Quality-aware data analytics

Hong-Linh Truong
Distributed Systems Group, TU Wien

truong@dsg.tuwien.ac.at
http://dsg.tuwien.ac.at/staff/truong
@linhsolar

What this lecture is about

- Data analytics – general view
- Data analytics workflow structures and systems
- Enable quality of analytics (QoA) for data analytics
- Quality of data in data analytics workflows
- Data elasticity management

What this lecture is about

- After this lecture
  - Apply and revise the analytics part in your project
  - Deal with quality of analytics and see how you could offer quality-aware analytics in your project

Big Data

- Big volume, Big velocity, Big variety, Big Veracity
- Sources
  - Internet of Things, human participation, social networks, software services, environment monitoring, advanced science instruments, science discovery, etc.
- Several challenges in terms of data gathering, integration, and analytics

H. V. Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, and Cyrus Shahabi. 2014. Big data and its technical challenges. Commun. ACM 57, 7 (July 2014), 86-94. DOI=10.1145/2611567 http://doi.acm.org/10.1145/2611567

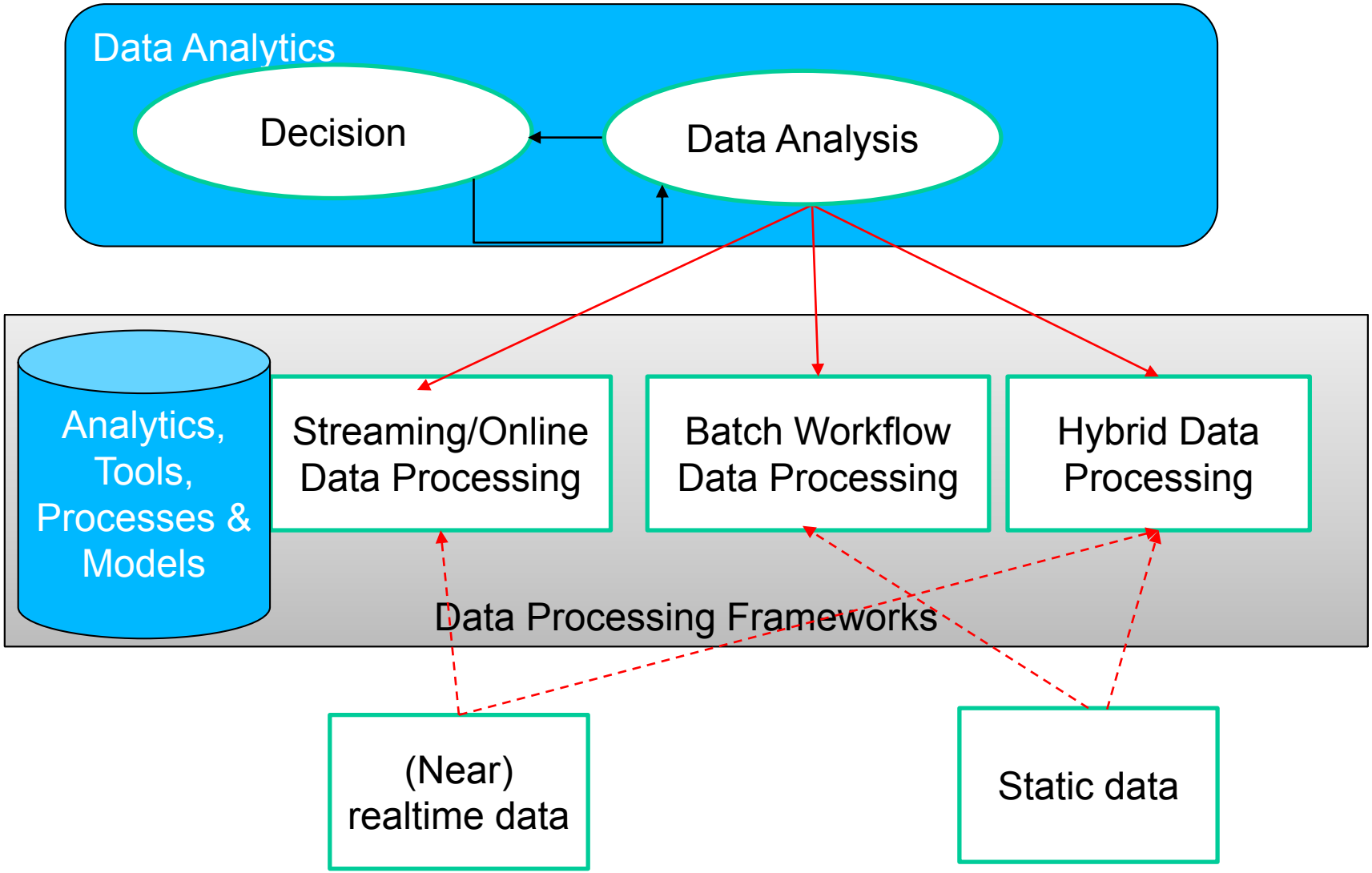
Data Management/Delivery Systems

- Static data – data at rest
  - Hadoop file systems
  - Large scale storage data systems
    - iRODS, NoSQL
  - Web services for Data-as-a-Service (e.g., GIS)
- Real time data – data in motion
  - Cloud data platforms, e.g. Xively
  - Several MOM (Message-oriented Middleware)
    - E.g., Apache Kafka
  - Domain-specific streamming systems (e.g., images)

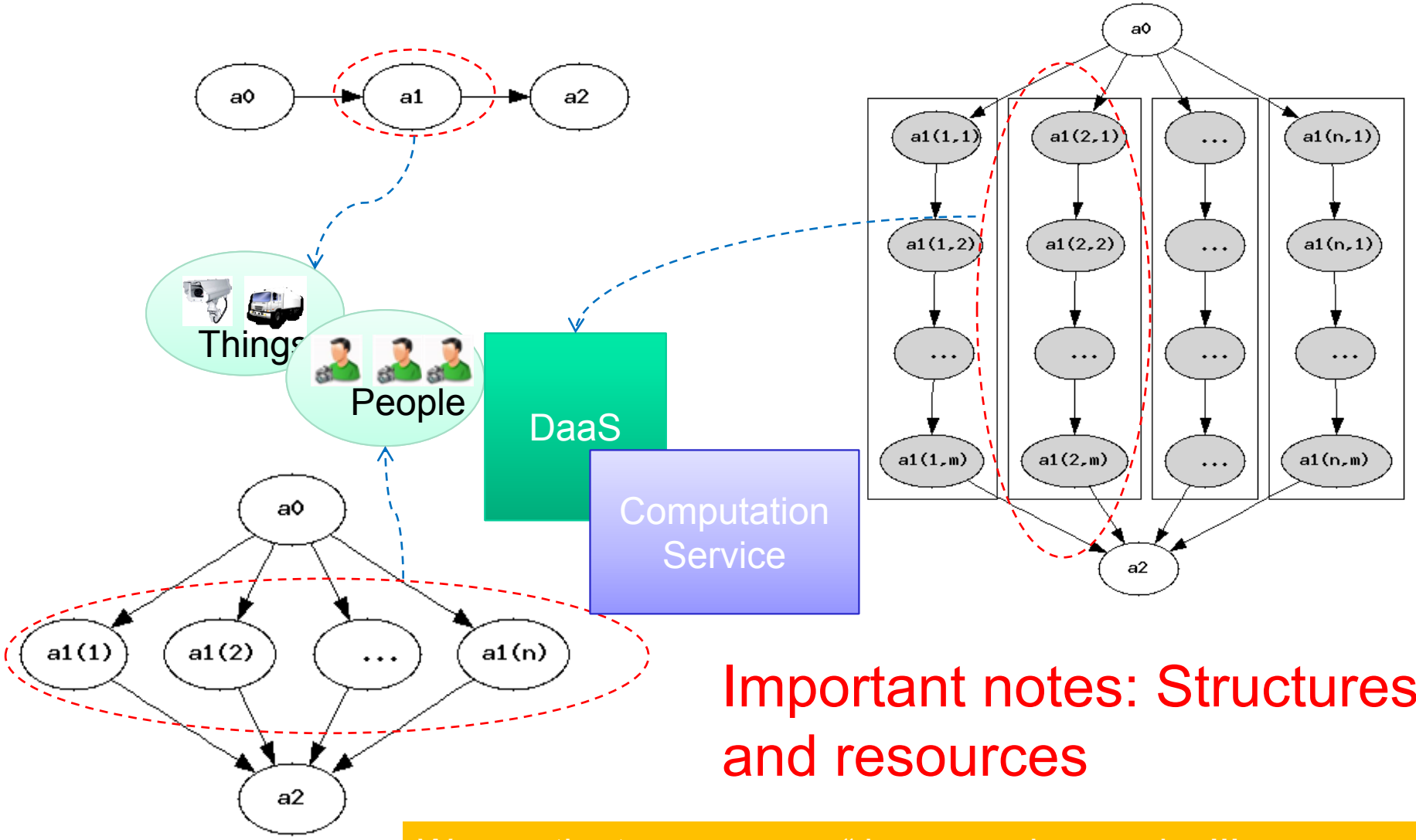
Data Processing Framework

- Batch processing
  - Mapreduce/Hadoop
  - Scientific workflows
- (Near) realtime streaming processing
  - S4 & Storm
- Hybrid data processing
  - Summingbird, Apache Kylin
  - Impala, Storm-YARN
  - Apache Spark

