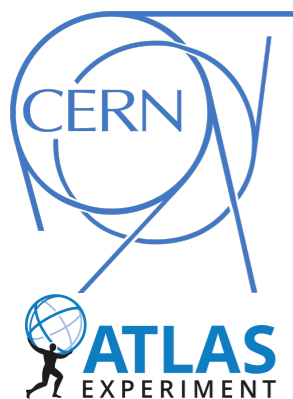


The 8th International Conference  
“Distributed Computing and Grid-technologies in Science and Education”  
JINR, Dubna, Russia, 10-14 September, 2018

# The ATLAS Production System

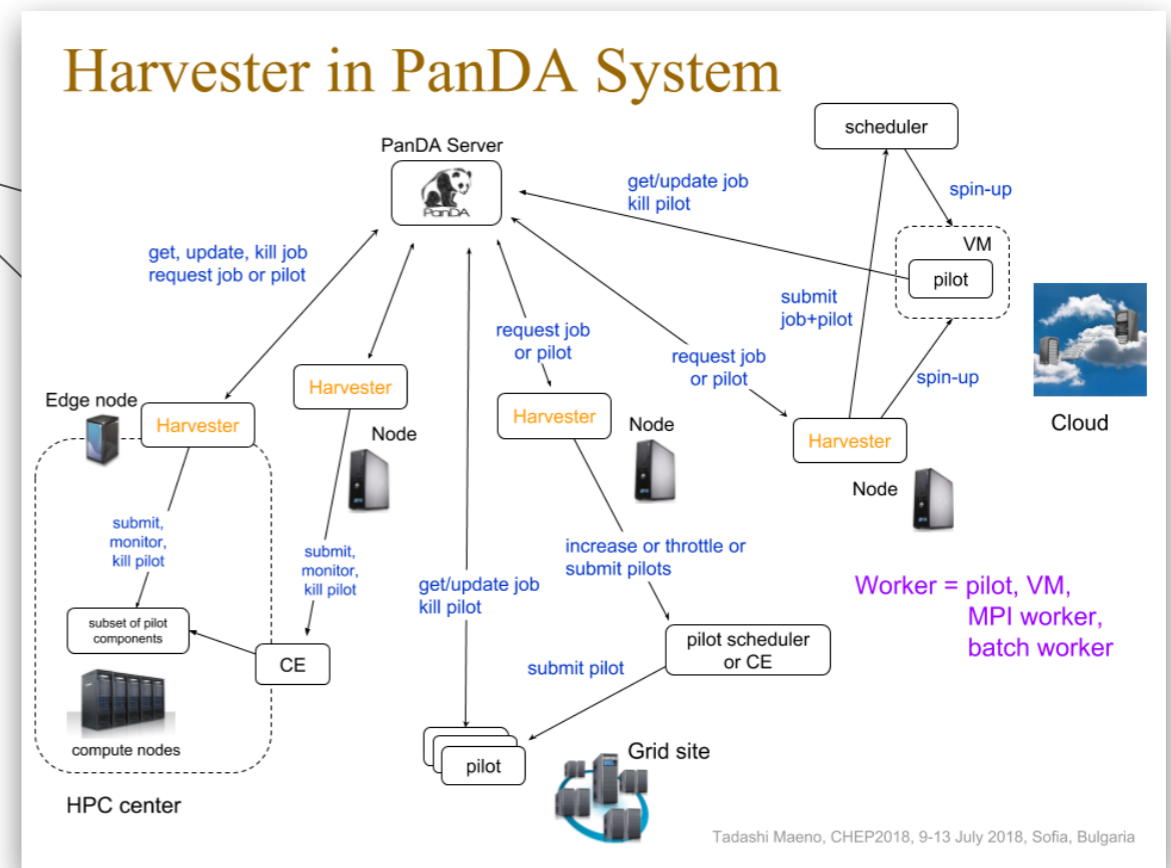
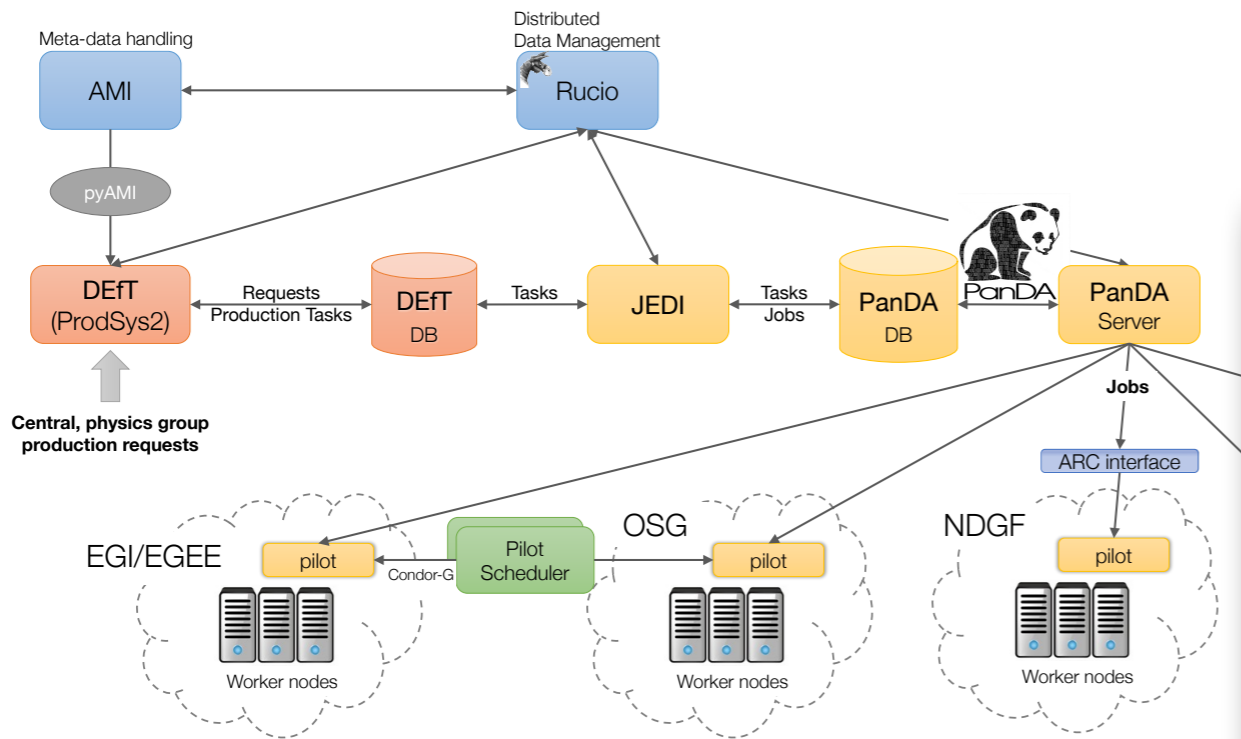
## Predictive Analytics service: an approach for intelligent task analysis

