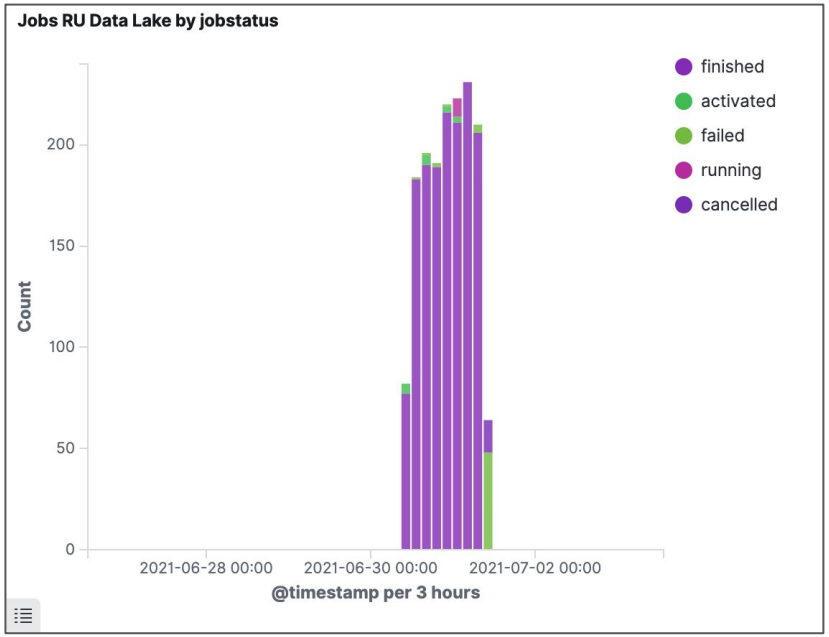
Jobs monitoring. General information

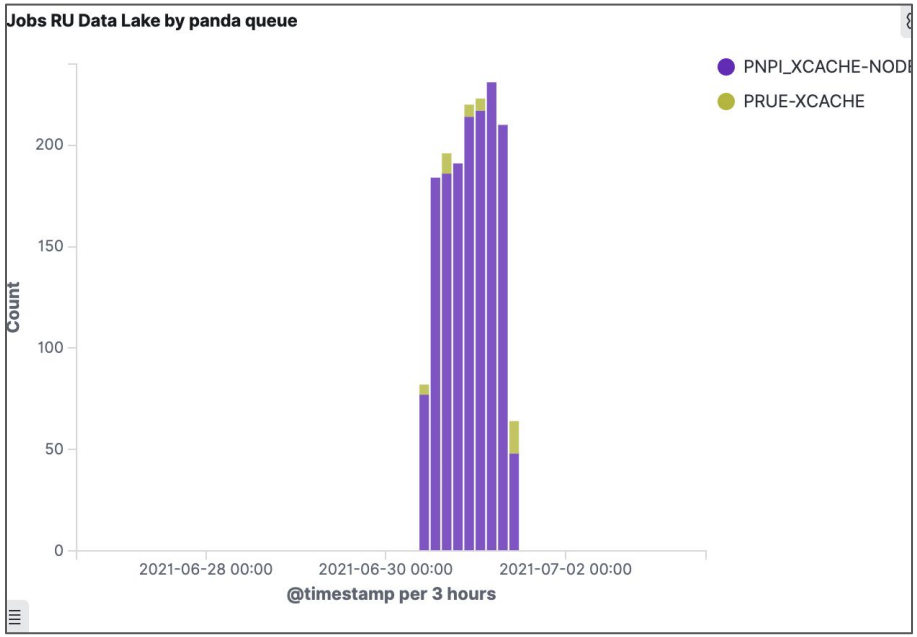
HC Tests

hctestid.keyword: Descending	Count
20204114	1,601

Jobs RU Data Lake by panda queue and processingtype (table)

computingsite.keyword: Descending	processingtype.keyword: Descending	Count
PNPI_XCACHE-NODE	gangarobot-deriv-datalake_copy_jinr	1,558
PRUE-XCACHE	gangarobot-deriv-datalake_copy_jinr	43





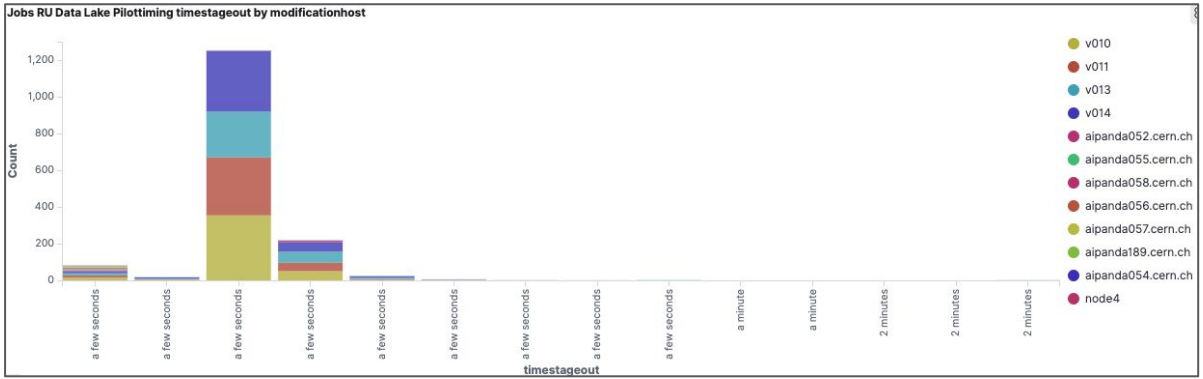
Jobs monitoring. Job processing efficiency

