DICE: Quality-Aware DevOps For Big Data Applications

GIULIANO CASALE, IMPERIAL COLLEGE LONDON
(On behalf of the DICE consortium)

G.CASALE@IMPERIAL.AC.UK

DevOpsDays Warsaw

DICE Research Project

Horizon 2020 research project (4M€, 2015-18)

Consortium:

XLAB NOT IDLE

Imperial College London

NETFECTIVE Creative Technology

Universidad Zaragoza 1542

@ IeAT

POLITECNICO DI MILANO

prodevelop Integración de tecnologías

flexiOPS

ATC ATHENS TECHNOLOGY CENTER

Our mission: supporting SMEs in developing high-quality Big data applications

DevOpsDays Warsaw

"…AND AFTER BIG DATA SOLVES ALL OUR PROBLEMS, WE'LL RIDE AWAY ON MAGIC FLYING UNICORNS"

Quote: Marketoonist.com

A story from our consortium....

Company

core market:
Legacy software
modernization

A story from our consortium....

New market:
Analysis of legacy
Big data

The barrier: a steep learning curve!

Example: configuring a Big data system...

```
102
103 drpc.port: 3772
104 drpc.worker.threads: 64
105 drpc.max_buffer_size: 1048576
106 drpc.queue.size: 128
107 drpc.invocations.port: 3773
108 drpc.invocations.threads: 64
109 drpc.request.timeout.secs: 600
110 drpc.childopts: "-Xmx768m"
111 drpc.http.port: 3774
112 drpc.https.port: -1
113 drpc.https.keystore.password: ""
114 drpc.https.keystore.type: "JKS"
115 drpc.http.creds.plugin: org.apache.storm.security.auth.DefaultHttpCredentialsPlugi
116 drpc.authorizer.acl.filename: "drpc-auth-acl.yaml"
117 drpc.authorizer.acl.strict: false
118
119 transactional.zookeeper.root: "/transactional"
120 transactional.zookeeper.servers: null
121 transactional.zookeeper.port: null
122
123 ## blobstore configs
124 supervisor.blobstore.class: "org.apache.storm.blobstore.NimbusBlobStore"
125 supervisor.blobstore.download.thread.count: 5
126 supervisor.blobstore.download.max_retries: 3
127 supervisor.localizer.cache.target.size.mb: 10240
128 supervisor.localizer.cleanup.interval.ms: 600000
129
```

!

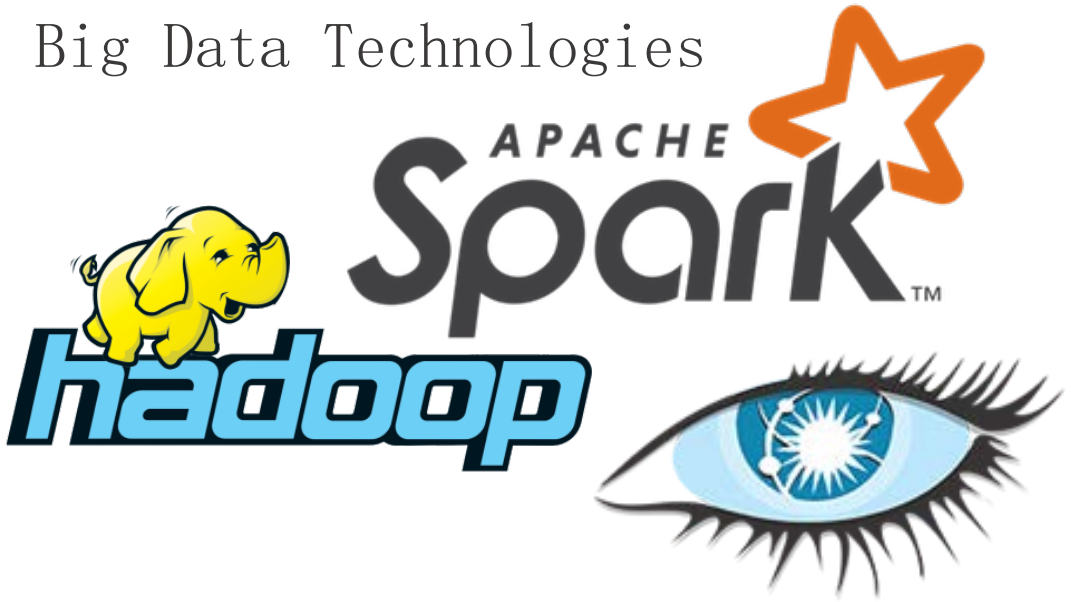
Riddles for the brave

- Will my Big data application s
- How do I select the right Big technology?
- How do I choose the best architecture? And how to verify that I am right?
- What if service-level agreements change?