**Mikhail TITOV**, Mikhail BORODIN, Dmitry GOLUBKOV, Alexei KLIMENTOV  
on behalf of the ATLAS Collaboration

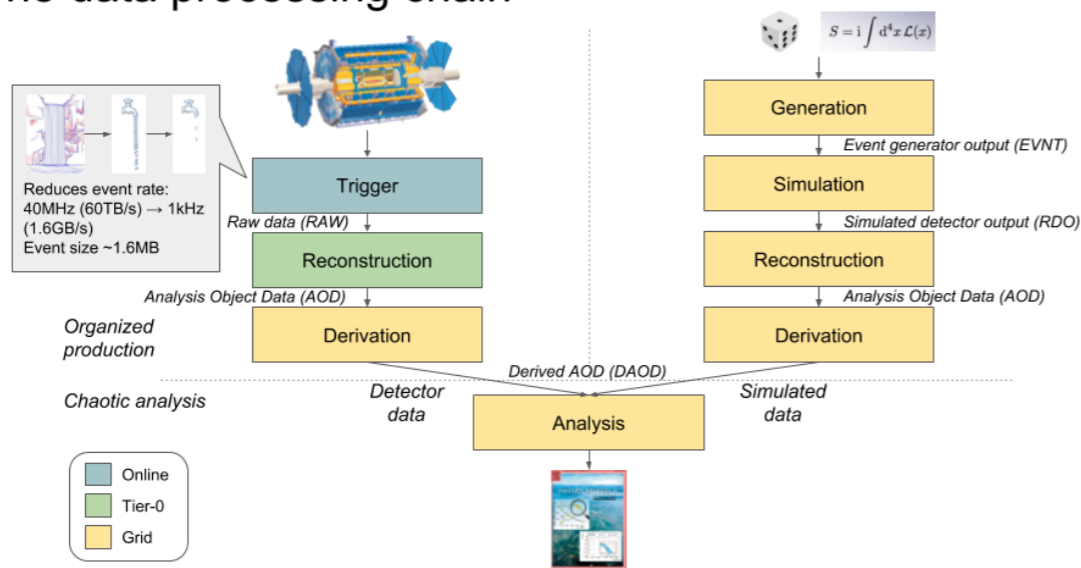


# Introduction

# ATLAS Workflow Management



## The data processing chain



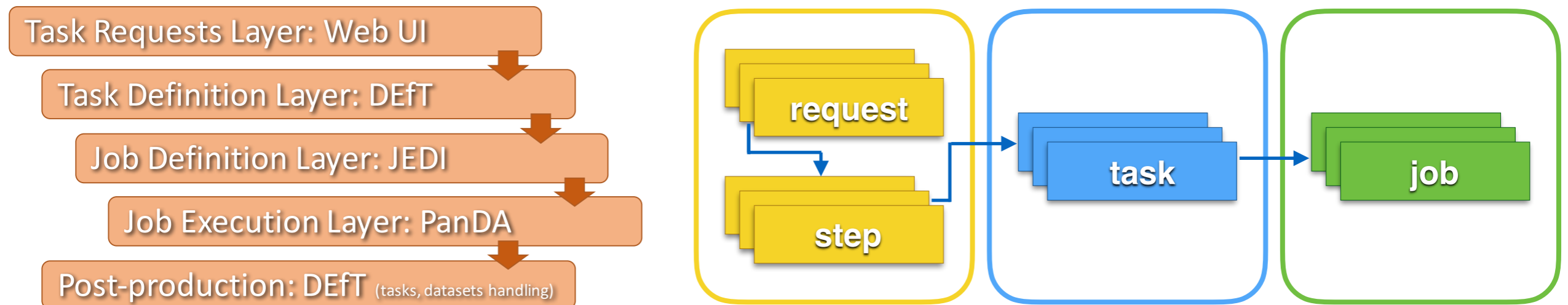
# ATLAS Production System [ProdSys2]

## Database Engine for Tasks (DEfT)

- Formulate the tasks, chains of tasks and task groups (production request)
  - *Task represents a logical grouping of computing jobs, that are responsible of the execution of program/transformation on input files and generate output files*
- Complete with all necessary parameters

## Job Execution and Definition Interface (JEDI)

- Task-level workload management (i.e., brokerage and execution)
- Dynamic job definition and execution (optimization of the resources usage)



Problem statement

# Analytical service focused on tasks

- Analysis of task processing
  - Example: selection and regulation of key task features that affect its processing the most
- Modeling of processed data lifecycles for deep task analysis
  - Example: generate guidelines for particular stage of data processing
- Forecasting processes with focus on data and tasks states as well as on the management system itself
  - Example: detect the source of any potential malfunction

