yak: WebAssembly-based Data Analysis Platform with Data Providers Driven Analysis Control

COMPSAC 2025 SEPT
July 10, 2025

Naoki Matsumoto (Kyoto University, Japan)
Tomoya Kawanishi (Seera Networks, Inc, Japan)
Kenji Ohira (The University of Osaka, Japan)
Masahiro Kozuka (Okayama University, Japan)

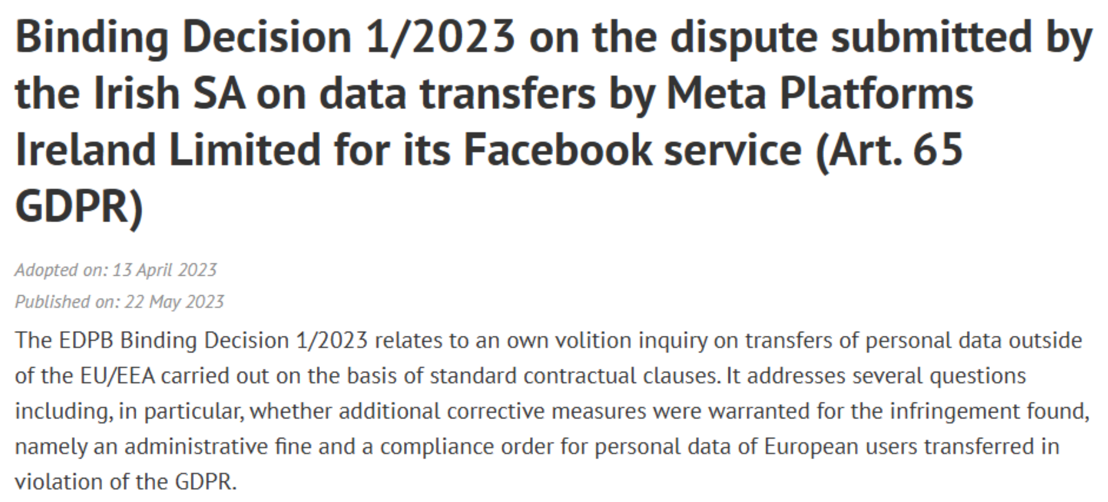
京都大学

Background: More and more data utilization

Privacy is the important factor to utilize collected data.

General Data Protection Regulation (GDPR) regulates the usage.

However, **some companies sometimes violate the contracts.**

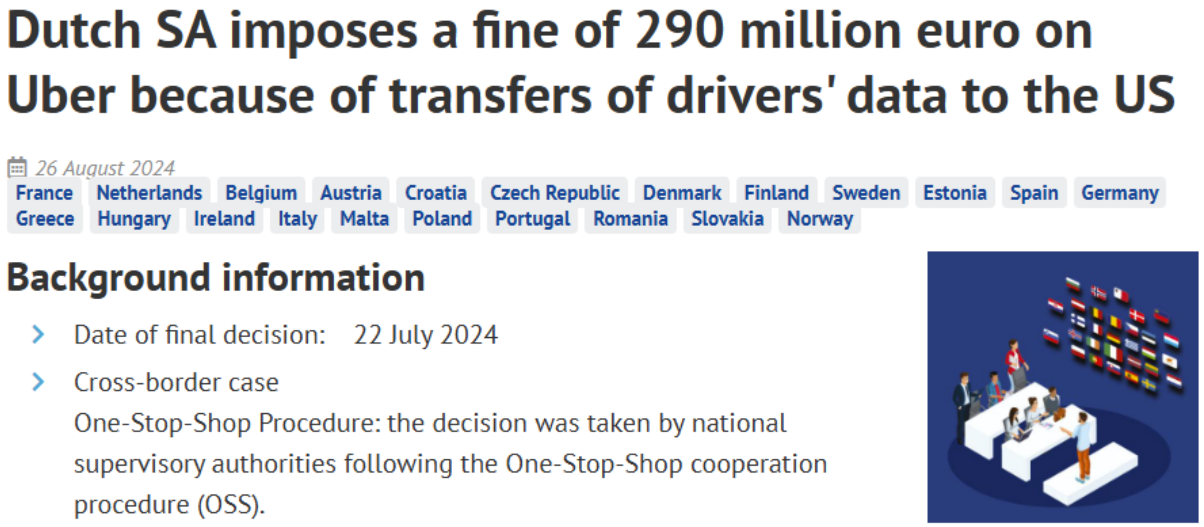


Binding Decision 1/2023 on the dispute submitted by the Irish SA on data transfers by Meta Platforms Ireland Limited for its Facebook service (Art. 65 GDPR)

Adopted on: 13 April 2023

Published on: 22 May 2023

The EDPB Binding Decision 1/2023 relates to an own volition inquiry on transfers of personal data outside of the EU/EEA carried out on the basis of standard contractual clauses. It addresses several questions including, in particular, whether additional corrective measures were warranted for the infringement found, namely an administrative fine and a compliance order for personal data of European users transferred in violation of the GDPR.



Dutch SA imposes a fine of 290 million euro on Uber because of transfers of drivers' data to the US

26 August 2024

France Netherlands Belgium Austria Croatia Czech Republic Denmark Finland Sweden Estonia Spain Germany Greece Hungary Ireland Italy Malta Poland Portugal Romania Slovakia Norway

Background information

- Date of final decision: 22 July 2024
- Cross-border case
  One-Stop-Shop Procedure: the decision was taken by national supervisory authorities following the One-Stop-Shop cooperation procedure (OSS).

Issue: How to protect sensitive data?

Approaches to conceal "**analysis itself**" have been proposed.

- Cryptographic approaches (Homomorphic Encryption etc.)
- Trusted Execution Environment (TEE)-based systems

- They lack the consideration about the "**leakage via result**".
- Contract violations on usage must be considered.

A method to protect sensitive data including results under data providers' control is required.