Our Concerns

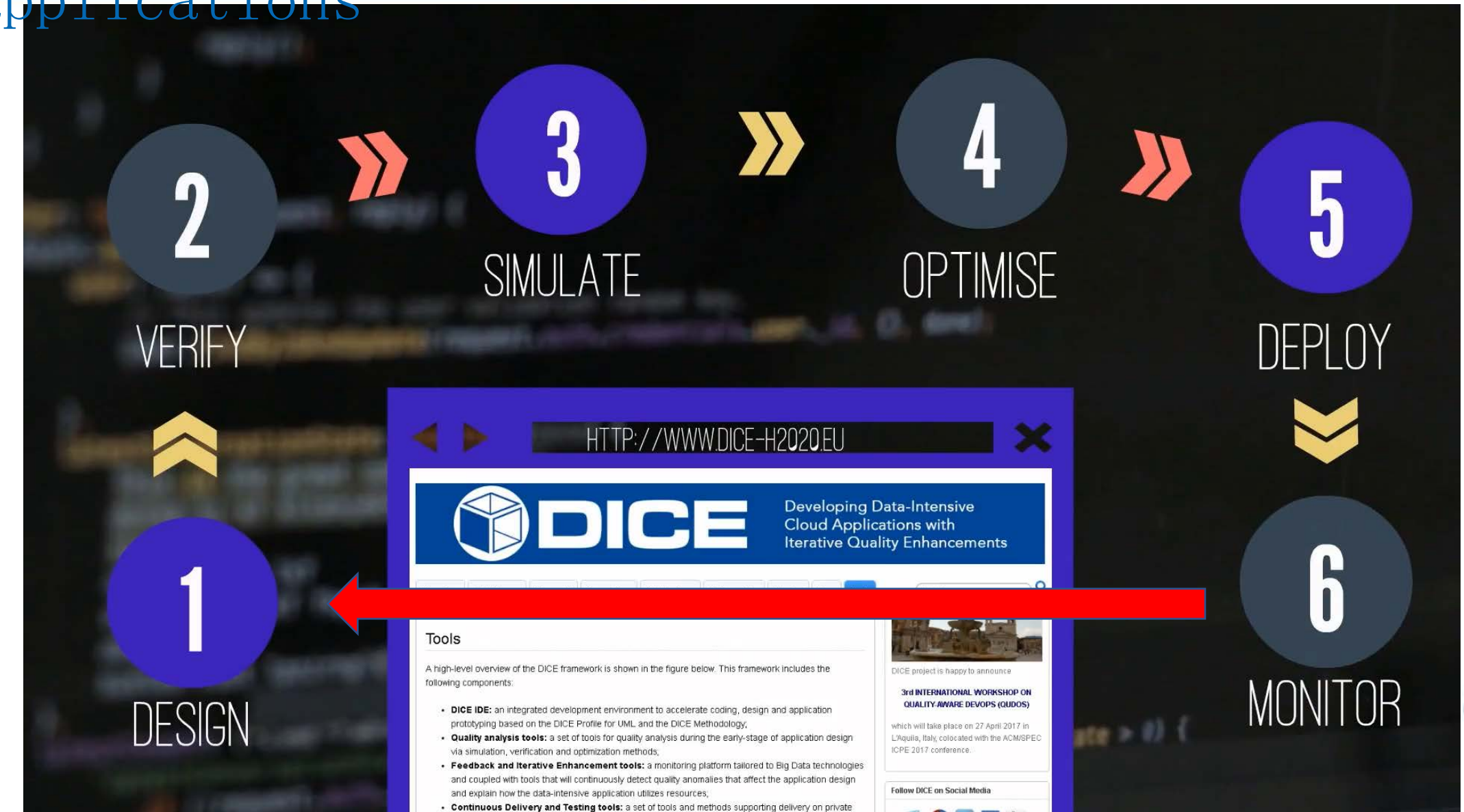
Characterize Data Properties

DevOps tools for popular Big Data Technologies



Quality-Aware Reasoning Tools

DICE: a DevOps toolchain for Big data applications



Different perspectives

Developers

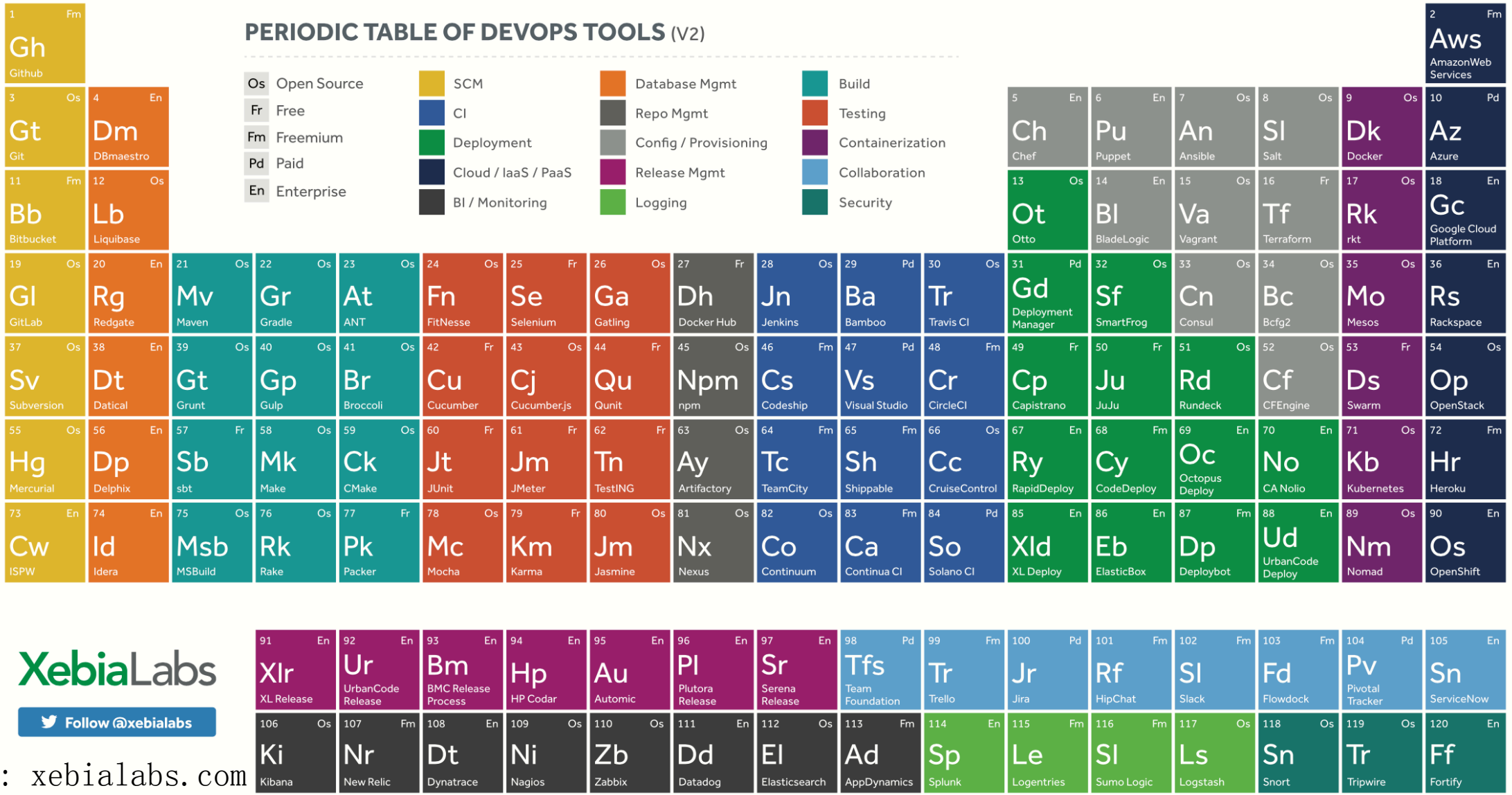
Ops

UNIFIED MODELING LANGUAGE™

Architects

Testers

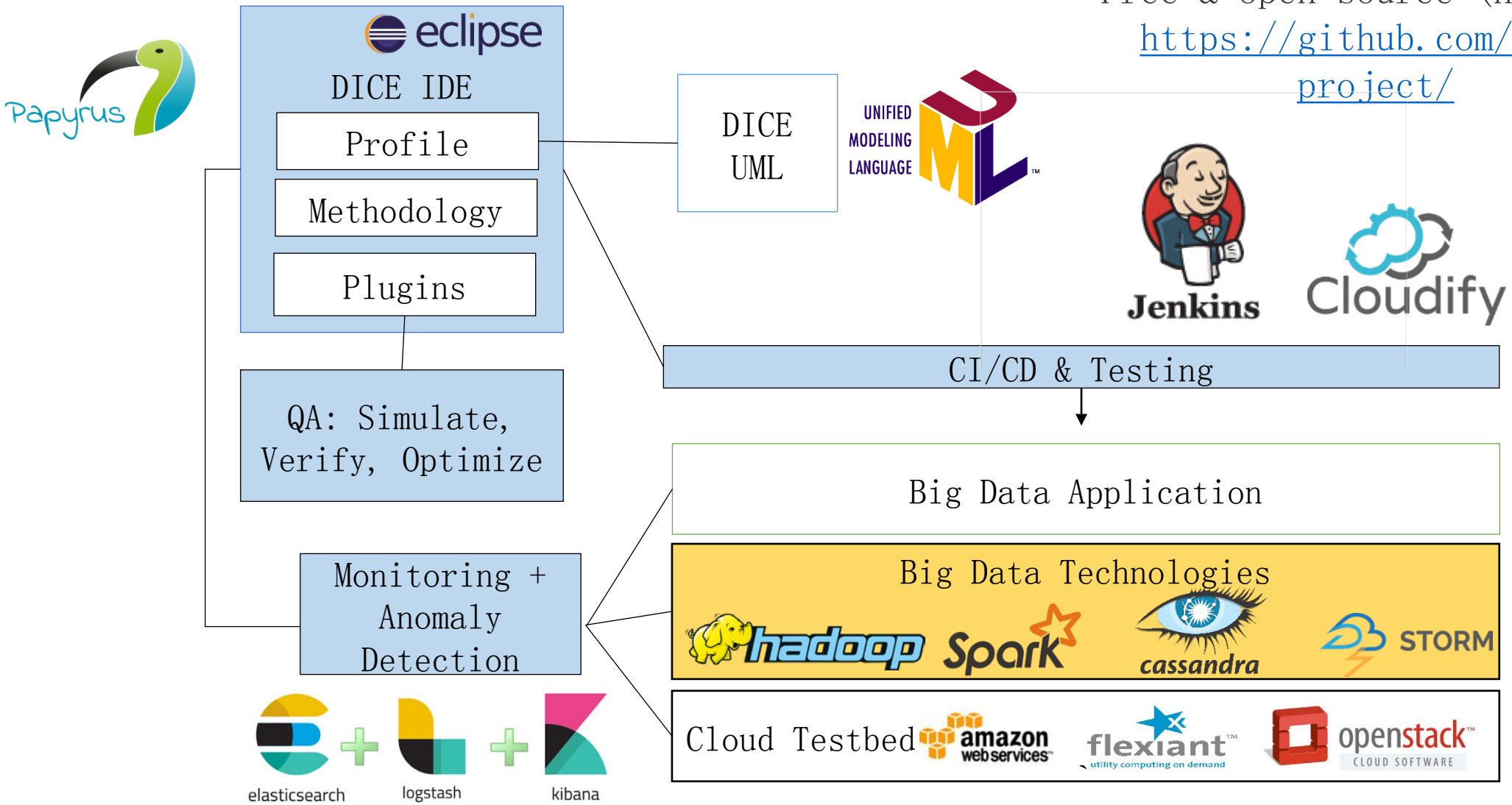
The problem with creating a unified view