ProdSys2 Predictive Analytics (PA) service

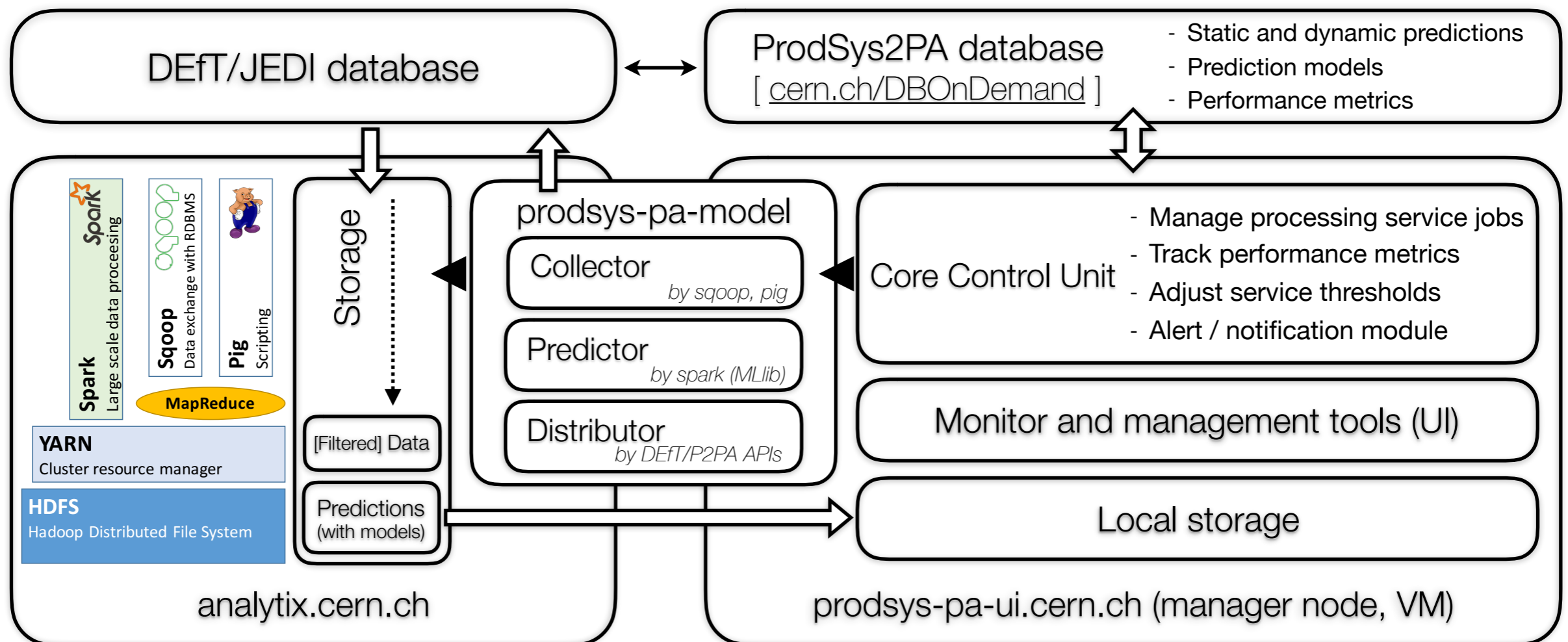
# ProdSys2 PA service

Key components:

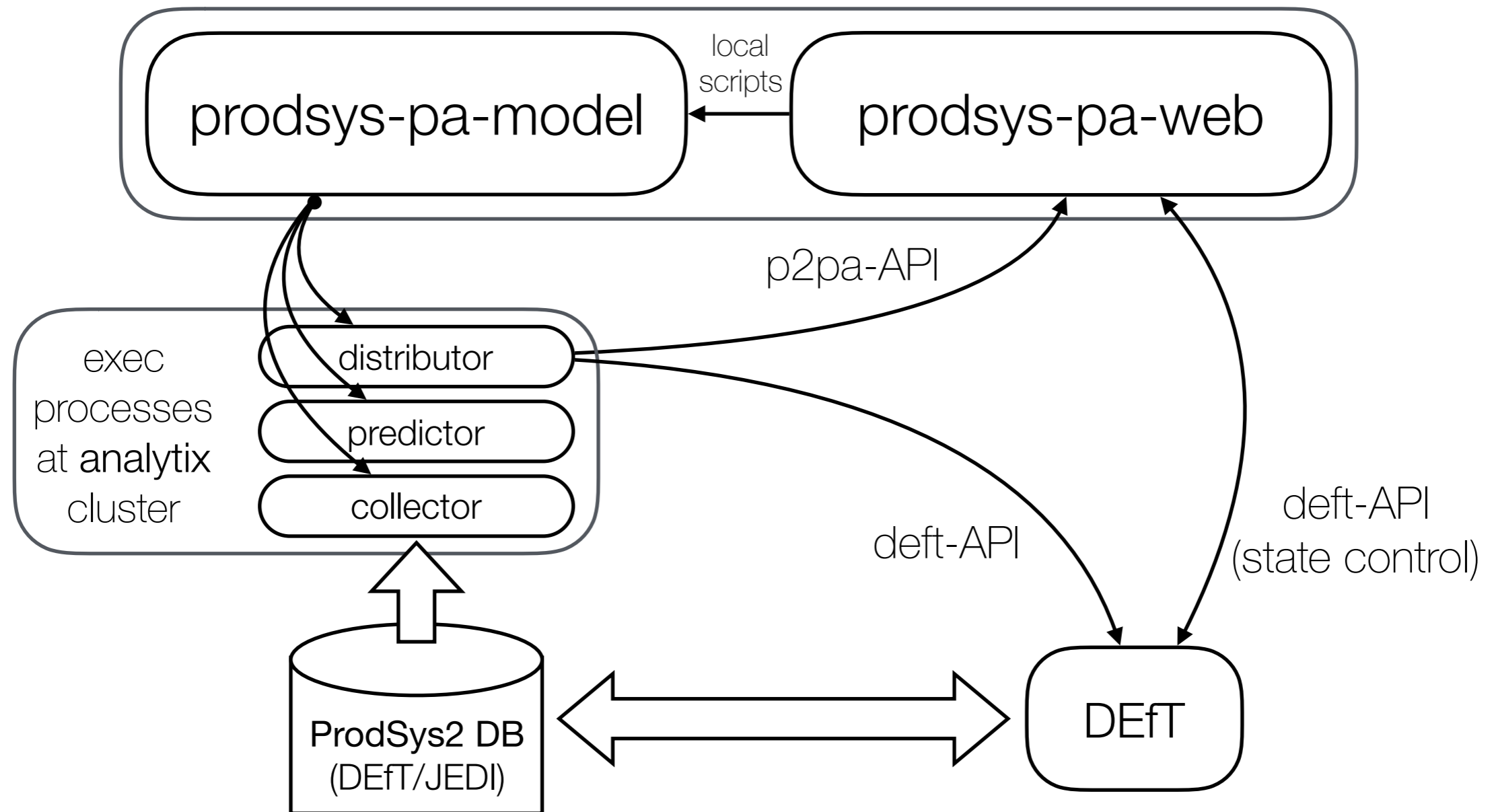
- **Predictive model handling** (*prodsys-pa-model*) - an independent package that is adjusted and integrated into P2PA
  - Data Collection
    - Source - data/mc-tasks from DEFT/JEDI
  - Data Analysis
    - Predictive model creation and usage for the process of TTC (Time-To-Complete) predictions generation
- **Web application** (*prodsys-pa-web*) - a central operation hub
  - Monitor (e.g., exec processes that are historic data of runs, evaluation of estimated durations of task executions)
  - Control (e.g., selection parameters for training and input data collections, method / technology and set of features for prediction process)



