PF monitoring (P2P-191)

[P2P-273] PF monitoring redesign

Created: 26-09-2023 Updated: 06-05-2024 Resolved: 06-05-2024

Status:	deployed
Project:	P2P / ROM
Components:	None
Affects versions:	None
Fix versions:	v2024_3
Parent:	PF monitoring

Type:	Task	**Priority:**	P5 / Low
Reporter:	Stepan	**Assignee:**	Tomas Korec
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Attachments:	P2P-20240305-233455.png P2P.ipynb P2P_Trendlines.ipynb PF_Trendlines.ipynb PatternFeed-20240305-225434.png PatternFeed.ipynb Preprocessed-20240305-215552.png Preprocessed.ipynb Trendlines.ipynb image (1)-20231127-215008.png image-20231127-215004.png image-20240318-141525.png image-20240318-142308.png
Rank:	2|i00h05:

Description

This ticket is overall ticket for PF/P2P monitoring redesign. It should include redesign of trend lines and parsed values criteria report to be automatized for cases, when we will start receiving more and more domains. Currently monitoring is based only on human input and is prone to human errors.

Comments

Comment by _Stepan [03-10-2023]

After the discussion we have decided that the final results should be a report, which will be sent on daily basi
trendlines report and also will contain errors from quality criteria report, but this will be based on rules, whic

Comment by Tomas Korec [27-11-2023]

Hi Stepan,

after a few first weeks during which I have been pursuing some formalities for the school regarding the proje
detection, and trying to figure out how to implement them to expand the current monitoring, I came with the

- Rework the Adam Malecek's current monitoring on counting metrics for preprocessing, PF, and P2P
- Create a reporting of the results, so Vitek doesn't have to go through the produced JSONs manually e
- Use Hana's trend lines as the base for anomaly detection, so the actual anomalies can be recognized v
- Add anomaly detection on trend lines to report. It will include extensive changes in current monitorin
- Text based metrics were also mentioned, but I haven't pursued it more yet. Can you please tell me abo

There might be overall between anomaly detection using linear regression and trend lines especially for PF a

Tomas

Comment by [Tomas Korec] [27-11-2023]

Hi Stepan,

I started with reworking the Adam's monitoring and creating the actual anomaly detection from it, first on Pr

The current monitoring engine could be divided to 3 parts.

1. Loading historic metrics, current data and counting metrics from it
2. Creating model using the historic data for predicting expected values, applying this model on current
3. Saving metrics of current data between historic data, forming message about results and report the res

To be able to start working on the models, I need to load the data first – part 1. However, the whole monitori
whole engine also isn't possible because it would cause generating new files and saving them in S3 + I would

So, to have up-to-date data accessible, the best way is to replicate part 1 on the side, which I am currently wo

As we will have email report instead of JSON files and Slack channel that need to be checked manually, port

I created [this Figma Board] with the structure of the engine and will update it as progressing, removing the un

Tomas

Comment by Stepan [28-11-2023]

Thanks for the report Tomas Korec . Everything looks fine, let me know the progress in the future.

Also regarding the string metrics - we were discussing it and currently it should wait until the counting metri

Comment by Tomas Korec [18-12-2023]

Hi Stepan,

I continued on the replicating of the part 2 like described in my previous comment. When I got almost to the
will be matched against the prediction of the model trained on the 30 days backward not including the latest

I completely missed this because of lack of experience, I made up my mind I have to do it the way I describe
can see I am wrong, just tell me it please.

However, it wasn't lost time completely since I know how the whole app of the current monitoring work, so

In summary, I left this for now and started working on the model itself.