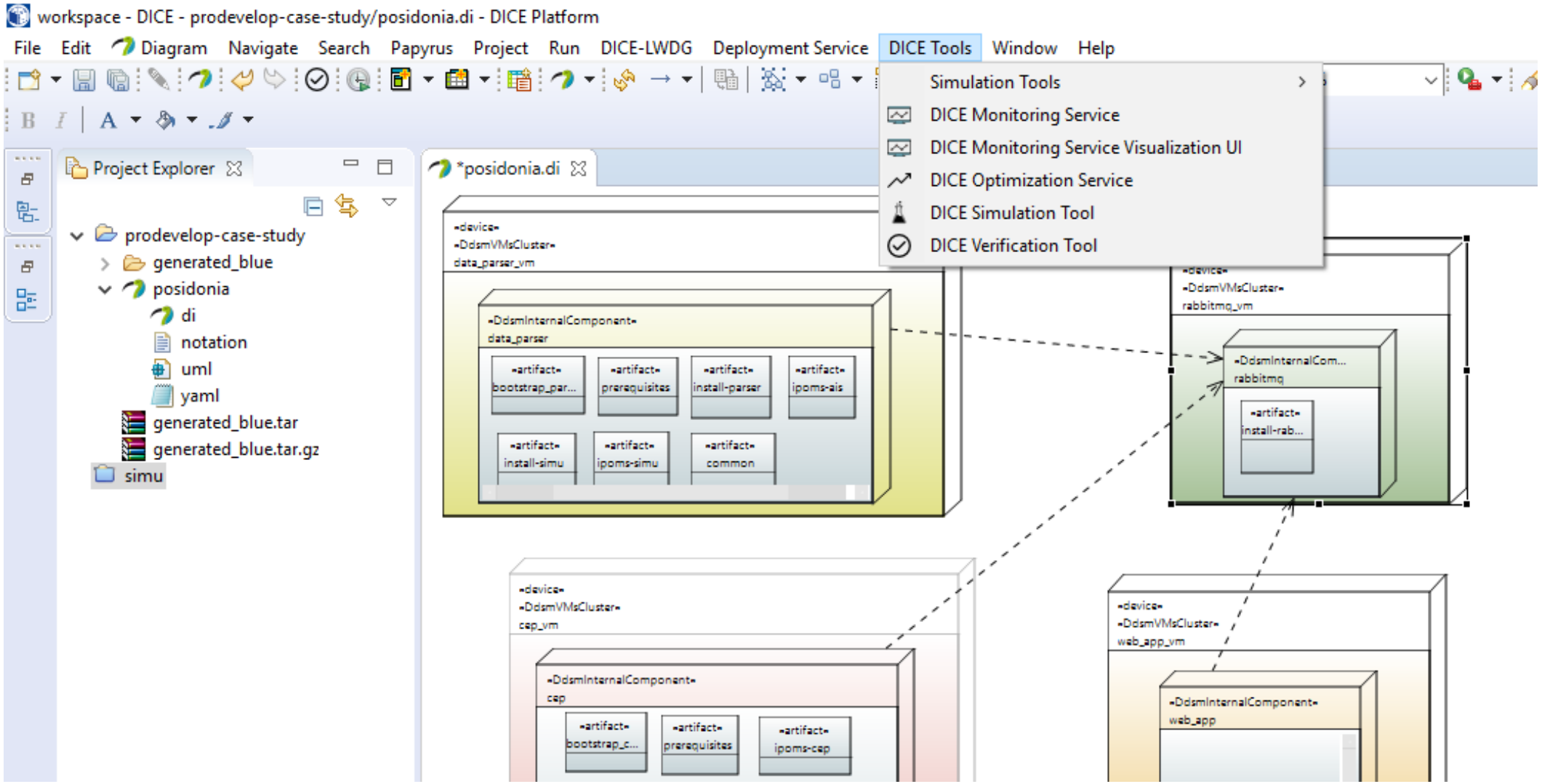
source: xebialabs.com

The DICE Framework: Model-Driven DevOps

Free & open source (non-GPL)
https://github.com/dice-project/

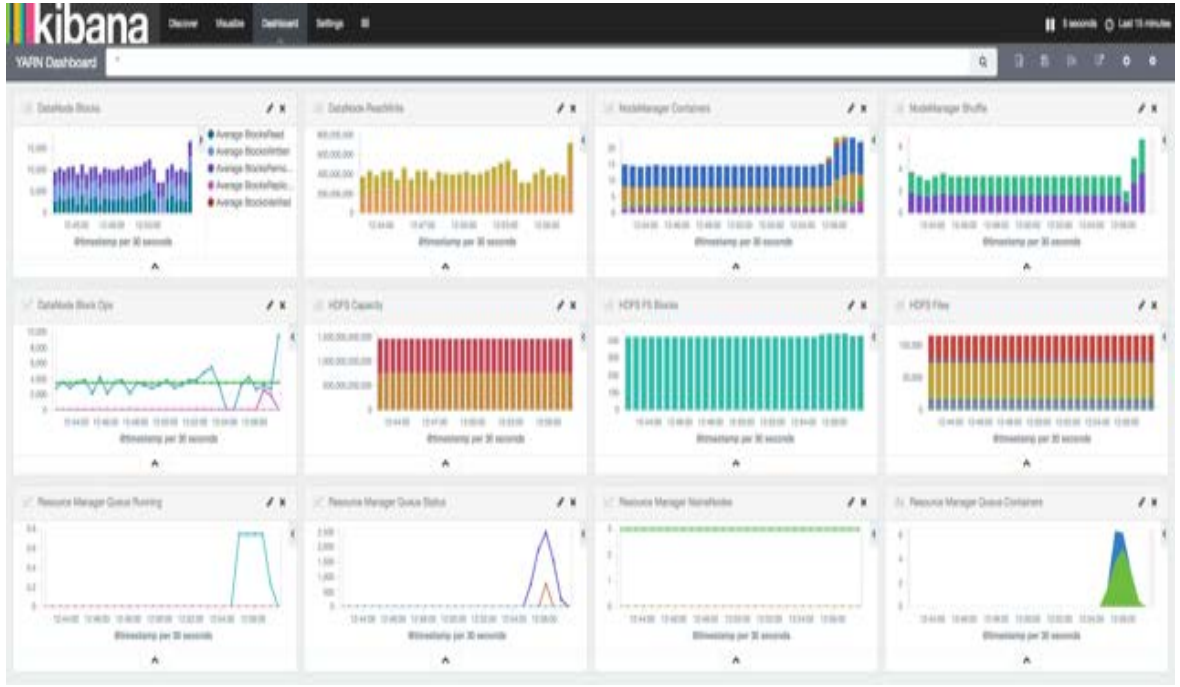


Dev View

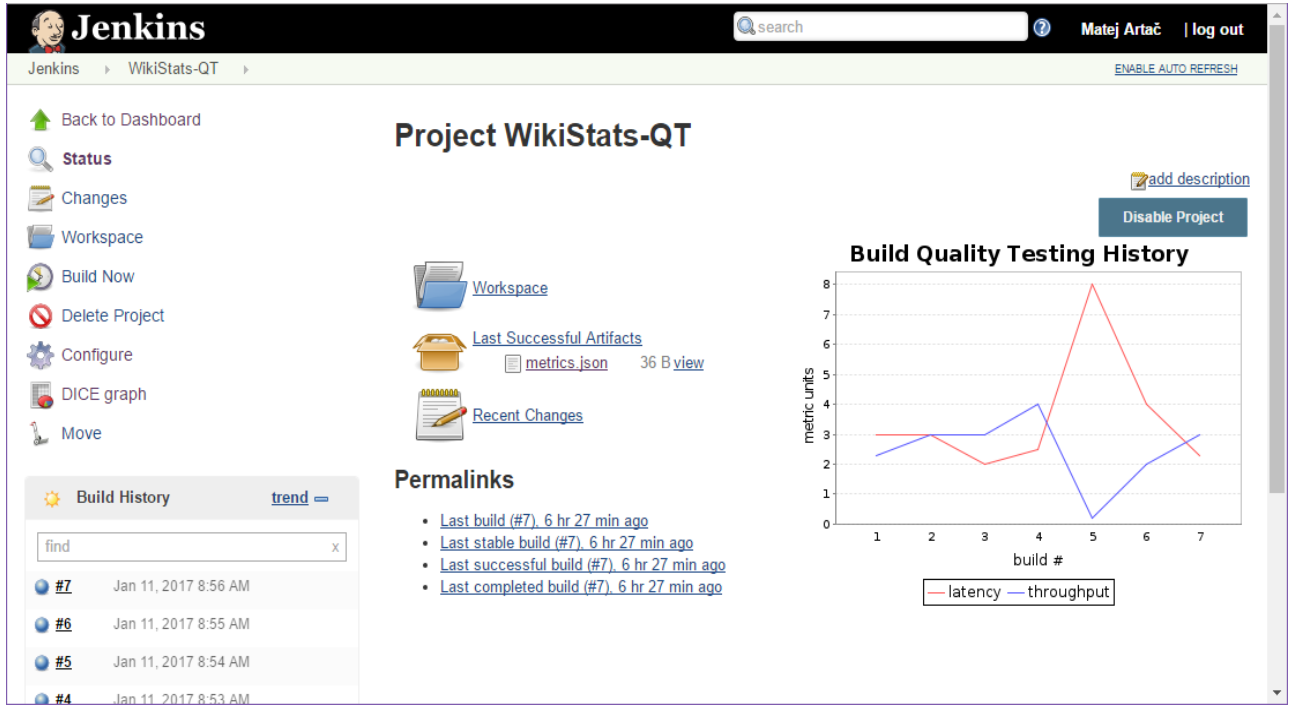


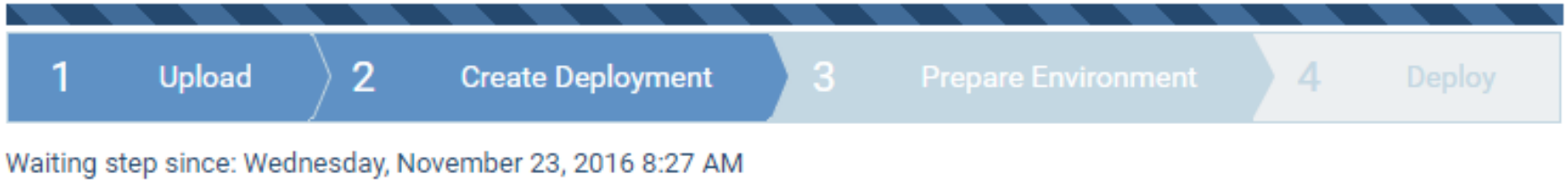
Ops Views

Kibana



Jenkins testing

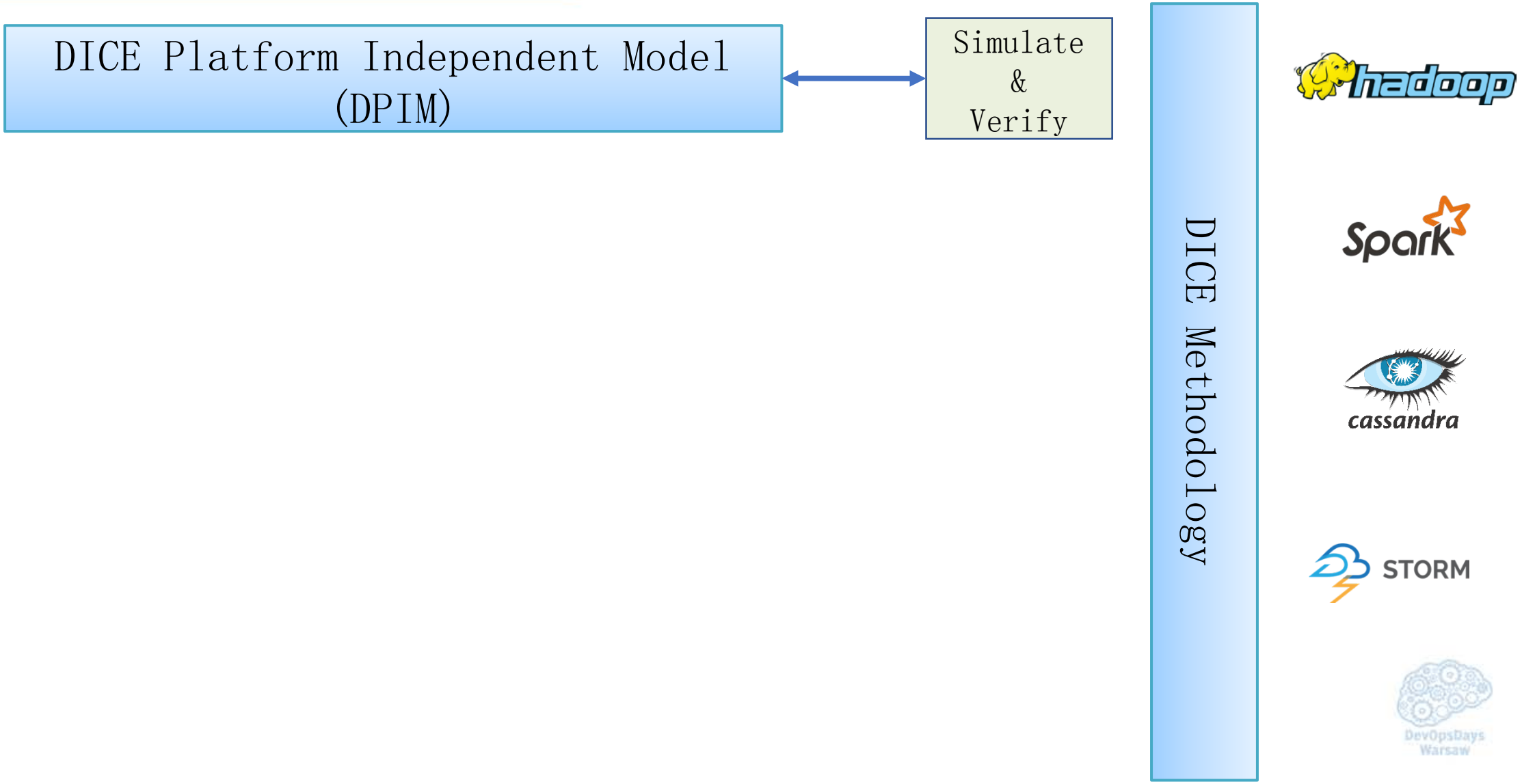




Cloudify

