PRIVAaaS: Privacy Approach for a Distributed Cloud-based Data Analytics Platforms

Tania Basso
Regina Moraes
Nuno Antunes
Marco Vieira
Walter Santos
Wagner Meira Jr.

Context

- **Big Data** → datasets with large amount of data
- **Big Data Analytics** → use advanced analytic techniques and algorithms against these datasets
- Individuals provide *Personally Identifiable Information (PII)* → **privacy concerns**

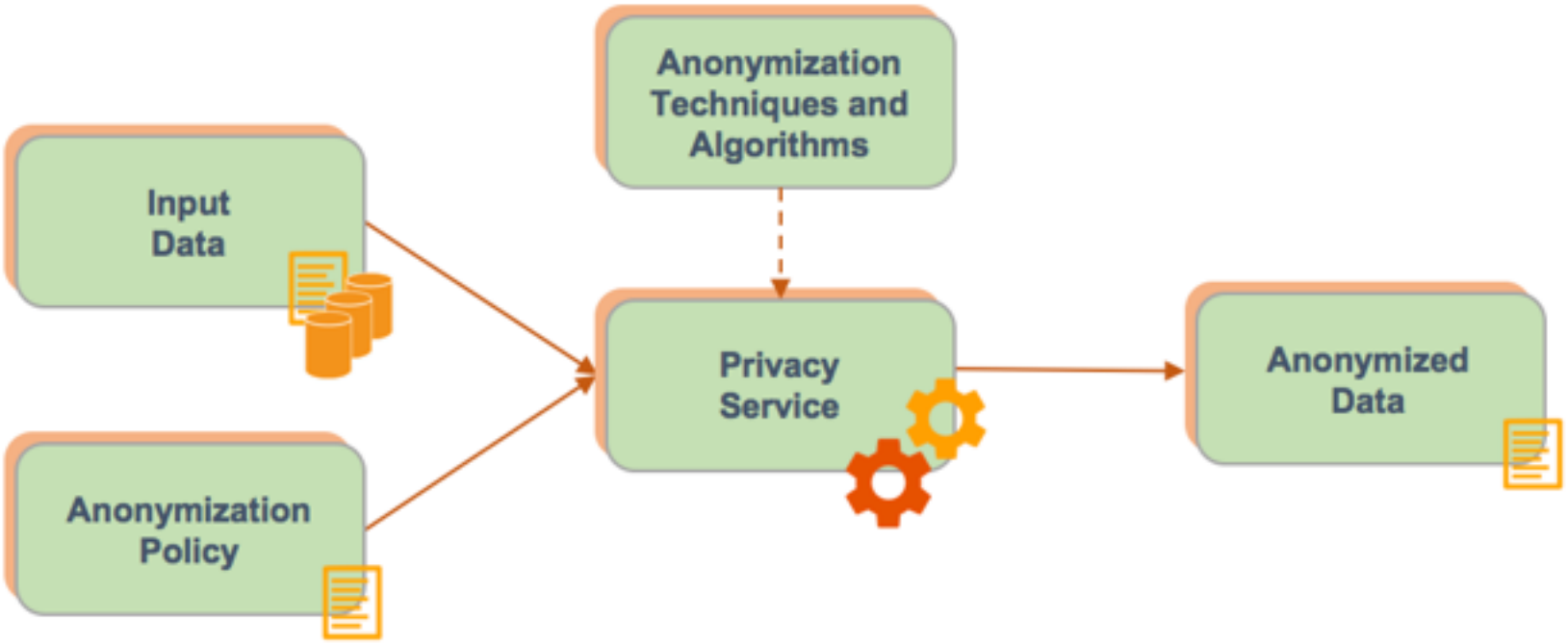
Context

- Approach for protecting data privacy based on anonymization policies – **PRIVAaaS**
- Discusses PRIVAaaS integration in the LEMONADE
- LEMONADE (**L**ive **E**nvironment for **M**ining **o**f **N**on-trivial **A**mount of **D**ata from **E**verywhere) - web-based platform for data analytics

