Predictive Analytics in Database Activity Monitoring

Context22 OpenSource Project Update #1

Release of the FIRST Excessive Extractions of Sensitive Data Detection Module by Predictive Analytics in Database Activity Monitoring

Context22 OpenSource Project Update #1

ALL MAJOR DATA LEAKS START AT THE DATABASES

Taking the Data Leak issue at its source/origin : the Database

Catching Data Leak with DLP is important, crucial, but it’s late and it’s putting an unnecessary burden on it.

By regulations, companies are required to monitor their Databases. They have the infrastructure to monitor the Database traffic, in particular what gets OUT of the Databases. The more you detect excessive extractions at that level, the more companies can : 1/ tighten their Databases against leaks 2/ catch them early and provide early warning to DLP “Large extraction in progress”/”Excessive Extraction in progress”

Process

Enrich Traffic

Build Daily TS by SQL

Locate the Top 10 SQLs or so, leveraging Pareto’s law of 80/20, 20% of the SQLs extract 80% of the Sensitive Data

Compute Predictions per day of week

For 80% of the SQLs not monitored under Predictive Analysis, take as Threshold the amount extracted by the smallest of the 20% monitored

Specific treatment of the new SQLs including the SQL Injection

The Necessity Of Bringing In Data Analysis Skills

First Release : The Core Computation

LSTM implementation is on pause for now

Here is a more manageable option : Holt-Winters, in the Triple Exponential Smoothing version with Multiplicative seasonality

Exponential Smoothing

Benefit of Predictive Analytics (PA) : Heavy computation to generate the Pred is done ahead of time

In RT, check of the Pred vs. Actual is light

Demo

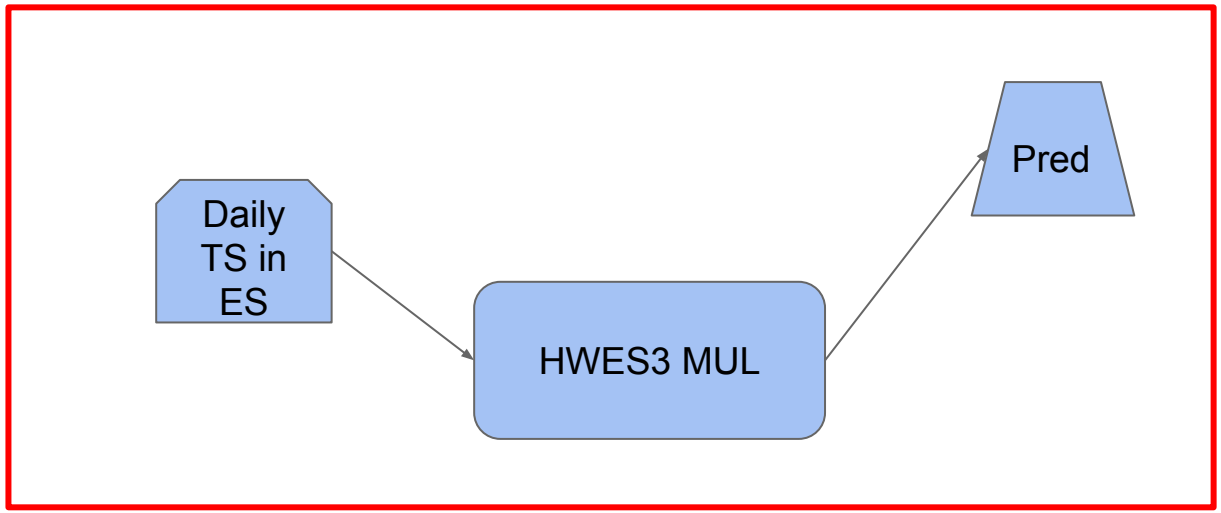
Demo Outline

- Presentation of the TS (Kibana)
- Presentation of the Program :
  - Based on Etqad Khan Python code
  - With a different TS, daily instead of monthly, taken I don't remember where
- Run the program
- Presentation of the Test version to come (Cloud)