Conceptual View



Data analytics processes – a bird view



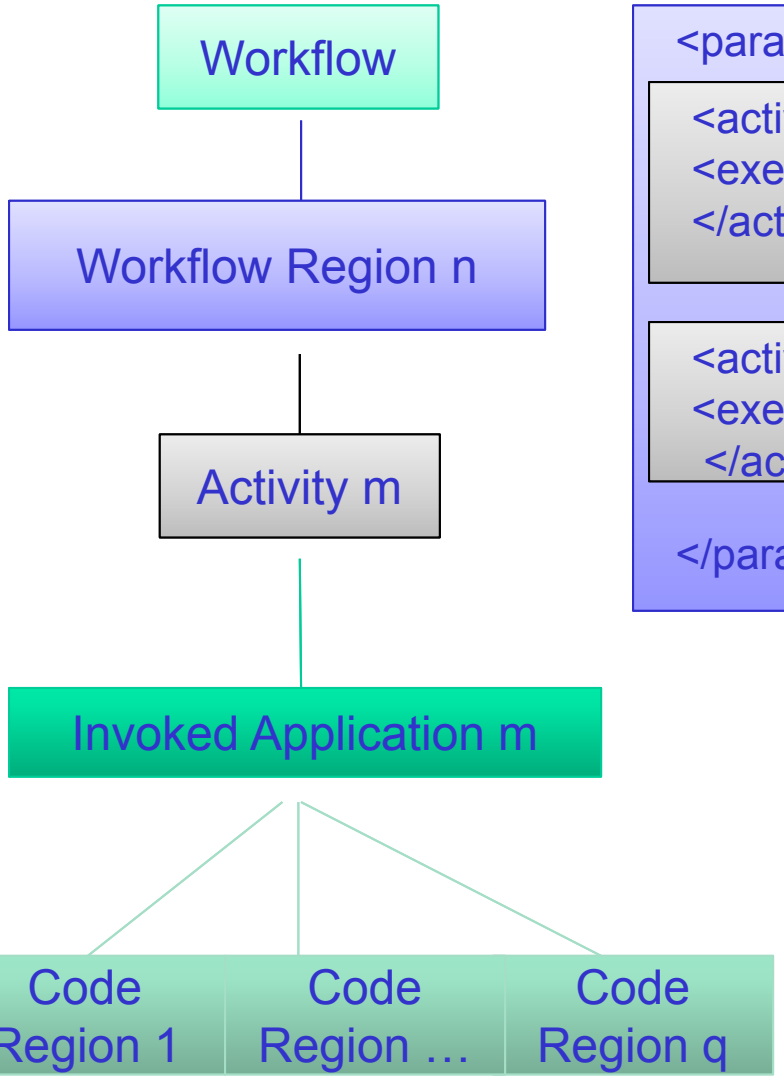
Important notes: Structures and resources

We use the term „process“ in a generic meaning!!!

Data analytics processes

- Main categories
  - (Batch) workflow-based processing
  - Stream data processing
  - Hybrid data processing

Workflow-based processing

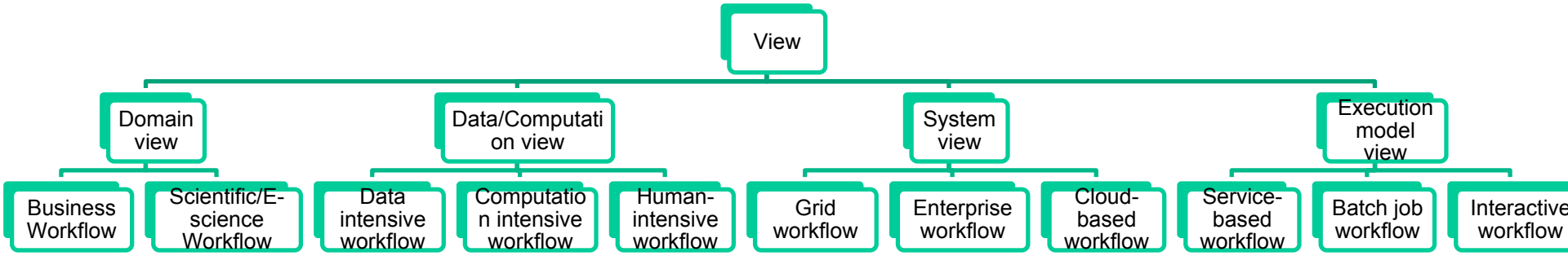



```
<parallel>
  <activity name="mProject1">
  <executable name="mProject1"/>
  </activity>

  <activity name="mProject2">
  <executable name="mProject2"/>
   </activity>

</parallel>
```

```
mProject1Service.java
    public void mProject1() {
        A();
        while () {
            ...
        }
}
```

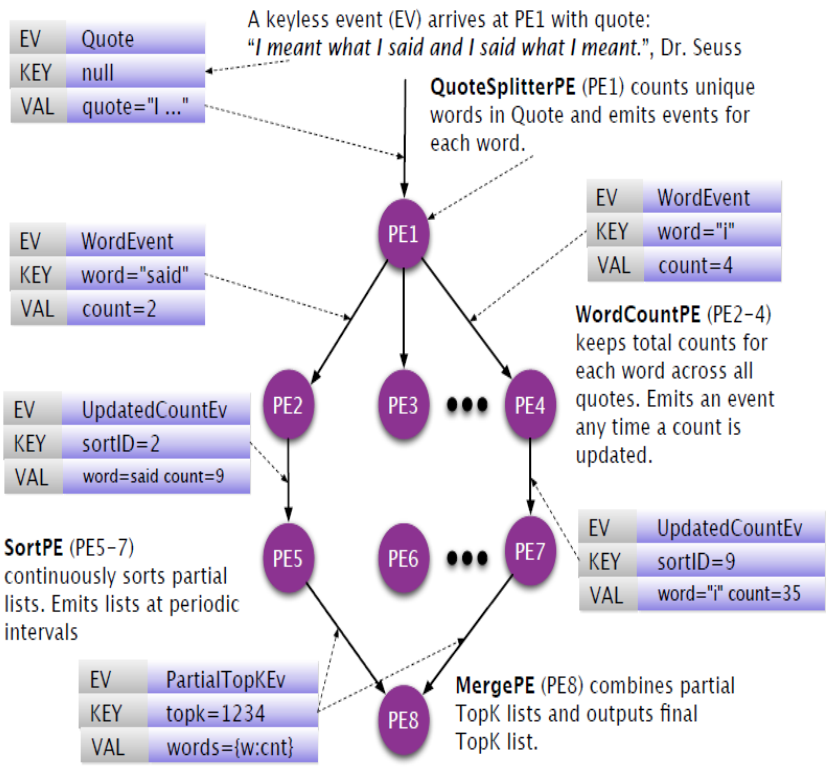

Hong Linh Truong, Schahram Dustdar, Thomas Fahringer: Performance metrics and ontologies for Grid workflows. Future Generation Comp. Syst. 23(6): 760-772 (2007)

Different views of (data analytics) workflow systems



Stream data processing

- Processing elements/operators are arranged in graphs
- Streaming data comes to processing elements
- Results from an element are passed to another



PE ID	PE Name	Key Tuple
PE1	QuoteSplitterPE	null
PE2	WordCountPE	word="said"
PE4	WordCountPE	word="i"
PE5	SortPE	sortID=2
PE7	SortPE	sortID=9
PE8	MergePE	topK=1234

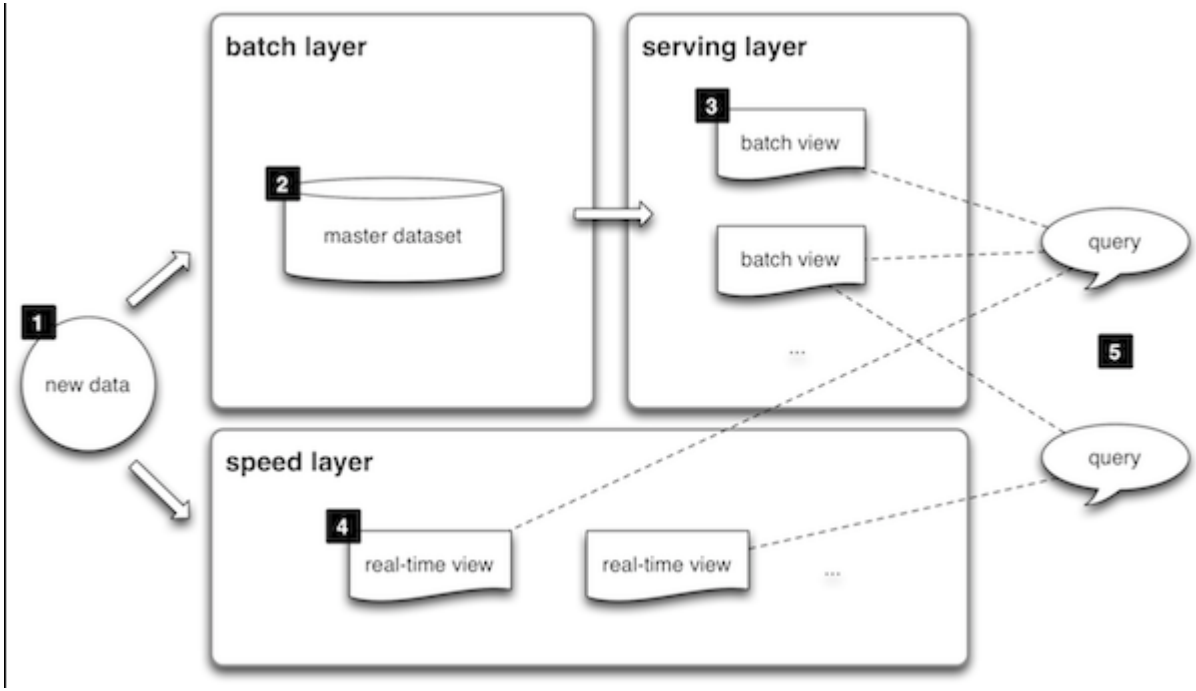
Figure 1. Word Count Example

Source: Neumeyer, L.; Robbins, B.; Nair, A.; Kesari, A., "S4: Distributed Stream Computing Platform," Data Mining Workshops (ICDMW), 2010 IEEE International Conference on , vol., no., pp.170,177, 13-13 Dec. 2010

Check also: http://www.infosys.tuwien.ac.at/teaching/courses/socloud/ws2011/slides/streamprocessing.pdf

Hybrid data processing

Combine batch processing and streaming processing
e.g., https://spark.apache.org/