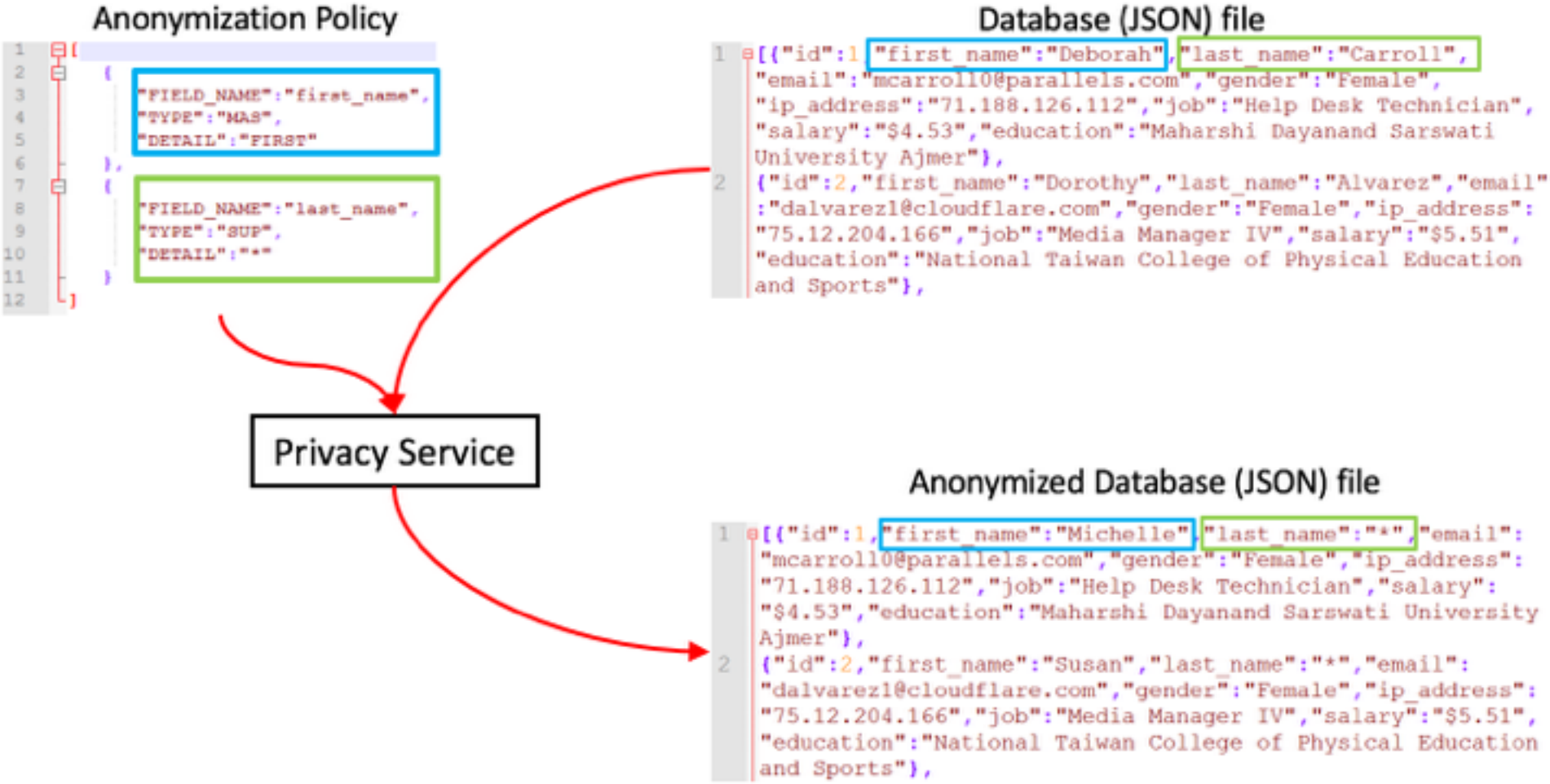
PRIVAaaS

- Set of tools to protect sensible information processed by data analytics
- Based on anonymization techniques and policies
- Free, open source, developed in Java
- Implements generalization, suppression, masking (replacement) and encryption.

PRIVAaaS


Anonymization Techniques and Algorithms
Input Data
Privacy Service
Anonymized Data
Anonymization Policy