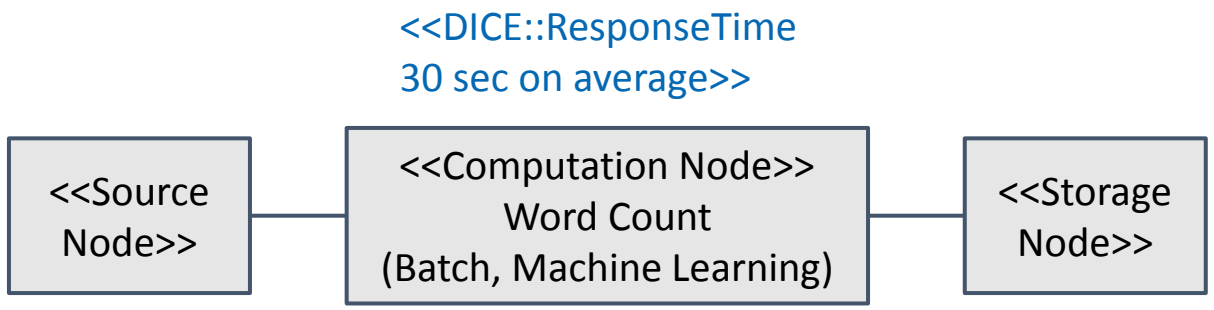
DevOpsDays Warsaw

Behind the Scenes: Dev side

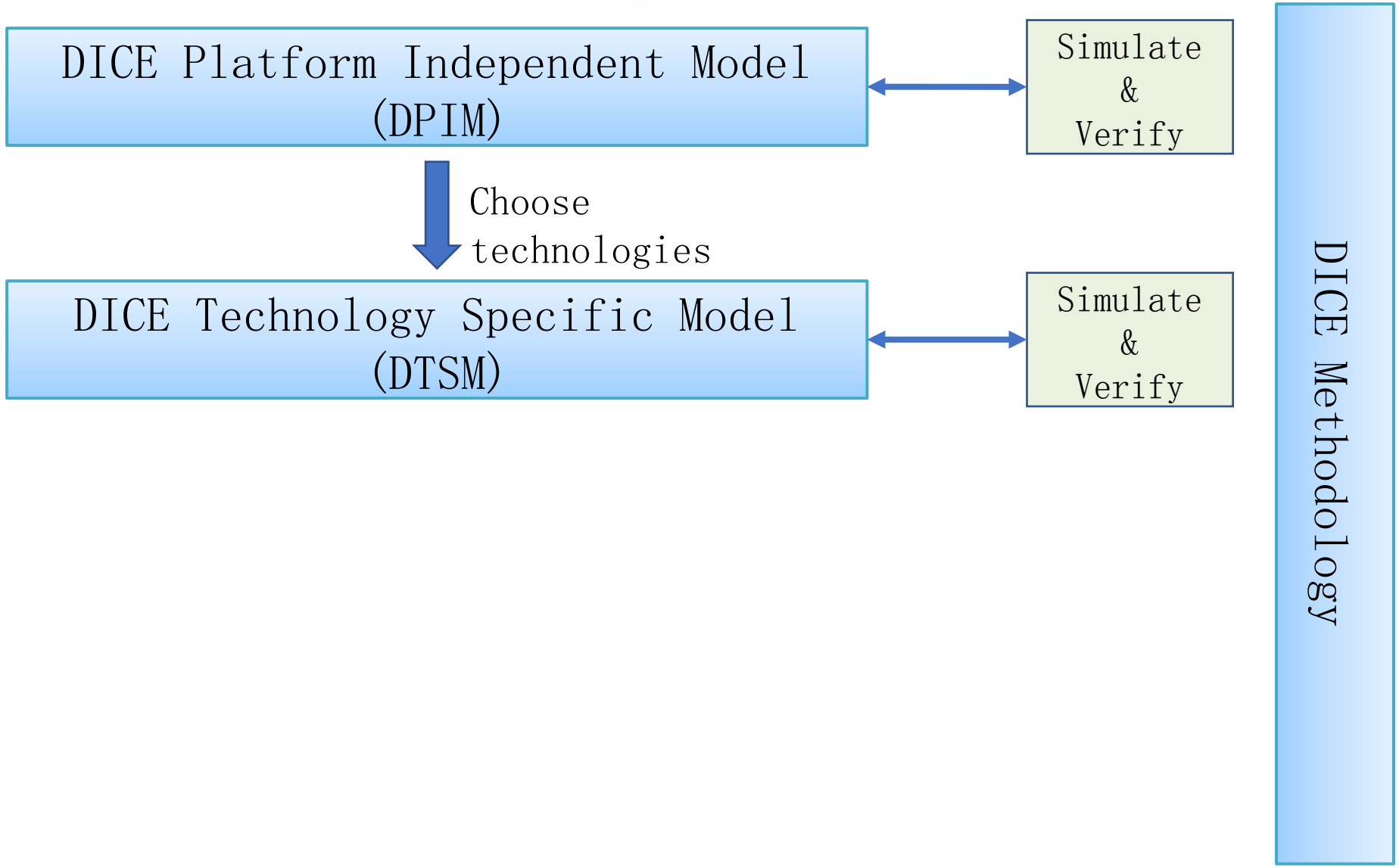


Wordcount example: platform independent model

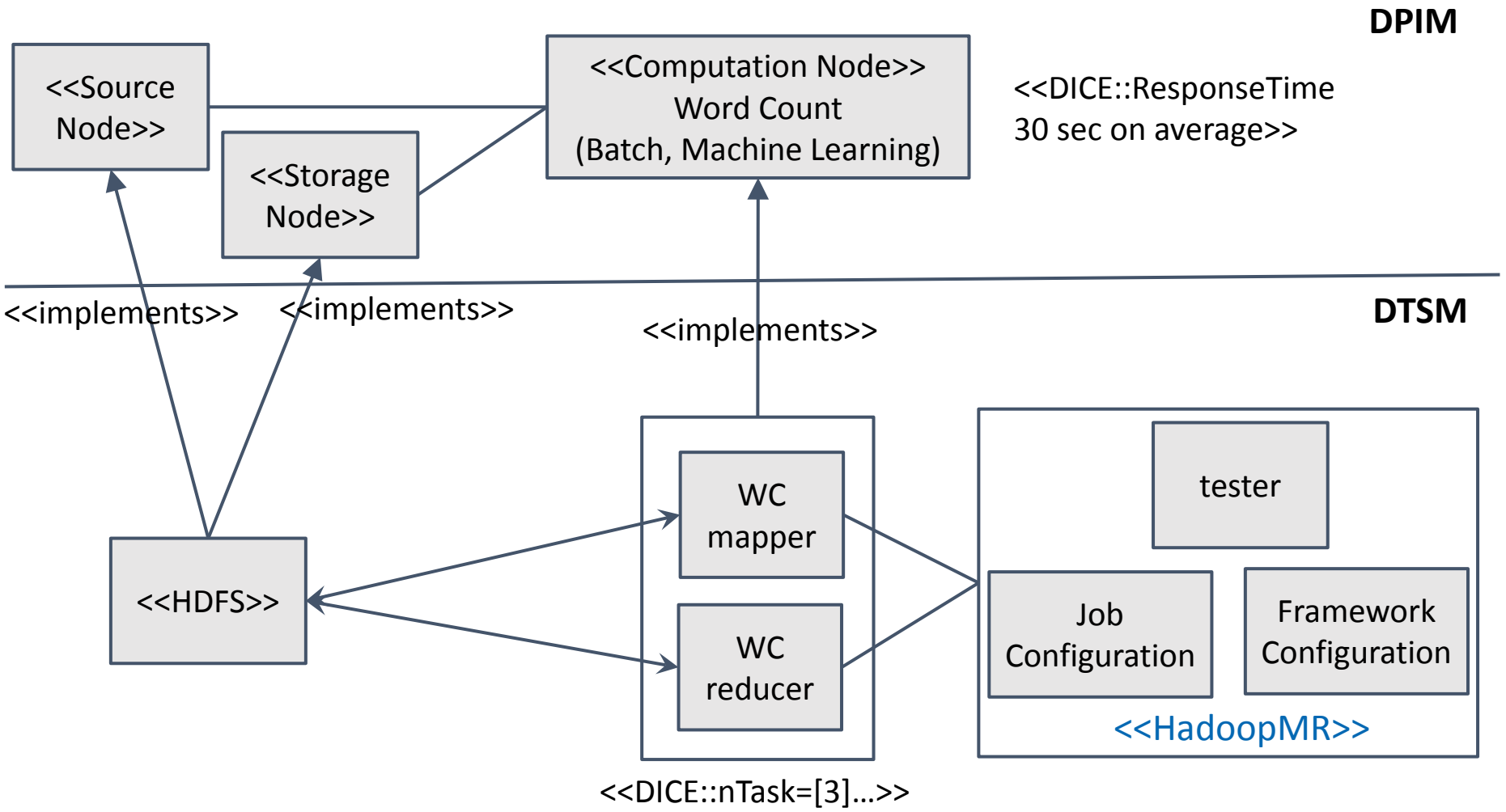
DPIM



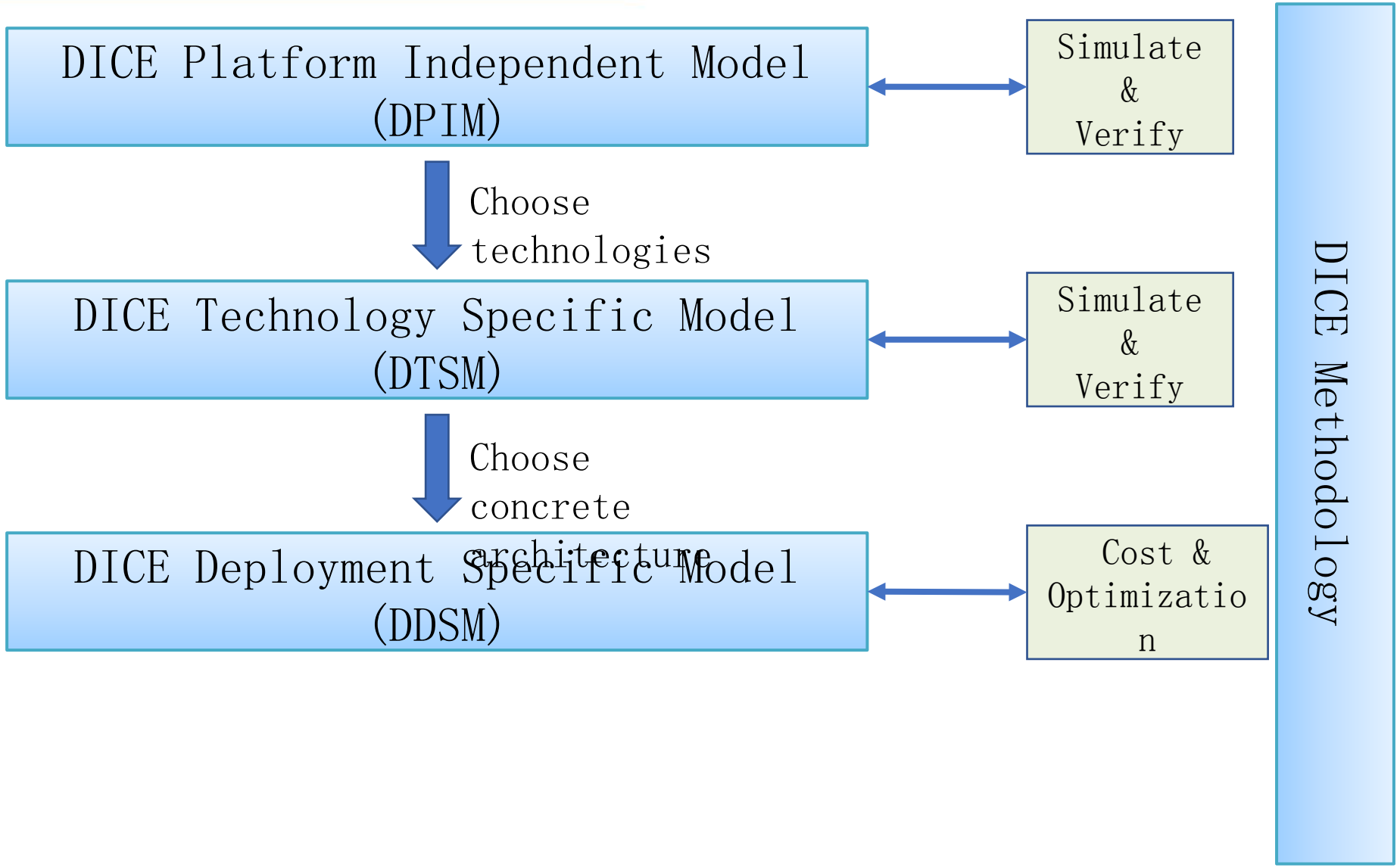
Behind the Scenes: Dev side



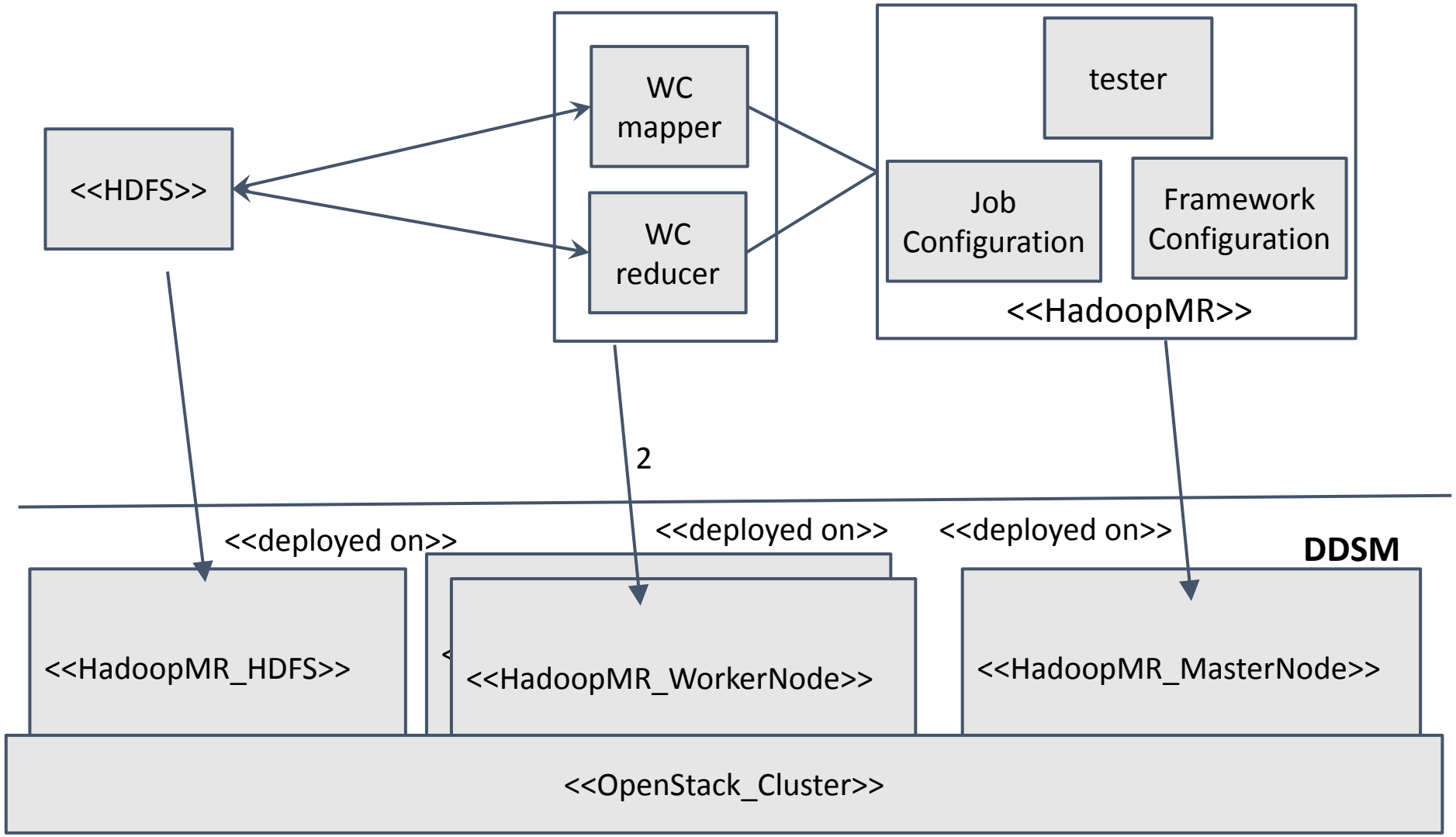
Wordcount example: technology specific model