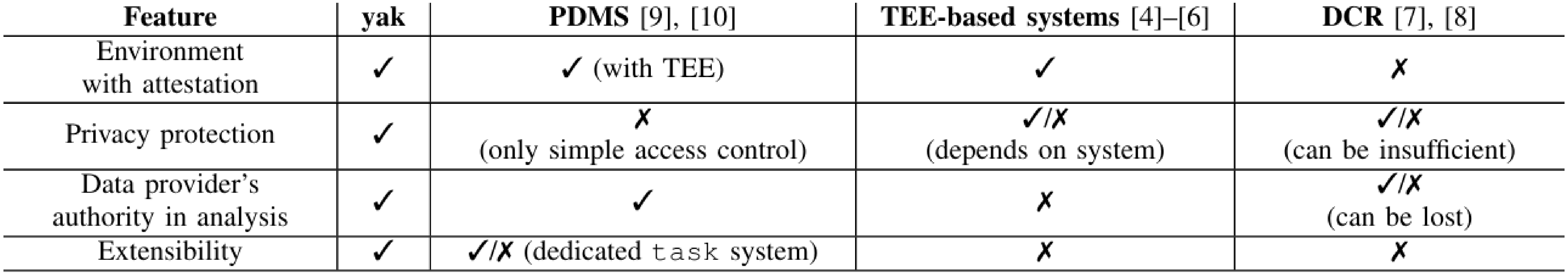
Related Works/Approaches

PDMS(Personal Data Management System)

- ✓ Data providers(owners) can manage access to their data.
- ✗ Provides only simple access control and lacks extensibility.

TEE-based systems

- ✗ Does not provide data provider’s authority in analysis.



Feature	**yak**	**PDMS** [9], [10]	**TEE-based systems** [4]–[6]	**DCR** [7], [8]
Environment with attestation	✓	✓ (with TEE)	✓	✗
Privacy protection	✓	✗ (only simple access control)	✓/✗ (depends on system)	✓/✗ (can be insufficient)
Data provider’s authority in analysis	✓	✓	✗	✓/✗ (can be lost)
Extensibility	✓	✓/✗ (dedicated `task` system)	✗	✗

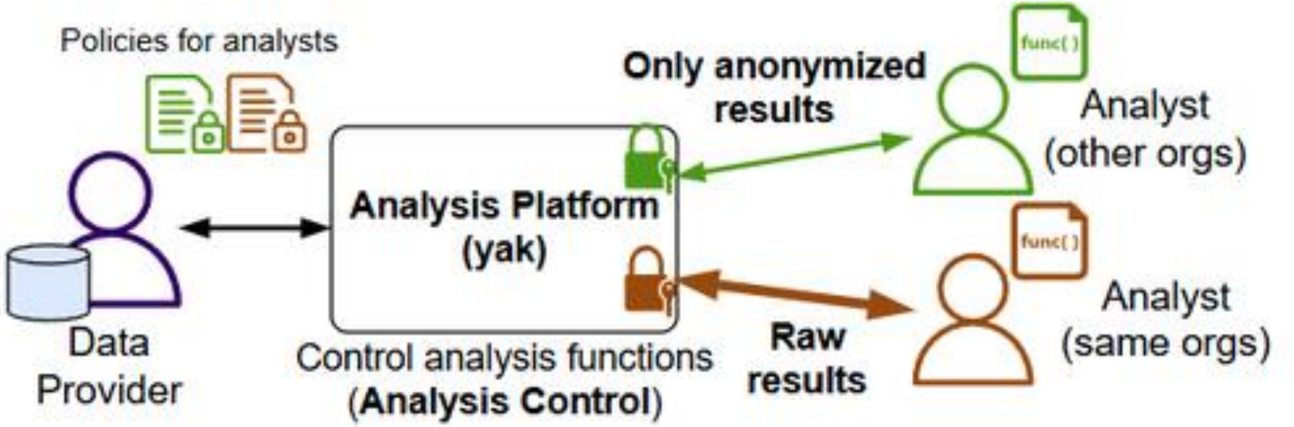
Solution: Analysis Control by Data Providers

<u>Analysis Control: data providers govern entire analysis</u>.

- Data providers define policies to control execution of analysis.
- Both “prohibiting” and “forcing” analysis are supported.

e.g.) Results are forcibly anonymized if it is defined in the policy.

Data providers can utilize policies based on analysts.
→ Giving gradation on data and analysis results.



Design of Analysis Control

Q: How to control the analysis with various datasets?

A: With **function specific validations** and **“reject and require” based policy**.

Function specific validations

The data leakage depends on the input and the function.

A) Calculating mean of 1 value.   B) Calculating variance of 100 values.

“A” leaks the actual values, but “B” is hard to predict actual values and should be allowed.

Analysis control **checks the input and output whether satisfies the policy**.

→ Enabling fine-grained control based on the both data and analysis.

“Reject” and “Require” for Analysis Functions

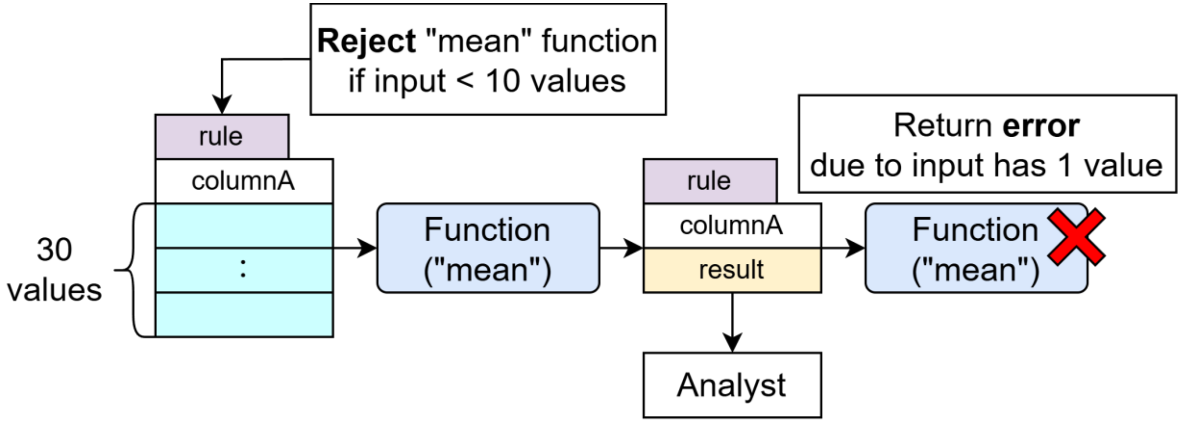
Reject: prohibits the function execution.