PRIVAaaS

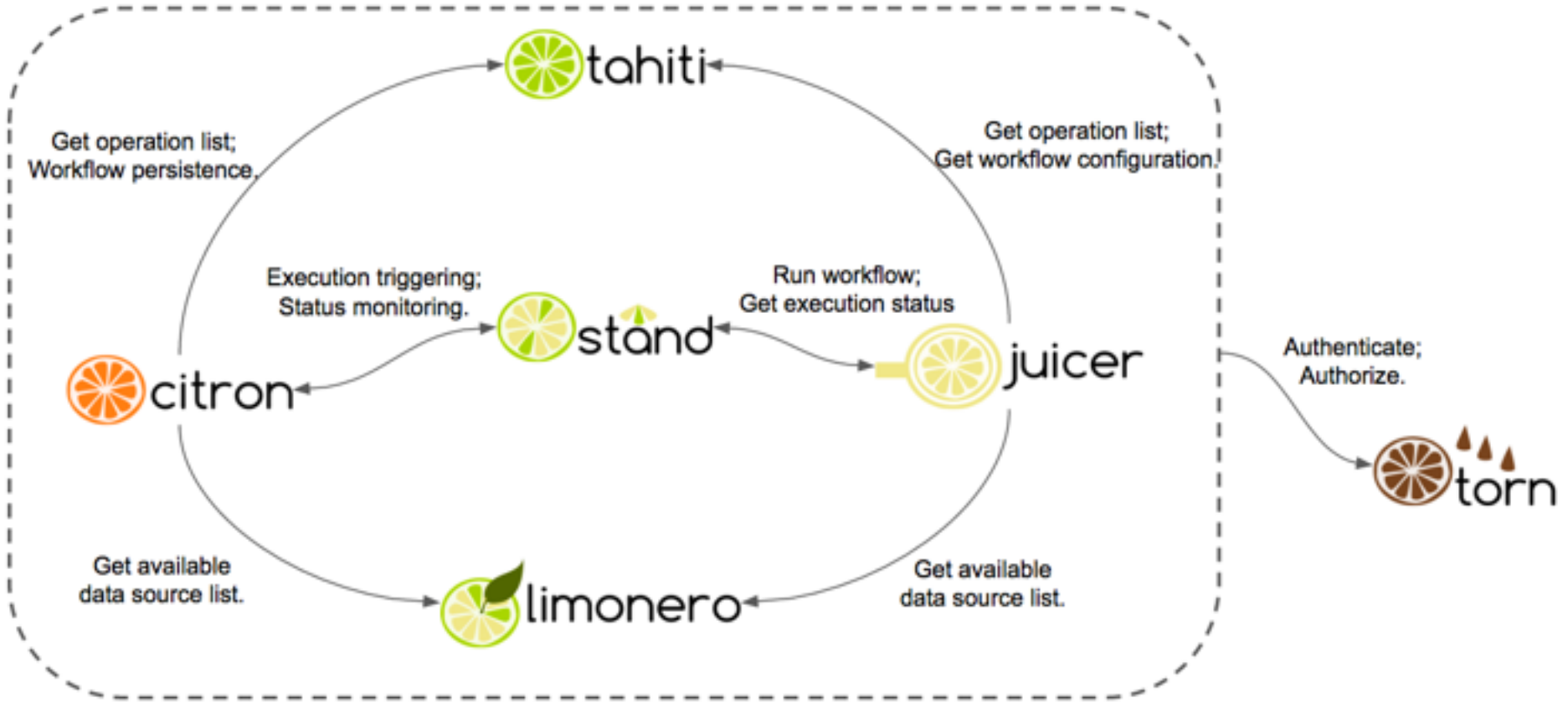

Anonymization Policy
{
"FIELD_NAME":"first_name",
"TYPE":"MAS",
"DETAIL":"FIRST"
},
{
"FIELD_NAME":"last_name",
"TYPE":"SUP",
"DETAIL":"*"
}
]
Privacy Service
Database (JSON) file
[{"id":1,"first_name":"Deborah","last_name":"Carroll",
"email":"mcarroll0@parallels.com","gender":"Female",
"ip_address":"71.188.126.112","job":"Help Desk Technician",
"salary":"$4.53","education":"Maharshi Dayanand Sarswati
University Ajmer"},
{"id":2,"first_name":"Dorothy","last_name":"Alvarez","email"
:"dalvarez1@cloudflare.com","gender":"Female","ip_address":
"75.12.204.166","job":"Media Manager IV","salary":"$5.51",
"education":"National Taiwan College of Physical Education
and Sports"},
Anonymized Database (JSON) file
[{"id":1,"first_name":"Michelle","last_name":"*","email":
"mcarroll0@parallels.com","gender":"Female","ip_address":
"71.188.126.112","job":"Help Desk Technician","salary":
"$4.53","education":"Maharshi Dayanand Sarswati University
Ajmer"},
{"id":2,"first_name":"Susan","last_name":"*","email":
"dalvarez1@cloudflare.com","gender":"Female","ip_address":
"75.12.204.166","job":"Media Manager IV","salary":"$5.51",
"education":"National Taiwan College of Physical Education
and Sports"},

LEMONADE

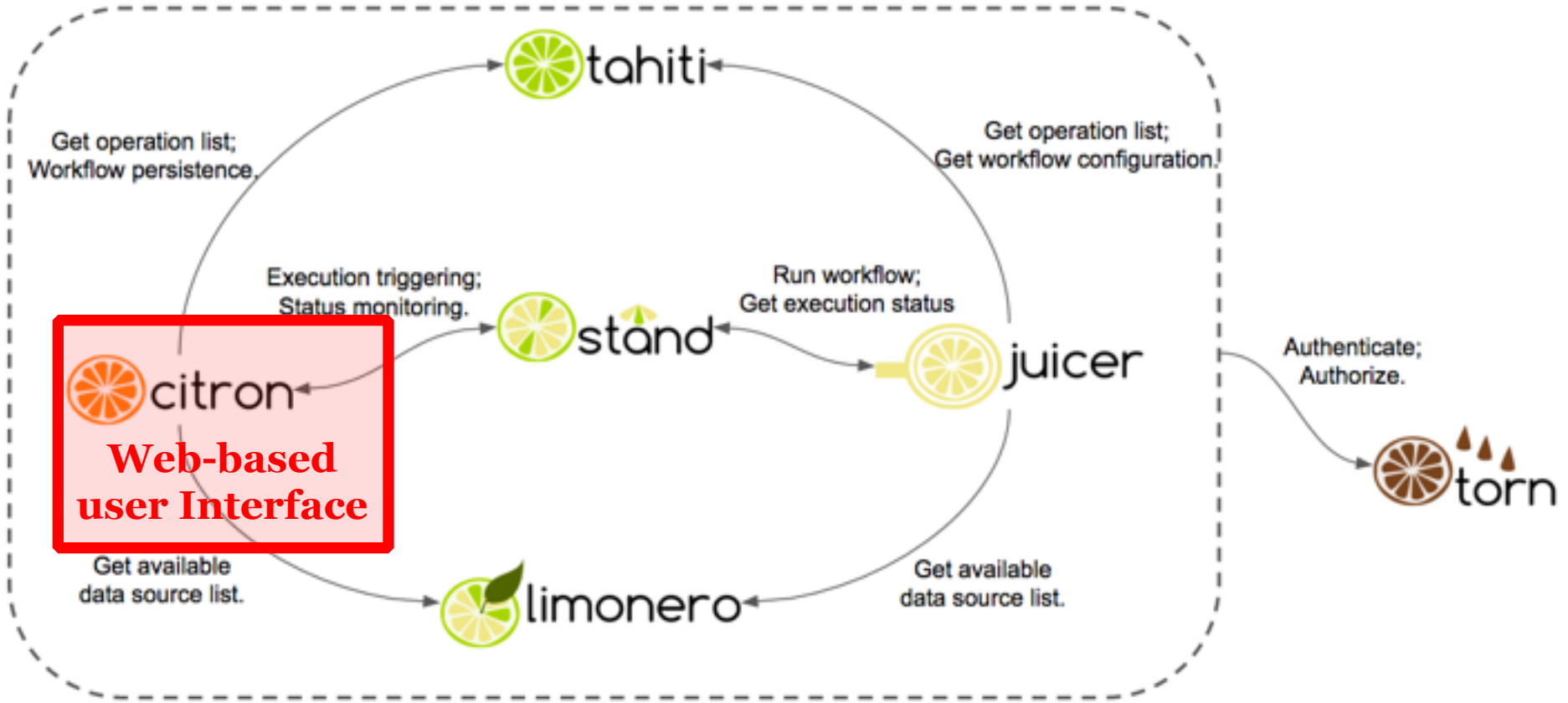
Live **E**nvironment for **M**ining **o**f **N**on-trivial **A**mount of **D**ata from **E**verywhere