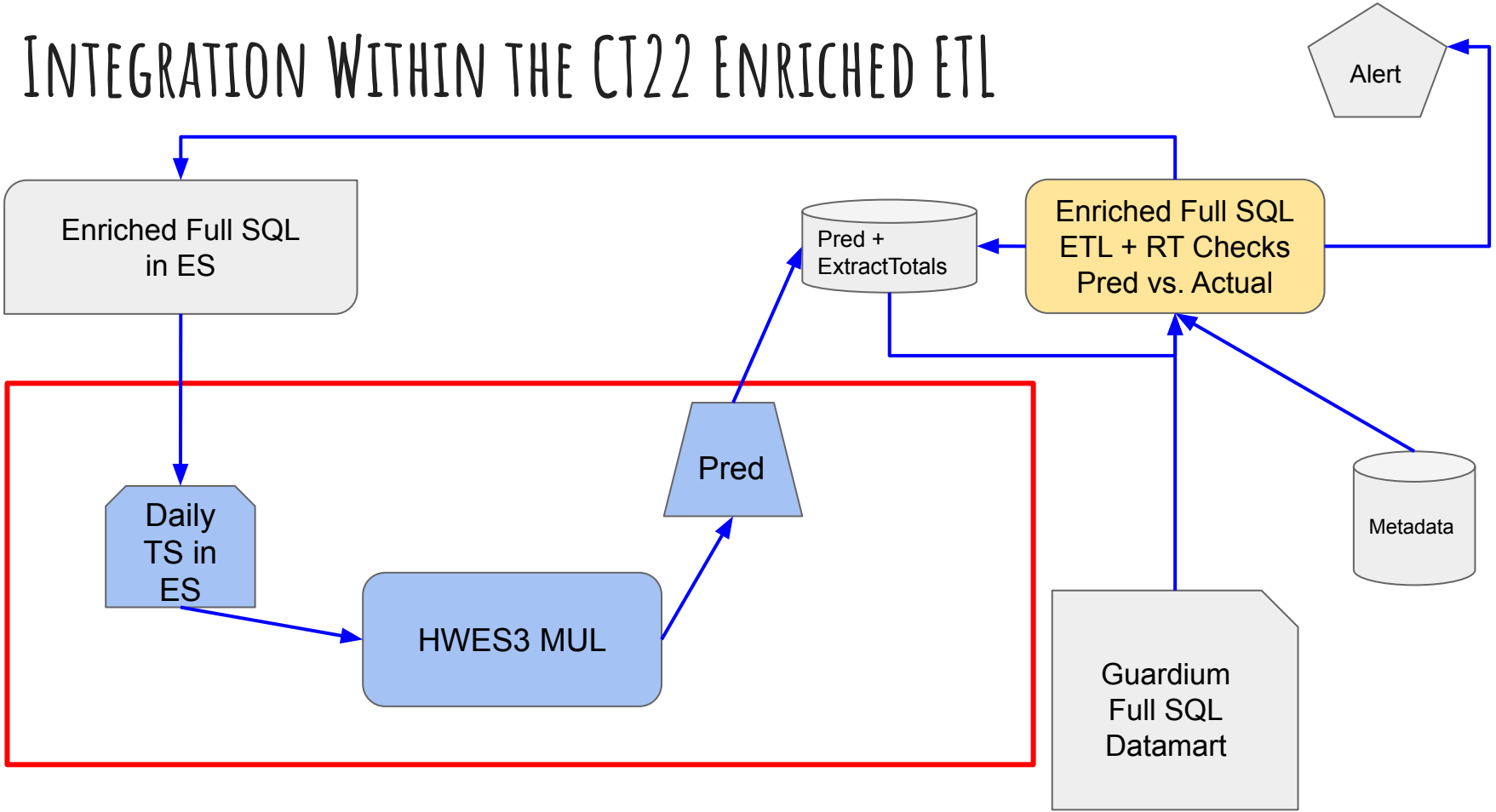
ES Index in input

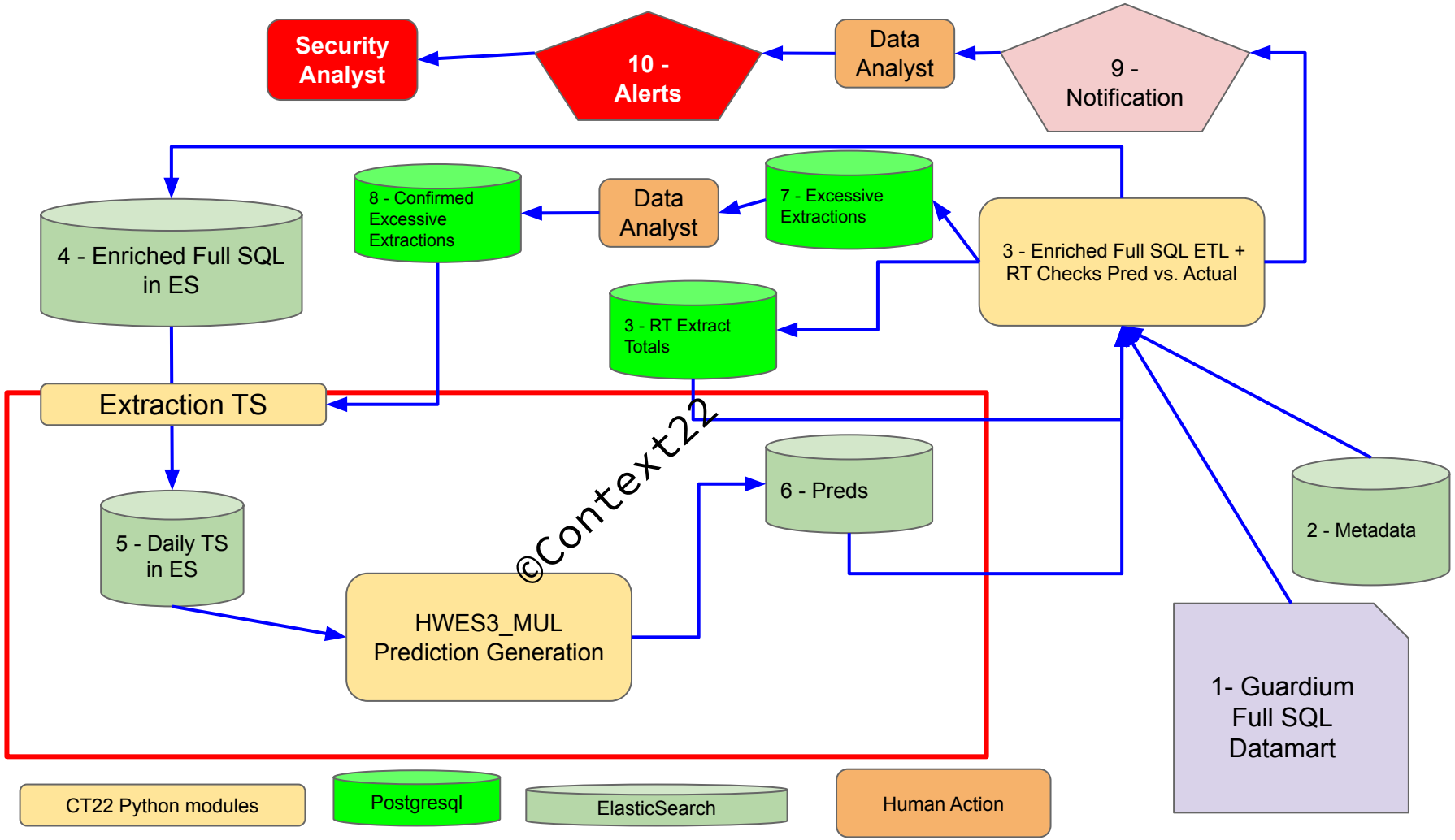
Next Steps

Current Set-up : Core Module

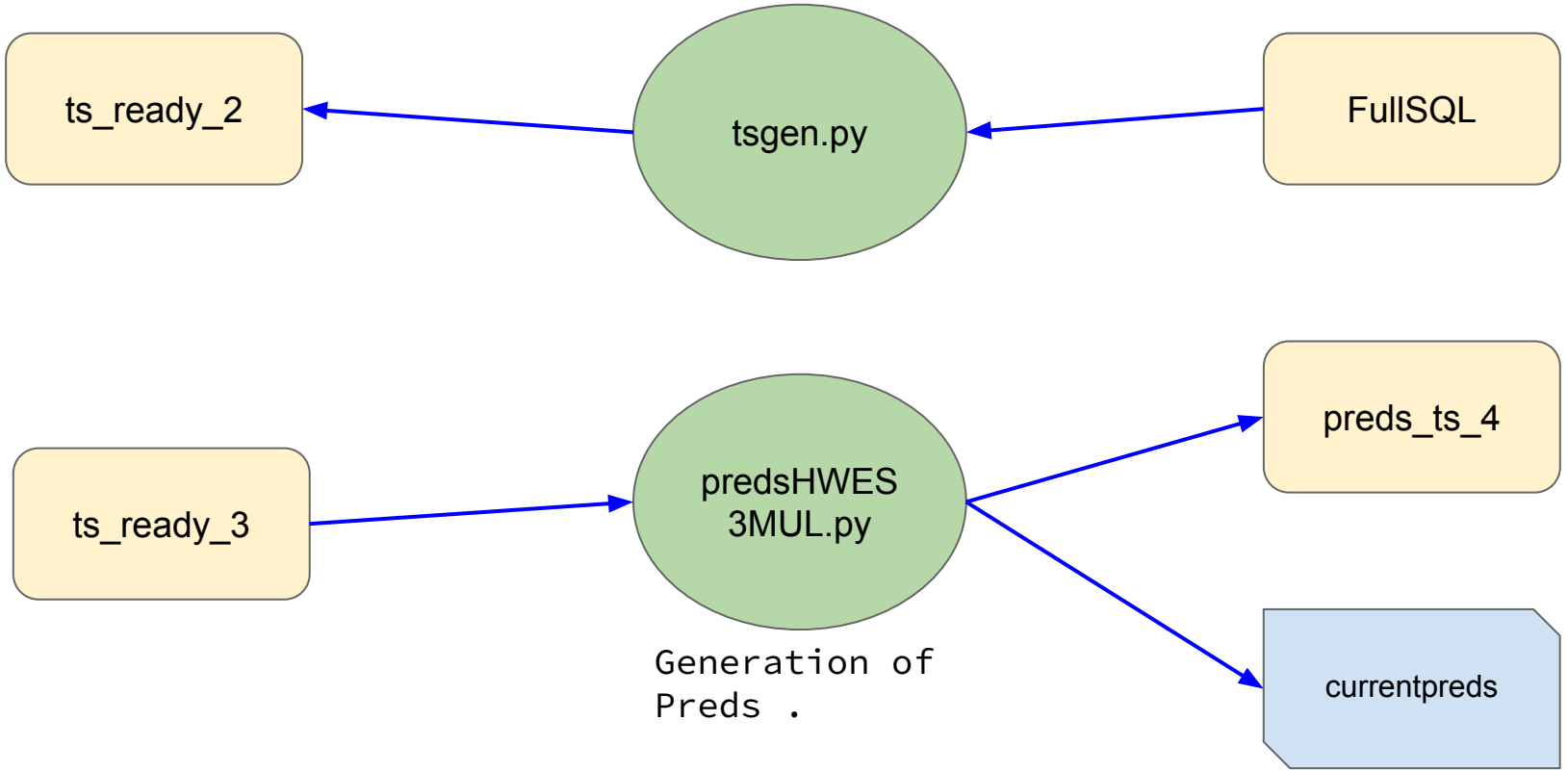


Integration Within the CT22 Enriched ETL




Security Analyst
10 - Alerts
Data Analyst
9 - Notification
4 - Enriched Full SQL in ES
8 - Confirmed Excessive Extractions
Data Analyst
7 - Excessive Extractions
3 - Enriched Full SQL ETL + RT Checks Pred vs. Actual
3 - RT Extract Totals
Extraction TS
©Context22
6 - Preds
2 - Metadata
5 - Daily TS in ES
HWES3_MUL Prediction Generation
1- Guardium Full SQL Datamart
CT22 Python modules
Postgresql
ElasticSearch
Human Action

Preds Gen Data Architecture and Tests



Generation of
Preds .

Preds in ES vs. Preds in PostGreSQL

Preds in ES

1 record for ALL preds types for a given day and a given SQL

It's needed for use in Kibana and format is easy to change since it's NoSQL

Preds in PostGreSql

1 record per Type of Preds and per day and SQL

That way I can add new type of pred without having to change the format which is very difficult in SQL.