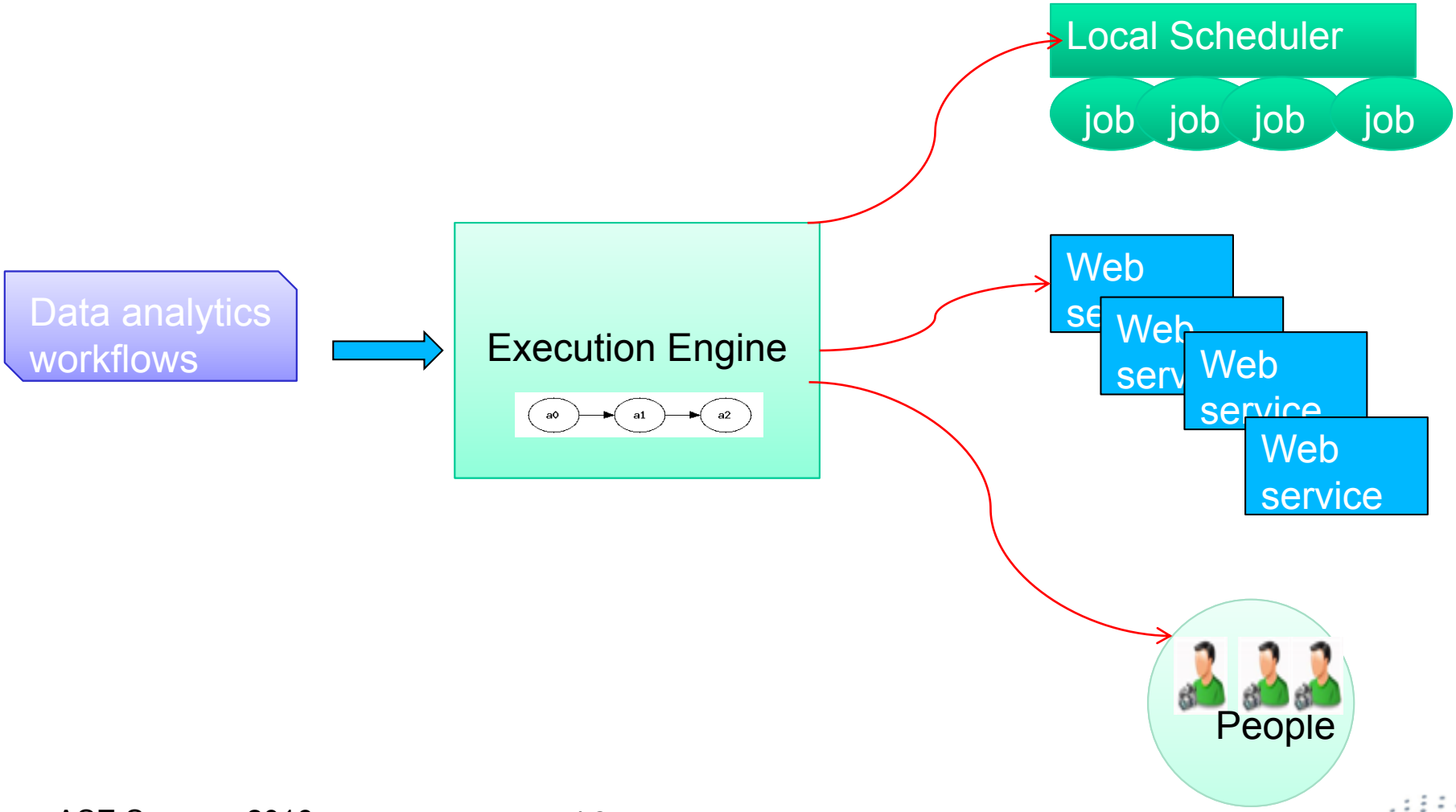
Source:http://lambda-architecture.net/

Applications

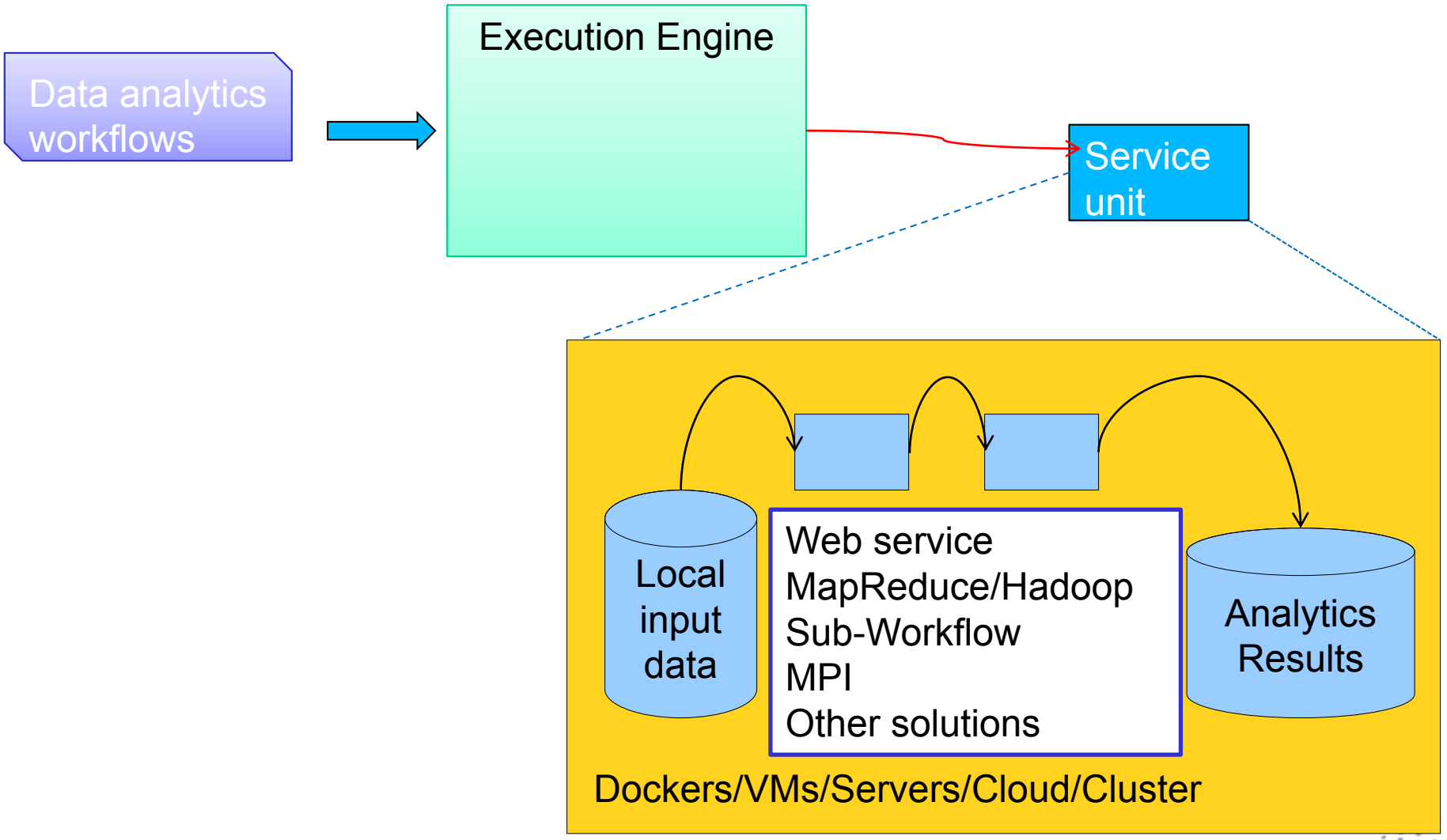
- When we have different problems required different data processing models for different workload/performance
  - Near realtime monitoring + predictive analytics
- Support many phases in data integration and analytics with the same framework
- Dealing with static and realtime data in decision making

WORKFLOWS

Data analytics workflow execution models



Data analytics workflow execution models



Representing and programming data analytics workflows/processes

- Programming languages
  - General- and specific-purpose programming languages, such as Java, Python, Swift
- Programming models
  - such as MapReduce, Hadoop, Complex event processing, Spark
- Descriptive languages
  - BPEL and several languages designed for specific workflow engines
- They can also be combined

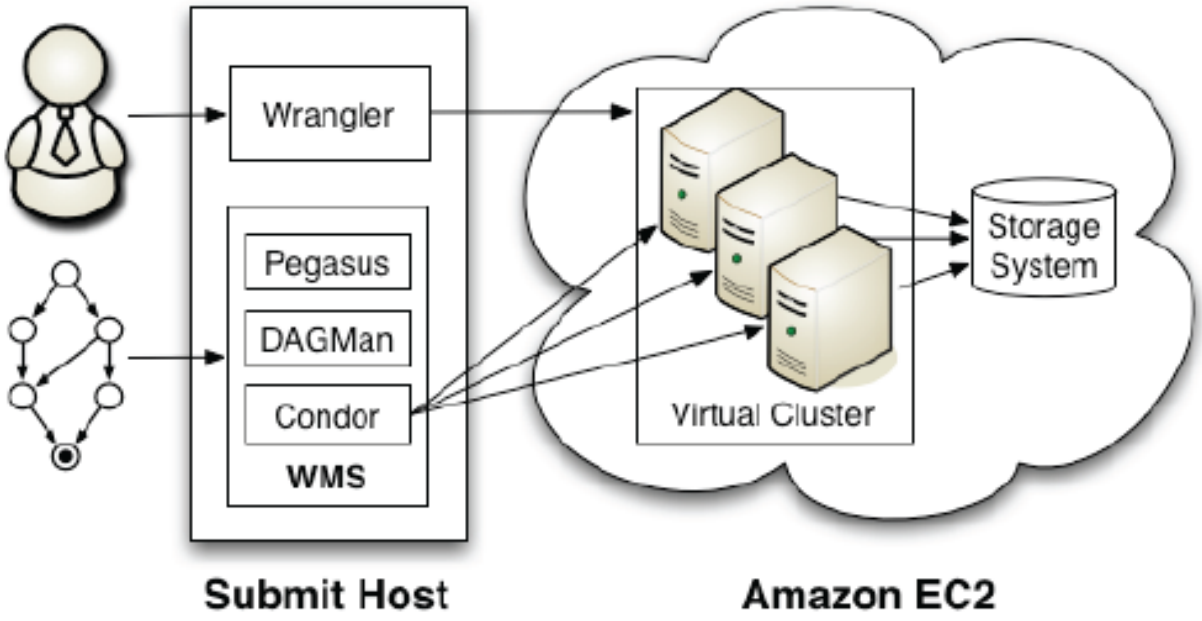
Examples of systems and frameworks for data analytics workflows