- Interoperable, scalable and open platform for data analytics
- Users can drag and drop operations and data sources to compose different ETL processes and Machine Learning Workflows
- Runs on Spark, while providing resource allocation and QoS in a cloud-based system

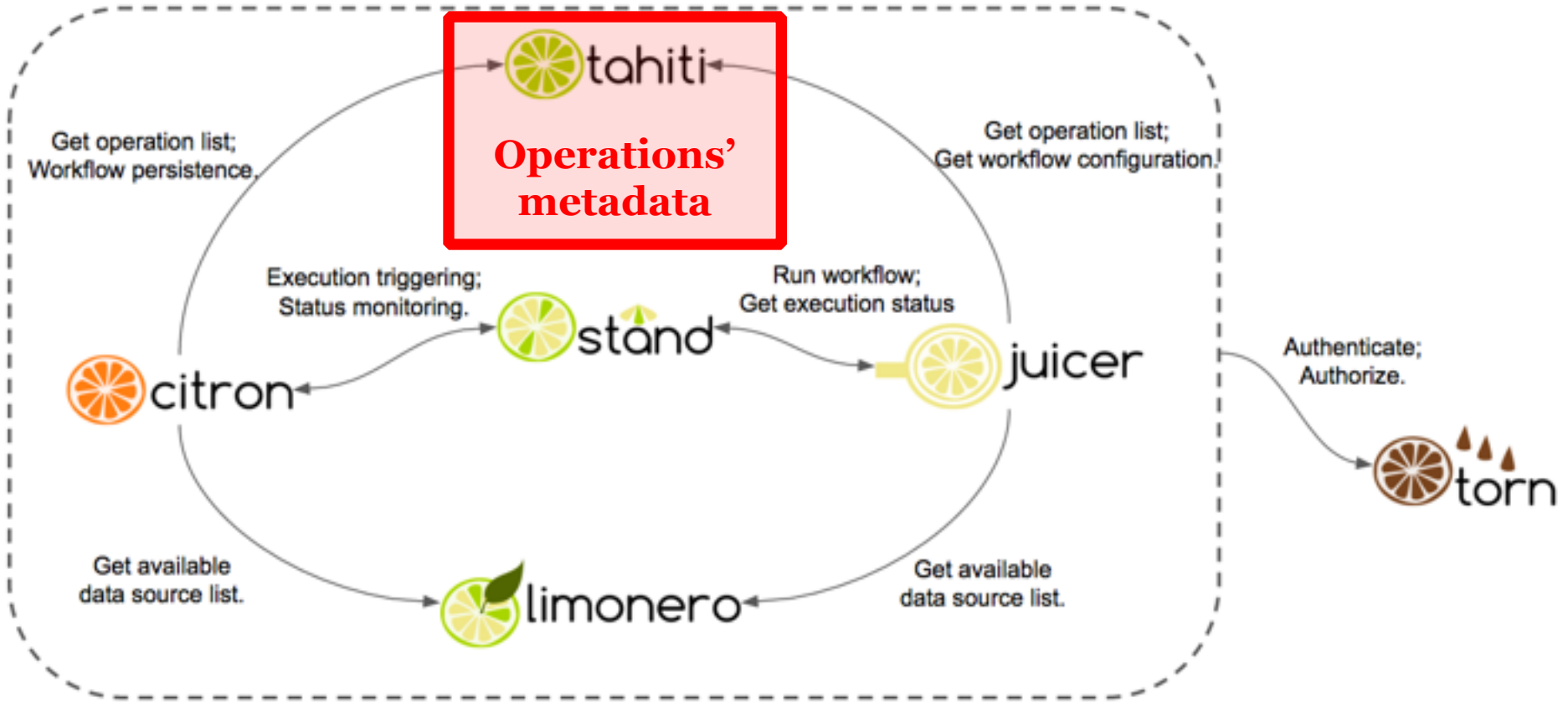
LEMONADE - Architecture



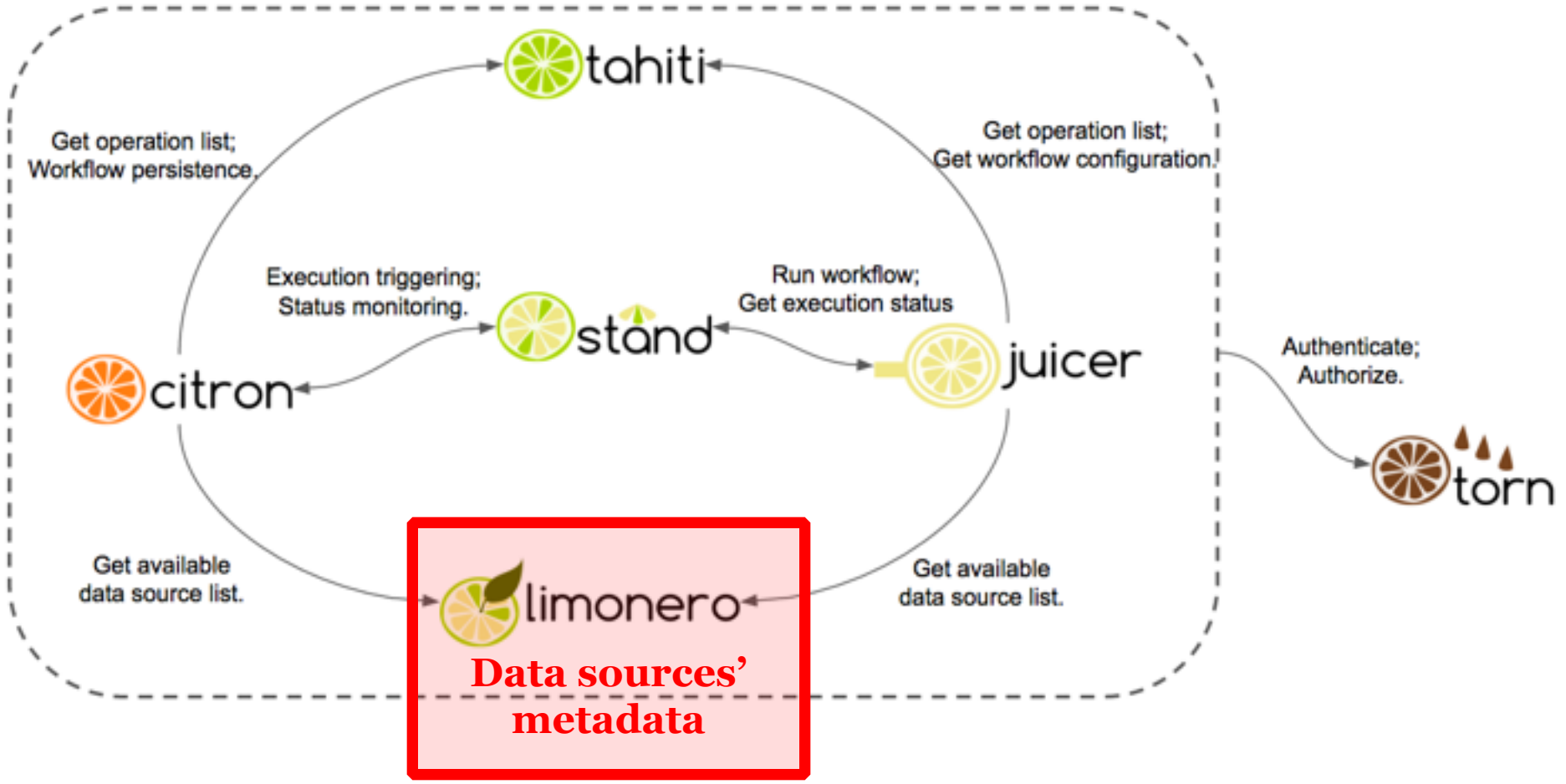
LEMONADE


tahiti
Get operation list;
Workflow persistence.
Get operation list;
Get workflow configuration.
Execution triggering;
Status monitoring.
Run workflow;