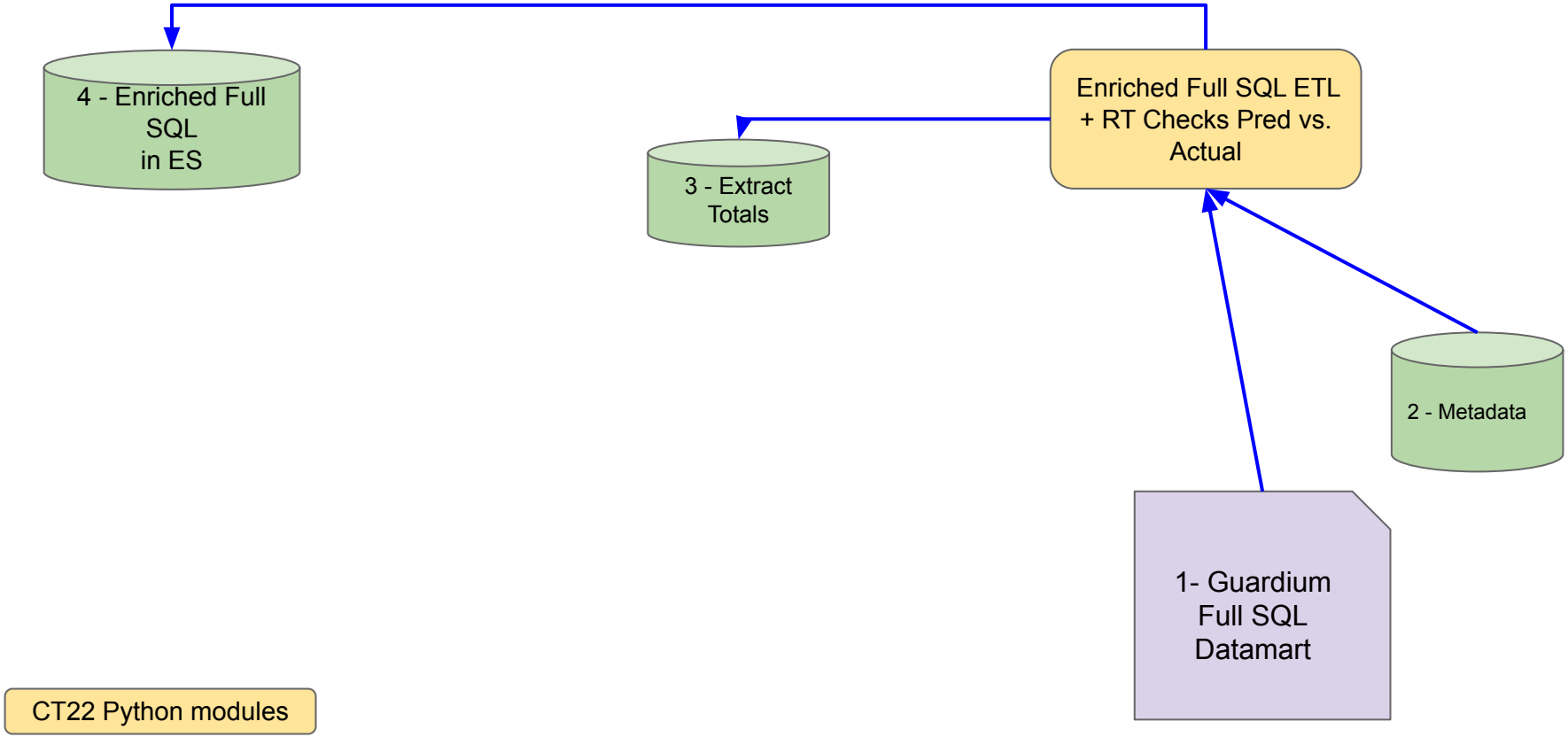
Actually not needed as this table will get ONLY the active pred, the one that will be used in RT checking

Examples of records

hash/year/dayofyear/predHWES3MUL/predHWES1 …
hash/year/dayofyear/predHWES3MUL/predHWES1 …
hash/year/dayofyear/predHWES3MUL/predHWES1 …