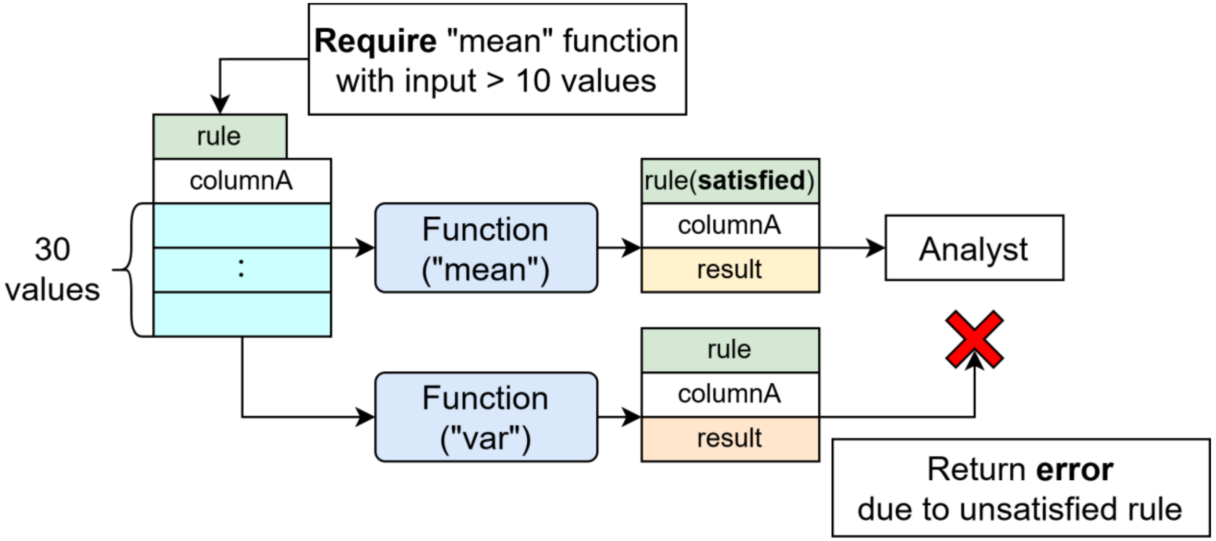
Used to explicitly ban risky functions.

Require: forces the function execution.

Used to force to execute functions before the result returned. Used in anonymization too.

“Table Validator” is a special require feature to check table formatted data and modify it as needed.

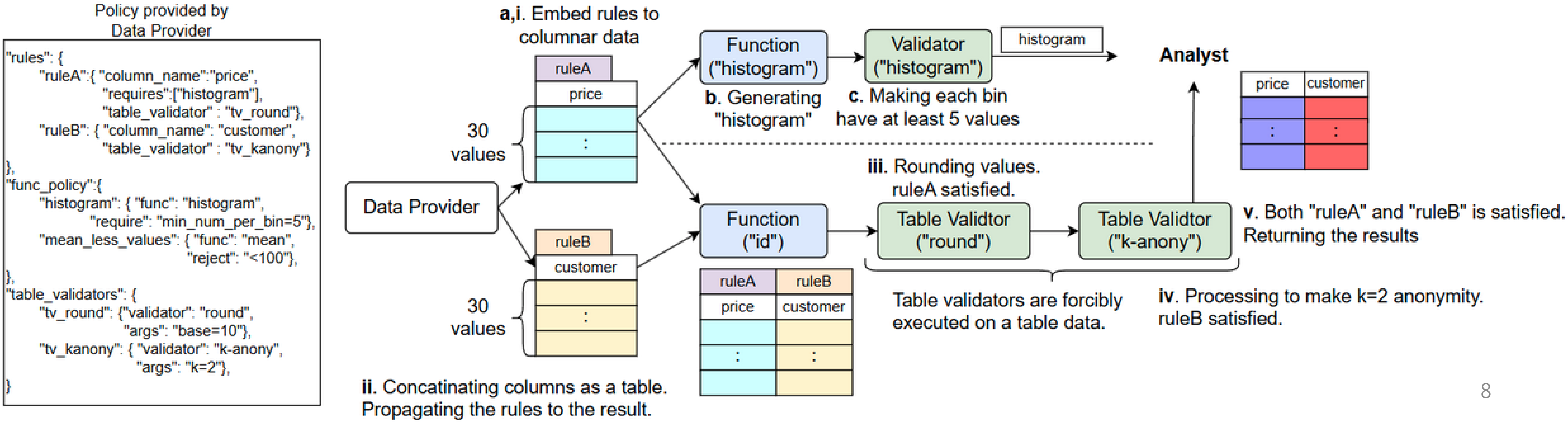




Example of Analysis Control

a-c): As for “price” column, it needs to execute “histogram” before return values.
→ The histogram executed and the result is modified according to the policy.

i-iii): “price” and “customer” columns has table validator “tv_round” and “tv_kanony”.
→ The result is table formatted, so performing round and k-anonymization on the result table.



Policy provided by Data Provider

```
"rules": {
      "ruleA":{ "column_name":"price",
                 "requires":["histogram"],
                 "table_validator" : "tv_round"},
      "ruleB": { "column_name": "customer",
                 "table_validator" : "tv_kanony"}
},
"func_policy":{
      "histogram": { "func": "histogram",
               "require": "min_num_per_bin=5"},
      "mean_less_values": { "func": "mean",
                         "reject": "<100"},
},
"table_validators": {
      "tv_round": {"validator": "round",
                  "args": "base=10"},
      "tv_kanony": { "validator": "k-anony",
                   "args": "k=2"},
}
```

Data Provider

a,i. Embed rules to columnar data

ruleA
price
30 values

ruleB
customer
30 values

Function ("histogram")

Validator ("histogram")

histogram

Analyst

b. Generating "histogram"

c. Making each bin have at least 5 values

Function ("id")

Table Validtor ("round")

Table Validtor ("k-anony")

ruleA	ruleB
price	customer
:	:

price	customer
:	:

ii. Concatinating columns as a table. Propagating the rules to the result.

iii. Rounding values. ruleA satisfied.

Table validators are forcibly executed on a table data.