Get execution status
stand
juicer
citron
Web-based
user Interface
Authenticate;
Authorize.
torn
Get available
data source list.
limonero
Get available
data source list.

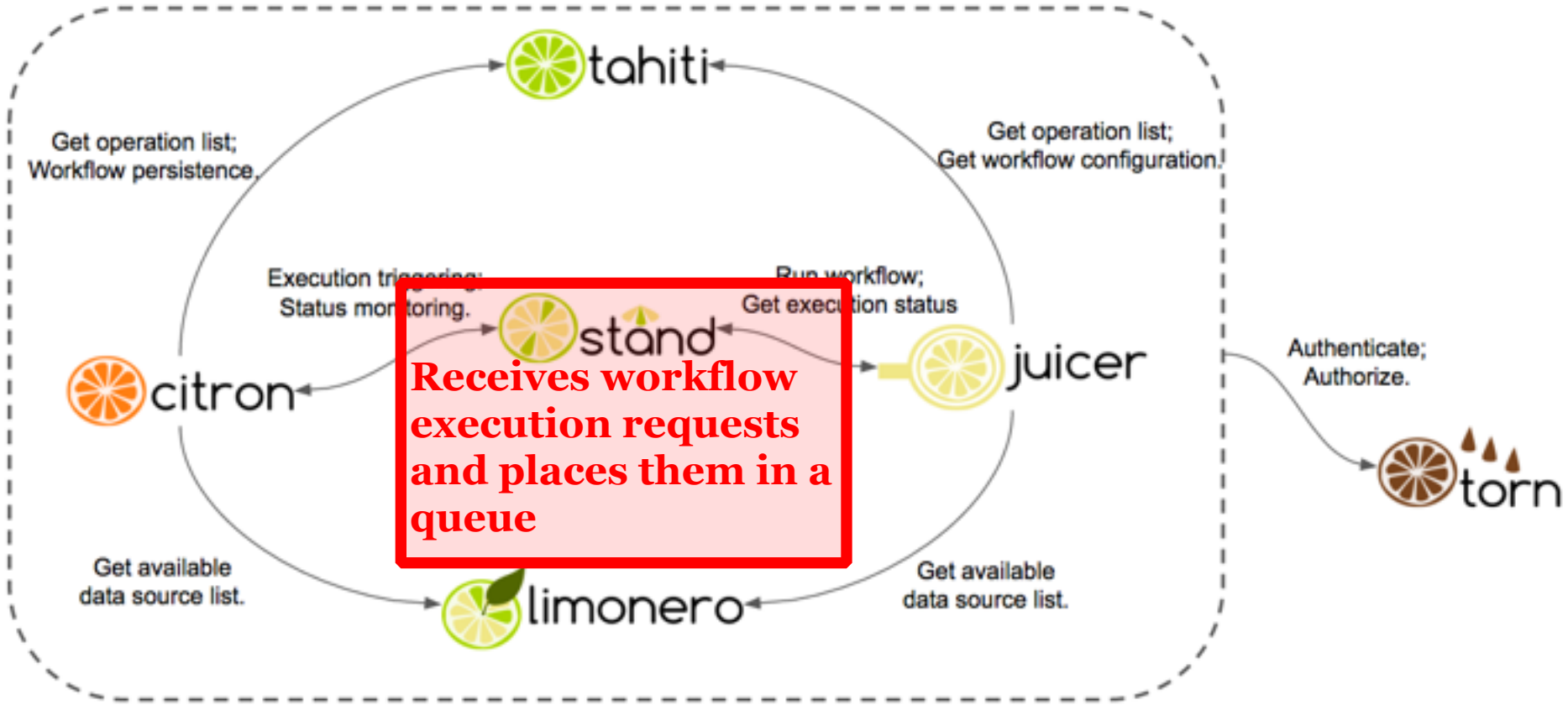
LEMONADE


tahiti
Operations' metadata
Get operation list;
Workflow persistence.
Get operation list;
Get workflow configuration.
Execution triggering;
Status monitoring.
Run workflow;
Get execution status
stand
citron
juicer
Authenticate;
Authorize.
torn
Get available
data source list.
limonero
Get available
data source list.