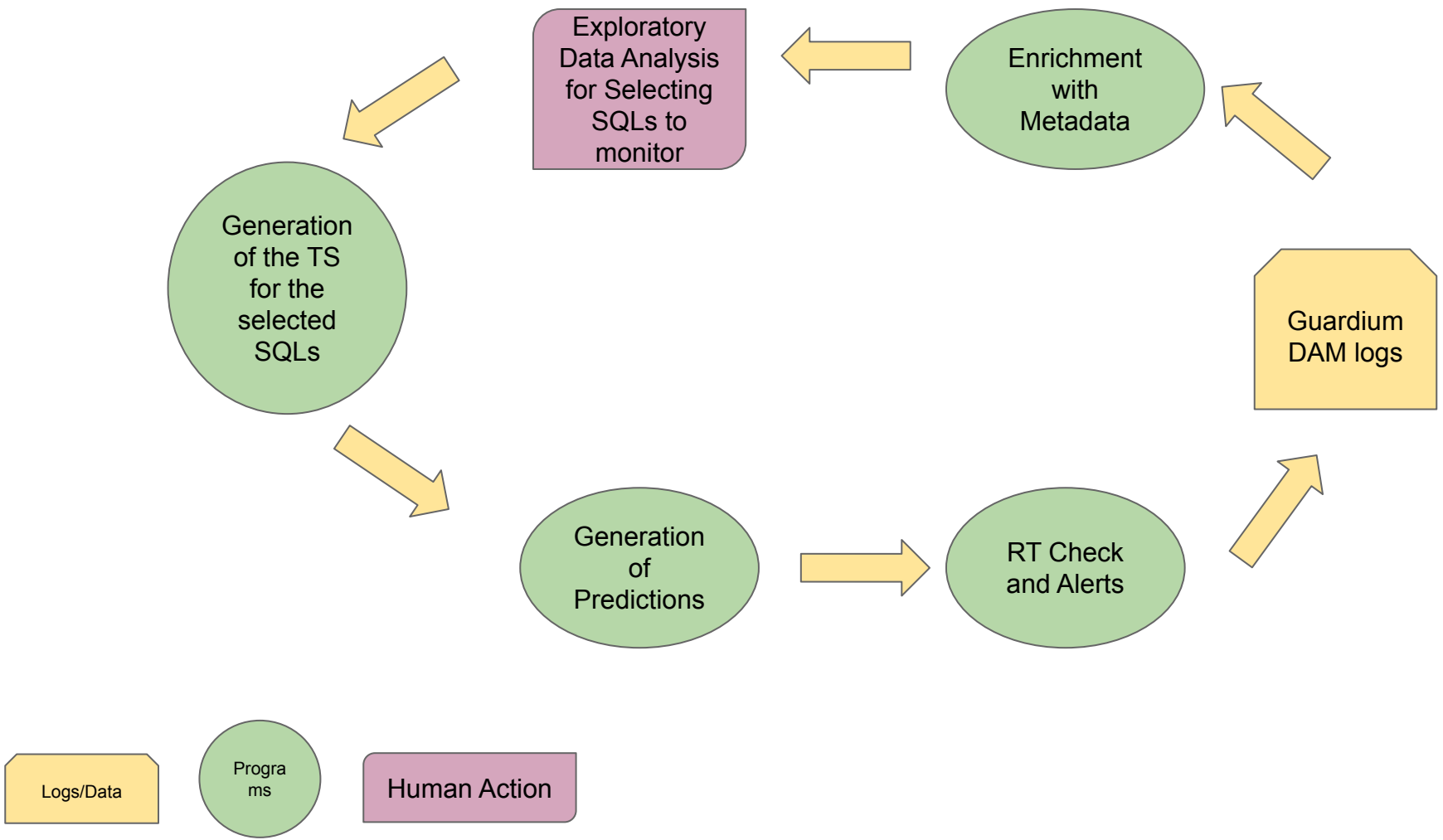
hash/year/dayofyear/predstype/pred
hash/year/dayofyear/predstype/pred
hash/year/dayofyear/predstype/pred