Behind the Scenes: Dev side

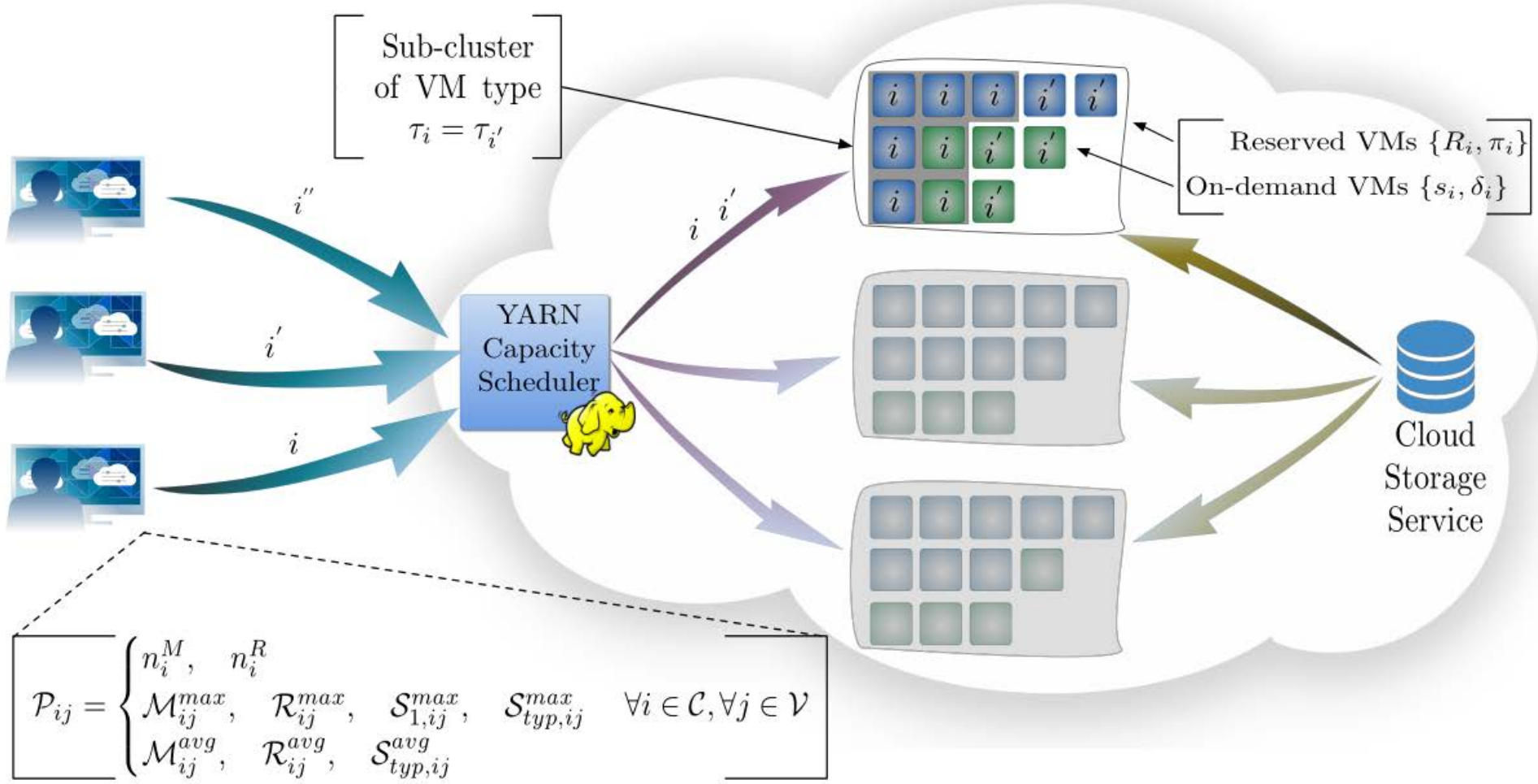


Wordcount example: deployment specific model

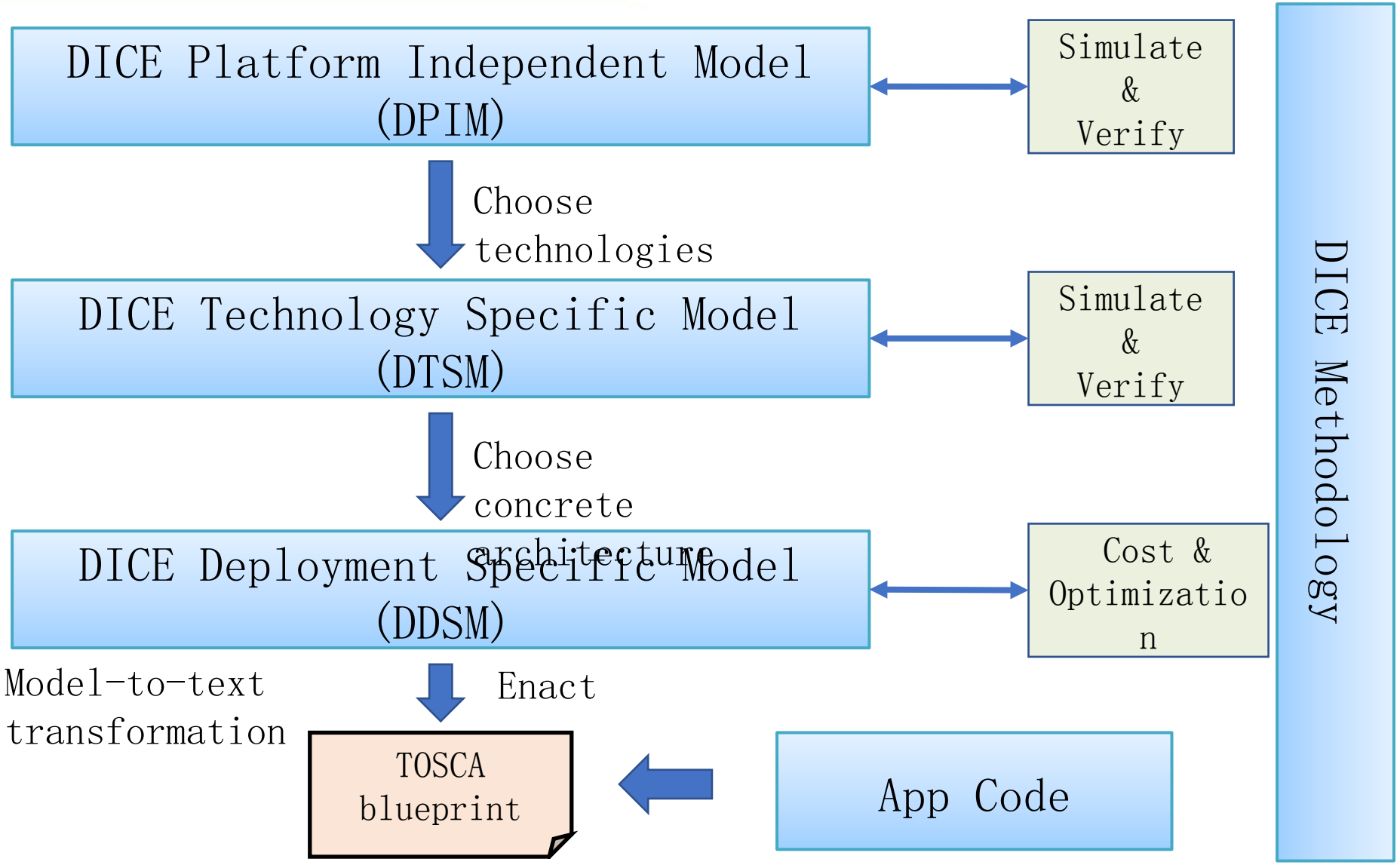


Simulation, Optimization, Verification

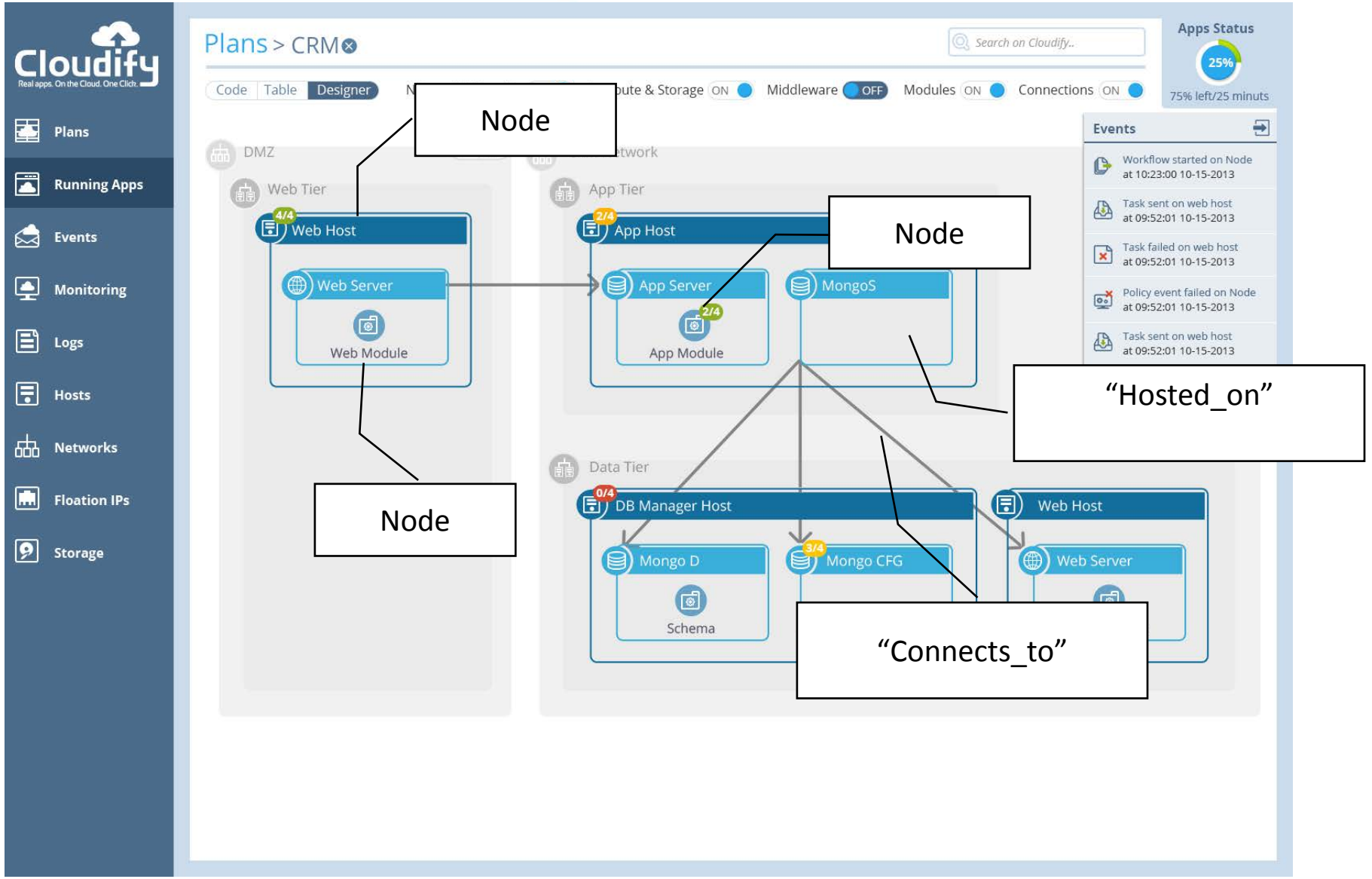
Example: Cost Optimization



Behind the Scenes: Dev side



Deployment: OASIS TOSCA & Cloudify

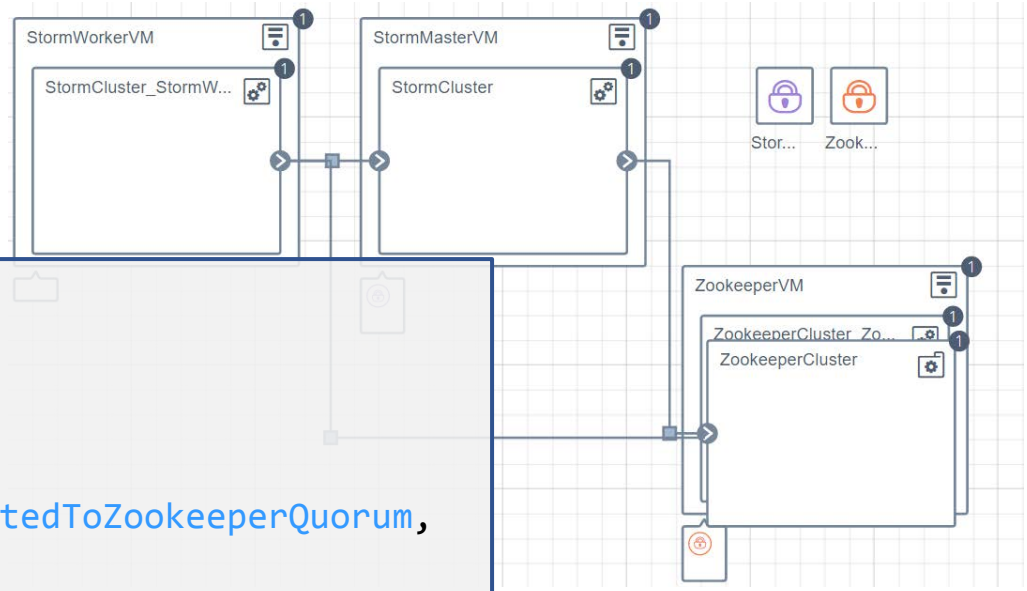


DICE TOSCA blueprints

Powered by Cloudify

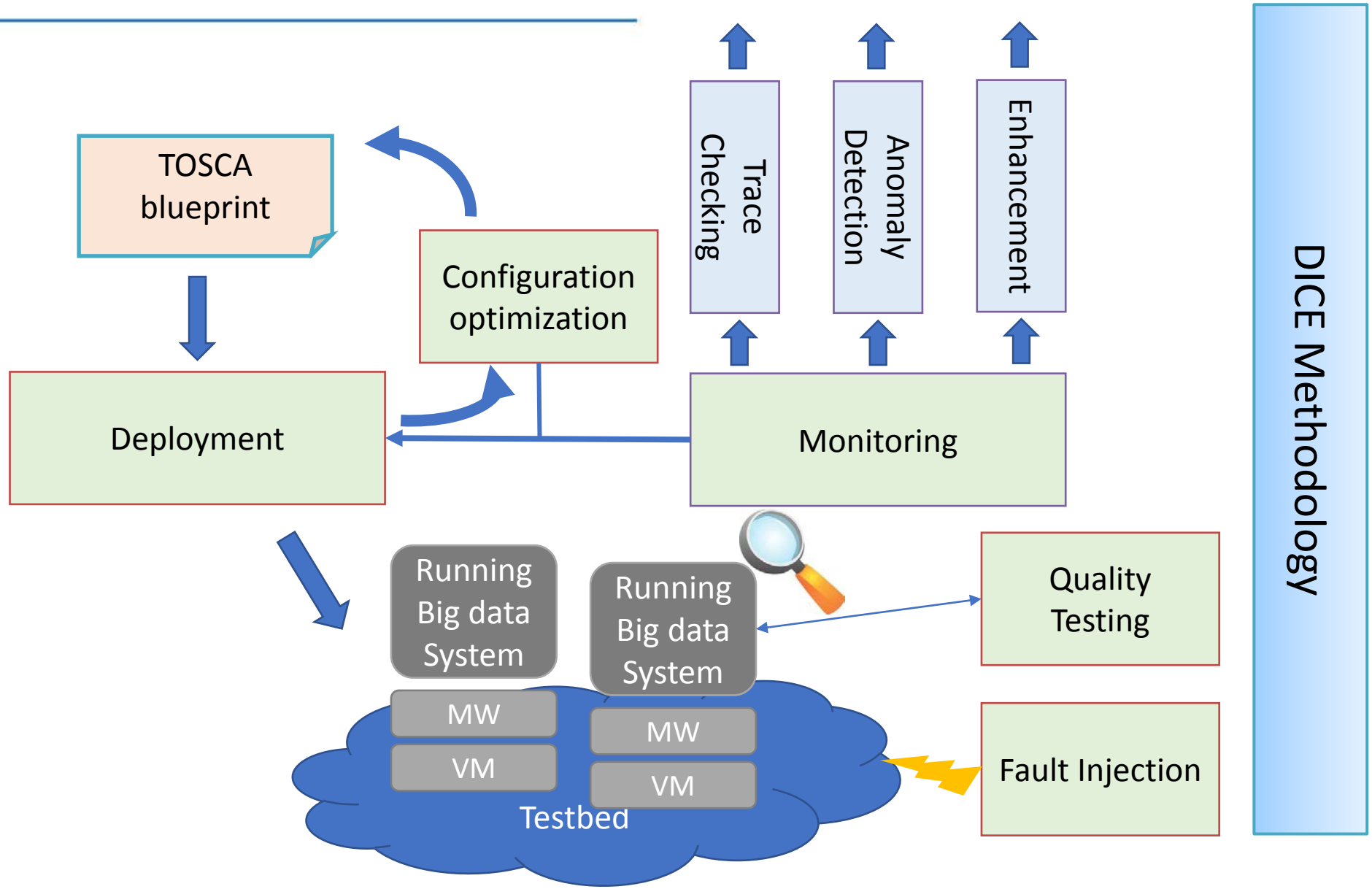
```
tosca_definitions_version: cloudify_dsl_1_3
imports: [
  'http://dice-project.github.io/DICE-Deployment-Cloudify/spec/openstack/0.1.4/plugin.yaml']
node_templates:
  ZookeeperVM:
    type: dice.hosts.Small
    instances: {deploy: 1}
    relationships:
    - {type: dice.relationships.ProtectedBy, target: ZookeeperCluster_ZookeeperVM_worker_firewall}
  StormMasterVM:
    type: dice.hosts.Small
    relationships:
    - {type: dice.relationships.ProtectedBy, target: StormCluster_firewall}
  StormWorkerVM:
    type: dice.hosts.Small
    instances: {deploy: 1}
  StormCluster:
    type: dice.components.storm.Nimbus
    relationships:
    - {type: dice.r
    - {type: dice.r
    properties:
      configuratio
        queueSize:
  StormCluster_firewall:
    type: dice.firewall_rules.storm.Nimbus
  StormCluster_StormWorkerVM_slave:
    type: dice.components.storm.Worker
    relationships:
    - {type: dice.r
    - {type: dice.r
    - {type: dice.r
    properties:
      configuratio
        memoryCapac
  ZookeeperCluster_ZookeeperVM_worker:
    type: dice.components.zookeeper.Server
    relationships:
    - {type: dice.r
    - {type: dice.r
    properties:
      configuration: {tickTime: '1500', initLimit: '10', syncLimit: '5'}
  ZookeeperCluster_ZookeeperVM_worker_firewall:
    type: dice.firewall_rules.zookeeper.Server
  ZookeeperCluster:
    type: dice.components.zookeeper.Quorum
    relationships:
    - {type: dice.relationships.zookeeper.QuorumContains, target: ZookeeperVM}
```

```
StormCluster:
    type: dice.components.storm.Nimbus
    relationships:
    - {type: dice.relationships.ContainedIn,
       target: StormMasterVM}
    - {type: dice.relationships.storm.ConnectedToZookeeperQuorum,
       target: ZookeeperCluster}
    properties:
      configuration: {taskTimeout: '30', supervisorTimeout: '60',
                      monitorFrequency: '10', queueSize: '100000',
                      retryTimes: '5', retryInterval: '2000'}
```