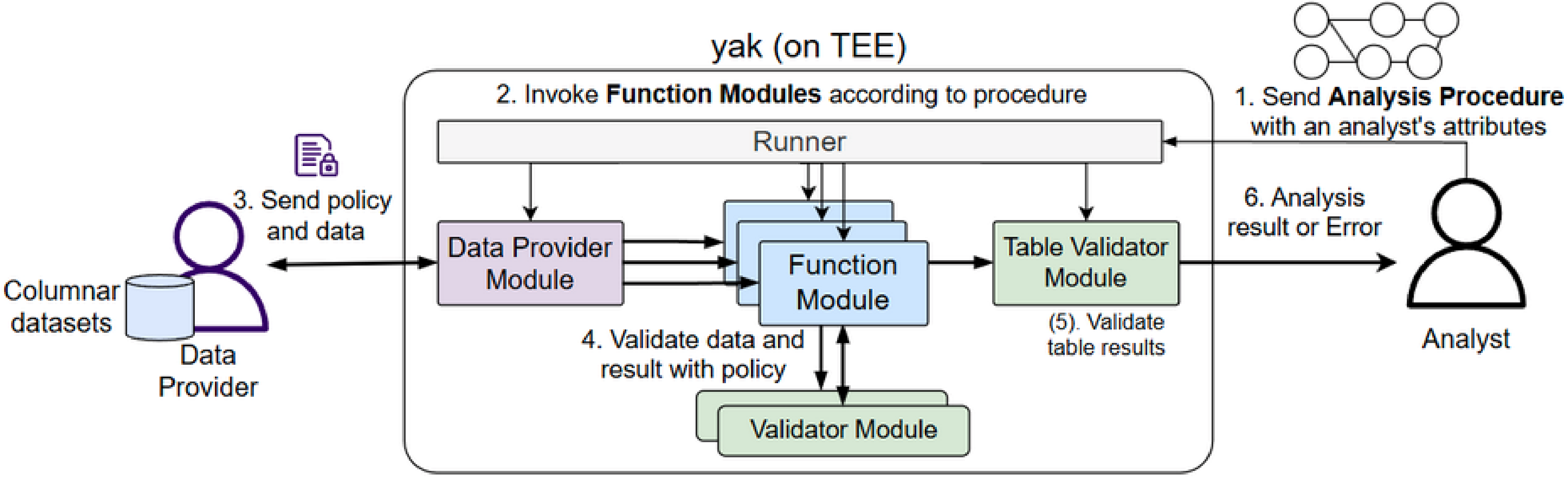
iv. Processing to make k=2 anonymity. ruleB satisfied.

v. Both "ruleA" and "ruleB" is satisfied. Returning the results

yak: Data Analysis Platform with Analysis Control

yak is a platform implementing analysis control.

- **Analysts**: define their analysis procedures with DAG.
- **Data Providers**: provide their data and control its use via a module.
- **Platformers**: provide a TEE to protect sensitive data from malicious people.



yak (on TEE)
2. Invoke **Function Modules** according to procedure
Runner
1. Send **Analysis Procedure** with an analyst's attributes
3. Send policy and data
Data Provider Module
Function Module
Table Validator Module
6. Analysis result or Error
Columnar datasets
Data Provider
4. Validate data and result with policy
(5). Validate table results
Validator Module
Analyst

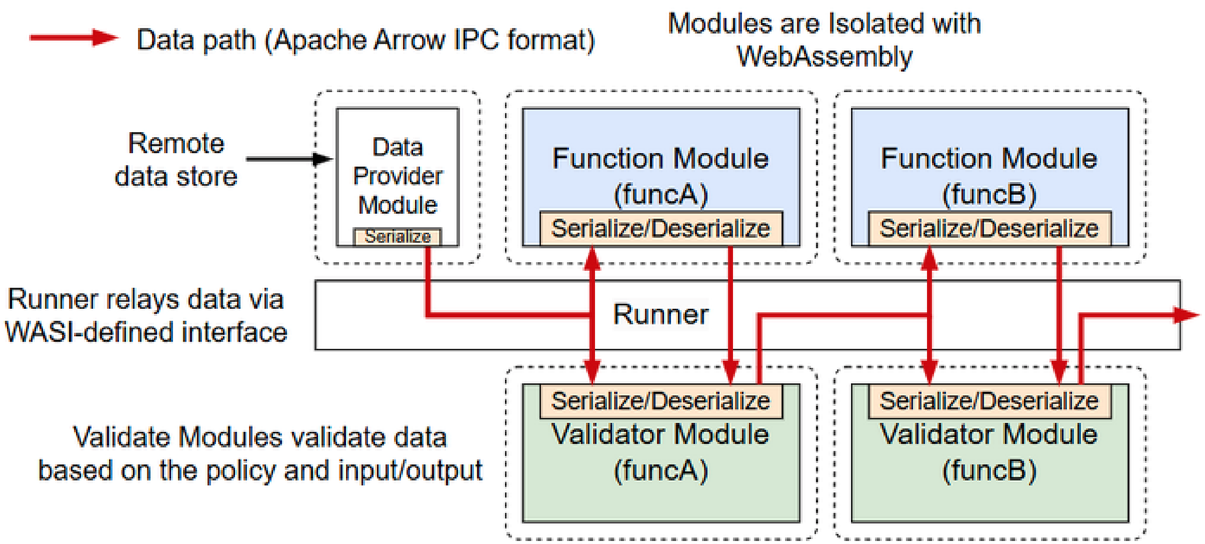
WebAssembly-based Module System

yak is designed to be **highly extensible** with WebAssembly modules.

WebAssembly provides **lightweight isolation** with a runtime.

- Developers can develop their modules with popular languages like Rust.
- Modules can utilize interfaces for filesystem and sockets called WASI.

yak defines modules per analysis function and connects them with Apache Arrow IPC format.



Data Protection for Data Providers

yak protects its platform and data with TEE and Authentication.

<u>TEE(AMD SEV-SNP) based platform</u>

- AMD SEV-SNP ensures the “Confidentiality” and “Integrity” of the environment.
- yak provides a protocol to verify the environment and yak are genuine.

<u>Authentication for analysts</u>

- yak authenticates analysts with data providers and processes policy.
- JWT token is used to provide authentication

Remote Attestation for Data Providers

“Attestation” verifies the integrity of the environment and code.