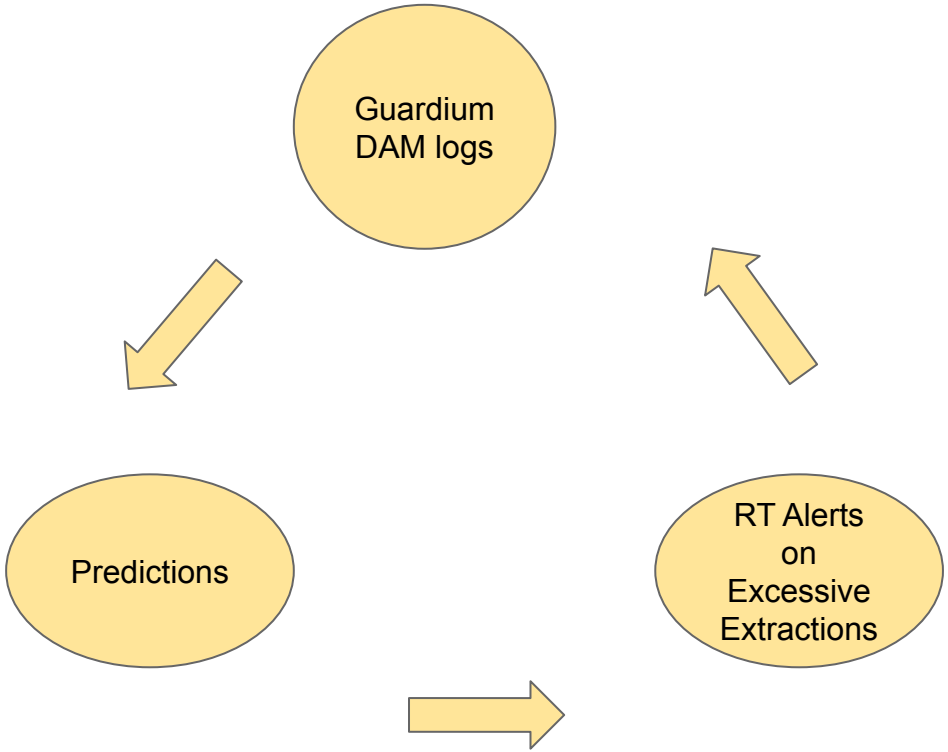
Step #1 : Enrichment and Totaling Extraction of the Day



The Time Series, the Model, the Fit, The Prediction

The Time Series, the Model, the Fit, The Prediction

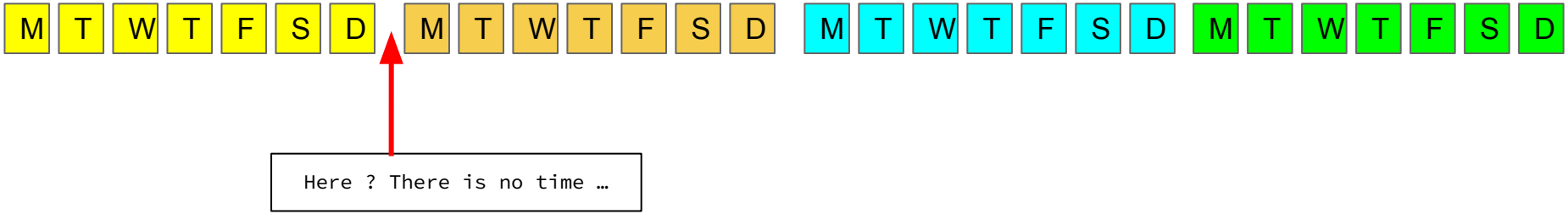

Exploratory Data Analysis for Selecting SQLs to monitor
Enrichment with Metadata
Guardium DAM logs
Generation of the TS for the selected SQLs
Generation of Predictions
RT Check and Alerts
Logs/Data
Programs
Human Action


Guardium
DAM logs
RT Alerts
on
Excessive
Extractions
Predictions

Others

- TZ issue
- At What Time to Generate the Preds ?
  - Issue : If you have to wait for Sunday to complete to have the data to generate the Pred for Monday, you basically have no time to compute, as the computation should happen in the instant at midnight, which is not possible
  - Solution :
    - Split daily TS into TS per day of week (TS for the Mondays, the Tuesday etc…)
    - Perform the computation for each day of week : On Wednesday compute for the next Tuesday

When To Perform The Prediction Computation ?



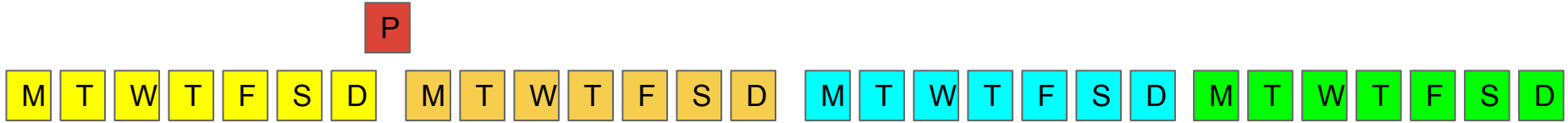
If you have to wait for Sunday (D) to have all the data needed to compute the pred for Monday (M), you have to do that between Sunday 11:59 PM and Monday 12:00 AM, meaning that basically you have no time.