- ASKALON
- KEPLER
- TAVERNA
- ADEPT
- MapReduce/Hadoop
- R
- TRIDENT
- Apache ODE + WS-BPEL
- JOpera
- Pegasus
- Swift
- Airflow

Pros and cons of (data analytics) workflow systems

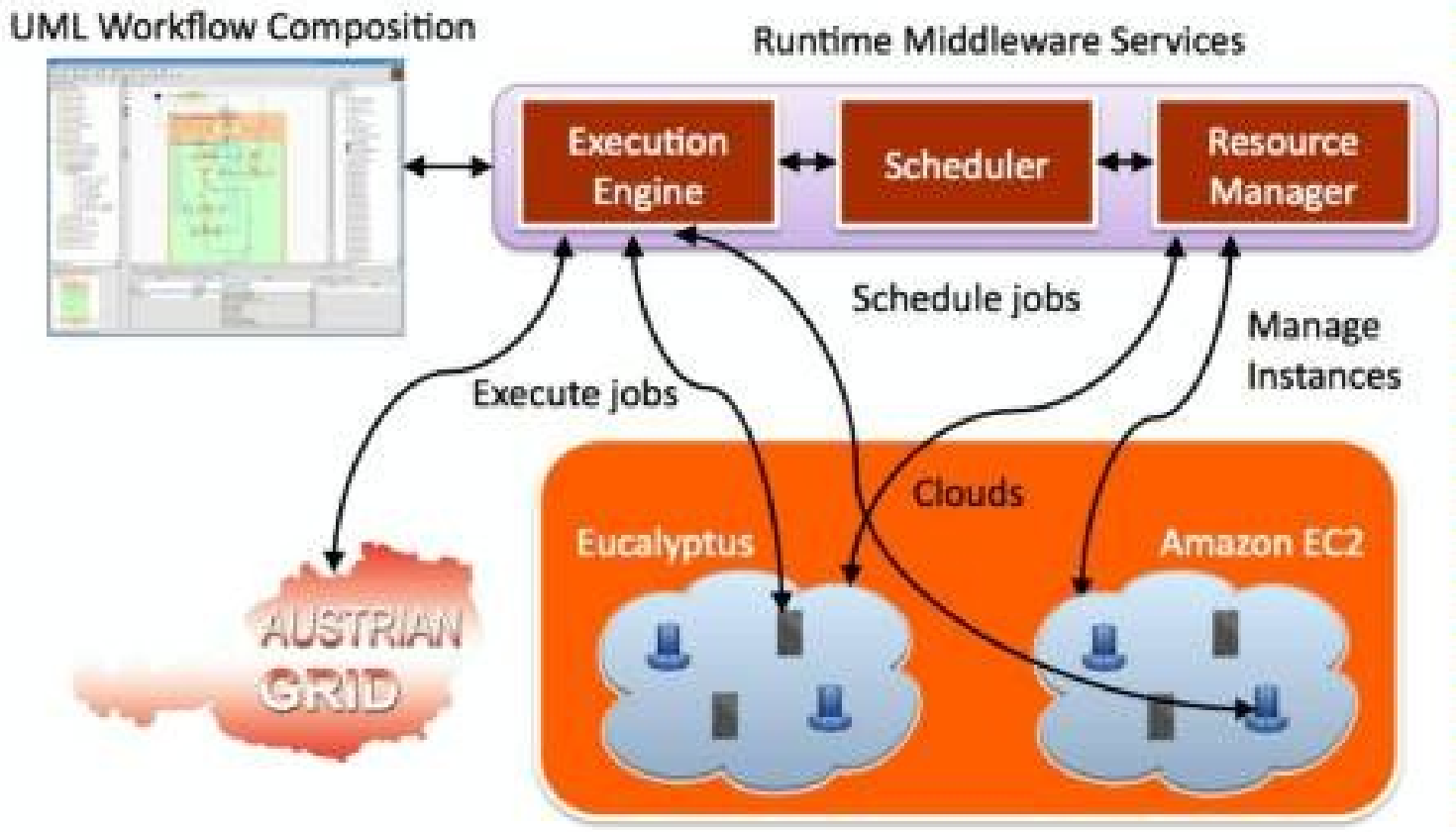
- Ian J. Taylor, Ewa Deelman, Dennis B. Gannon, and Matthew Shields. 2006. Workflows for E-Science: Scientific Workflows for Grids. Springer-Verlag New York, Inc., Secaucus, NJ, USA.
- Bertram Ludäscher, Mathias Weske, Timothy M. McPhillips, Shawn Bowers: Scientific Workflows: Business as Usual? BPM 2009: 31-47
- Mirko Sonntag, Dimka Karastoyanova, Frank Leymann: The Missing Features of Workflow Systems for Scientific Computations. Software Engineering (Workshops) 2010: 209-216
- Lavanya Ramakrishnan and Beth Plale. 2010. A multi-dimensional classification model for scientific workflow characteristics. In Proceedings of the 1st International Workshop on Workflow Approaches to New Data-centric Science (Wands '10). ACM, New York, NY, USA, , Article 4 , 12 pages. DOI=10.1145/1833398.1833402 http://doi.acm.org/10.1145/1833398.1833402
- Jia Yu and Rajkumar Buyya. 2005. A taxonomy of scientific workflow systems for grid computing. SIGMOD Rec. 34, 3 (September 2005), 44-49. DOI=10.1145/1084805.1084814 http://doi.acm.org/10.1145/1084805.1084814

Some examples (1)



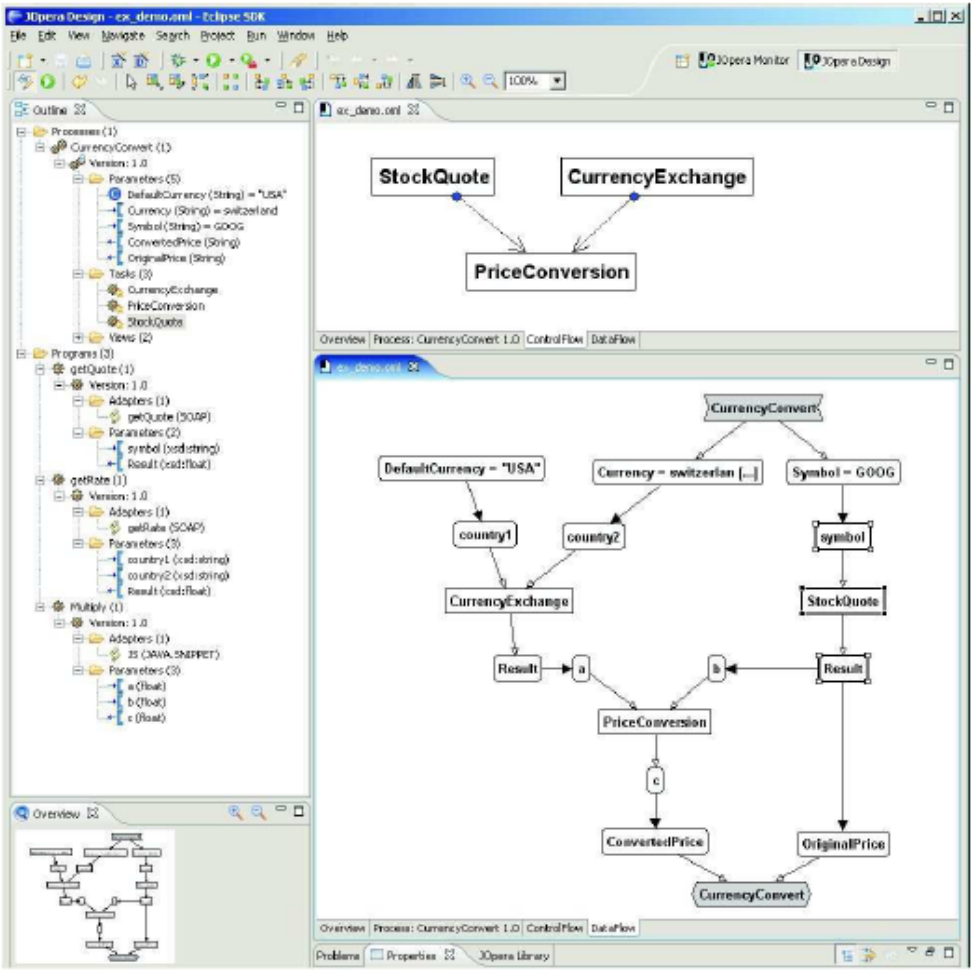
Source: Gideon Juve, Ewa Deelman, G. Bruce Berriman, Benjamin P. Berman, Philip Maechling: An Evaluation of the Cost and Performance of Scientific Workflows on Amazon EC2. J. Grid Comput. 10(1): 5-21 (2012)

Some examples (2)



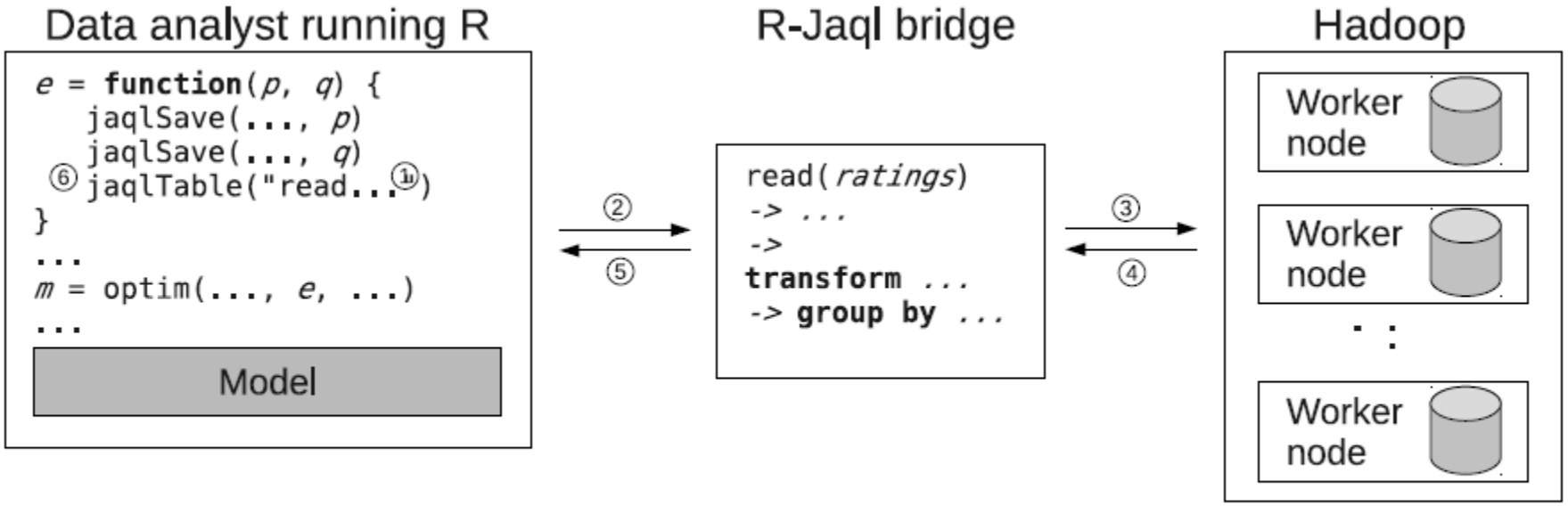
Source: http://www.dps.uibk.ac.at/projects/brokerage/