DevOpsDays Warsaw

Behind the Scenes: Ops side



DICE Ops Tools: configuration optimization for Big data

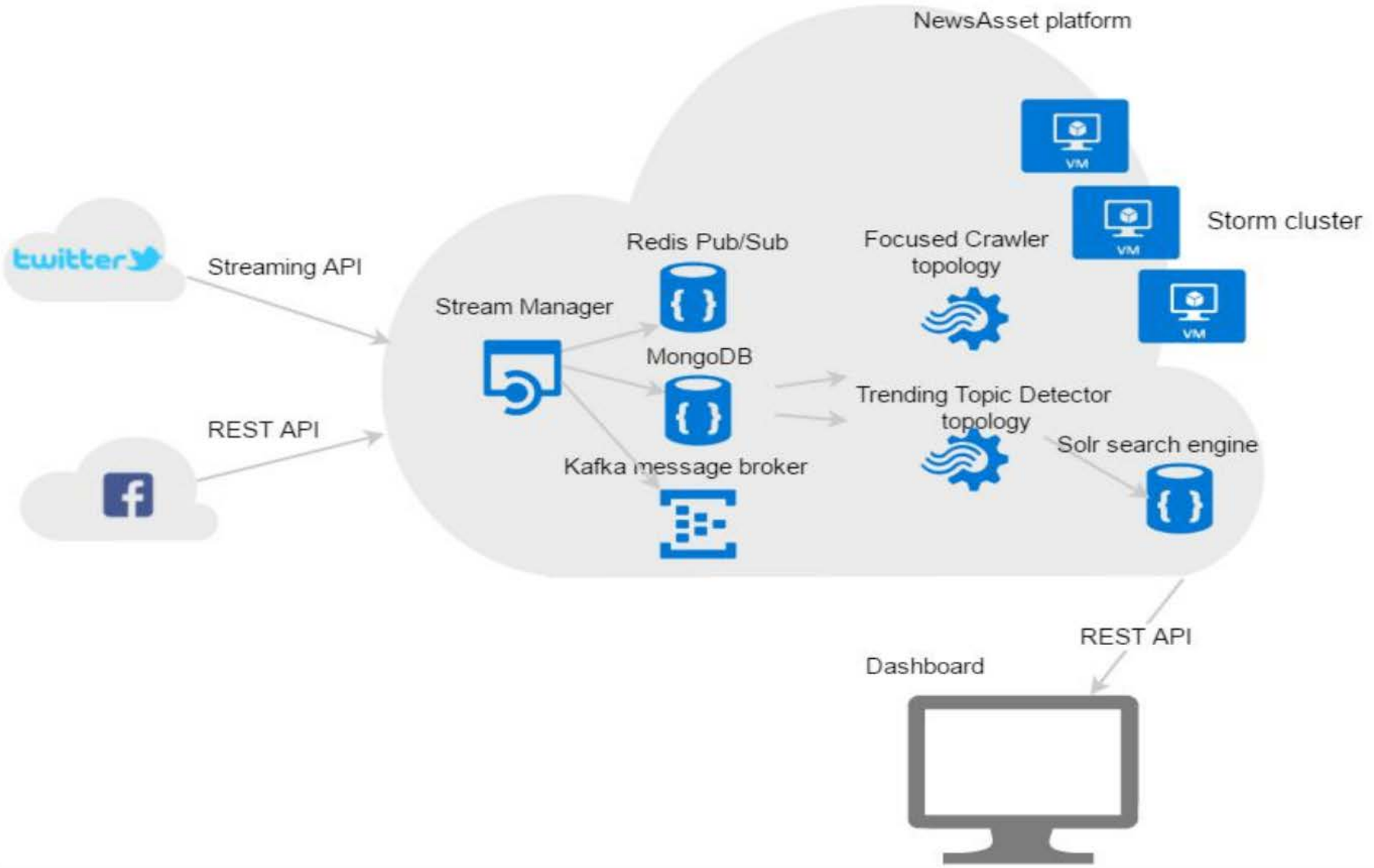
Example: configuring a Big data system...

```
102
103 drpc.port: 3772
104 drpc.worker.threads: 64
105 drpc.max_buffer_size: 1048576
106 drpc.queue.size: 128
107 drpc.invocations.port: 3773
108 drpc.invocations.threads: 64
109 drpc.request.timeout.secs: 600
110 drpc.childopts: "-Xmx768m"
111 drpc.http.port: 3774
112 drpc.https.port: -1
113 drpc.https.keystore.password: ""
114 drpc.https.keystore.type: "JKS"
115 drpc.http.creds.plugin: org.apache.storm.security.auth.DefaultHttpCredentialsPlugi
116 drpc.authorizer.acl.filename: "drpc-auth-acl.yaml"
117 drpc.authorizer.acl.strict: false
118
119 transactional.zookeeper.root: "/transactional"
120 transactional.zookeeper.servers: null
121 transactional.zookeeper.port: null
122
123 ## blobstore configs
124 supervisor.blobstore.class: "org.apache.storm.blobstore.NimbusBlobStore"
125 supervisor.blobstore.download.thread.count: 5
126 supervisor.blobstore.download.max_retries: 3
127 supervisor.localizer.cache.target.size.mb: 10240
128 supervisor.localizer.cleanup.interval.ms: 600000
129
```

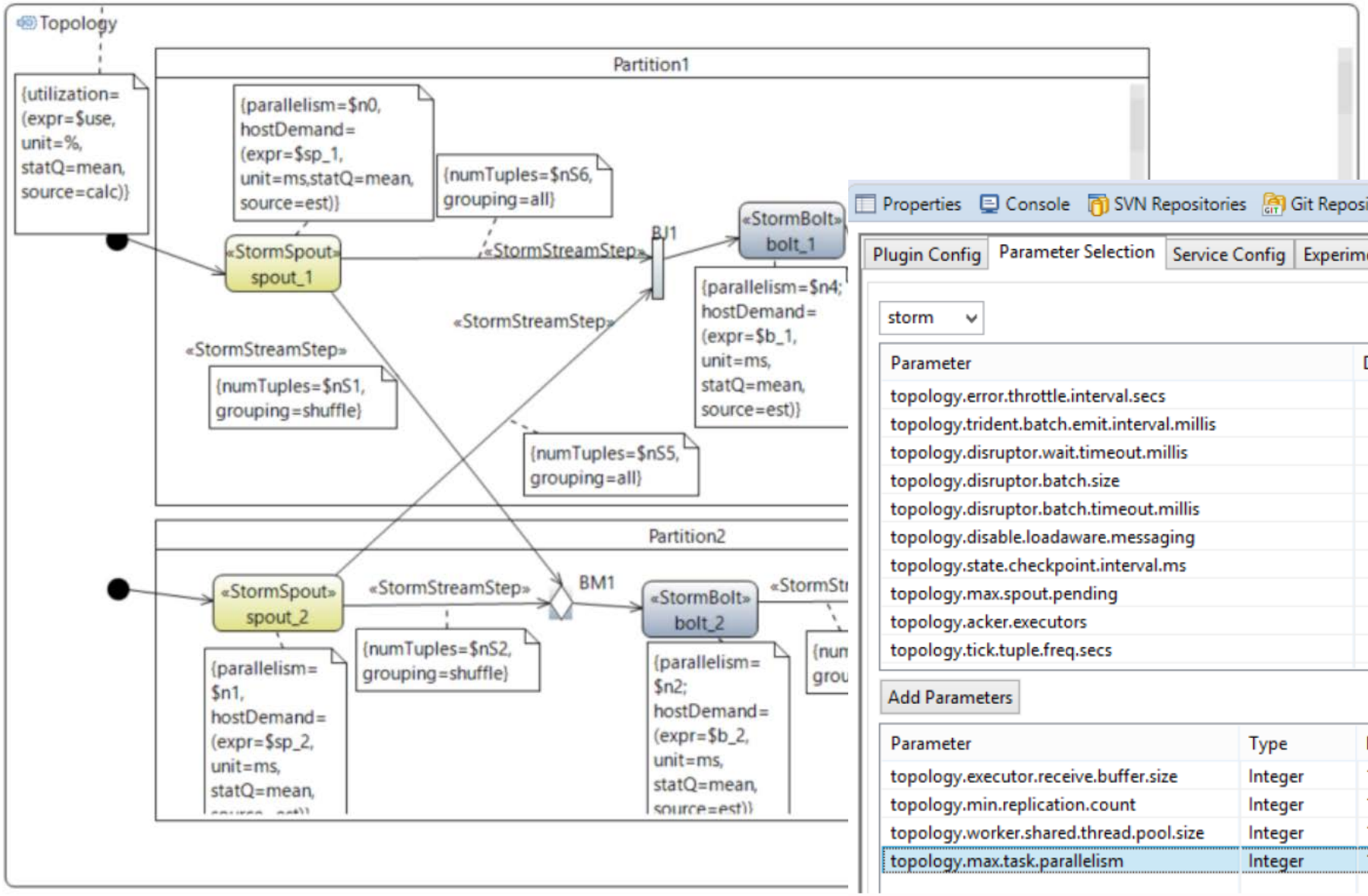
!

NewsAsset platform

News & social media analysis



DICE configuration optimization



Properties | Console | SVN Repositories | Git Repositories | Invocations Registry | DICE Configuration View

Plugin Config | Parameter Selection | Service Config | Experiment Config | App Config | Experiments

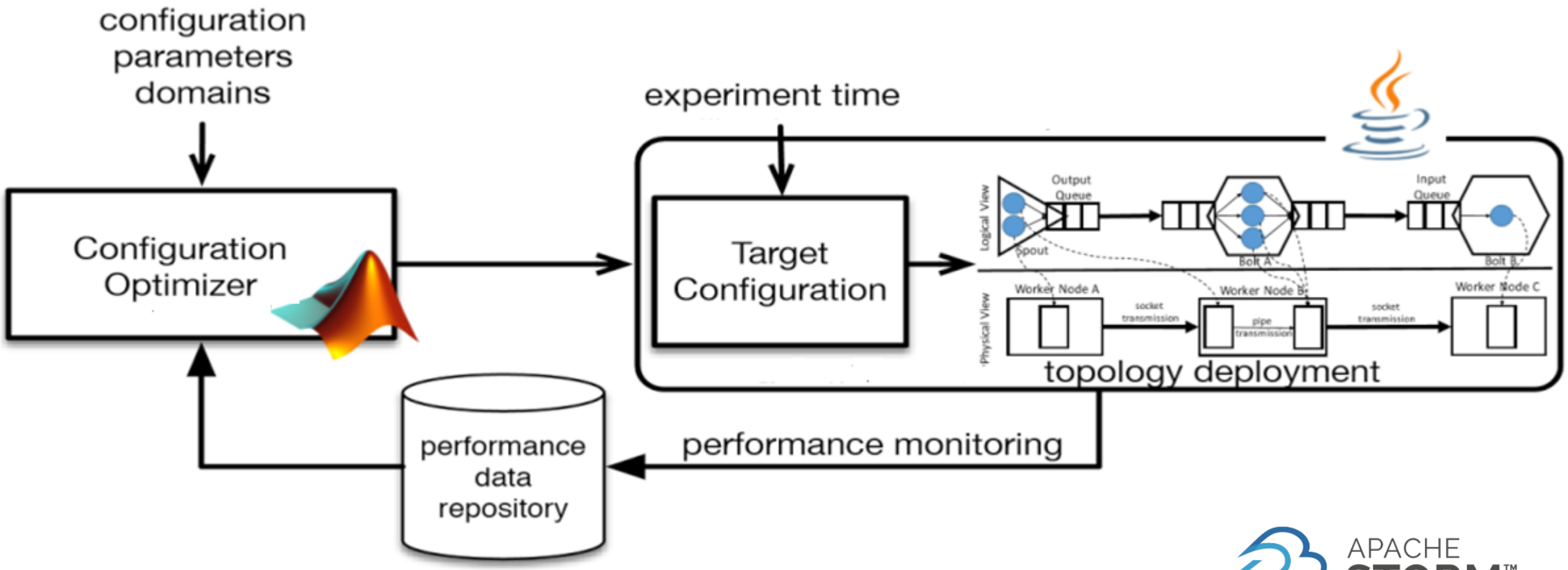
storm

Parameter	Description
topology.error.throttle.interval.secs	
topology.trident.batch.emit.interval.millis	
topology.disruptor.wait.timeout.millis	
topology.disruptor.batch.size	
topology.disruptor.batch.timeout.millis	
topology.disable.loadaware.messaging	
topology.state.checkpoint.interval.ms	
topology.max.spout.pending	
topology.acker.executors	
topology.tick.tuple.freq.secs	

Add Parameters

Parameter	Type	Min	Max	Step	Options
topology.executor.receive.buffer.size	Integer	1024	2048	1	
topology.min.replication.count	Integer	1	10	1	
topology.worker.shared.thread.pool.size	Integer	1	20	1	
topology.max.task.parallelism	Integer	1	50	1	