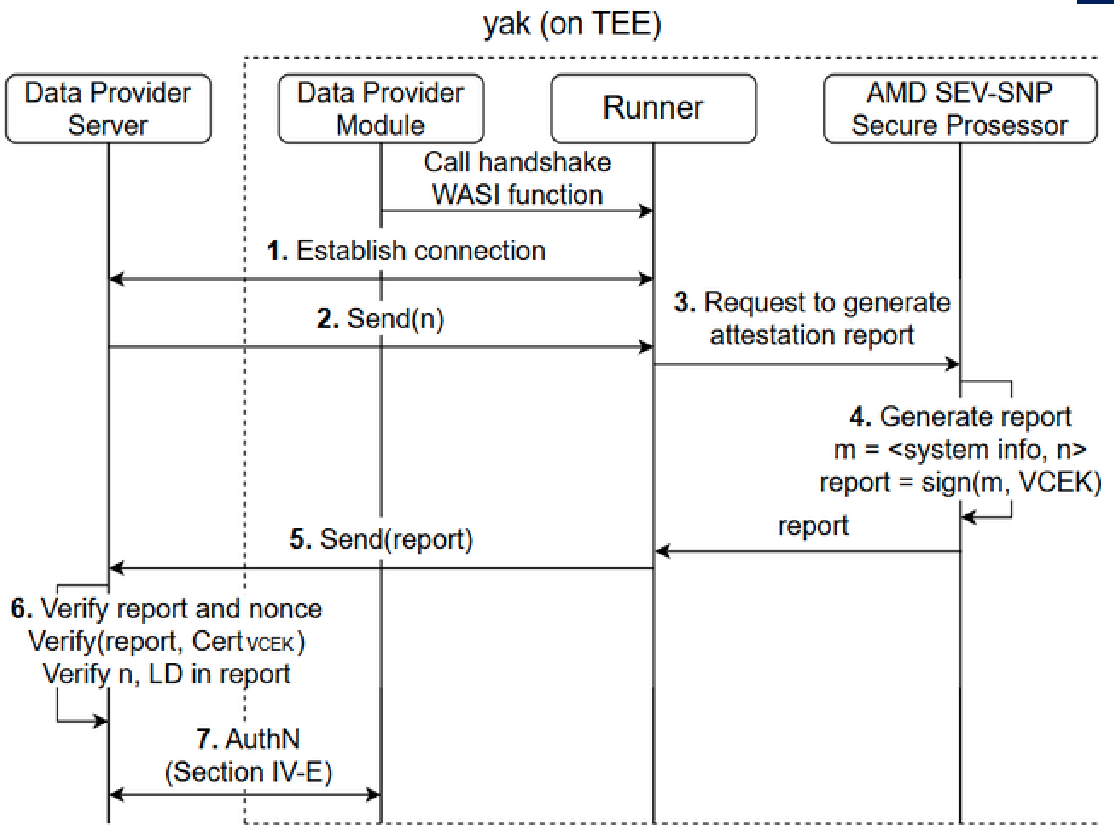
yak’s attestation protocol ensures the “<u>**Confidentiality**</u>” and “<u>**Integrity**</u>” of yak.

1. A runner establishes a secure E2E connection with data provider’s server

2. The server sends nonce n.

4. AMD SEV-SNP SP signs the report including n with VCEK(a key embedded in SP).

6. The server verifies the report’s signature and LD(memory content including yak).



JWT-based Authentication and Policy Provisioning

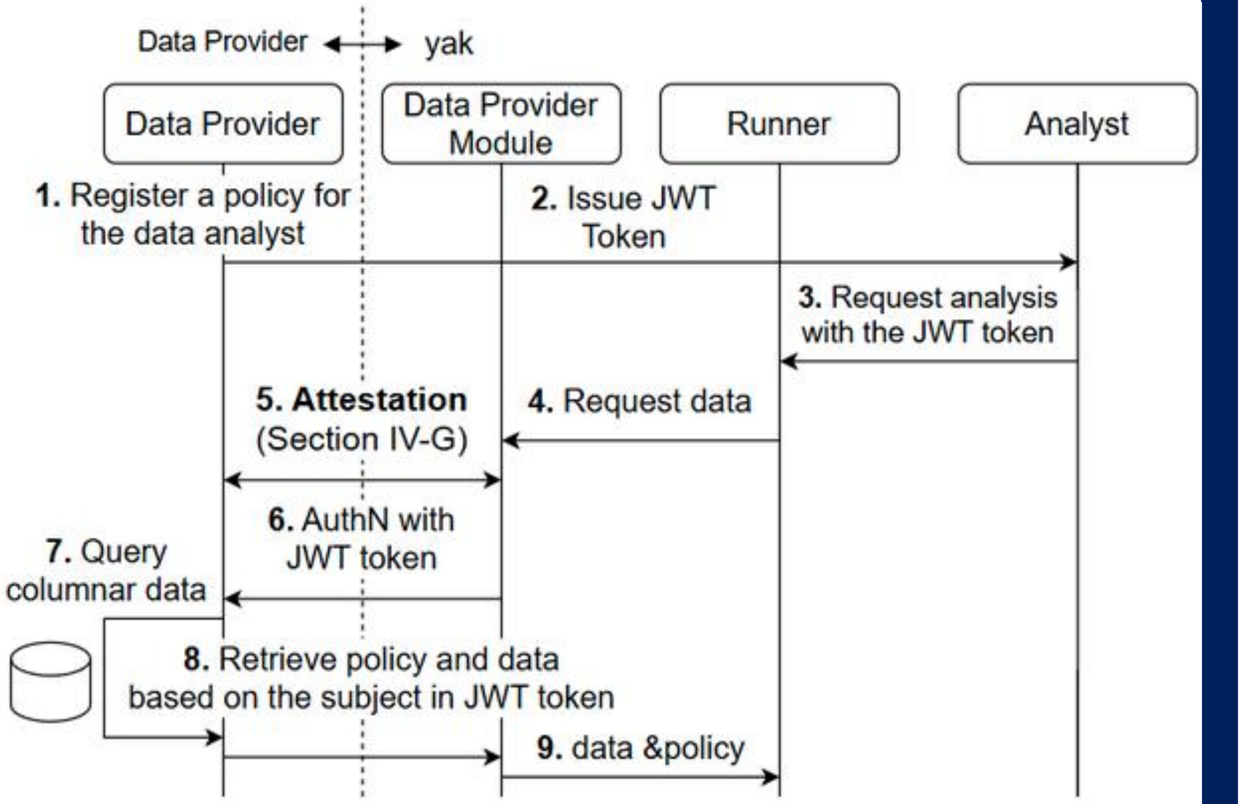
yak utilizes “JWT token” to authenticate analysts.

1,2. A data provider issues a token to an analyst.

3,4. The analyst sends analysis requests with the token.

6. The data provider authenticates the analyst with the token.

7,8,9. The data provider provide a policy according to the subject in the token.