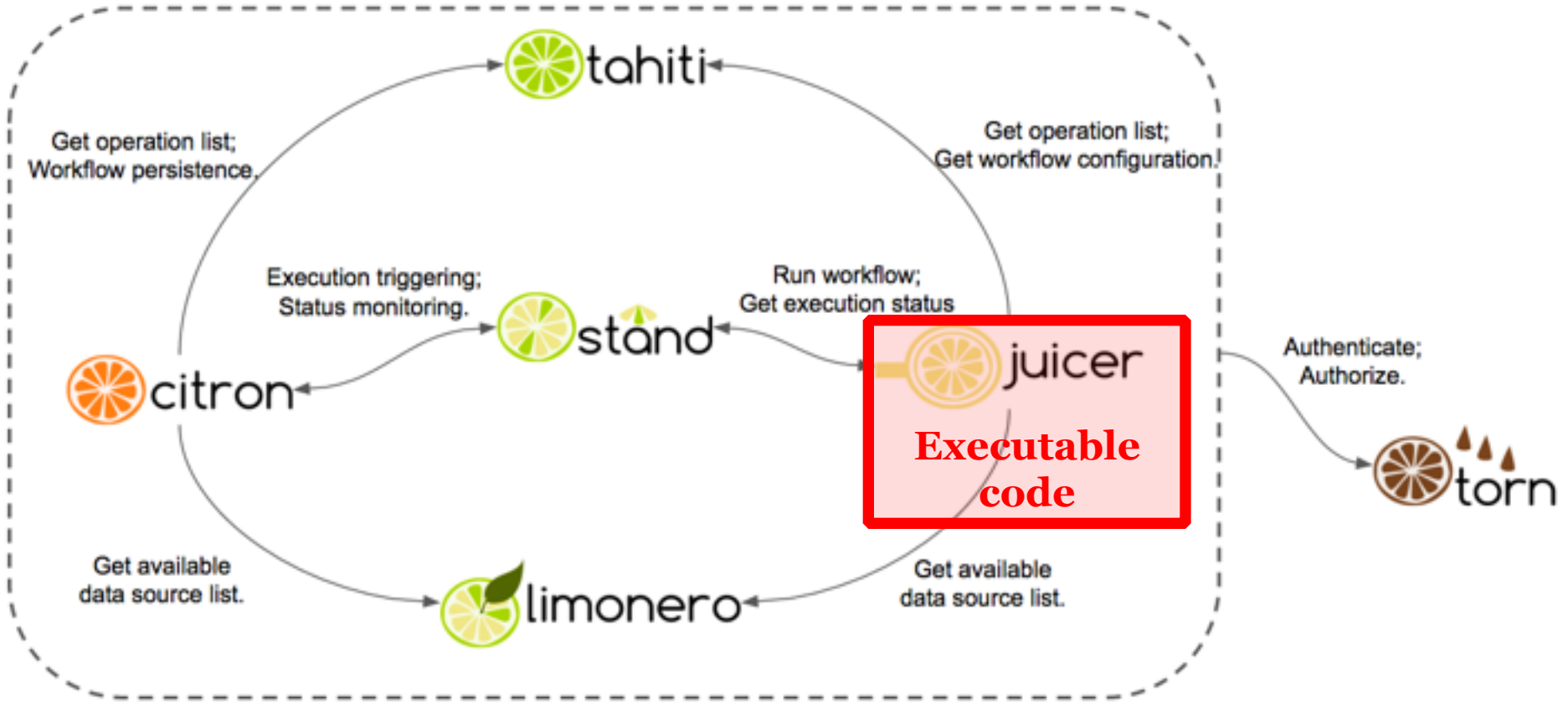
LEMONADE


tahiti
Get operation list;
Workflow persistence.
Get operation list;
Get workflow configuration.
Execution triggering;
Status monitoring.
stand
Run workflow;
Get execution status
citron
juicer
Authenticate;
Authorize.
torn
Get available
data source list.
limonero
Get available
data source list.
Data sources' metadata

