On Developing and Operating of Data Elasticity Management Process

(Ongoing work under submission)

Tien-Dung Nguyen, Hong-Linh Truong, Georgiana Copil, Duc-Hung Le, Daniel Moldovan, and Schahram Dustdar

Distributed System Group, TUWien

http://www.infosys.tuwien.ac.at/research/viecom/

Content

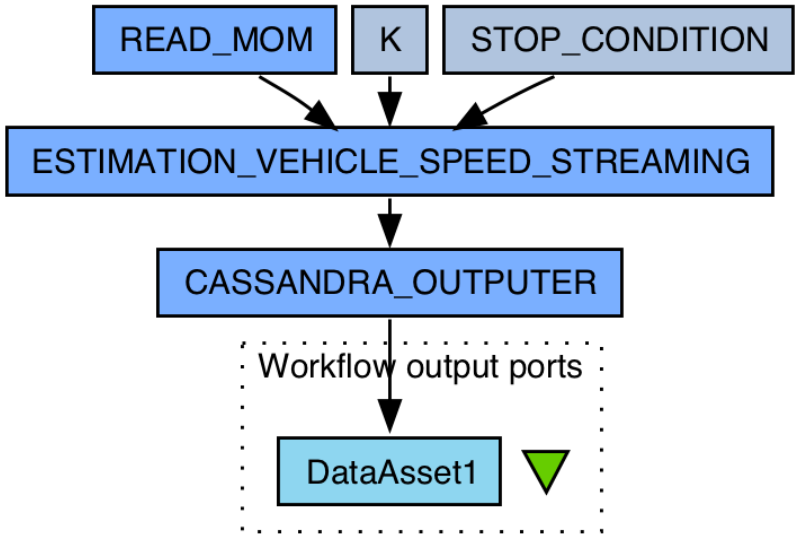
- Motivation & Contributions
- Elasticity Model for Data Asset
- Generating and Operating Data Elasticity Management Process
- Evaluation
- Conclusion

Motivation

- Provider
  - has data analytic workflows
  - execution of a data analytic workflow results a data asset
- The objective of the provider
  - quality of data
  - performance of the execution of data analytic workflow
  - cost for computation and data resource.
- The objective of data consumers
  - data asset with expectation of quality of data, performance and cost

Example: Ensuring Quality of GPS Data Asset

- GPS data of motorbikes in Ho Chi Minh city
- GPS Data-as-a-Service (DaaS) provider and several DaaS consumers (e.g., Taxi company)



GPS Data Asset Example

Timestamp	DeviceID	Longitude	Latitude	Speed	Local Area	Estimated Speed in local area
Wed Sep 10 07:45:00 ICT 2014	51B00552	10.660332	106.779396	0	CMT8-VTS	10.25
Wed Sep 10 07:45:23 ICT 2014	51C29797	10.749635	106.67208	24	CMT8-DBP	30.5
Wed Sep 10 07:46:24 ICT 2014	51B01907	10.877548	106.64205	0	CMT8-AC	21.1

- Incorrect data of vehicle speed, location is used in data analytics can lead to bad quality data asset

Example: Ensuring Quality of GPS Data Asset

- Example of GPS data consumer objective
  - Quality of data:
    - Vehicle location accuracy>= 91%
    - Vehicle speed accuracy>= 81%
  - Performance
    - deliveryTime <55 s
  - Cost <= €0.05