# ProdSys2 PA architecture



# ProdSys2 PA packages



# P2PA | Model handling | Analysis processes

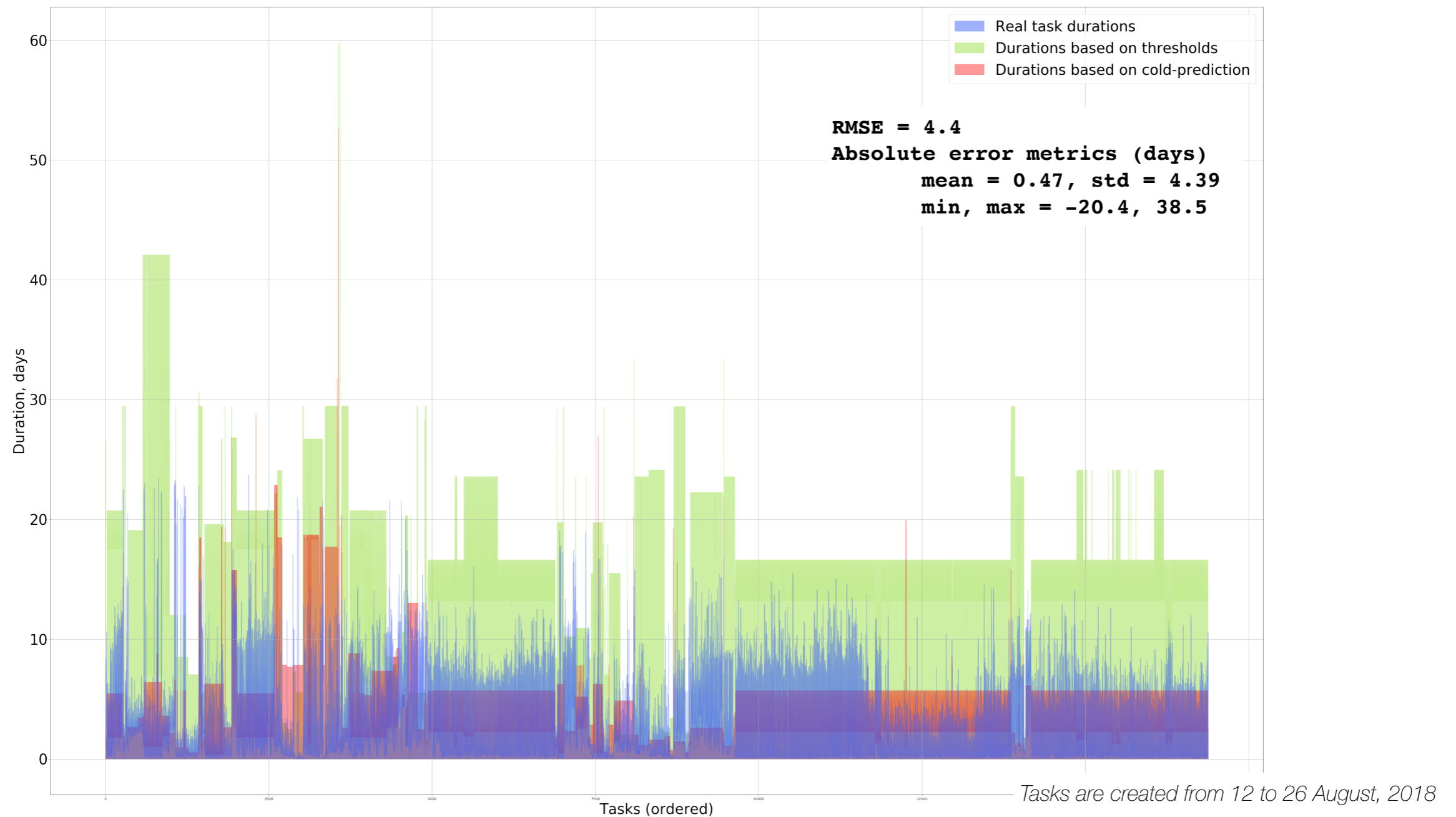
## Threshold definition

- Calculates the upper limit of the duration of tasks execution process in such a way that 95% of all tasks of the corresponding type and for the defined time period (the last 180 days) are executed not longer than the calculated value.