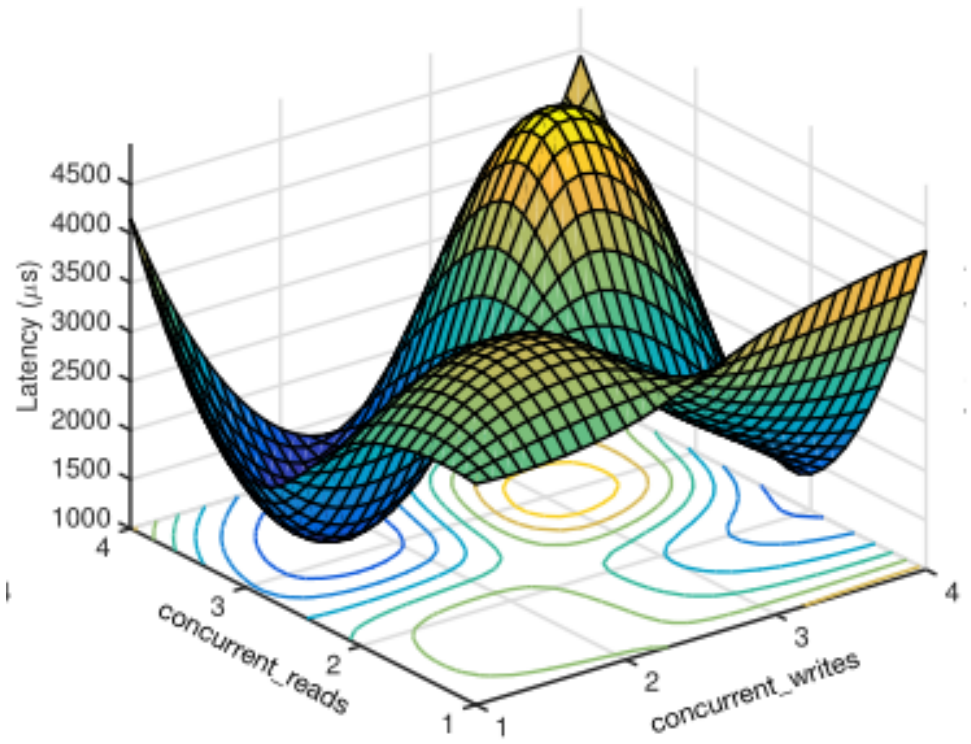
DICE configuration optimization

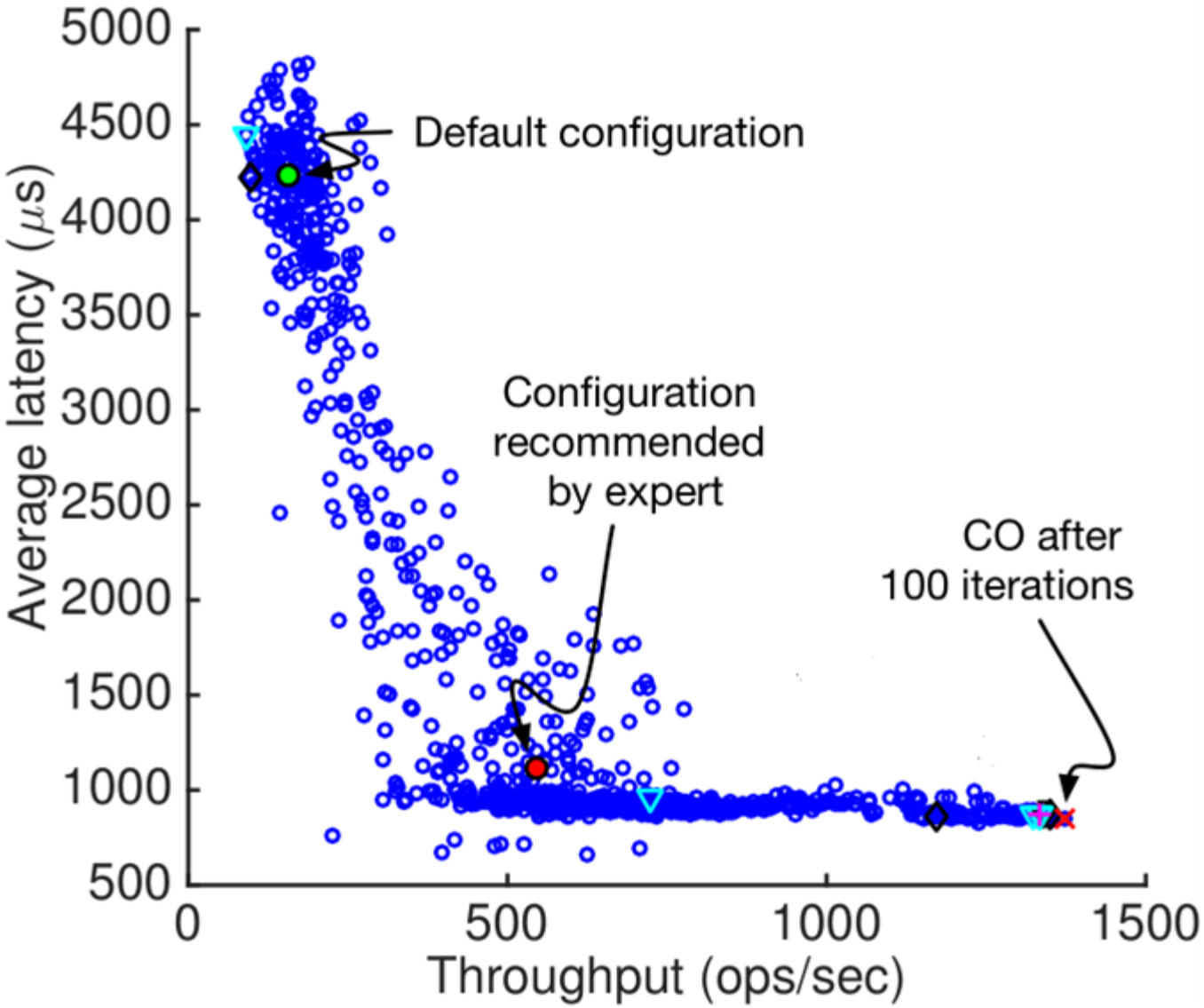


APACHE STORM™
Distributed • Resilient • Real-time

DICE configuration optimization

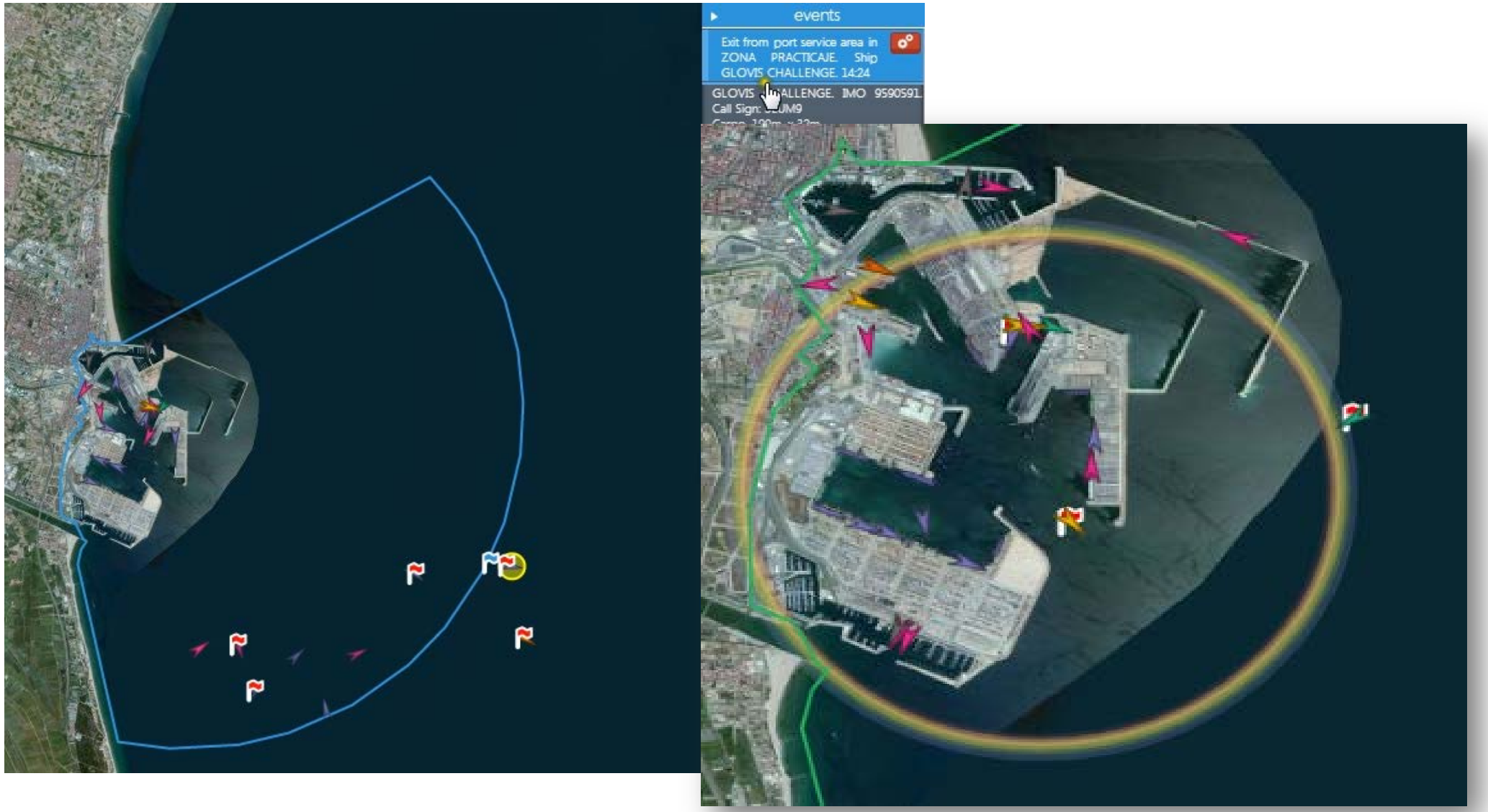
Applying Bayesian Optimization to DevOps