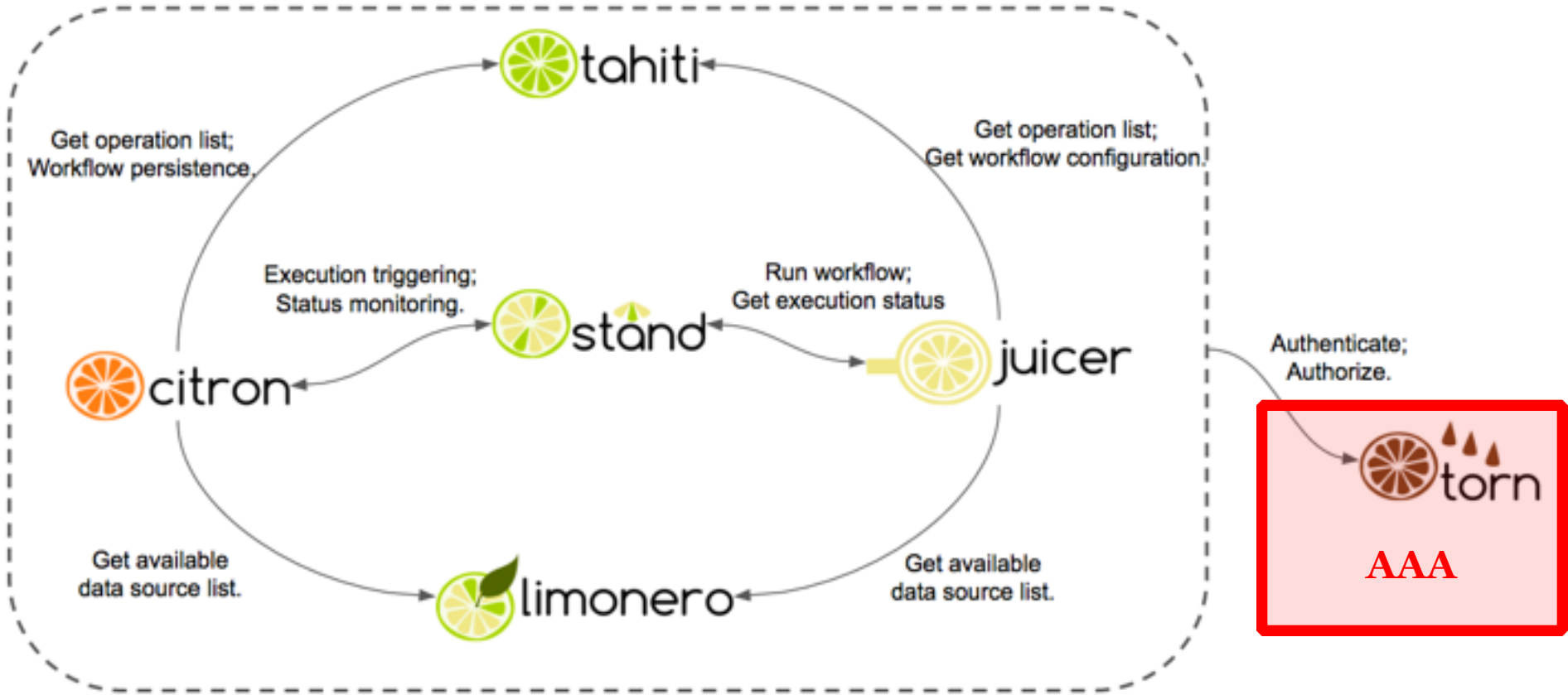
LEMONADE


tahiti
Get operation list;
Workflow persistence.
Get operation list;
Get workflow configuration.
Execution triggering;
Status monitoring.
Run workflow;
Get execution status
stand
citron
juicer
Receives workflow execution requests and places them in a queue
Authenticate;
Authorize.
torn
Get available data source list.
limonero
Get available data source list.

LEMONADE


tahiti
Get operation list;
Workflow persistence.
Get operation list;
Get workflow configuration.
Execution triggering;
Status monitoring.
stand
Run workflow;
Get execution status
citron
juicer
Executable code
Authenticate;
Authorize.
torn
Get available
data source list.
limonero
Get available
data source list.