  - Tasks are grouped by `<project.productionStep.workingGroup>`

## Cold-prediction generation

- Estimates task duration during the task formation process (uses descriptive data and created earlier a predictive model).
  - Apache Spark.MLlib - Random Forests regression method

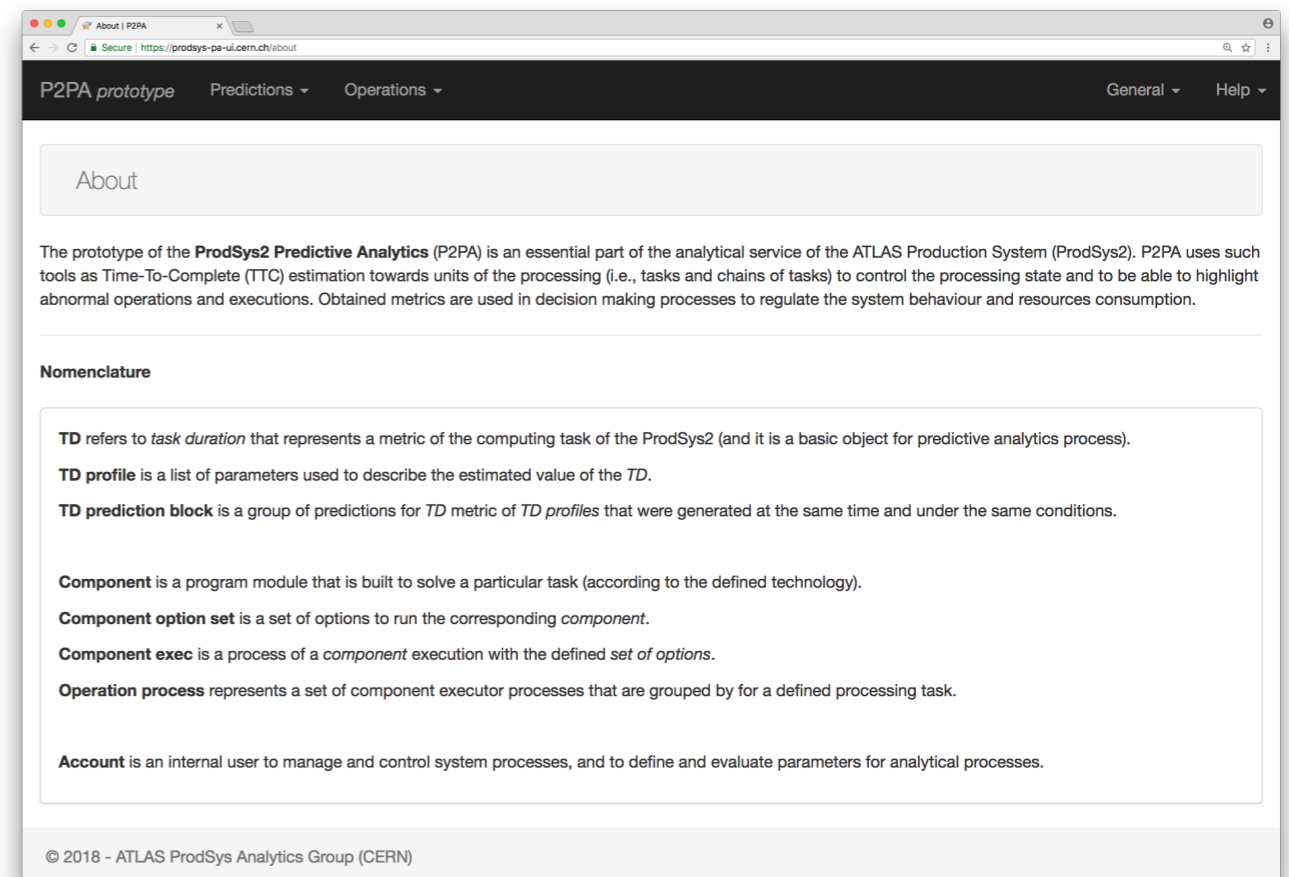
# P2PA | Model handling | Task TTC estimation