Tomas

Comment by Stepan [18-12-2023]

Based on our discussion we will divide the implementation into 4 steps:

1. implement/update current prediction model for specific metrics etc., with this updated model rewrite
2. implement new algorithm that will keep the first step the same (counting of monitoring metrics), but
3. enrich counting of monitoring metrics with current metrics presented in trendlines - more metrics bei
4. add monitoring of string metrics (currently presented in quality criteria report)

Comment by Tomas Korec [14-01-2024]

Hi Stepan,

in the end, I didn't do much over Christmas holidays. The first week of the new year was pretty busy in the D

If we want to try to have "model" for each counting metrics (meaning trying to find the most suitable type in
time to this project? Please remember I am working on it beyond my usual responsibilities in the Data Lake a

I researched linear regression a bit deeper to have the understanding of the mechanics beyond the Python fun

Tomorrow, let's talk about trying to find the most suitable model type for each metrics or just have the optim

Tomas

Comment by Tomas Korec [16-01-2024]

Hi Stepan,

based on our yesterdays discussion, I continue in amending/developing the basic anomaly detection model (l
so many of them for this) with its own threshold, we will have only one type/model like we have now, but I w
considering the time I can dedicate to it, we wouldn't get to the anomaly detection on time series and string m

When I am done with this model, I will code the report, and it will be implemented to the existing Adam's ap

Tomas

Comment by Tomas Korec [19-01-2024]

Hi Stepan,

at this moment, based on the fixed (constant) threshold, the decision about anomaly/novelty is done. I guess w
based on the historic data.

Tomas

Comment by Stepan [22-01-2024]

I would suggest if possible to add this information about probability into the anomaly detection as well. It mi

Comment by Tomas Korec [04-02-2024]

Hi Stepan,

while working on counting metrics, I looked at the time series for PF and P2P as well. The reason is that at th
observed automatically by anomaly detection model and receive only the report. It brings the following quest

1/ The monitoring on these time series is currently separated from app that monitors the counting metrics. Th
results is triggered immediately after the monitoring finishes, or do we need to separate counting metrics and
at the same time when counting metrics for PF and P2P data are counted? This is crucial for deciding th

2/ The time series in Jupyter notebook are created and displayed for domains. They are even stored in S3 for
and results of monitoring are presented by metric and domain or dataset. If we have higher number of domain
results of anomaly detection by domain and datasets primarily and then report on what metric the anomaly w
experiences?

Tomas

Comment by Stepan [05-02-2024]

Hi Tomas Korec ,

thanks for these good questions.

Add 1) As you are saying currently, monitoring results and timeseries results are counted separately in differe

Other thing is that metrics in monitoring are probably subset (or should be in the future) of the metrics in tren
metrics and put it into the trendlines part of calculation and store it in the same way and from this calculate th

Or ideally, as this is a new approach into the monitoring, we should create a new approach on how to work/c
(raw/preproc/pf/p2p) and save it somewhere on the S3 where the report will later consume these data.

Hopefully this makes sense to you? I would suggest to somehow clean the currently monitoring and present i

One last thing that I am thinking in regards with this issue - there will probably be two ways of reporting as w
days for PF/P2P and probably also an alerting email if the preproc/PF/P2P calculation was not started and lat

Add 2)

The issue here is that monitoring is presented for preproc/PF/P2P data where on preprocessing you do not cu
all domains etc.

I would suggest for preprocessing to keep it the way as it is and for the PF/P2P to put it into the same "forma
it the way the monitoring is functioning.

Currently PF/P2P monitoring metrics are only subset of the metrics that are used in trendlines monitoring, bu
to keep it the way the monitoring should work so I would suggest to push it towards this way of implementin

Comment by Stepan [05-02-2024]

Next steps after todays discussion

- there is a lot of duplication presented (monitoring/counting metrics, trendlines), but we need to dedup
- deduplicate also list of monitoring metrics - present it in a presentable format, where we can discuss,

- create new version of monitoring - disregard previous versions and create your own partitioning etc., trendlines and saving it into new S3 folders etc.)
- output format and reporting will be discussed later, but here I think we have overall idea how it shoul

Overall this step should be described → deduplicate metrics, put them together, count them on daily basis wi

Comment by Tomas Korec [17-02-2024]

Hi Stepan,

hope you're doing well on Monday 🙂

I am attaching two tables for PF and P2P trend-line metrics with proposed names and description how I unde