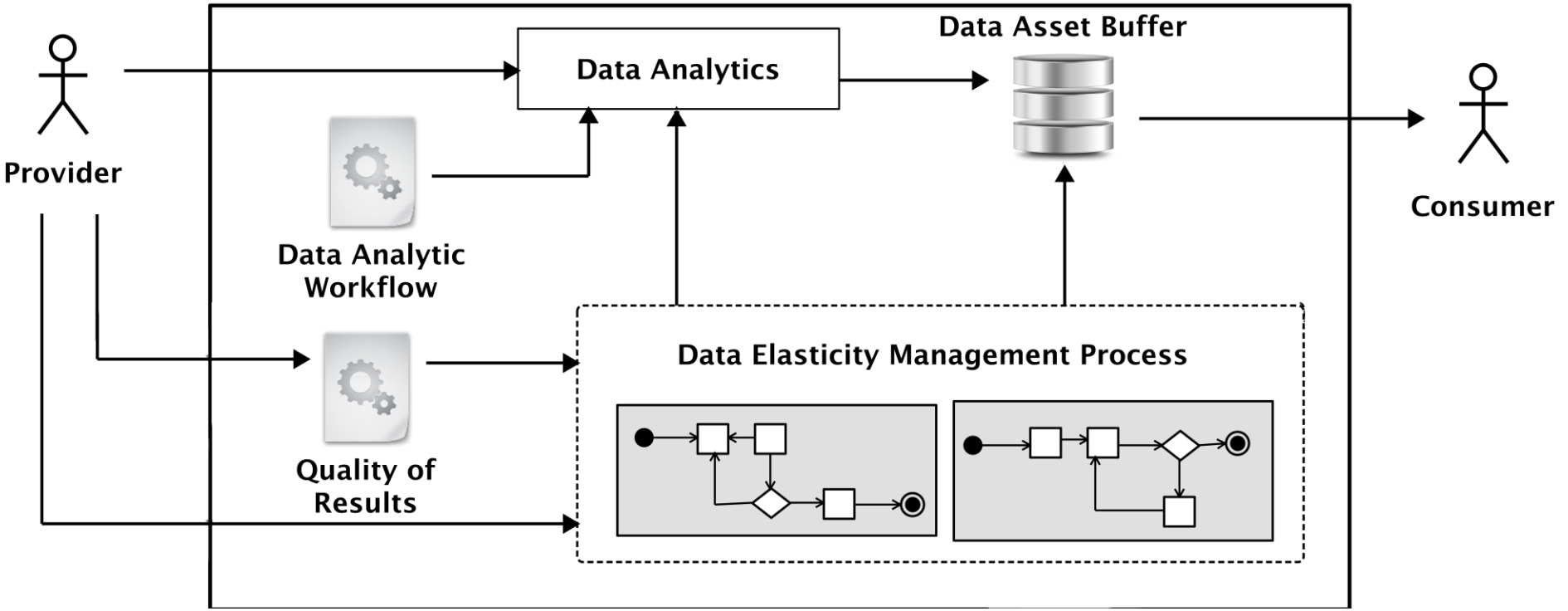
Problem Statement

- Provider needs **Data Elasticity Management Process** to ensure quality of data, performance and cost
  - Improve quality of data: monitor/adjust quality of data
  - Guarantee performance: scaling in/out services to monitor/adjust quality of data to adapt with changes of number of consumers
  - Cost: minimize computation cost
- Existing solutions deal with
  - quality of services and cost when selecting services for service composition [1]
  - Improve quality of data asset by refining/replacing/extending analytic tasks in data analytics workflow [2]

[1] Lijuan Wang, Jun Shen, Junzhou Luo, Fang Dong:**An Improved Genetic Algorithm for Cost-Effective Data-Intensive Service Composition**. SKG 2013: 105-112

[2] Michael Reiter, Uwe Breitenbücher, Schahram Dustdar, Dimka Karastoyanova, Frank Leymann, Hong Linh Truong:**A Novel Framework for Monitoring and Analyzing Quality of Data in Simulation Workflows**. eScience 2011: 105-112

Approach



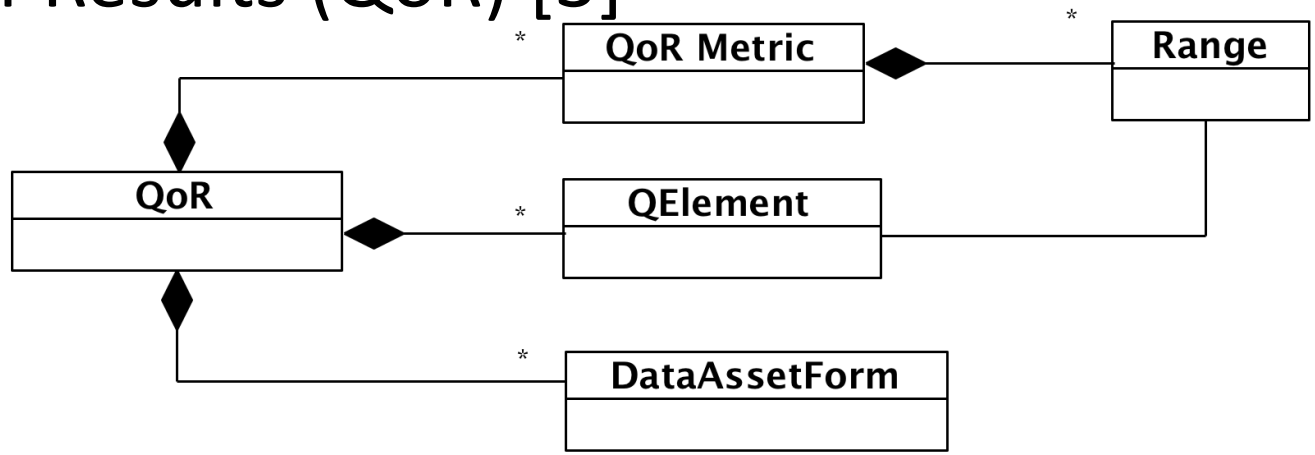
- Generating **Data Elasticity Management Process** from information **in data analytics workflow** and expected **quality of results** of the data asset.

Approach

- Algorithm to generate data elasticity management process
  - Inputs:
    - Data Analytics Workflow
    - Quality of Results
    - Primitive Actions
  - Output
    - Data Elasticity Management Process
- Runtime Environment for Data Elasticity Management Process

Quality of Results

- Provider specifies the expectation of quality of data, performance and cost of data asset they want to sell
- Quality of Results (QoR) [3]