Some examples (3)



Source: Cesare Pautasso, Thomas Heinis, Gustavo Alonso: JOpera: Autonomic Service Orchestration. IEEE Data Eng. Bull. 29(3): 32-39 (2006)

Some examples (4)



① Issue query to compute gradients
② Forward query / parameters to Jaql
③ Execute the query in parallel on cluster
④ Fetch result
⑤ Format result as R data frame
⑥ Use the result in R

Source: Sudipto Das, Yannis Sismanis, Kevin S. Beyer, Rainer Gemulla, Peter J. Haas, and John McPherson. 2010. Ricardo: integrating R and Hadoop. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (SIGMOD '10). ACM, New York, NY, USA, 987-998. DOI=10.1145/1807167.1807275 http://doi.acm.org/10.1145/1807167.1807275

Airflow from Airbnb

- Workflow is a DAG (Direct Acyclic Graph)
- Task/Operator:
  - BashOperator, PythonOperator, EmailOperator, HTTPOperator, SqlOperator, Sensor,
  - DockerOperator, HiveOperator, S3FileTransferOperator, PrestoToMysqlOperator, SlackOperator

```
with DAG('my_dag', start_date=datetime(2016, 1, 1)) as dag:
    (
        dag
        >> DummyOperator(task_id='dummy_1')
        >> BashOperator(
            task_id='bash_1',
            bash_command='echo "HELLO!"')
        >> PythonOperator(
            task_id='python_1',
            python_callable=lambda: print("GOODBYE!"))
    )
```

Source: http://pythonhosted.org/airflow

Mapreduce

```
map(String key, String value):
    // key: document name
    // value: document contents
    for each word w in value:
        EmitIntermediate(w, "1");

reduce(String key, Iterator values):
    // key: a word
    // values: a list of counts
    int result = 0;
    for each v in values:
        result += ParseInt(v);
    Emit(AsString(result));
```

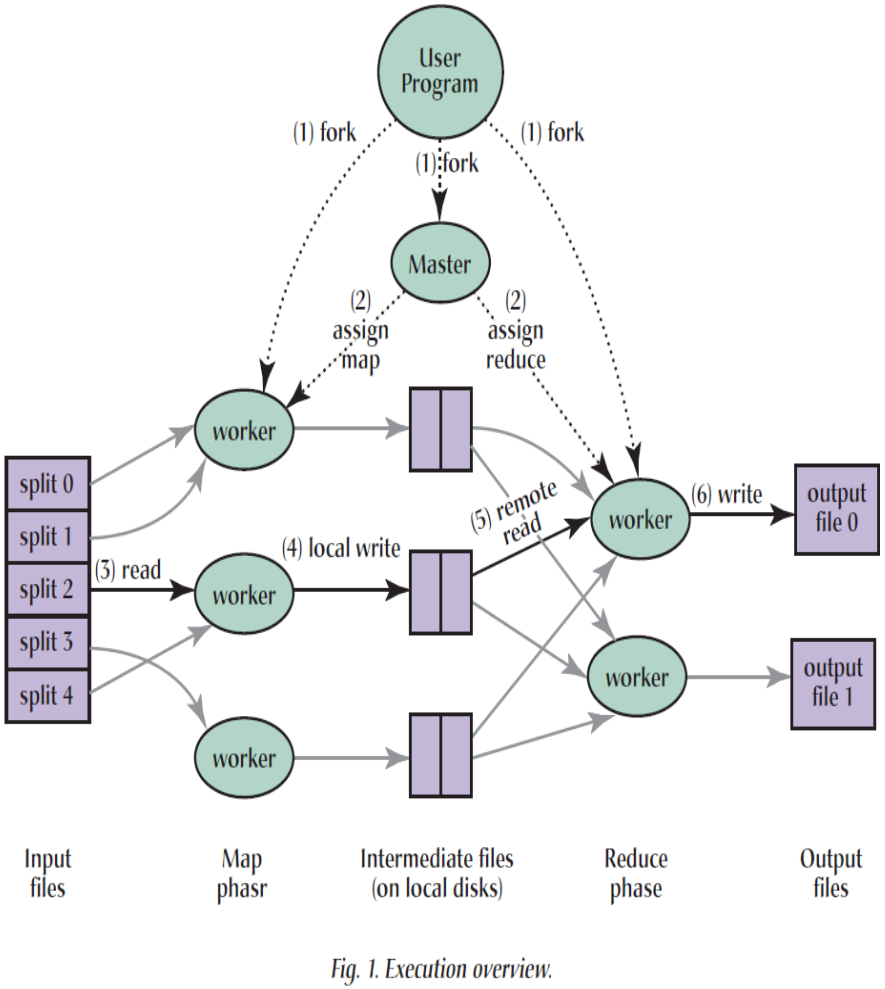


Fig. 1. Execution overview.

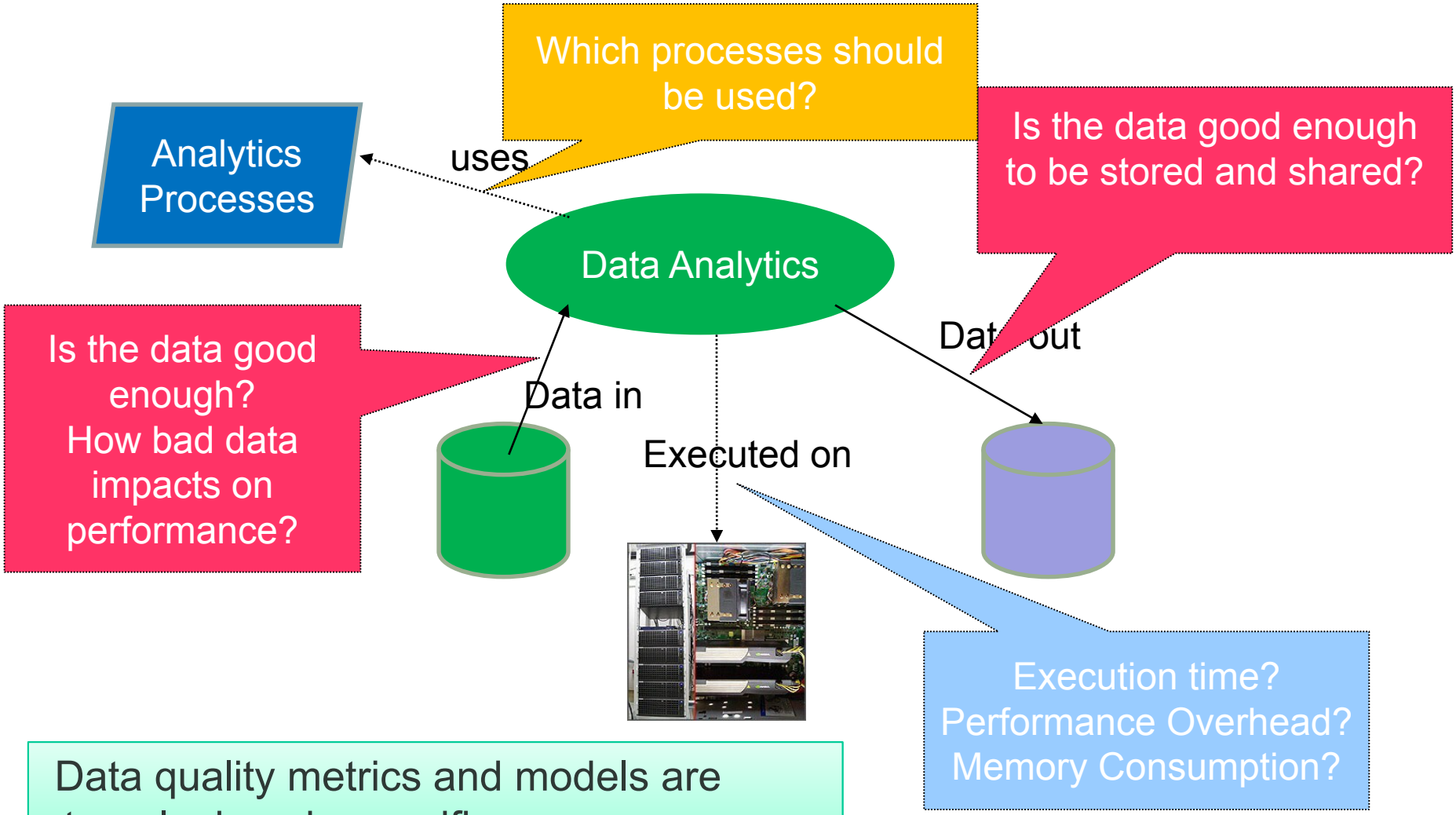
Source: Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. Commun. ACM 51, 1 (January 2008), 107-113. DOI=10.1145/1327452.1327492 http://doi.acm.org/10.1145/1327452.1327492

QUALITY OF ANALYTICS

Quality of Analytics (QoA)