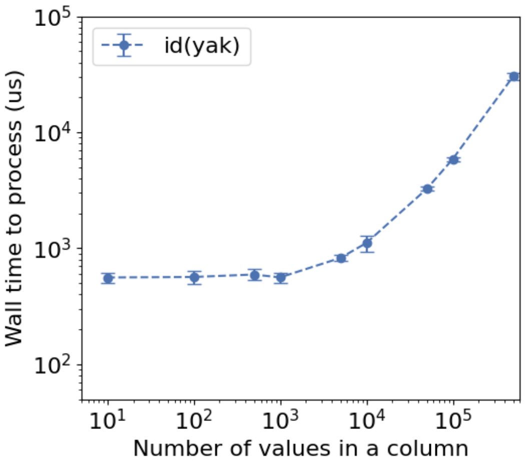

Data Provider
yak
Data Provider
Data Provider Module
Runner
Analyst
1. Register a policy for the data analyst
2. Issue JWT Token
3. Request analysis with the JWT token
4. Request data
5. Attestation (Section IV-G)
6. AuthN with JWT token
7. Query columnar data
8. Retrieve policy and data based on the subject in JWT token
9. data &policy

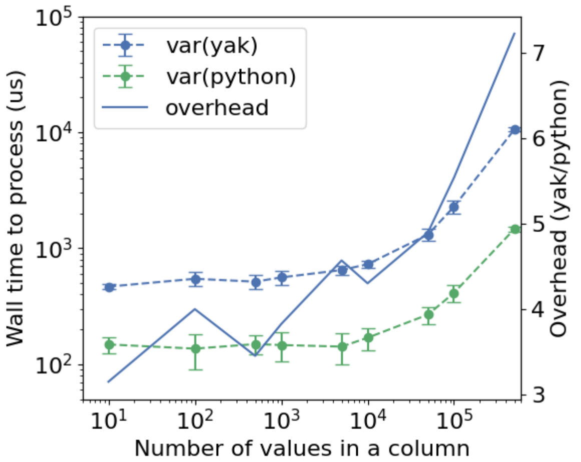
Performance Evaluation

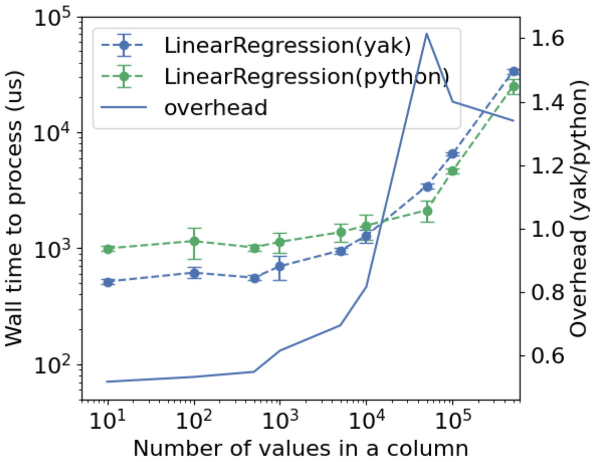
Compared the performance with usual(insecure) Python code.

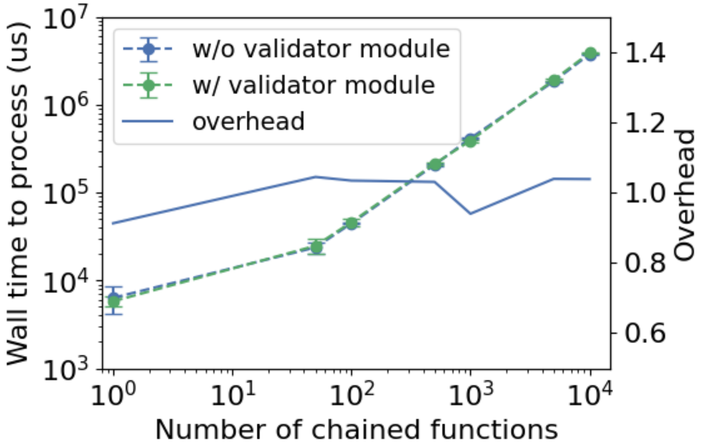
→ The results showed **<u>some but negligible overheads</u>**.

- Number of values: Longer(7x at most) and memory inflation(15x with FP32)
- Number of modules: Linear increase and 400 KB memory/function module







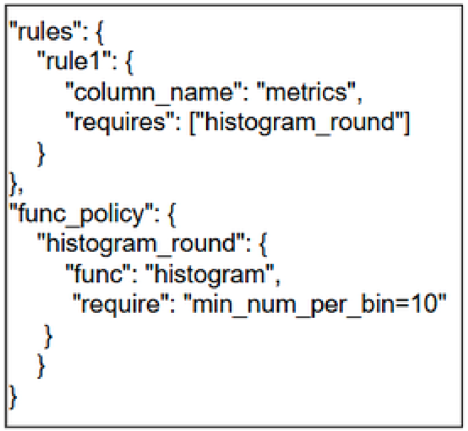


Case Study: Anonymization with Policy

Data providers can **anonymize results forcibly by defining policies**.

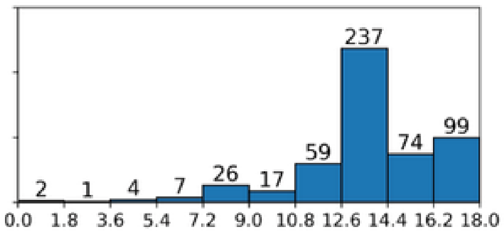
- Validator modules: result anonymization (e.g. rounding)
- Table validators: table anonymization (e.g. k-anonymization)

Policy

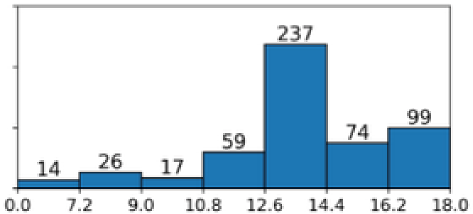


```
"rules": {
   "rule1": {
      "column_name": "metrics",
      "requires": ["histogram_round"]
   }
},
"func_policy": {
   "histogram_round": {
      "func": "histogram",
       "require": "min_num_per_bin=10"
    }
   }
}
```