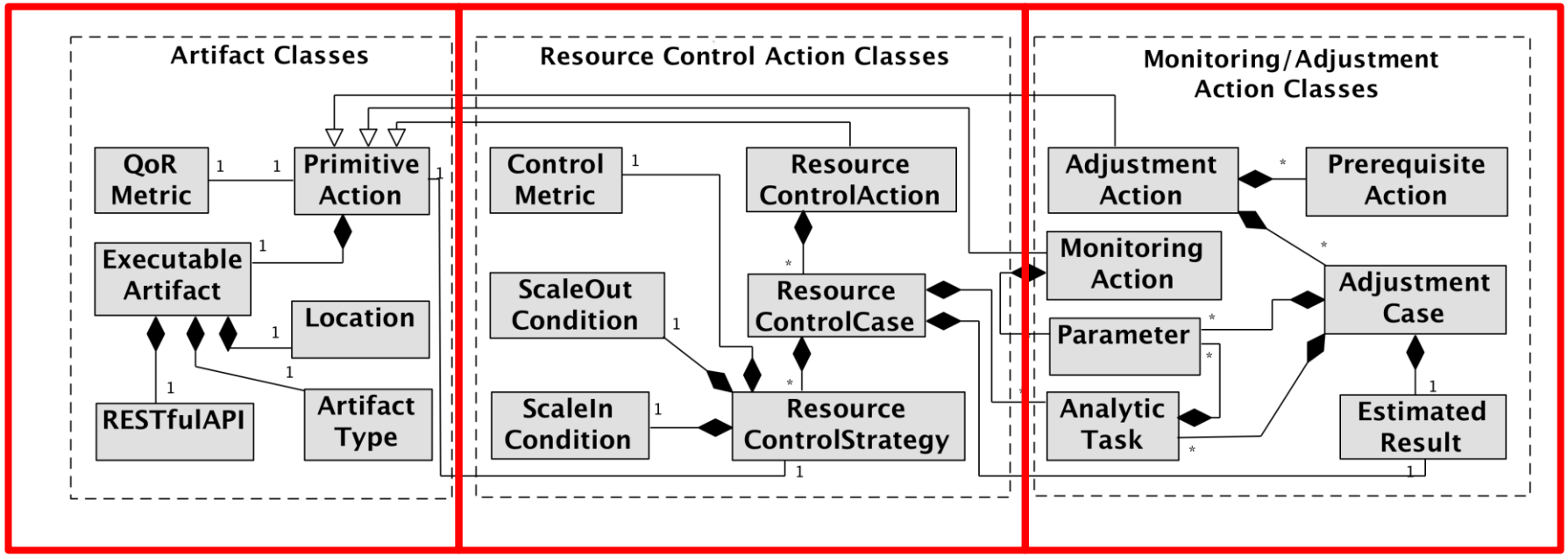
[3] Hong Linh Truong, Schahram Dustdar: **Principles of Software-Defined Elastic Systems for Big Data Analytics**. IC2E 2014: 562-567

Example: Quality of Results

```
!at.ac.tuwien.dsg.depic.common.entity.qor.QoRModel
dataAssetForm: CSV
listOfMetrics:
- !at.ac.tuwien.dsg.depic.common.entity.qor.QoRMetric
   name: vehicleArc
   listOfRanges:
   - !at.ac.tuwien.dsg.depic.common.entity.qor.Range
      rangeID: vehicleArc_co1
     fromValue: 80.0
      toValue: 100.0
   unit: '%'
…
listOfQElements:
- !at.ac.tuwien.dsg.depic.common.entity.qor.QElement
   listOfRanges:
   - speedArc_co1
   - vehicleArc_co1
   - deliveryTime_co2
   cost: 0.05
   qElementID: qElement1
```

Primitive Action for Data Elasticity Management

- Objective: to capture information of action to adjust quality of data, performance
- Primitive actions are filled by data experts, profiling tool, benchmarking (e.g., Talend[1]])



1. https://www.talend.com

Example: Primitive Action

```
- !at.ac.tuwien.dsg.depic.common.entity.primitiveaction.ResourceControlAction
   associatedQoRMetric: deliveryTime
   listOfResourceControlStrategies:
   - !at.ac.tuwien.dsg.depic.common.entity.primitiveaction.ResourceControlCase
      estimatedResult:
         conditionID: deliveryTime_c1
         metricName: deliveryTime
         upperBound: 55.0
      listOfResourceControlStrategies:
      -
!at.ac.tuwien.dsg.depic.common.entity.primitiveaction.ResourceControlStrategy
         controlMetric: cpuUsage
         primitiveAction: VehicleLocationAccuracyMeasurement
         scaleInCondition:
            conditionID: c_in
            metricName: cpuUsage
            upperBound: 20.0
         scaleOutCondition:
            conditionID: c_out
            lowerBound: 50.0
            metricName: cpuUsage
            upperBound: 100.0
```