- Characterize the results of analytics processes
- Different elements of QoA
  - Performance
  - Data quality
  - Cost
  - Form/data format of output results
  - Etc.
- Customer: expects QoA
- Provider: offers QoA and enforces QoA

Performance and Data Quality Aspects



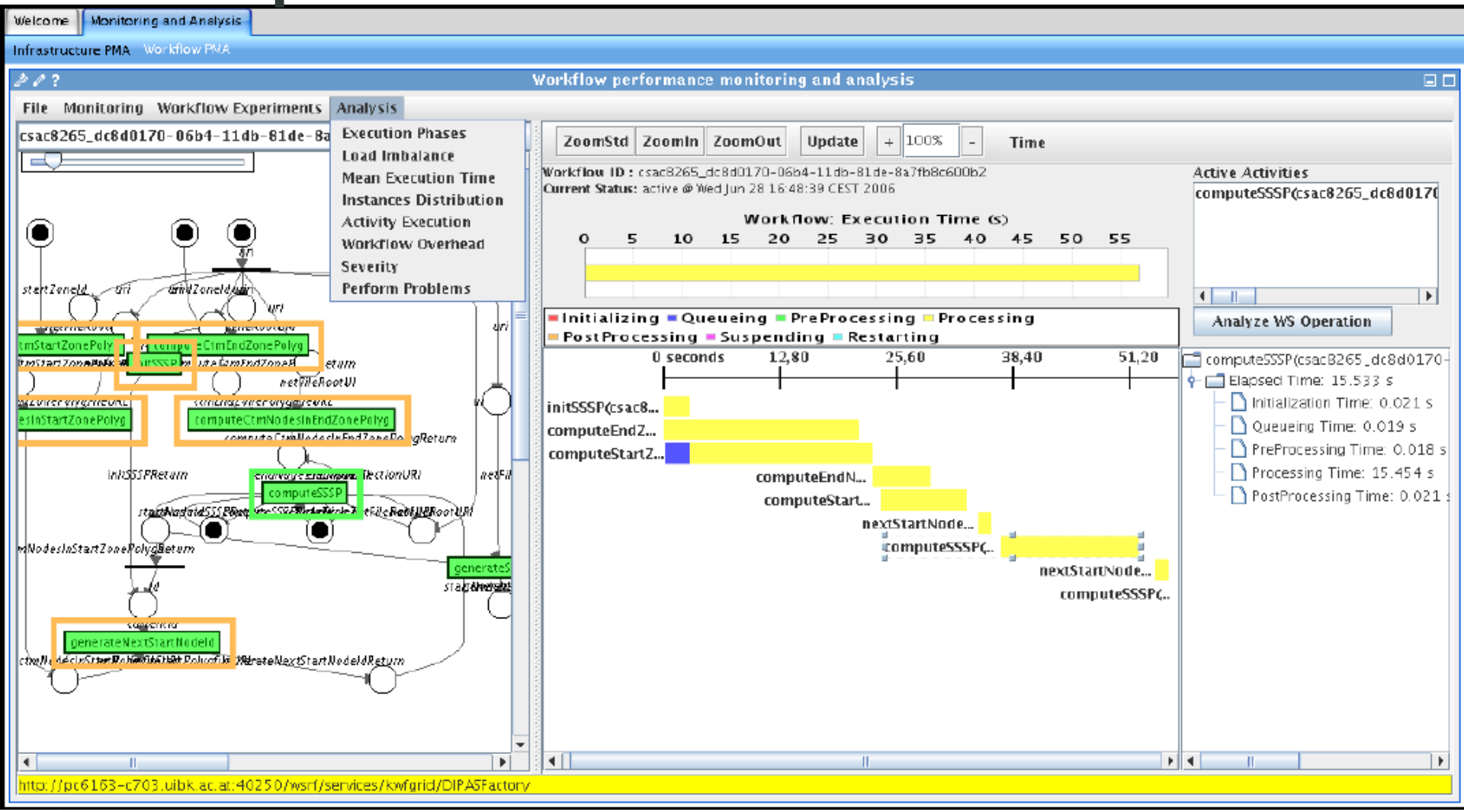
Data quality metrics and models are strongly domain-specific

SO HOW DO WE ENABLE QOA-AWARE ANALYTICS?

Solutions

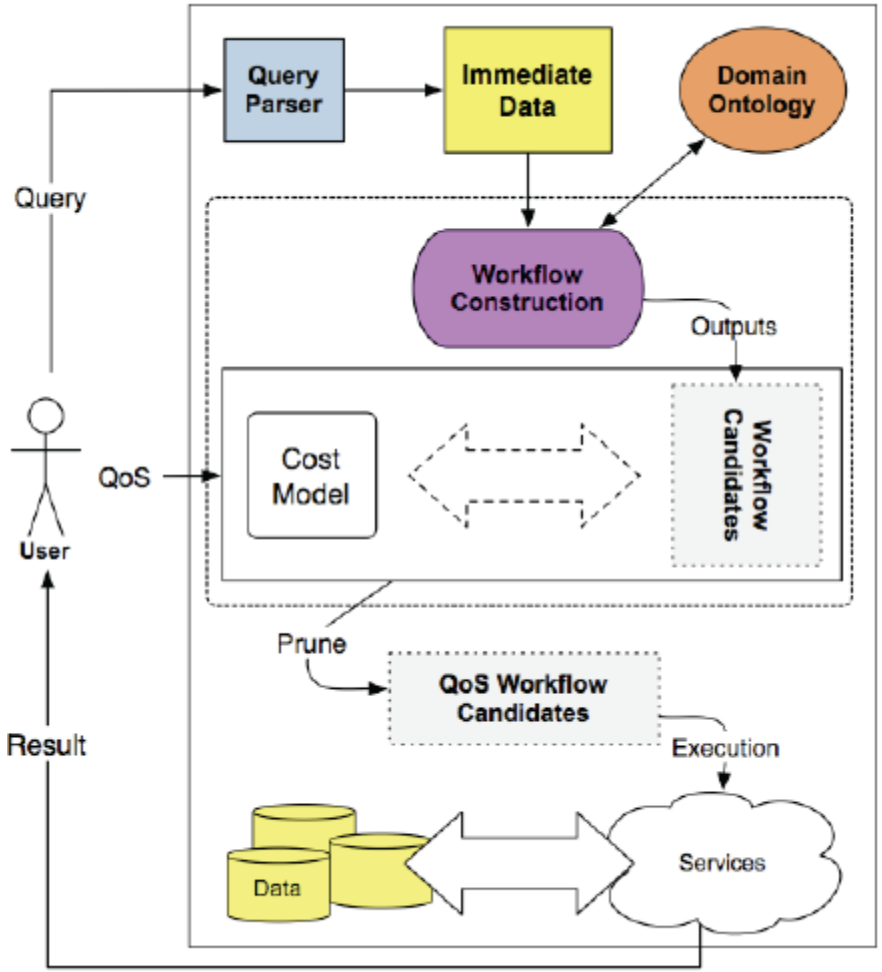
- Computational resources provisioning?
- Replication of analytics ?
- Performance and cost measurement and optimization?
- Improve quality of input data ?
- Improve the quality of output data?

Well-addressed concerns -- performance



Hong Linh Truong, Peter Brunner, Vlad Nae, Thomas Fahringer: DIPAS: A distributed performance analysis service for grid service-based workflows. Future Generation Comp. Syst. 25(4): 385-398 (2009)

Well-addressed concerns – performance/cost



Source: David Chiu, Sagar Deshpande, Gagan Agrawal, Rongxing Li: Cost and accuracy sensitive dynamic workflow composition over grid environments. GRID 2008: 9-16

QUALITY OF DATA IN DATA ANALYTICS WORKFLOWS

Very little support

- Qurator workbench
  - “Personal quality models” can be expressed and embedded into query processors or workflows.
  - Assume that quality evidence is presented
- Kepler
  - A data quality monitor allows user to specify quality thresholds.
  - Expect that rules can be used to control the execution based on quality.

P Missier, S M Embury, M Greenwood, A D Preece, & B Jin, Managing Information Quality in e-Science: the Qurator Workbench, Proc ACM International Conference on Management of Data (SIGMOD 2007), ACM Press, pages 1150-1152, 2007.

Aisa Na’im, Daniel Crawl,Maria Indrawan, Ilkay Altintas, and Shulei Sun. Monitoring data quality in kepler. In Salim Hariri and Kate Keahey, editors, HPDC, pages 560–564. ACM, 2010.

Research questions

- What are main QoD metrics, what are the relationship between QoD metrics and other service level objectives, and what are their roles and possible trade-offs?
- How to support different domain-specific QoD models and link them to workflow structures?
- How to model, evaluate and estimate QoD associated with data movement into, within, and out to workflows? When and where software or scientists can perform automatic or manual QoD measurement and analysis
- How to optimize the workflow composition and execution based on QoD specification?
- How does QoD impact on the provisioning of data services, computational services and supporting services?