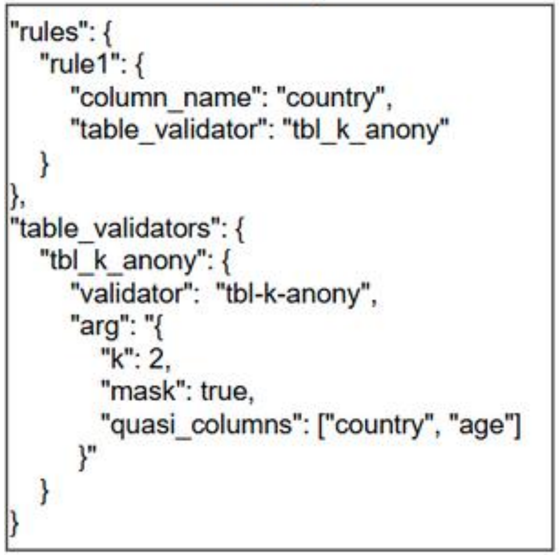
Raw Result (Policy is not enforced)



Anonymized Result

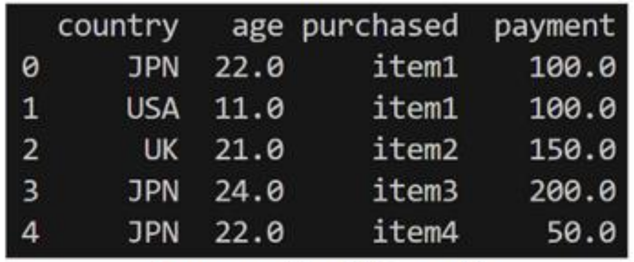


Policy



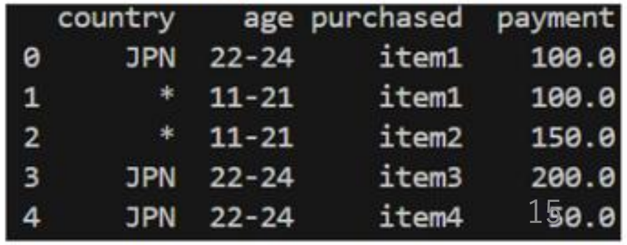
```
"rules": {
   "rule1": {
      "column_name": "country",
      "table_validator": "tbl_k_anony"
   }
},
"table_validators": {
   "tbl_k_anony": {
      "validator":  "tbl-k-anony",
      "arg": "{
         "k": 2,
         "mask": true,
         "quasi_columns": ["country", "age"]
      }"
   }
}
```

Raw Result (Policy is not enforced)



```
  country   age purchased  payment
0     JPN  22.0     item1    100.0
1     USA  11.0     item1    100.0
2      UK  21.0     item2    150.0
3     JPN  24.0     item3    200.0
4     JPN  22.0     item4     50.0
```

Anonymized Result



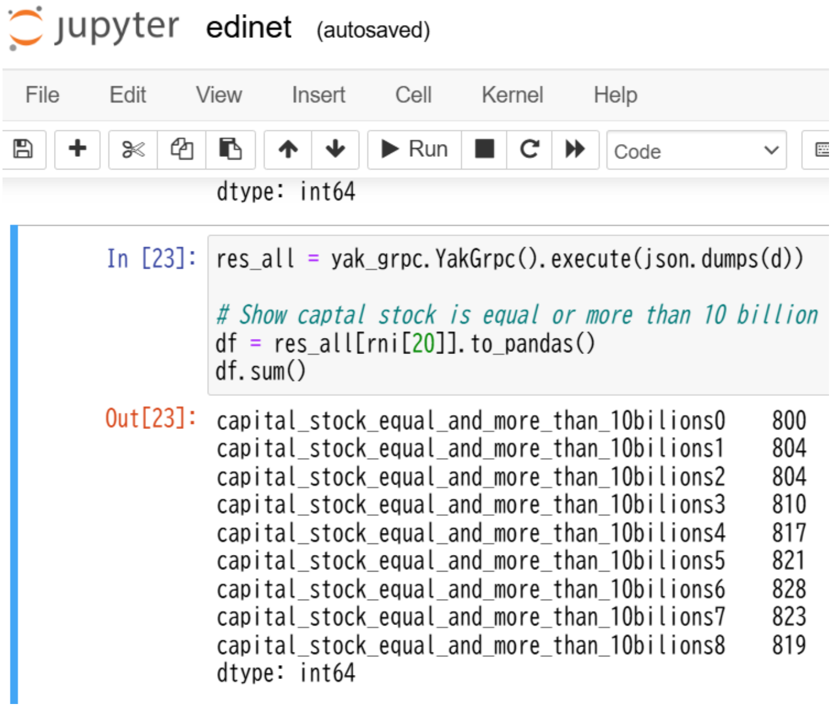
```
  country    age purchased  payment
0     JPN  22-24     item1    100.0
1       *  11-21     item1    100.0
2       *  11-21     item2    150.0
3     JPN  22-24     item3    200.0
4     JPN  22-24     item4     50.0
```

Case Study: Analysis with Jupyter Notebook

“<u>**Usability**</u>” is important factor for a usable analysis platform.

yak provides **a dedicated library** for Jupyter Notebook.

→ Analysts can use yak in a familiar environment.



```
dtype: int64
```

In [23]:

```
res_all = yak_grpc.YakGrpc().execute(json.dumps(d))

# Show captal stock is equal or more than 10 billion
df = res_all[rni[20]].to_pandas()
df.sum()
```

Out[23]:

```
capital_stock_equal_and_more_than_10bilions0    800
capital_stock_equal_and_more_than_10bilions1    804
capital_stock_equal_and_more_than_10bilions2    804
capital_stock_equal_and_more_than_10bilions3    810
capital_stock_equal_and_more_than_10bilions4    817
capital_stock_equal_and_more_than_10bilions5    821
capital_stock_equal_and_more_than_10bilions6    828
capital_stock_equal_and_more_than_10bilions7    823
capital_stock_equal_and_more_than_10bilions8    819
dtype: int64
```

Discussion: Validity and Risk of Analysis Control

Validity: Depending on a policy configuration

“Which data should be concealed or anonymized” relies on a policy.

Risk: Insufficient or misconfiguration in a policy

The configuration requires understanding on data = Sometimes difficult.

Solution: profile-based approach or starting from small.

- Policy sets by experts for specific format datasets bring out-of-box data use.
- “Reject” policies and “table validators” enables start from small steps.

Future work: Assisting tools like the data leakage estimation.

- Differential privacy approaches will be helpful.