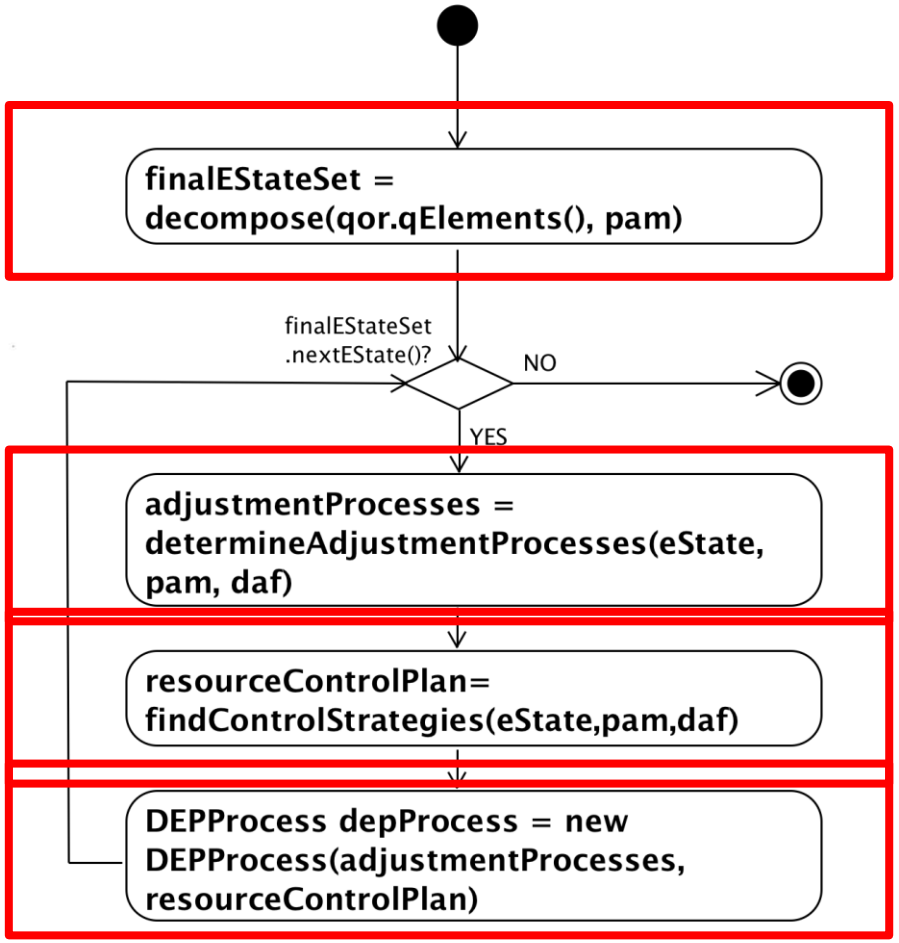
Generating Data Elasticity Management Process

Inputs:

- QoR
- Data analytic workflow
- Primitive Actions

Output:

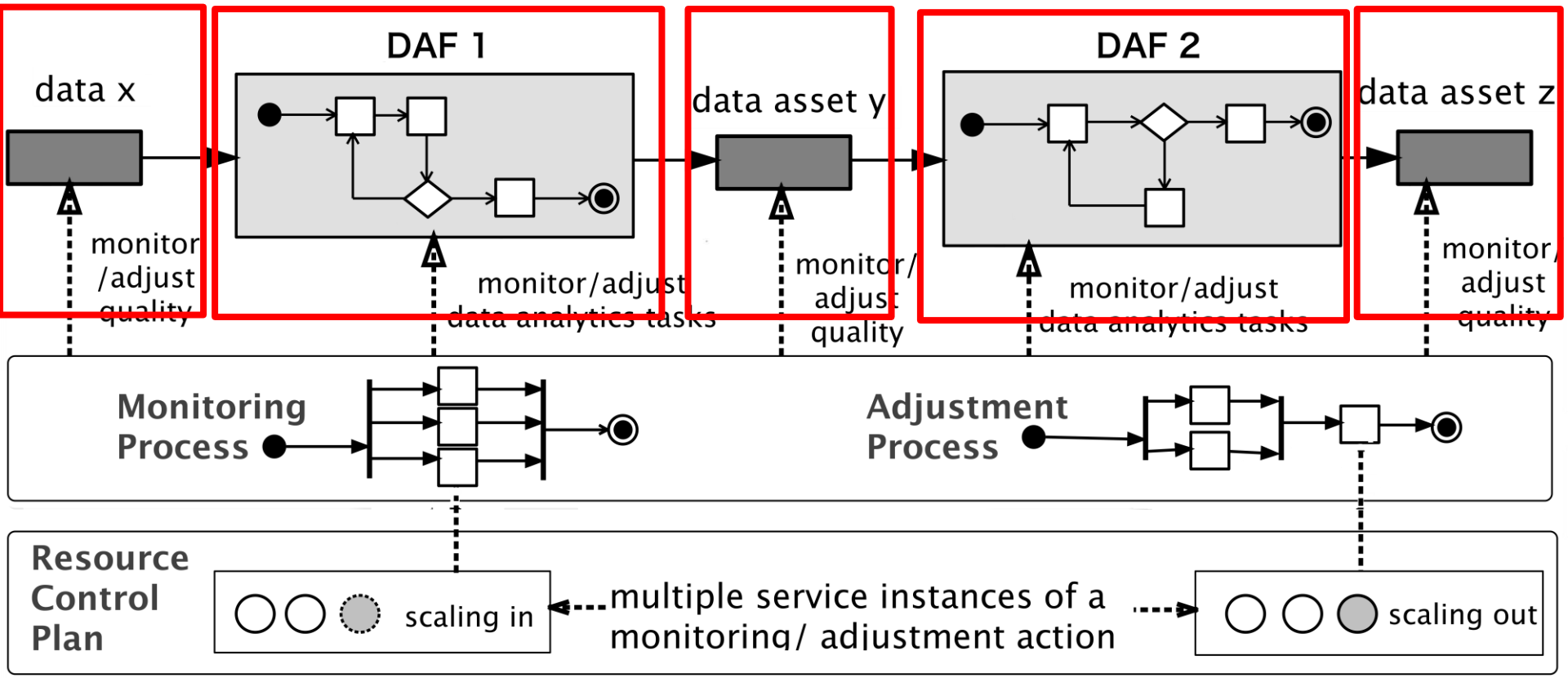
- Data Elasticity Management Process



Generating Data Elasticity Management Process

- Final eState Set of Data Asset
  - **decompose** ranges in **qElements** and **estimated results** in primitive actions
  - combine conditions to determine eState
- Adjustment Process
  - for each condition in an eState, check if the QoR metric is associated with an adjustment action
  - if true, **determine adjustment case** of a primitive action based on **estimated results and analytic tasks**
  - building a workflow of primitive actions based on their **prerequisite actions**
- Resource Control Plan
  - for each condition in an eState, check if the QoR metric is associated with an resource control action
  - if true, **determine resource control case** based **on estimated results and analytics tasks**

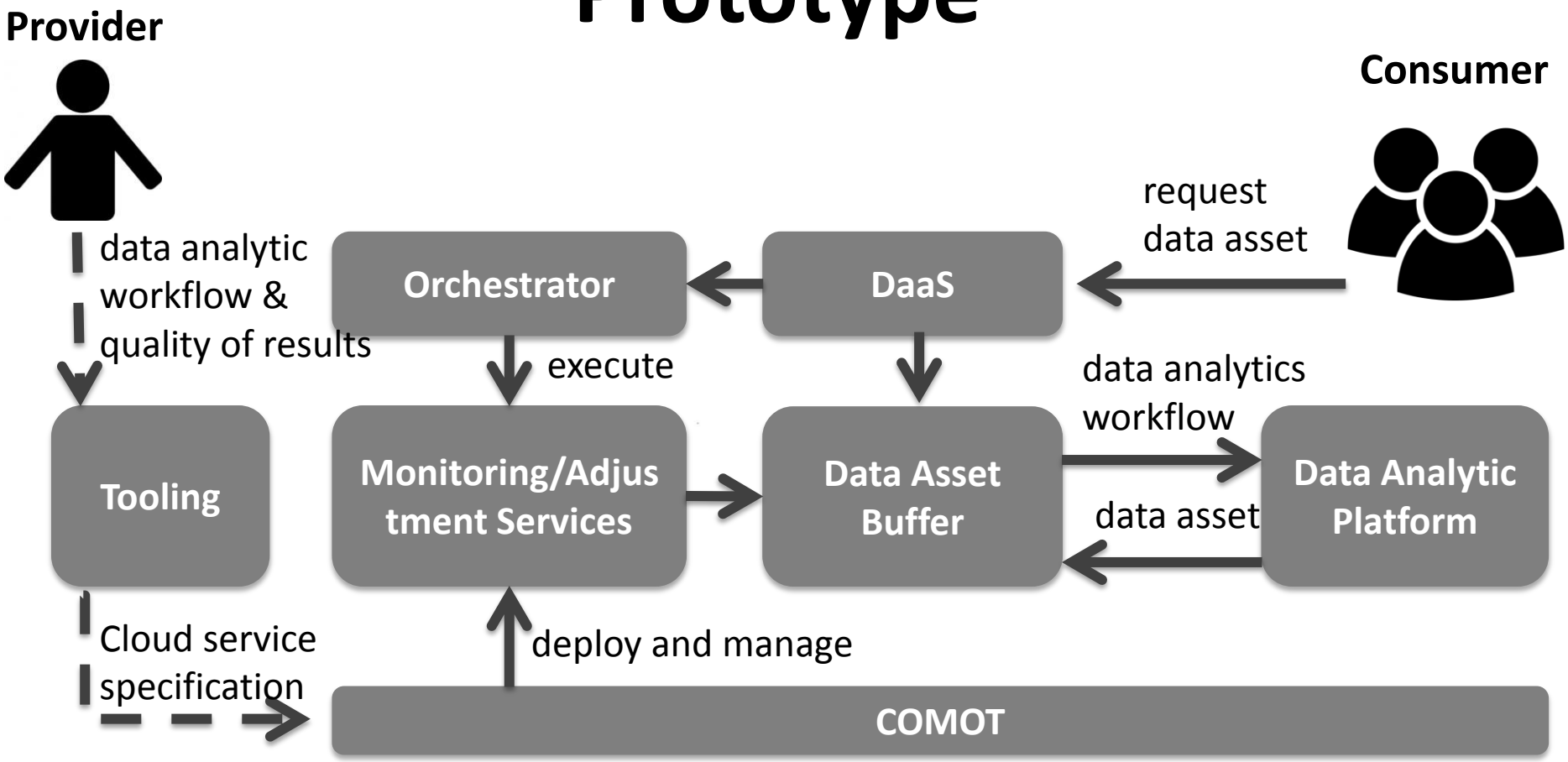
Using Data Elasticity Management Process to ensure QoR