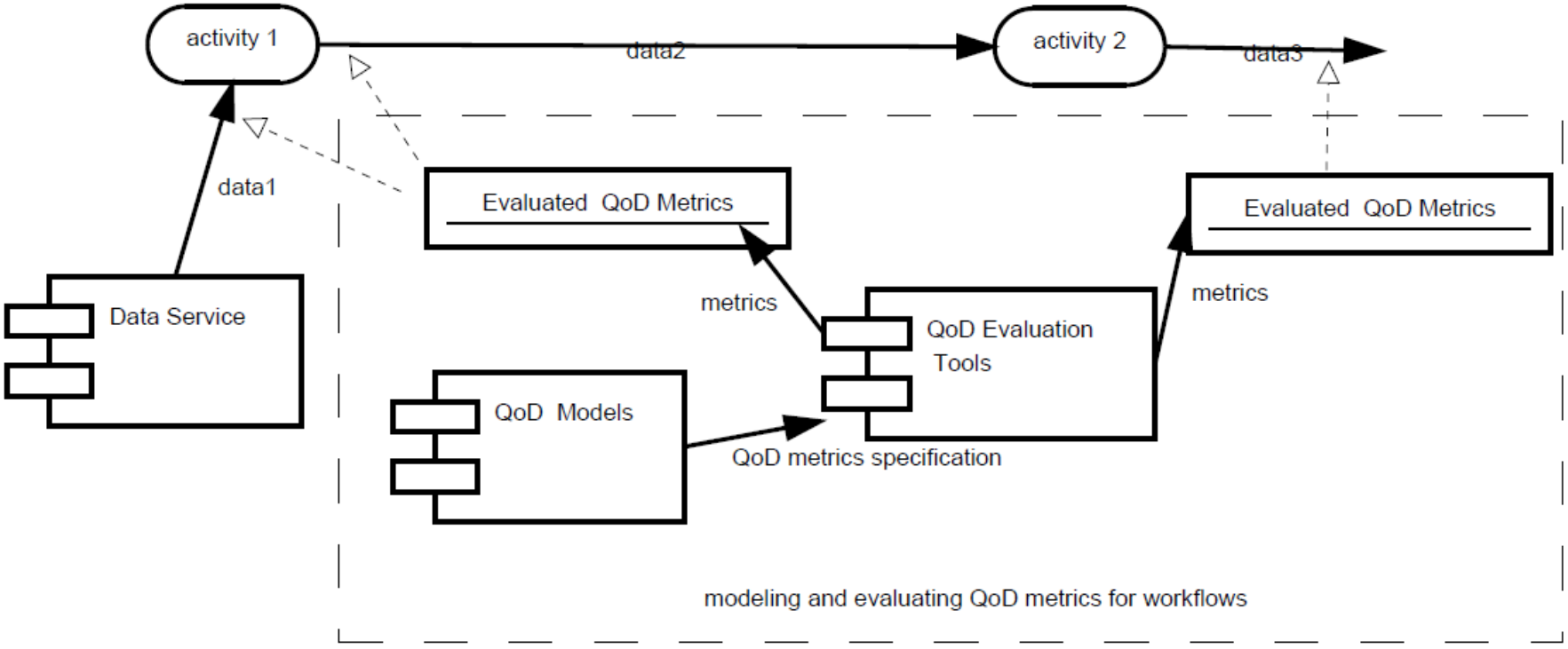
Approach

Core models, techniques and algorithms to allow the modeling and evaluating QoD metrics

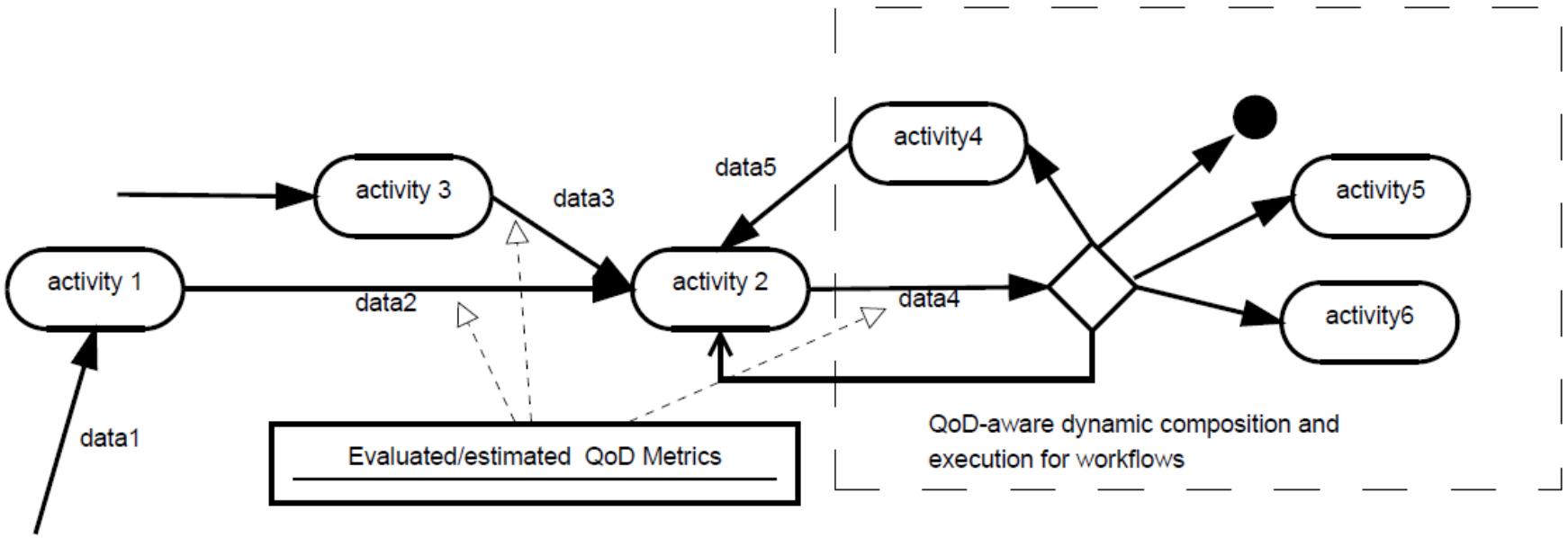
QoD-aware composition and execution

QoD-aware service provisioning and infrastructure optimization

Modeling and evaluating QoD metrics for data analytics workflows



QoD-aware optimization for data analytics workflow composition and execution



HOW TO INTEGRATE QOD EVALUATORS? AND WHICH CONCERNS NEED TO BE CONSIDERED?

QoD metrics evaluation

- Domain-specific metrics
  - Need specific tools and expertise for determining metrics
- Evaluation
  - Cannot done by software only: humans are required
- Complex integration model
  - Where to put QoD evaluators and why?
  - How evaluators obtain the data to be evaluated?
- Impact of QoD evaluation on performance of data analytics workflows

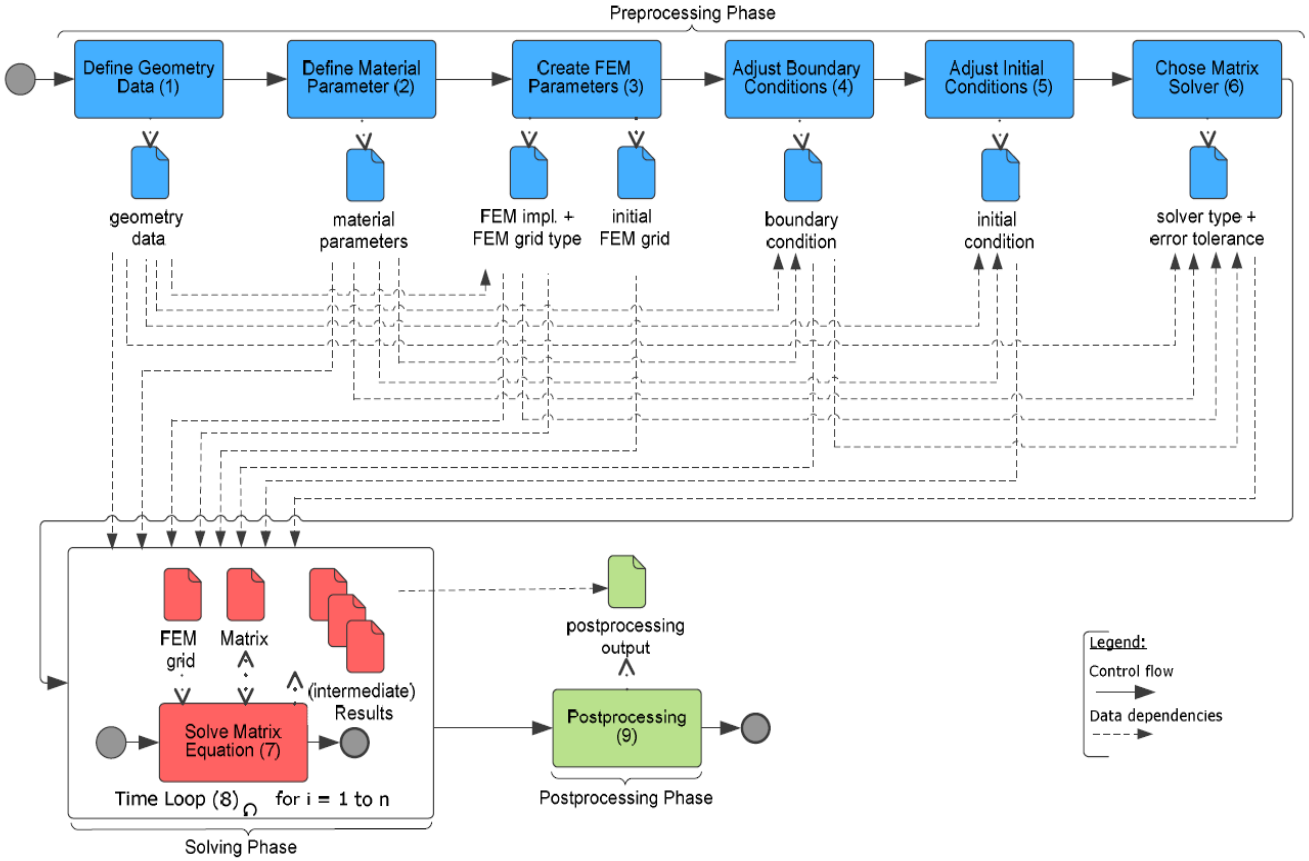
WHAT KIND OF OPTIMIZATION CAN BE DONE?

QoD-aware optimization for data analytics workflows

- Improving quality of analytics
- Reducing analytics costs and time
- Enabling early failure detection
- Enabling elasticitiy of services provisioning
- Enabling elastic data analytics support
- Etc.