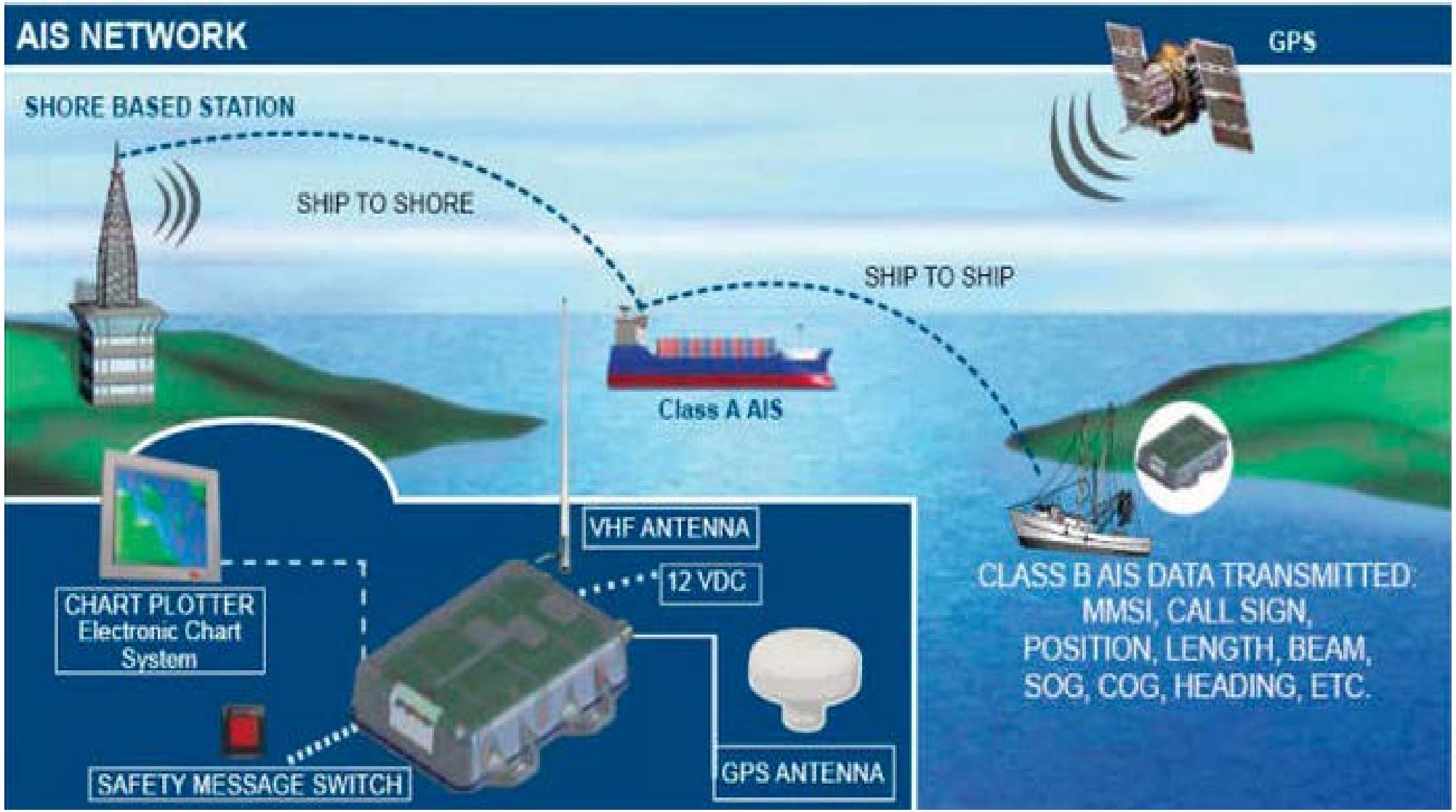


DICE Ops Tools: deployment

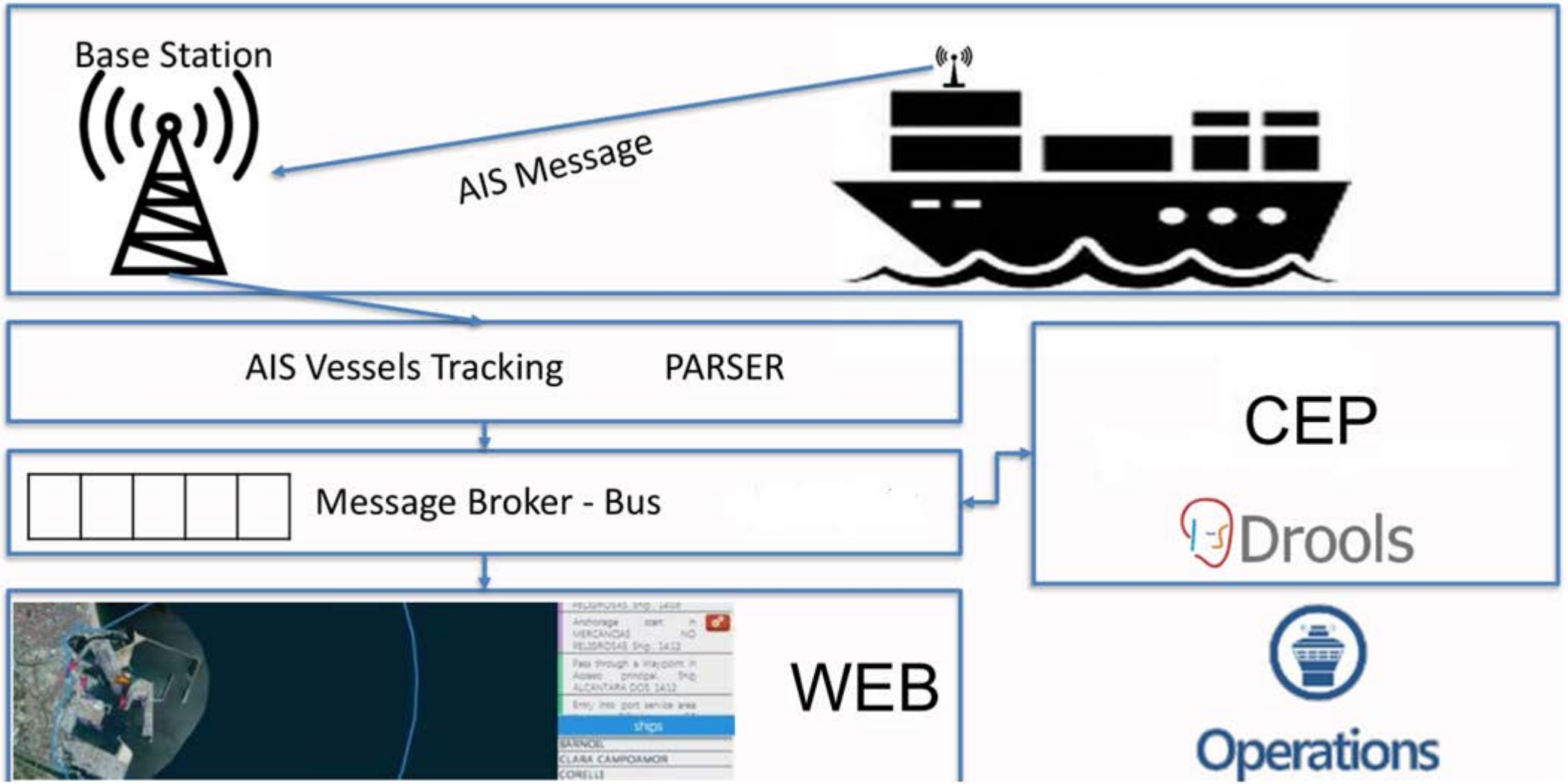
Guiding ships in ports with geofencing



The data



Posidonia Operations



In numbers

Production testbed deployment

8h
Before

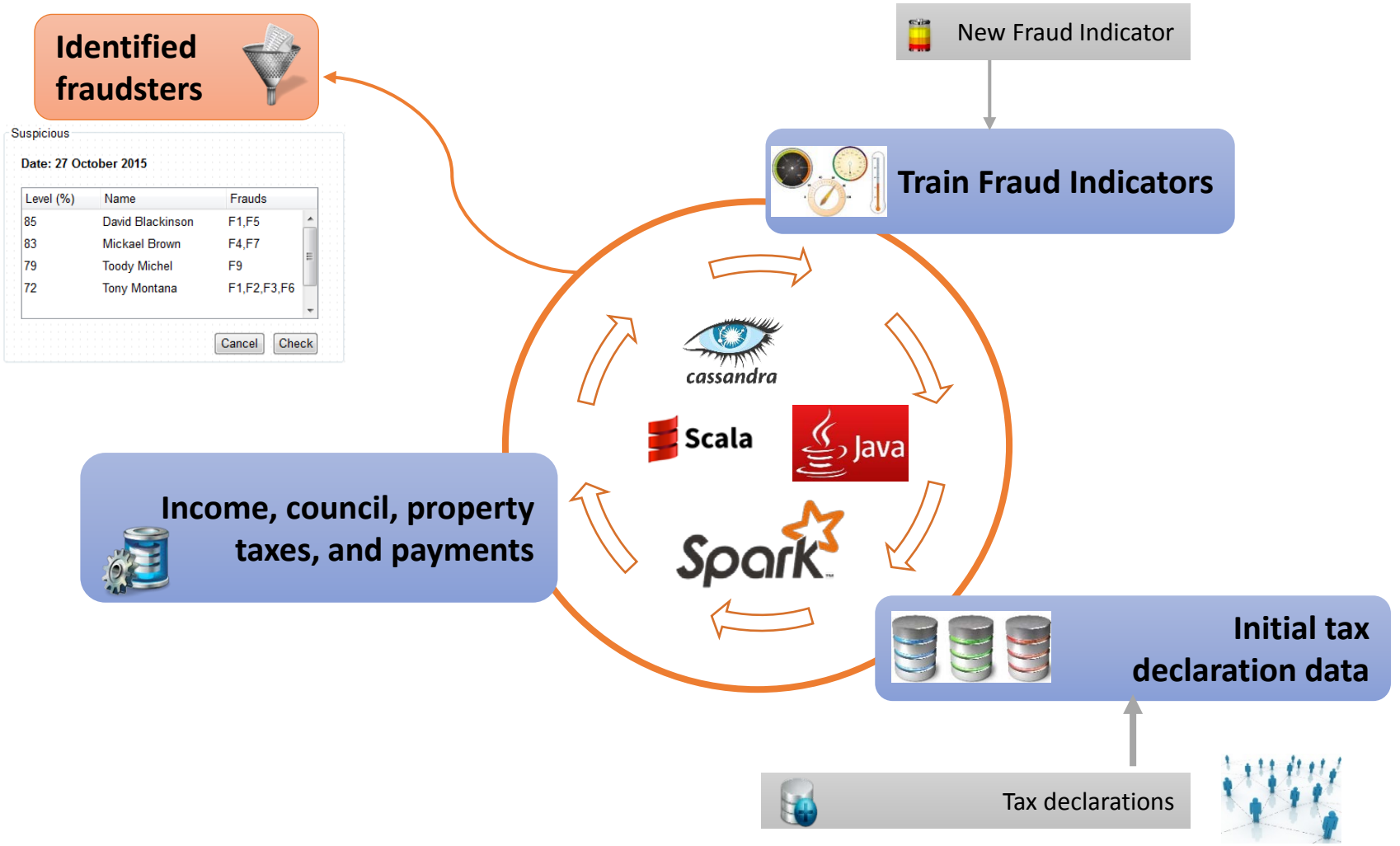
5h
1st attempt

20min
Later attempts

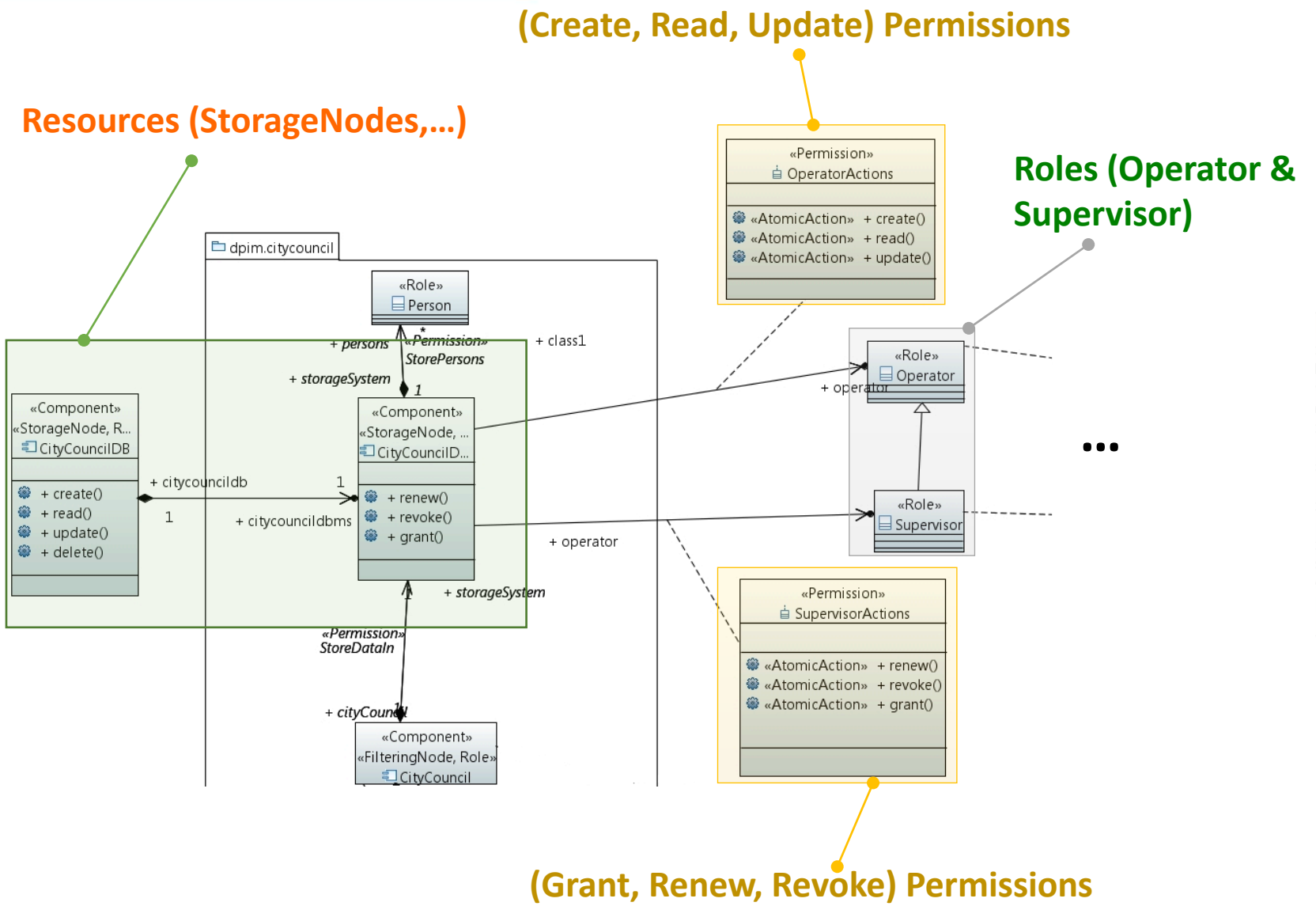
Cross-cutting concerns: data privacy

Application Domain: Tax Fraud Detection

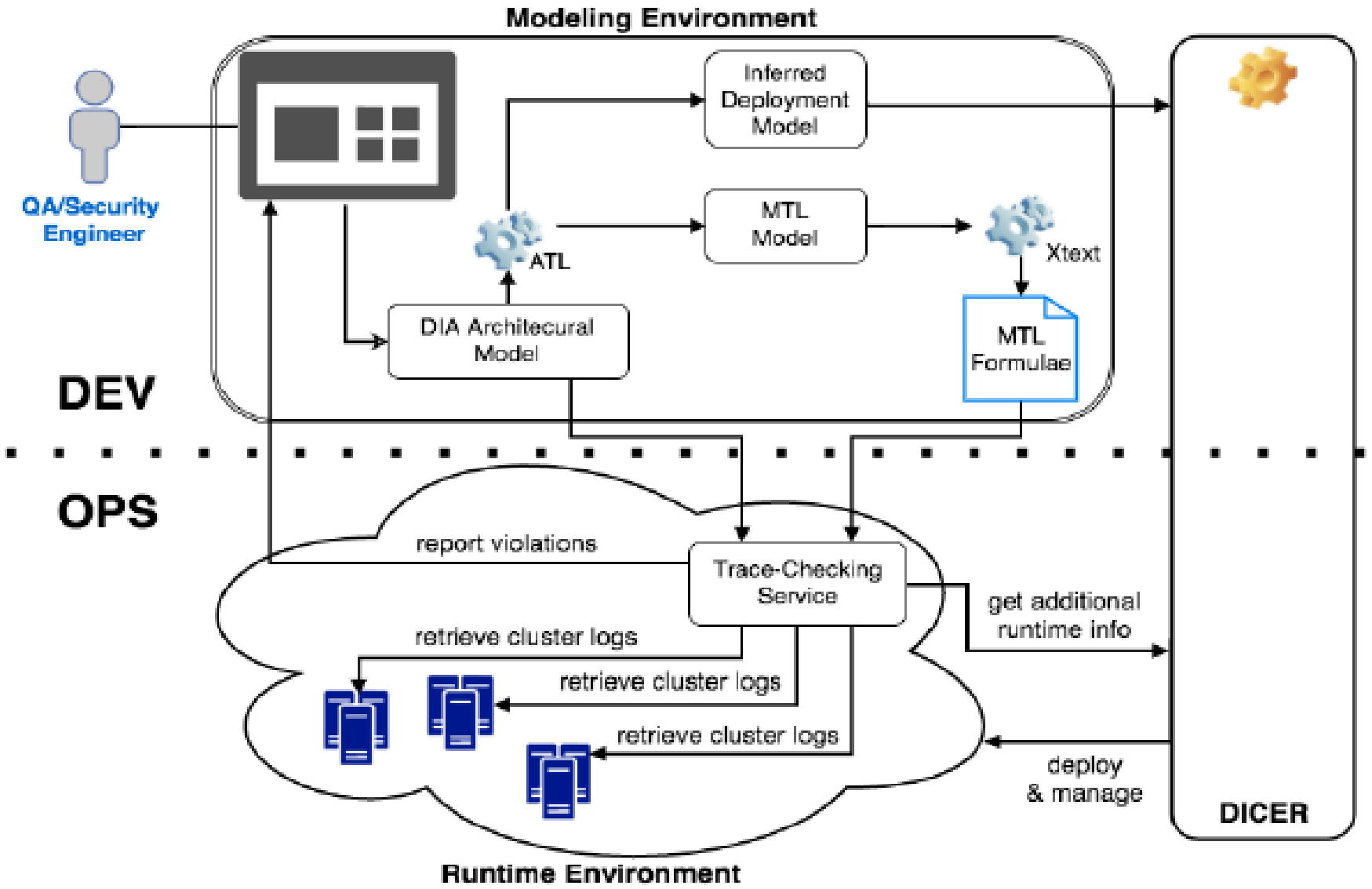
Using Big data for Tax Fraud detection



Correctness of data privacy rules



Runtime verification of data privacy rules



Conclusion

- Model-driven helps “gluing” Dev, Ops, & QA
  - Shared view of the system
  - Facilitates reasoning on and enforcing quality concerns such as privacy, performance, …
- Automated model-driven deployment:
  - Valuable to SMEs new to DevOps and/or Big data

DICE

WWW.DICE-H2020.EU

DEVOPS FOR BIG DATA

Q&A

GIULIANO CASALE, IMPERIAL COLLEGE LONDON

G.CASALE@IMPERIAL.AC.UK

DevOpsDays Warsaw