PATTER FEED

Proposed name	Position in Google Sheet	
number_of_events_per_domain	A26	how many behaviors (events cour
number_of_events_per_domain_per_dataset	A27	how many behaviors (events cour
number_of_events_per_domian_per_patternId		
count_bot_panelists_per_domain	A28	count of bot_panelists, group by c
count_bot_panelists_per_domain_per_dataset	A28	count of bot_panelists, group by c
bot_events_per_domain	A29	
bot_events_per_domain_per_dataset	A29	
duplicated_search_term_events_per_domain	A30	
events_with_PIDs_per_domain	A31	
events_with_PIDs_per_domain_per_dataset	A31	

PATH-TO-PURCHASE

Proposed name	Position in Google Sheet	
number_of_events_per_domain	A39	is it event_per_prod
number_of_events_per_domain_per_dataset	A40	is it event_per_prod
number_of_events_per_domain_per_patternId		
bot_panelists_per_domain	A46	
bot_panelists_per_domain_per_dataset	A46	
bot_events_per_domain	A47	is it bot_events_per_
bot_events_per_domain_per_dataset	A47	is it bot_events_per_
events_by_product_notnull_PID_per_domain	A48	
events_by_product_notnull_deterministic_PID_per_domain	A49	
products_in_catalog_per_domain	A50	

events_by_product_in_catalog_per_domain	A51	

What about these, do we want them too? panelists, user_session, products

---

I put those metrics to the Augmented Dickey-Fuller test (plotted ACF/PACF too, same result) and realized th
However, not all data is stationary, some showed up to be non-stationary and it even comes from the same m

To handle this, stationary and non-stationary time series for same metrics, I propose using ARIMA model. H
would make model to fit all the time series. I am therefore thinking about using auto ARIMA that might not b
will achieve the better overall results.

[P2P_Trendlines.ipynb]
[PF_Trendlines.ipynb]

Tomas

Comment by [Tomas Korec] [18-02-2024]

Hi Stepan,

don't miss the previous comment, this's the second one this weekend

Let's have a talk regarding the report how we (you, DevOps) want it to work, look like, and be structure: by r

I am attaching the link to prototype. I just created it based on my understanding without any consultation, we

https://www.figma.com/proto/Q9rIPyNw4lpqazZhHCKaCO/Untitled?type=design&node-id=2-4&t=gVxxvI

Cheers

Tomas

Comment by Stepan [19-02-2024]

Hi [Tomas Korec] ,

I have checked the metrics and there are few points I wanted to make:

- number_of_events_per_domain_per_behavior metric in the table does not correspond to the descripti
  number_of_events_per_domain_per_dataset.
- metric above number_of_events_per_domain_per_behavior should be number_of_events_per_domai
  to have it over each of the dataset. Please add it into the table for PF metrics.
- both points above are also presented in P2P metrics and need to be presented there as well so please c
- For both PF/P2P metrics I am missing metrics for both rows and panelists counts per datasets and per
- for P2P I am missing metrics count_rows_per_metadatalink_per_domain, count_rows_per_pidsource
- lastly there are some important missing metrics in P2P table - count_deterministic_pids_per_domain,
  count_behaviors_with_extracted_pids_per_domain. Can you please add these into the table as well?
- Also can you please add comments into P2P table so I am sure that you understand the metrics and w
- Lastly regarding comments in P2P - bot_events_per_product and event_per_product are something li

Please revise these comments and update the comment with them in mind. Also let me know if something is

After this will be done we can go through the report and finalize the structure etc. However, for now it looks

Thanks!

Comment by Tomas Korec [19-02-2024]

Hi Stepan,

All the bullet points here are answering your bullet points in the same order:

- You are right in the first comment, thank you. 🙂 Fixed
- So we will add metric number_of_events_per_domain_per_patternId, right?
- Done
- Yeah, let's talk about this on call. I focused on metrics available in calculated trendlines metrics files.
- count_rows_per_metadatalink_per_domain & count_rows_per_pidsource_per_domain P2P metrics a
- The same case, these are also part of counting metrics monitoring.
- Yes please

All the metrics mentioned in the tables are meant for the trendlines/time series. I know we talked about moni will not miss the ones you mentioned since they're in counting metrics monitoring already, and if desired, ad too chaotic and extensive. 🙂

Tomas

Comment by Stepan [20-02-2024]

Hi Tomas Korec thanks for the quick answers.