Using Data Elasticity Management Process

- A general principle:
  - store data into a data buffer
  - perform actions on data in the buffer before delivering the data to customer
- Data buffers can have different plugins interfacing to different types of databases

Prototype



- Data analytics Platform: Taverna, Apache ActiveMQ, Spark, Cassandra
- Data Asset Buffer: MySQL, Cassandra, PostgreSQL

[4] Prototype Source. https://github.com/tuwiendsg/EPICS/tree/master/depic

[5] Hong Linh Truong, Schahram Dustdar, Georgiana Copil, Alessio Gambi, Waldemar Hummer, Duc-Hung Le, Daniel Moldovan: **CoMoT - A Platform-as-a-Service for Elasticity in the Cloud**. IC2E 2014: 619-622

Evaluation

Evaluation: Setup & Assumptions

- Scenario:
  - Near-real time GPS data of vehicles in HoChiMinh City. Data size 1.17GB.
  - Emulating data source by sending historical GPS data to scalable message oriented middleware (MOM)
  - 5 concurrent DaaS consumers
- Infrastructure
  - 1 VM (7GB RAM, 4 vCPUs, 40GB Disk) for Tooling, Orchestrator, Data Asset Loader and Data Analytics Workflow Management
  - 4 VMs (1GB RAM, 1 vCPU, 40GB) for monitoring/adjustment services at the beginning

Evaluation: Inputs

Provider:

- Vehicle location accuracy (%):
  [0 20], [21 40] [41 60], [61 80], [81 100]
- Vehicle speed accuracy (%):
  [0 20], [21 40] [41 60], [61 80], [81 100]
- Delivery time (s): [0 54], [55 120]

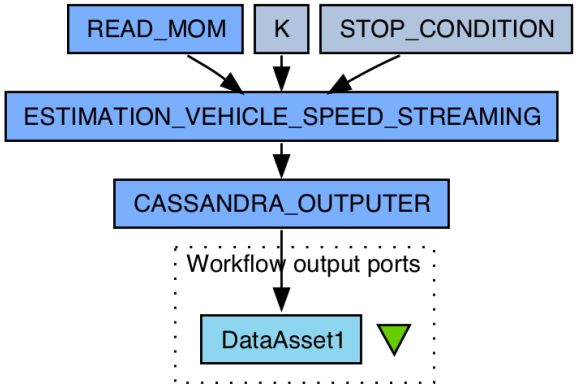


Figure 6: Data Analytics Workflow for provisioning GPS data

- Estimated cost function:

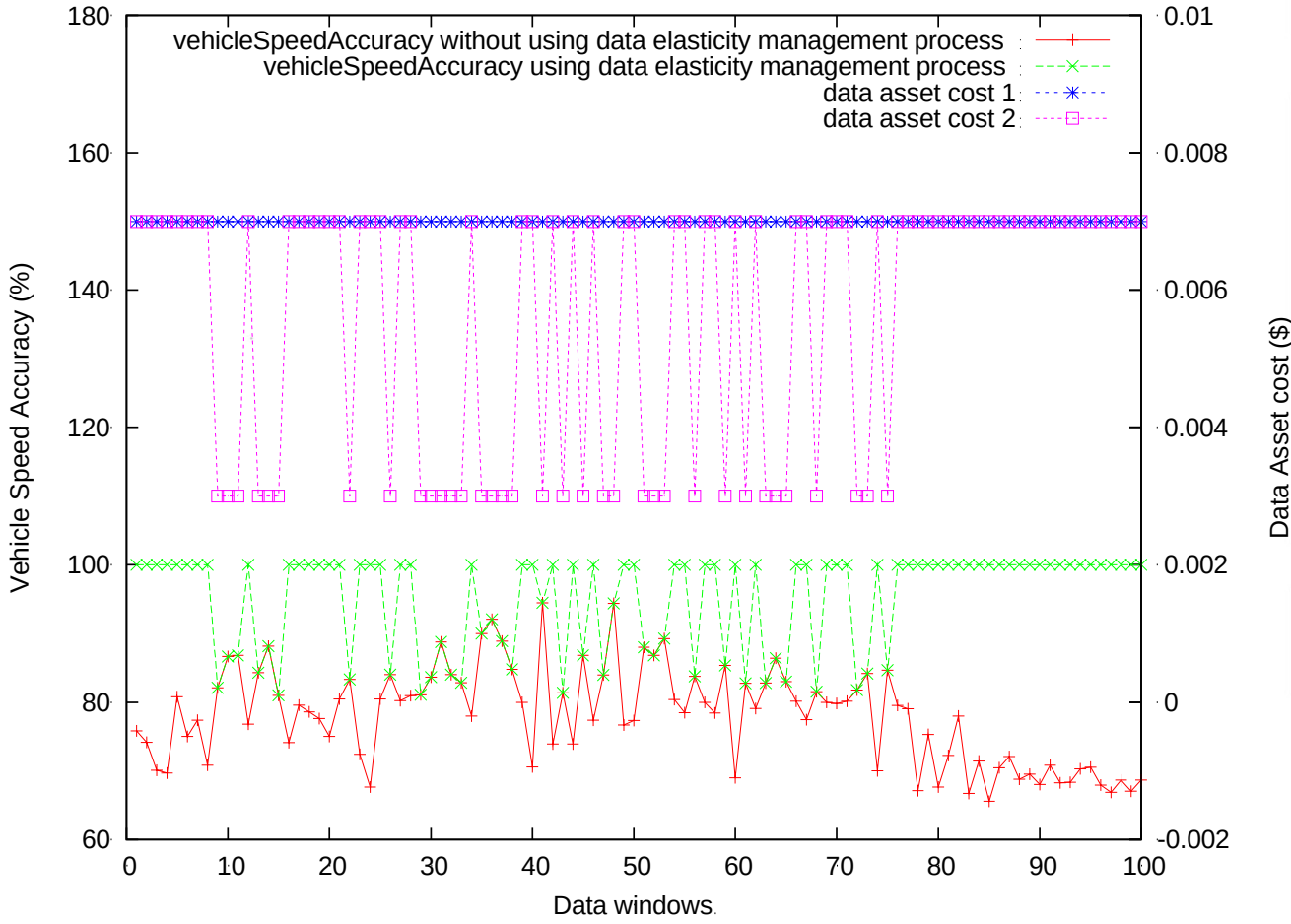
$$cost_{qElement} = \sum_{i=1}^{nbMetrics} unitCost(qorMetric_i) * \sum_{j=1}^{nbCond_i} (j * qElement_{cond_j})$$