EXAMPLE: QOD-AWARE SIMULATION WORKFLOWS

QoD-aware simulation workflows

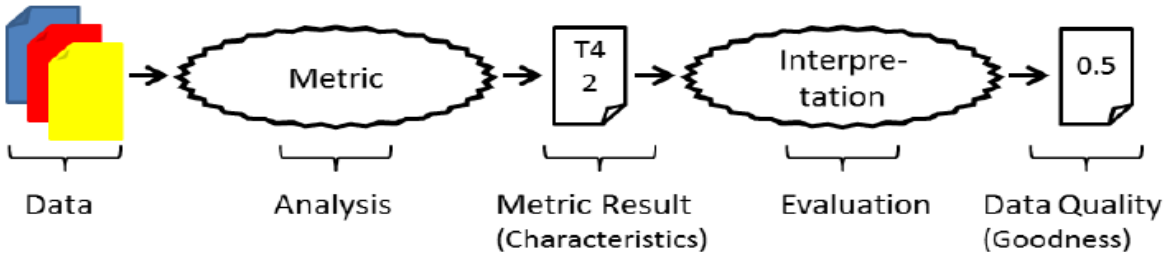


Michael Reiter, Hong Linh Truong, Schahram Dustdar, Dimka Karastoyanova, Robert Krause, Frank Leymann, Dieter Pahr: On Analyzing Quality of Data Influences on Performance of Finite Elements Driven Computational Simulations. Euro-Par 2012: 793-804

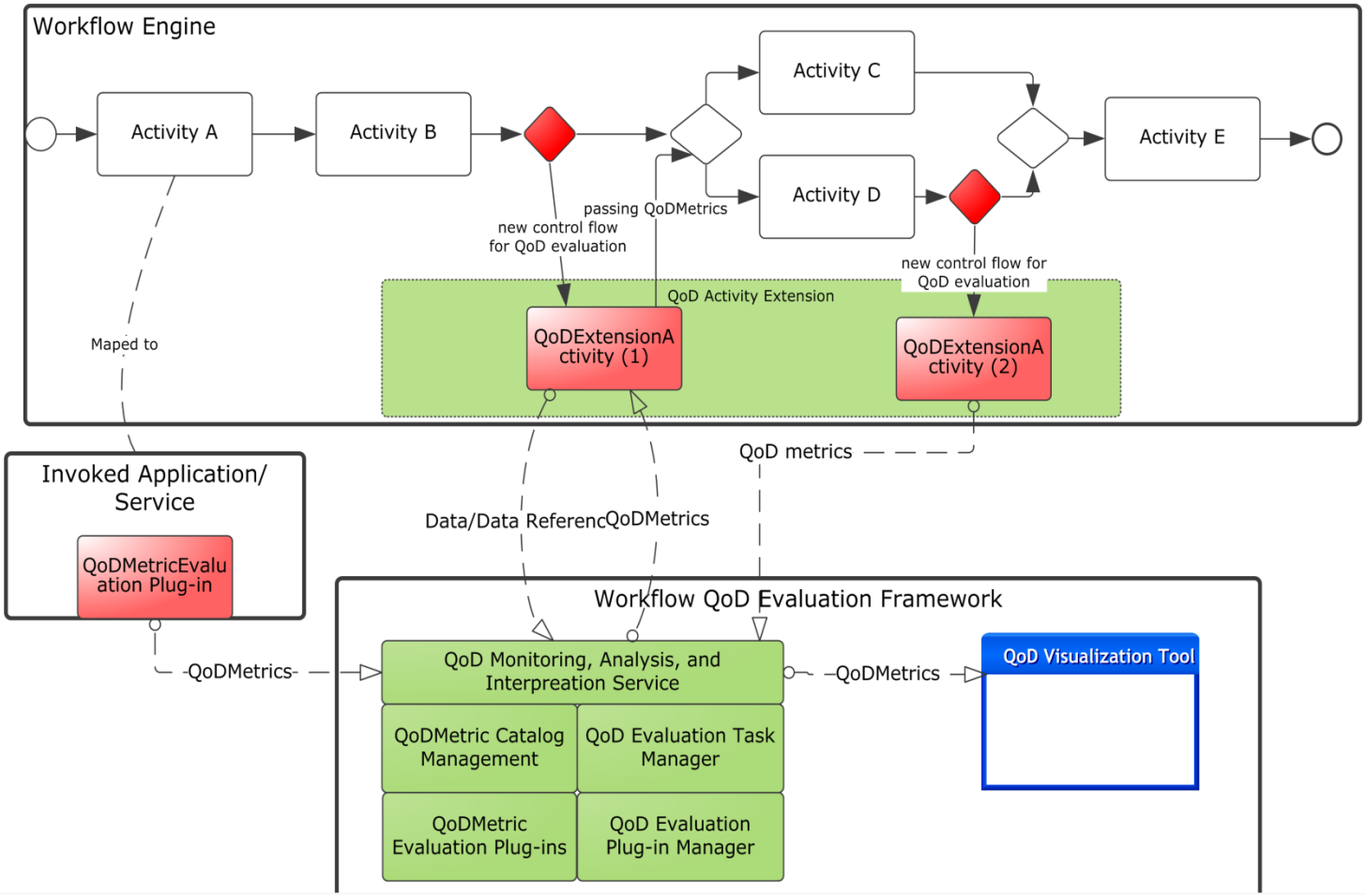
Michael Reiter, Uwe Breitenbücher, Schahram Dustdar, Dimka Karastoyanova, Frank Leymann, Hong Linh Truong: A Novel Framework for Monitoring and Analyzing Quality of Data in Simulation Workflows. eScience 2011: 105-112

Hybrid resources needed for quality evaluation

- Challenges:
  - Subjective and objective evaluation
  - Long running processes
- Our approach
  - Different QoD measurements
  - Human and software tasks



Evaluating quality of data in workflows



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

QoD Evaluator

- Software-based QoD evaluators
  - Can be provided under libraries integrated into invoked applications
  - Web services-based evaluators
- Human-based QoD evaluators
  - Built based on the concept human-based services
  - Can be interfaces via Human-Task
  - Simple mapping at the moment
    - Human resources from clouds/crowds

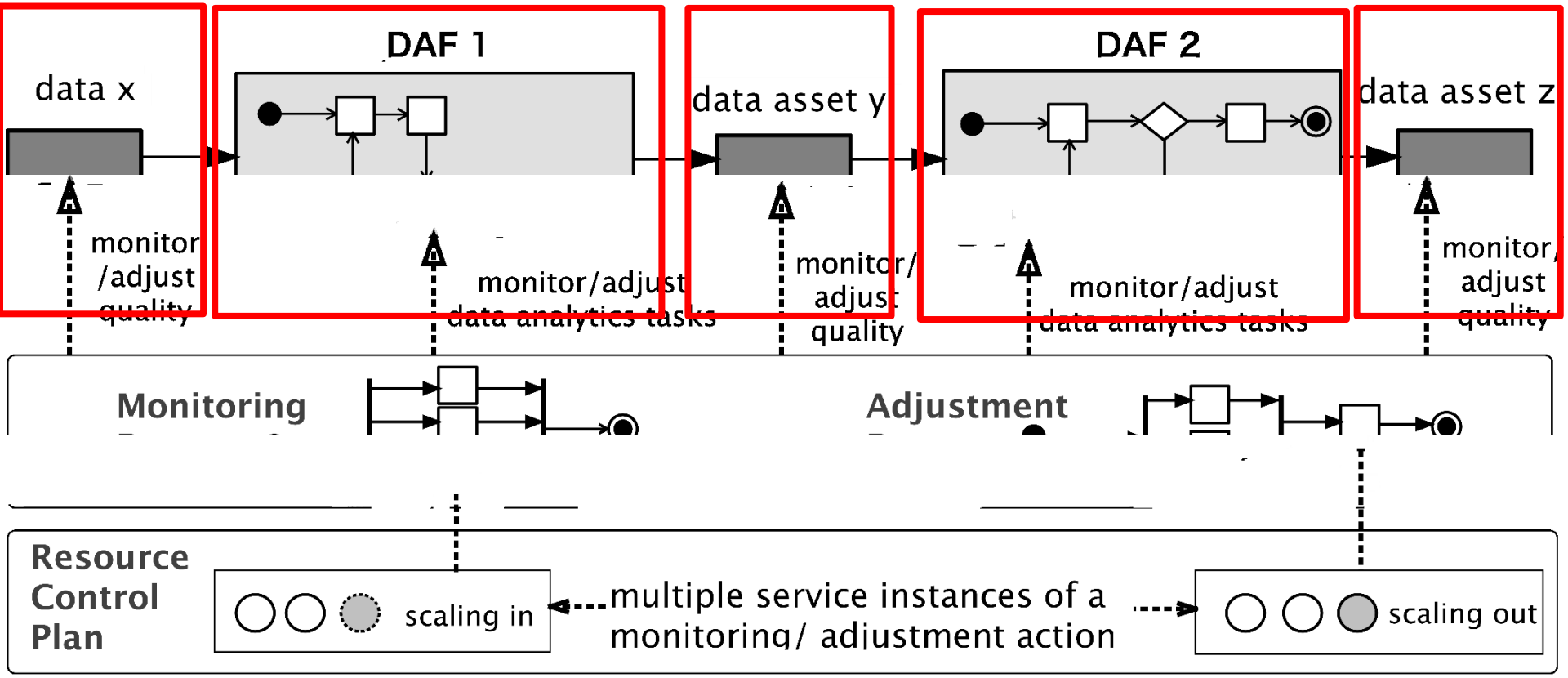
How to support QoA driven analytics with tradeoffs of multiple criteria?

QoA: QoD, performance, cost, etc.

Quality-of-analytics driven workflows

- How to support QoA driven analytics?
- Some basic steps
  - Conceptualize expected QoA
  - Associate the expected QoA with workflow activities
  - Use the expected QoA
    - to match/select underlying services (e.g., data sources, cloud IaaS, etc
  - Utilize the expected QoA and the measured QoA and apply elasticity principles for
    - Refine the workflow structure
    - Provision computation, network and data

Using Data Elasticity Management Process to ensure QoA



Tien-Dung Nguyen, Hong Linh Truong, Georgiana Copil, Duc-Hung Le, Daniel Moldovan, Schahram Dustdar:
On Developing and Operating of Data Elasticity Management Process. ICSOC 2015: 105-119

Exercises

- Read mentioned papers
- Discuss pros and cons of descriptive languages - and programming languages – based data analytics workflows
- Examine how QoD evaluators can be integrated into different programming models for QoA-aware data analytics workflows
- Implement some QoD evaluators
- Develop techniques for determining places where QoD evaluators can be performed in your mini projects
- Support data elasticity management in your mini project

Thanks for your attention

Hong-Linh Truong
Distributed Systems Group, TU Wien
truong@dsg.tuwien.ac.at
http://dsg.tuwien.ac.at/staff/truong
@linhsolar