Summary

Problem: Proper use of data with data provider's control

Solution: Analysis control and its implementation "yak"

- Analysis control forbids or forces analysis function execution.
- yak provides analysis control with TEE including attestation.

Evaluation: Case studies showed the ability for analysis.

- Overheads(7x longer, 15x memory inflation) were acceptable.

Discussion: Improvement of policy definition is required.

Appendix

Discussion: Usability for Data Providers

The requirements for protection varies on datasets.

- Ideal: Data providers have control.
- Real: Data providers have responsibility for their control too.

Understanding their own datasets is essential.

yak requires a policy to control the analysis on their data.

- Estimation on the data leakage
- More usable anonymization like differential privacy

will be helpful to define policies appropriately.

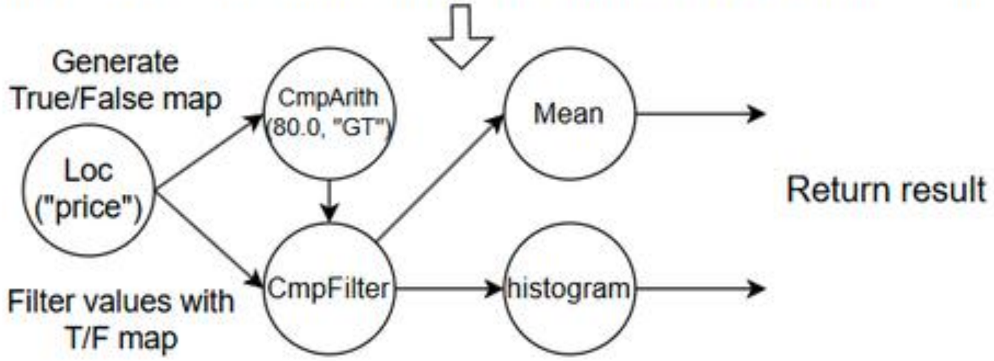
DAG-based Analysis Procedure

yak controls function invocations per module-basis.
→ it requires an <u>**explicit data dependency among modules**</u>.

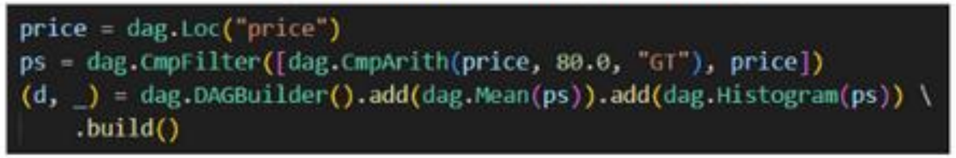
Analysts define their analysis with “**DAG-based procedure**”.

- Runner invokes modules according to the DAG.
- The only final results are returned to analysts if allowed.

Query: SELECT mean(price), histogram(price) WHERE price > 80.0



DAG generation with wrapper library

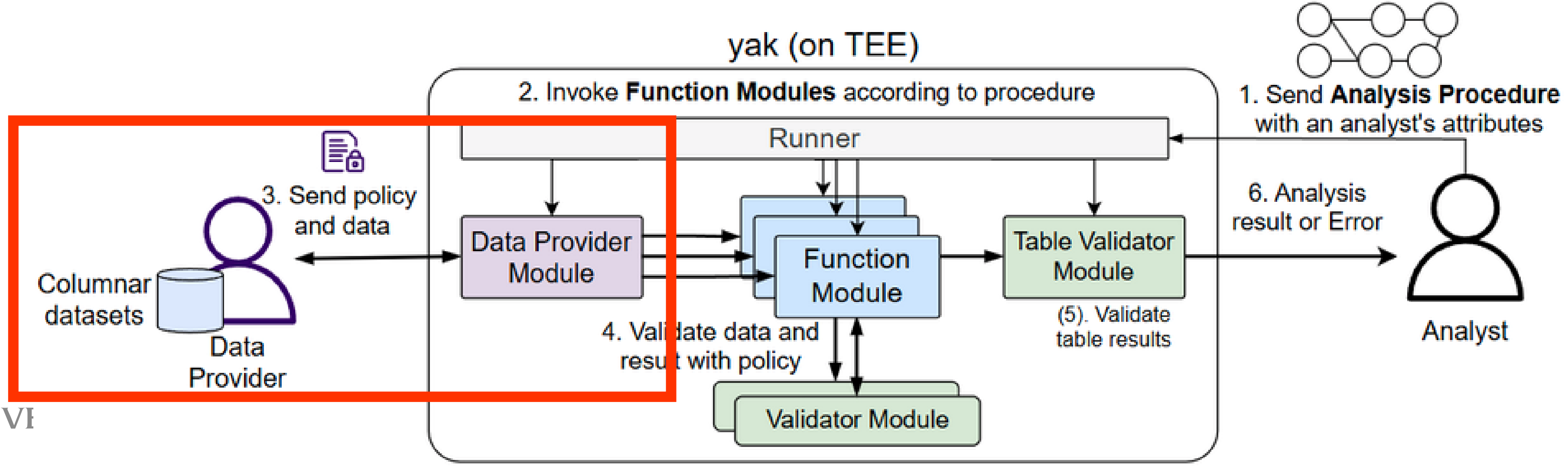


```
price = dag.Loc("price")
ps = dag.CmpFilter([dag.CmpArith(price, 80.0, "GT"), price])
(d, _) = dag.DAGBuilder().add(dag.Mean(ps)).add(dag.Histogram(ps)) \
    .build()
```

Flexible Data Providing by Data Providers

Data providers provide their data via **"Data Provider Module"**. It is used to

- Connect to their server and retrieve data.
- Convert data format into requested columnar format.
- Verify the environment and Authenticate analysts.


yak (on TEE)
2. Invoke Function Modules according to procedure
1. Send Analysis Procedure with an analyst's attributes
Runner
3. Send policy and data
Columnar datasets
Data Provider
Data Provider Module
Function Module
Table Validator Module
(5). Validate table results
6. Analysis result or Error
Analyst
4. Validate data and result with policy
Validator Module

Threat Model

Components in yak are intended to work correctly.

- Runner (The policy is enforced as defined)
- Modules (Function and Validator do have vulnerabilities)

TEE protects data against the attacks from platformers.

- AMD SEV-SNP's TEE ensures confidentiality and integrity
  → Data providers can verify it via the attestation

Authentication prevents malicious analyst from stealing data.

- The management of JWT tokens relies on the data providers.