On Monday, you need the Prediction but you don't have it yet because you CANNOT do the computation before since you need the data from before to compute the prediction for Monday.

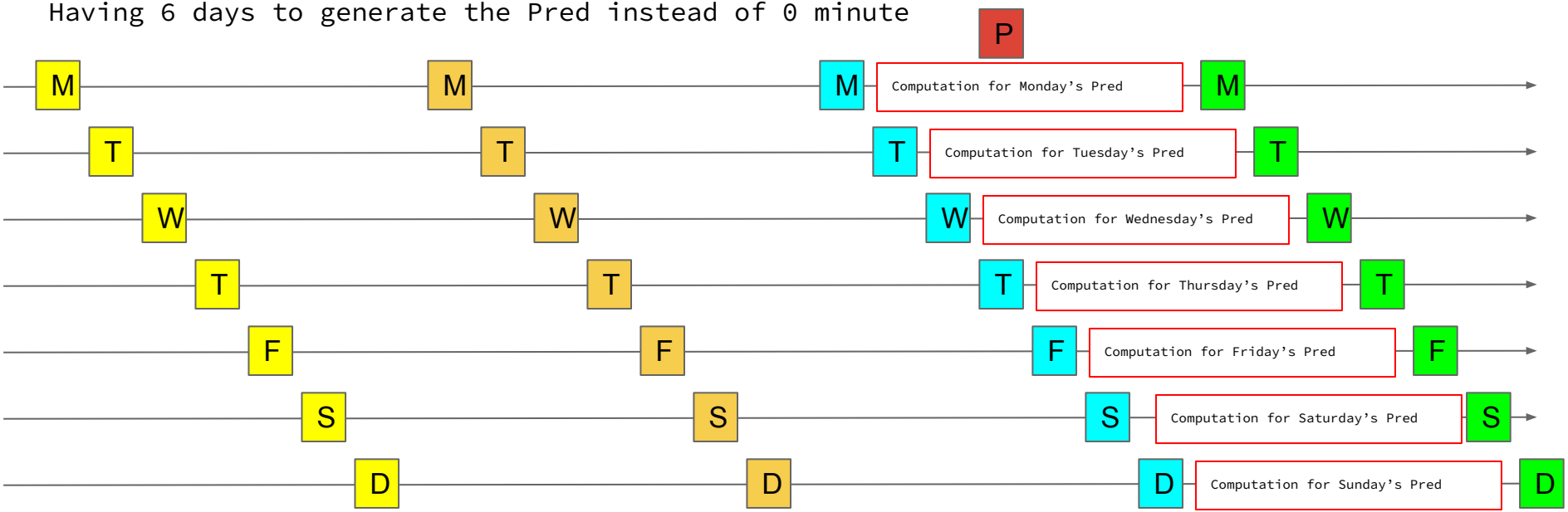
You may consider that the generation of the prediction can be done in the early hours of Monday and it will be sufficient. Yes, BUT, if you need the pred in the early hours of Monday, you may not get it in time and if your system had some trouble over the weekend (it will), you will be delayed even more.

Here is an option, based on a conclusion from experience.

When To Perform The Prediction Computation ?



Having 6 days to generate the Pred instead of 0 minute



Why this Approach Should Work For DAM ?

Mostly because :

- the Time unit is the day for DAM (by experience)- not the hour.

- Each day of the week in general is for specific tasks meaning that the traffic on a day doesn’t depend really of the traffic of the day before

This approach is still open for discussion

We are Looking for Collaboration

Our Proposal

- Start Small - Like a PoC - Focusing on Extractions of Sensitive Data
- Select 1 reasonable piece of traffic for example a Server, a DB or a User
- Decide on ES environment (on-prem vs. cloud)
- Install CT22 Enriched ETL & HWES3MUL modules
- Enrich the Extractions of Sensitive Data within the selected Traffic (the SQLs Select on Sensitive tables, collections, indexes)
- Analyze the different extractions - EDA : Exploratory Data Analyzis
- Select 1 or 2
- Perform Prediction generation on it/them
- Check in RT against actual, manually or by program
- Post-Mortem

How to contact us

info@context22.com
Support@context22.com
www.context22.com
(under construction)