# P2PA | Web UI | Infrastructure

## Web application (web UI) setup

- ▶ <https://prodsys-pa-ui.cern.ch> [accessible inside the CERN network]
- ▶ VM by CERN OpenStack IaaS (CC7 – x86\_64)
- ▶ Database On Demand (MySQL)
- ▶ Web server
  - ▶ nginx
  - ▶ gunicorn
- ▶ Web framework:
  - ▶ django
  - ▶ django REST framework
  - ▶ celery (using RabbitMQ)



# P2PA | Web UI | Operation processes

**Operation Processes List**

Category name starts:  Mode: All  Is active: No  2018-08-29 to 2018-09-05

Category	Account	Mode	Is active	Creation date	Timestamp
predictions generation (cold) [task]	P2PA Robot	auto	False	2018-09-05 14:00:02	2018-09-05 14:06:48
predictions generation (cold) [task]	P2PA Robot	auto	False	2018-09-05 02:00:02	2018-09-05 02:05:34
predictions generation (cold) [task]	P2PA Robot	auto	False	2018-09-04 14:00:01	2018-09-04 14:06:35
predictions generation (cold) [task]	P2PA Robot	auto	False	2018-09-04 02:00:01	2018-09-04 02:06:07
predictions generation (cold) [task]	P2PA Robot	auto	False	2018-09-03 14:00:02	2018-09-03 14:05:46
predictions generation (cold) [task]	P2PA Robot	auto	False	2018-09-03 02:00:02	2018-09-03 02:07:43
predictions generation (cold) [task]	P2PA Robot	auto	False	2018-09-02 14:00:01	2018-09-02 14:05:43
model creation (cold) [task]	P2PA Robot	auto	False	2018-09-02 05:00:02	2018-09-02 05:19:49
thresholds definition [task]	P2PA Robot	auto	False	2018-09-02 04:00:01	2018-09-02 04:06:28
predictions generation (cold) [task]	P2PA Robot	auto	False	2018-09-02 02:00:01	2018-09-02 02:12:44
predictions generation (cold) [task]	P2PA Robot	auto	False	2018-09-01 14:00:01	2018-09-01 14:07:06
predictions generation (cold) [task]	P2PA Robot	auto	False	2018-09-01 02:00:01	2018-09-01 02:06:46
predictions generation (cold) [task]	P2PA Robot	auto	False	2018-08-31 14:00:01	2018-08-31 14:05:42
predictions generation (cold) [task]	P2PA Robot	auto	False	2018-08-31 02:00:02	2018-08-31 02:06:23
predictions generation (cold) [task]	P2PA Robot	auto	False	2018-08-30 14:00:02	2018-08-30 14:05:06