Consumers' expectations of the data asset

- Case 1:
  - Vehicle location accuracy > 81%
  - Vehicle speed accuracy > 81 %
  - Delivery Time < 55s
- Case 2:
  - Vehicle location accuracy > 61%
  - Vehicle Speed accuracy > 61 %
  - Delivery Time < 55s

Vehicle Speed Accuracy and Estimated Data Asset Cost

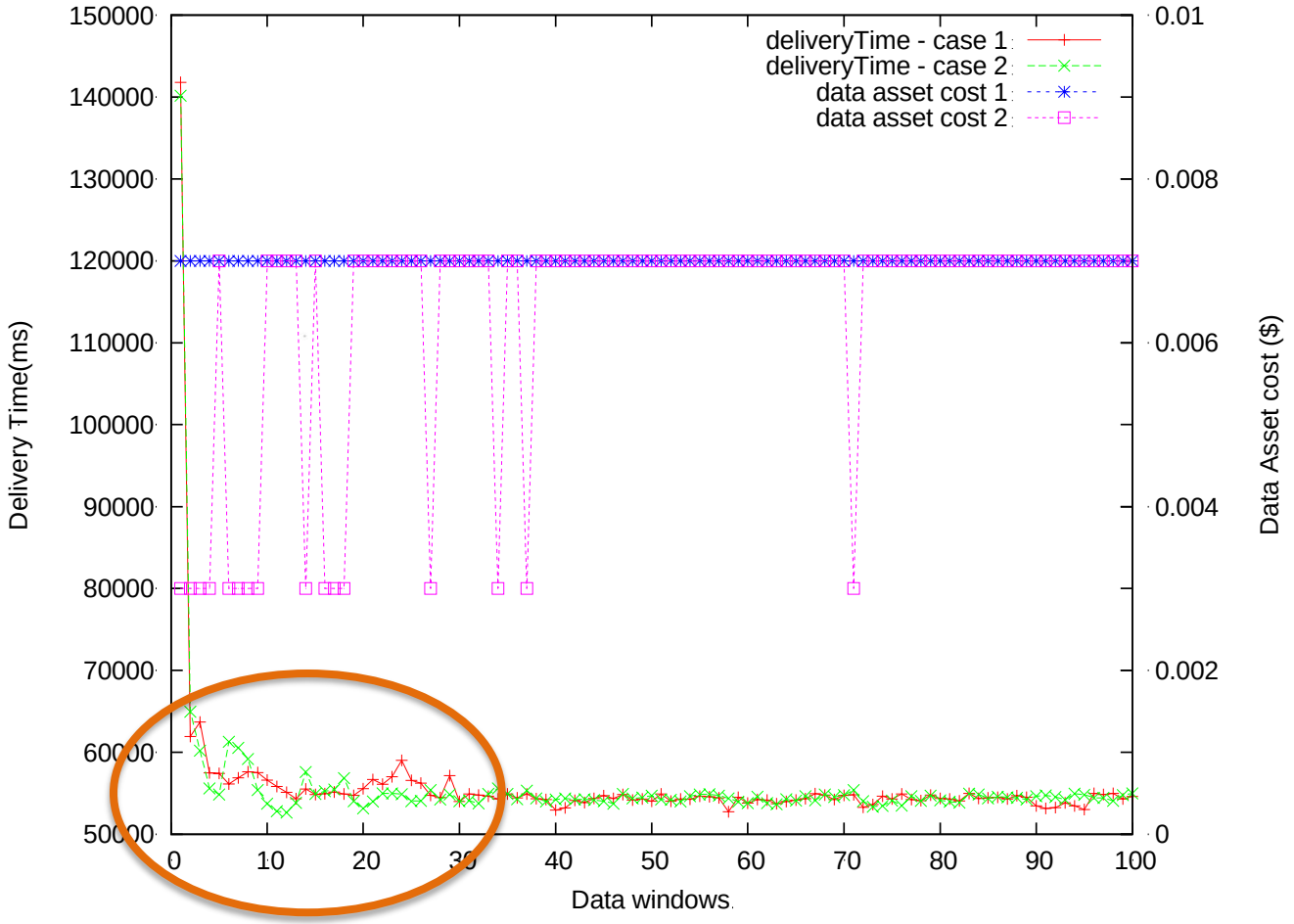
- 5 customers
- Vehicle Speed Accuracy meets expected QoR with data elasticity management process
- Trading off between cost and accuracy