With your comments:

- number_of_events_per_domain_per_patternId - yes this metric is needed and I agree with the naming
- the easiest way how to describe the differences between bot_events_per_domain, bot_events_per_dor
  - bot_events_per_domain - how many bot events are there per domain - are there specifically hi
  - bot_events_per_domain_per_dataset- how many bot events are there per domain/dataset - are
  - bot_events_per_product - how many bot events are there per product? (ratio of bot events/cou

Sorry about the comments regarding the missing metrics. I did not realized and did not know how you wante not forget about them in the future. With that I agree with the proposed way of implementation.

Comment by Stepan [20-02-2024]

I have checked the proposed monitoring report in the Figma and after discussion I have few points:

- the partitioning by metric and then domain/dataset is ok
- there should be the whole historical trendline so we can see the trend, but for current data also model
- verdicts should not be presented as only not ok verdicts will be displayed
- reports should be send for each of the platform (preproc, pf, p2p)
- reports should be send for each day, not any chunks

Comment by Tomas Korec [26-02-2024]

Hi Stepan,

the time series here, even within one metric, has different properties in terms of stationarity (presence of tren
developed every time we start observing new domain, but its number would increase exponentially.

We need to think from the perspective of anomaly detection engine, that is to work for all domains and metri
Therefore, we need to follow the same approach as in the case of counting metrics monitoring, having one al

From that reason, I chose to use Hyndman and Khandakar algorithm (auto ARIMA) that can differentiate tim
find the most suitable attributes. *By the next weeks, I will provide you some comparisons between fitting of m*

However, in order the Hyndman and Khandakar algorithm to work well in the anomaly detection engine, bef
to perform by itself, test for probability distribution (BoxCox) due to which we can decide whether to transfo
data, the final prediction was badly influenced by some sudden changes in data instead of them not being we

I tried to research if some methods for finding the optimal thresholds exists, but how we can, based on the his
within we have percentual probability of forecast correctness, i.e., standard deviation) and editing it later whi

Plan for this week:

Actually, as I was exploring the data's properties (applying various tests and plotting the data) and trying mo
metrics and time series within them + I want to provide you some comparison between fitted values of Hynd

Tomas

Comment by Tomas Korec [06-03-2024]

Hi Stepan,

Time Series

For time series that seems to have patterns in them (can be decomposed) or the value is to be constant, e.g. 0,
Sometimes, it even performs better. (I used two months data for model training).

For time series for which Hyndman and Khandakar algorithm decided only 1 differentiation based on KPSS
and 'q' attributes based on visual analysis of ACF/PACF plots than Hyndman and Khandakar algorithm has p

I believe there's a way how to recognize significant lags in ACF/PACF plots programatically which could all
theory.

For metrics whose granularity is for domain max, no for dataset, for behavior metrics, I tried founded models
within one domain but for different dataset or behavior. It proves the previous suggestion that for time series

The attached tables are only for illustration how auto ARIMA (Hyndman and Khandakar algorithm) perform

Patern Feed Time Series

Metric	Manually found model	AICc	MSE	auto ARIMA model	AICc
bot_panelists_per_domain	(1,0,1)	378.854	6.156	(1,0,0)	377.161
bot_events_per_domain	(2,0,2)	-82.502	0.037	(1,0,1)	-85.357

duplicated_search_term_events	(1,0,3)	-554.108	4.919	(1,1,1)	-548.68
events_with_pids_per_domain	(6,2,4)	-1453.78	1.489	(0,1,2)	-1552.961

P2P Time Series

Metric	Manually found model	AICc	MSE	auto ARIMA mo
number_of_events_per_domain	(2,2,3)	-1562.068	1.31	(0,1,0)
bot_panelists_per_domain	(1,0,1)	375.368	5.885	(1,0,0)
bot_events_per_domain	(1,0,1)	223.32	1.751	(1,0,0)
events_by_product_notnull_pid	(2,2,4)	-1486.06	1.528	(0,1,0)
events_by_product_notnull_deterministic_pid	(6,2,3)	-1565.39	2.954	(0,1,0)
products_in_catalog	(2,2,3)	-1543.61	2.693	(0,1,0)
events_by_product_in_catalog	(2,2,4)	-1517.07	1.755	(0,1,0)