**Operation Process**

**Id** 228

**Category** predictions generation (cold) [task]

**Account** p2pa-robot | service

**Mode** auto

**Is active** False

**Creation date** 2018-09-05 02:00:02

**Timestamp** 2018-09-05 02:05:34

**Component Exec[s]**

Component	Order	Options	Is active	Status	Started at	Finished at
p2pamodel.handlers.collector	1	--verbose --force --work-dir new_tasks_cold --config cfg_new_tasks_cold --days 1 --output-type input	False	succeeded	2018-09-05 02:00:02	2018-09-05 02:03:54
p2pamodel.handlers.predictor	2	--verbose --force --work-dir model_cold --data-dir new_tasks_cold --config cfg_cold	False	succeeded	2018-09-05 02:03:54	2018-09-05 02:04:52
p2pamodel.handlers.distributor	3	--verbose --data-dir new_tasks_cold --method set_ttcj_dict	False	succeeded	2018-09-05 02:04:52	2018-09-05 02:05:34

© 2018 - ATLAS ProdSys Analytics Group (CERN)



# P2PA | Web UI | Predictions generation

The image displays two overlapping screenshots of the P2PA Web UI. The background screenshot shows a 'TD Profiles List' table with columns for Task id, TTC predicted, TTC threshold, TTC real, Completeness (%), and Timestamp. The foreground screenshot shows a detailed view of a 'TD Profile' for task id 15308915, with a red arrow pointing from the task id in the list to the 'Task id' field in the detailed view.

**TD Profiles List [ per prediction block: 186 ]**

Task id	TTC predicted	TTC threshold	TTC real	Completeness (%)	Timestamp
15308950	2018-09-06 15:00:15	2018-09-09 11:14:53	-	0.00	2018-09-05 13:42:32
15308942	2018-09-09 18:19:04	2018-10-04 08:59:01	-	93.34	2018-09-05 13:42:32
15308939	2018-09-05 06:50:42	2018-09-09 11:14:31	-	0.00	2018-09-05 13:42:32
15308930	2018-09-05 12:20:34	2018-09-06 19:40:52	-	0.00	2018-09-05 13:42:32
15308921	2018-09-06 14:59:10	2018-09-09 11:13:48	-	0.00	2018-09-05 13:42:32
15308915	2018-09-09 18:18:06	2018-10-04 08:58:03	-	81.00	2018-09-05 13:42:32
15308911	2018-09-05 06:49:38	2018-09-09 11:13:26	-	0.00	2018-09-05 13:42:32
15308903	2018-09-09 18:17:37	2018-10-04 08:57:34	2018-09-05 03:52:22	100.00	2018-09-05 04:36:31
15308899	2018-09-05 12:19:30	2018-09-06 19:39:48	-	0.00	2018-09-05 13:42:32
15308891	2018-09-06 14:58:06	2018-09-09 11:12:44	-	0.00	2018-09-05 13:42:32
15308888	2018-09-09 18:17:08	2018-10-04 08:57:05	2018-09-05 05:32:05	100.00	2018-09-05 07:35:22

**TD Profile**

**Task id** 15308915 [external link]

**TTC predicted** 2018-09-09 18:18:06

**TTC threshold** 2018-10-04 08:58:03

**TTC real** -

**Completeness** 81.00%

**Timestamp** 2018-09-05 13:42:32

**TD Prediction Block**

Num elements	Confidence	MSE	Is active	Creation date	Timestamp
1344	0.00	0.00	True	2018-09-05 02:05:34	2018-09-05 13:42:58

© 2018 - ATLAS ProdSys Analytics Group (CERN)

Conclusion



# Summary

- ProdSys2 Predictive Analytics service is designed to enhance workflow control at the ATLAS Production System and to be able to detect and highlight abnormal operations and executions.
- It is planned to use obtained metrics in decision making processes to regulate the system behaviour and resources consumption.
- The quality of obtained metrics (estimated values of controlled parameters) is constantly improving and new evaluation parameters and metrics will be introduced.

\* If you have any questions/comments/suggestions regarding P2PA, please email: [atlas-adc-prodsys2-analytics@cern.ch](mailto:atlas-adc-prodsys2-analytics@cern.ch)

# Acknowledgements

- This work has been carried out using computing resources of the federal collective usage center Complex for Simulation and Data Processing for Mega-science Facilities at NRC "Kurchatov Institute", <http://ckp.nrcki.ru/>.
- Also, this work was funded in part by the Russian Ministry of Science and Education under contract No. 14.Z50.31.0024.