Delivery Time and Estimated Data Asset Cost

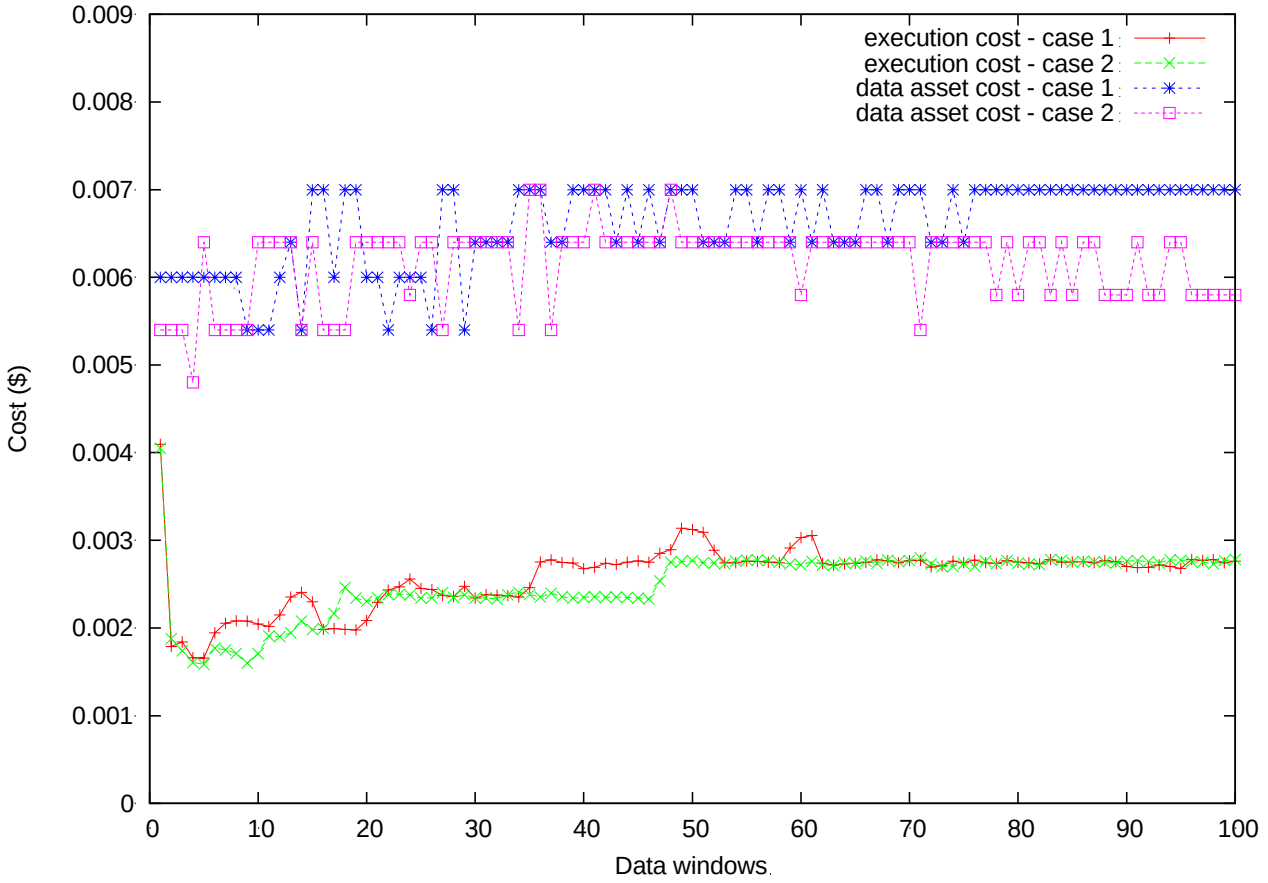
- 5 customers
- With data elasticity management process: delivery time is ensured from 28th window

- Trading off between cost and deliveryTime



Data Asset Cost Patterns

- **Data asset cost:** defined by data asset cost function
- **Execution cost:**
  - execution time,
  - number of VMs
  - unitCost of 1 VM (e.g., Amazon EC2)
- Support provider to determine appropriate cost function
- The pattern of data asset cost vs.. execution cost



Conclusions & Future Work

- Summary
  - Supporting provider to choose appropriate cost model for data asset
  - Elasticity of quality, cost and resource usage
- Future work
  - Optimizing the execution of data analytics workflow
  - A programming framework to support the DaaS provider to easily develop primitive actions

Thanks for your attention!