Are we good to start building the app around these models/algorithms? Within the rest of this and next week,
reporting. However, I will need you to remind me when the metrics are counted from data, because for count

Counting Metrics Models

I compared the other existing robust regression model types – Trimmed Mean, Andrew Wave, Tukey Biweig
Wave, and Tukey Biweight performs all very similarly, but more importantly, they all outperformed HuberT
between these three model types with just a small difference in performance.

The models in the following tables are sorted by highest to lowest performance by lowest to highest MAE.

-
    - The reason why I used MSE for time series and MAE for counting metrics is following:

Model time series are transformed, so the outliers are to be smoothened and should be within the normal dist

Counting metrics data is unchanged and MAE is more robust to outliers. MSE on the other hand enforces a h

Tomas

P.S. I am also attaching the jupyter notebook I used for looking for the models

[Preprocessed.ipynb]
[PatternFeed.ipynb]
[P2P.ipynb]
[Trendlines.ipynb]

Tomas

Comment by [Tomas Korec] [18-03-2024]

Hi Stepan,

I am sharing with you the steps I did while I was looking for the ARIMA models manually.

Step 1: Untransformed data doesn’t have to be normally distributed because of outliers, so I used Shapiro-W
of any subsequent statistical tests or models.

Step 2: I checked whether data is stationary (free of trend and seasonality; *mean*, *variance*, and *autocovarian*
it was stationary. ARIMA models needs data to be stationary.

Step 3: In step 3, I looked for ARIMA model’s attributes p, d, q.

ARIMA models combine two models and 1 method. Two models are Auto Regression(AR) and Moving Av

Auto Regression model presents the value of a variable at time *t* as a linear combination of its past values, plu
current value of a variable on its past values. p value can be found via PACF (partial autocorrelation function

As I can find p values via PACF plot, I can find q values via ACF plot which tells us how much moving aver
trend and seasonality; *mean*, *variance*, and *autocovariance* of the series are time invariant). In other words, th
number of lags that are significantly out of the defined boundaries.

Parameter d defines how many times I had to differentiate a particular time series to make it stationary.

Step 4: After I decided ARIMA model's attributes, I reviewed AICc value and by changing ARIMA model's
defined as AIC = 2K – 2_ln_(L). Therefore it might be often negative, but it doesn't influence anything. The

Step 5: I plotted residual component and ACF of the residuals and performed Portmanteau residuals test to v

Step 6: On the testing set, I counted Mean Square Error from the model's predicted values. At this point, I co

I hope this provides a bit more insight in how I looked for ARIMA models manually. If you have other quest

Cheers

Tomas

Comment by Tomas Korec [22-04-2024]

Hi Stepan,

hope you are doing well after your busy last month.

The current state is like this:

After implementing part of the anomaly detection for time series metrics (classes for data loading, model, an
are structured and how they will need to be parsed for reporting purposes, so I could create results of time se
merged easily.

This week, I should be able to finish most of the reporting part for all 3 instances, preprocessed, pattern feed,

However, I am developing it separated from AWS, so the code will require some changes before deployment

---

Can we then have a talk regarding type of results and their prioritization? Or maybe I could talk to a guy wh
(when tested value exceeds our expectation) or negative difference. Negative difference is pretty reasonable
the case. I just want to know what has been taken into consideration and on what results we have been focuse
am asking so my report is not to too overwhelming, especially if it will contain results from time series moni
monitoring will decide NOK result.

Tomas

Comment by Stepan [23-04-2024]

Hi Tomas Korec thanks for the updates, I will try to be more focused on this project from now on.

I would like to see the results first before trying to discuss the verdicts with person that is looking into it.

It would be great to prepare some testing version of the report so we can discuss what should be presented et

Comment by Tomas Korec [24-04-2024]

Hi Stepan,

yes, of course, sorting and including results can be done afterwards. Let’s have another Sync call in two wee

Tomas