Pilot timings:

- timepayload: Athena running time
- timestageout: time to upload output files
- timegetjob: time to get a payload
- timestagein: time to download input files

cpuconsumption

walltime: time a job was running

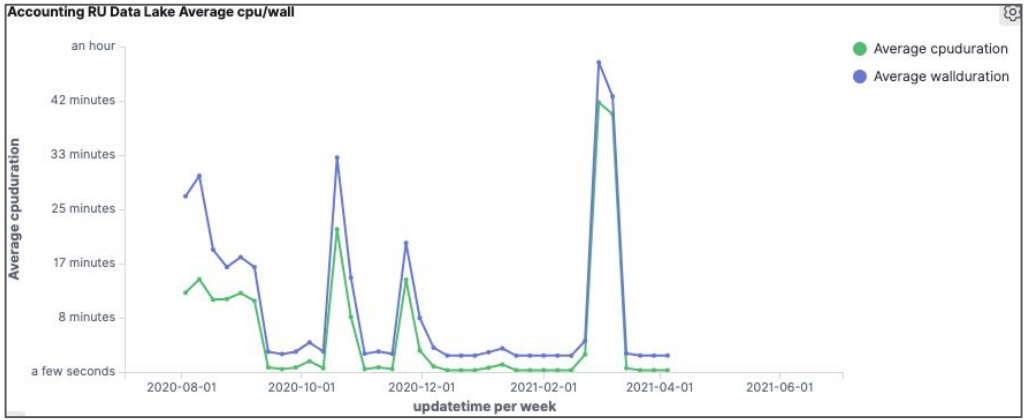


pandaid: Descending	jobstatus.keyword: Descending	batchid.keyword: Descending	computingsite.keyword: Descending	Count	Sum of cpu_eff_per_core_100	Sum of walltime_x_core	Sum of timegetjob	Sum of timepayload	Sum of timestagein	Sum of timestageout	Sum of timetotal_setup
5104857989	finished	576567.v012.pnpi	PNPI_XCACHE-NODE	1	80.7%	5 minutes	a few seconds	5 minutes	a minute	a few seconds	a few seconds
5104392698	finished	574734.v012.pnpi	PNPI_XCACHE-NODE	1	74.1%	6 minutes	a few seconds	6 minutes	a minute	a few seconds	a few seconds
5104869885	finished	576567.v012.pnpi	PNPI_XCACHE-NODE	1	72%	5 minutes	a few seconds	5 minutes	a minute	a few seconds	a few seconds
5104840556	finished	576436.v012.pnpi	PNPI_XCACHE-NODE	1	70.4%	7 minutes	a few seconds	7 minutes	a minute	a few seconds	a few seconds

Accounting monitoring. Overview

- This part of monitoring allows to monitor computational elements, linking the load of storage systems with the load on computational nodes
- Accounting database based on MySQL. The database contains information about accounting tasks, which describe the state of operation of computational elements
- Data from the database is obtained every 10 minutes using the developed JDBC Logstash filter, processed and exported to ElasticSearch index
- Accounting monitoring consists of 3 visualizations

localjobid.keyword: Descending	site.keyword: Descending	Count	Average memoryreal	Average memoryvirtual	Sum of wallduration	Sum of cpuduration
1507808.v012.pnpi	PNPI	1	1.8MB	8.8MB	an hour	28 minutes
1507812.v012.pnpi	PNPI	1	1.8MB	8.8MB	an hour	24 minutes
1508114.v012.pnpi	PNPI	1	2MB	8.6MB	an hour	15 minutes
1507801.v012.pnpi	PNPI	1	1.9MB	8.3MB	an hour	31 minutes
1507749.v012.pnpi	PNPI	1	1.9MB	8.3MB	an hour	35 minutes
1507604.v012.pnpi	PNPI	1	1.9MB	8.2MB	17 minutes	8 minutes
1507649.v012.pnpi	PNPI	1	1.9MB	8.2MB	an hour	37 minutes
1507986.v012.pnpi	PNPI	1	1.8MB	8.2MB	42 minutes	12 minutes



Custom monitoring and analytics

Due to limitations of Kibana visualisation features it was decided to create a custom web application for tests monitoring

Technology stack:

- Django framework
- ElasticSearch
- AngularJS + C3.js + DataTables

It allows a user to get advanced “ready for publication” plots of different test metrics, the most important are:

- Download input files time
- Athena running time
- Upload output files time
- Total time

Each computing queue represents one of RU Data Lake prototype configurations

Summary

- The unified monitoring system based on ELK stack was developed and deployed at PNPI
- It monitors all the components of Russian DataLake prototype including xCache, dCache, Accounting
- 4 dashboards were created in Kibana
- The custom web-application based on Django framework was developed to mitigate the lack of advanced visualisation features in Kibana
- Work to improve monitoring of the project continues

Acknowledgements:

This work was partially funded by the Russian Science Foundation under contract No.19-71-30008 (research is conducted in Plekhanov Russian University of Economics)

Thanks!