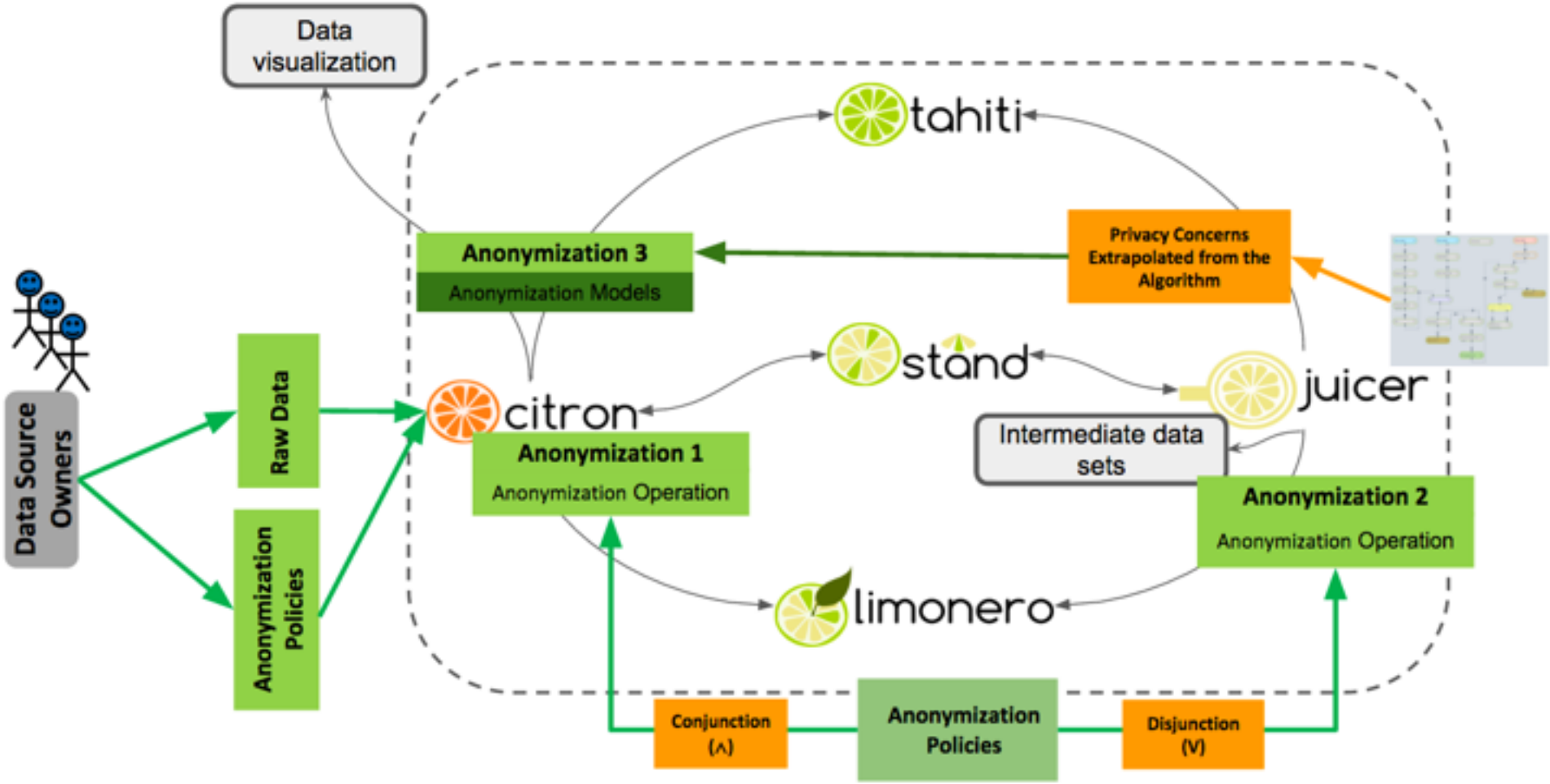
LEMONADE


tahiti
Get operation list;
Workflow persistence.
Get operation list;
Get workflow configuration.
Execution triggering;
Status monitoring.
stand
Run workflow;
Get execution status
citron
juicer
Authenticate;
Authorize.
torn
AAA
Get available
data source list.
limonero
Get available
data source list.

PRIVAaaS integration on LEMONADE

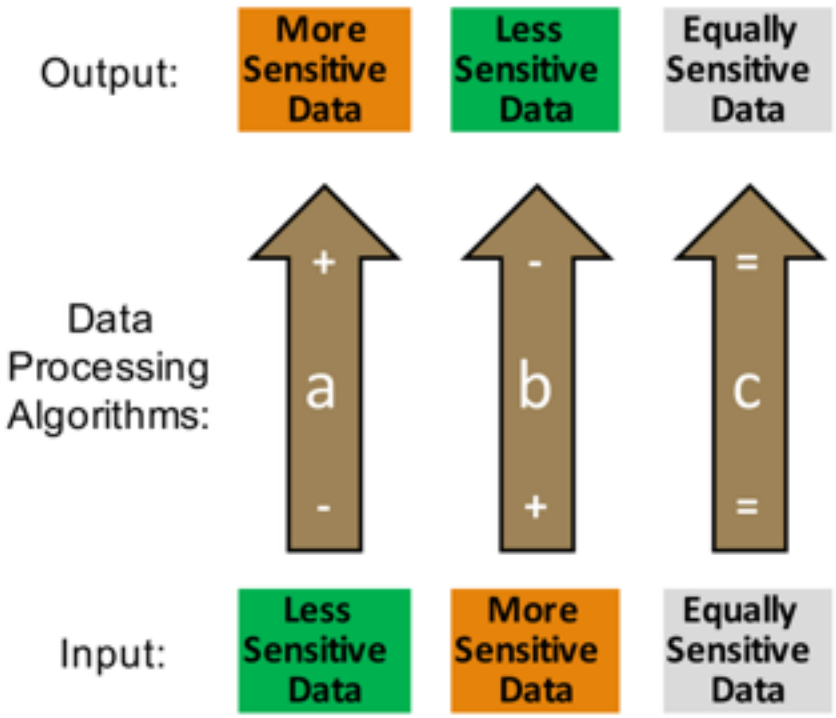
- Enforcement of anonymization policies → data visualization is performed with privacy protection.
- Protect sensible information while deals with data utility, keeping it at the highest possible level
- Divided into three levels, which model different aspects of the anonymization problem

PRIVAaaS integration on LEMONADE



PRIVAaaS integration on LEMONADE

- Control statistical disclosure of data analytics algorithms
- Three scenarios
- Reduce re-identification



Conclusions

- PRIVAaas is a 3-level integration approach
  - supports not only the privacy of native input data
  - supports also the privacy of output data generated by the operations
  - uses anonymization models prior to data visualization

Conclusions

- Use of anonymization polices allows:
  - data source owners to establish their own guidelines
  - applying conjunctive/disjunctive processes, preserving the data utility
  - enforcement guarantees that the rules established by data source owners will be respected

Conclusions

- Provide guarantees not only to data sources, but also to the operations and their outputs
- Go beyond static elements, but also on their dynamic behavior
- Implementation may raise issues that have not been addressed in the literature
  - disruptive platform
  - data analytics applications to an unprecedented level of service

Acknowledgments

- This work has been partially supported by the projects:
  - **EUBra-BIGSEA** (www.eubra-bigsea.eu), funded by the Brazilian Ministry of Science, Technology and Innovation (Project 23614 - MCTI/RNP 3rd Coordinated Call) and by the European Commission under the Cooperation Programme, Horizon 2020 grant agreement no 690116
  - **DEVASSES** (www.devasses.eu), funded by the European Union's FP7 under grant agreement no PIRSES-GA-2013-612569
  - UFMG team is also partially supported by **CNPq, FAPEMIG**, as well as projects **InWeb** and **MASWeb**.

Questions

regina@ft.unicamp.br

taniabasso@ft.unicamp.br
nmsa@dei.uc.pt
mvieira@dei.uc.pt
walter@dcc.ufmg.br
meira@dcc.ufmg.br