Quality-aware data analytics services

Hong-Linh Truong
Faculty of Informatics, TU Wien

hong-linh.truong@tuwien.ac.at
http://www.infosys.tuwien.ac.at/staff/truong
@linhsolar

What this lecture is about

- Big Data analytics services – general view
  - The meaning of quality-aware data analytics services
- Incident management for cloud-based big data analytics service systems
  - Concepts and approaches
- Quality of analytics (QoA) for data analytics services
  - Quality of data in data analytics workflows
  - Data elasticity management

What this lecture is about

- After this lecture
  - Make sure that you can monitor incidents in your systems
  - 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

Data: facts, responses, events, measurement, etc.

{"station_id":"1160629000","datapoint_id":122,"alarm_id":310,"event_time":"2016-09-17T02:05:54.000Z","isActive":false,"value":6,"valueThreshold":10}

What does it mean "Big data"?

NYC Taxi Data

The official TLC trip record dataset contains data for over 1.1 billion taxi trips from January 2009 through June 2015, covering both yellow and green taxis. Each individual trip record contains precise location coordinates for where the trip started and ended, timestamps for when the trip started and ended, plus a few other variables including fare amount, payment method, and distance traveled.

Open Big Data / Telecommunications - SMS, Call, Internet - MI

Description Tabular Preview API Resources

Schema

1. **Square id**: the id of the square that is part of the Milano GRID; TYPE: numeric
2. **Time interval**: the beginning of the time interval expressed as the number of millisecond elapsed from the Unix Epoch on January 1st, 1970 at UTC. The end of the time interval can be obtained by adding 600000 milliseconds (10 minutes) to this value. TYPE: numeric
3. **Country code**: the phone country code of a nation. Depending on the measured activity this value assumes different meanings that are explained later. TYPE: numeric
4. **SMS-in activity**: the activity in terms of received SMS inside the Square id, during the Time interval and sent from the nation identified by the Country code. TYPE: numeric
5. **SMS-out activity**: the activity in terms of sent SMS inside the Square id, during the Time interval and received by the nation identified by the Country code. TYPE: numeric
6. **Call-in activity**: the activity in terms of received calls inside the Square id, during the Time interval and issued from the nation identified by the Country code. TYPE: numeric
7. **Call-out activity**: the activity in terms of issued calls inside the Square id, during the Time interval and received by the nation identified by the Country code. TYPE: numeric
8. **Internet traffic activity**: the activity in terms of performed internet traffic inside the Square id, during the Time interval and by the nation of the users performing the connection identified by the Country code . TYPE: numeric

Big Data

- Sources
  - Internet of Things (IoT), 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

Characterize big data

- Big data is often characterized by the concepts of V*: Volume, Variety, Velocity, Veracity and Valence
  - Volume: size (big size, large-data set, massive of small data)
  - Variety: complexity (formats, types of data)
  - Velocity: speed (generating speed, data movement speed)
  - Veracity: quality is very different (bias, accuracy, etc.)
  - Valence: potential/possible relationships among different type of data w.r.t data combination

Data Management/Delivery Systems

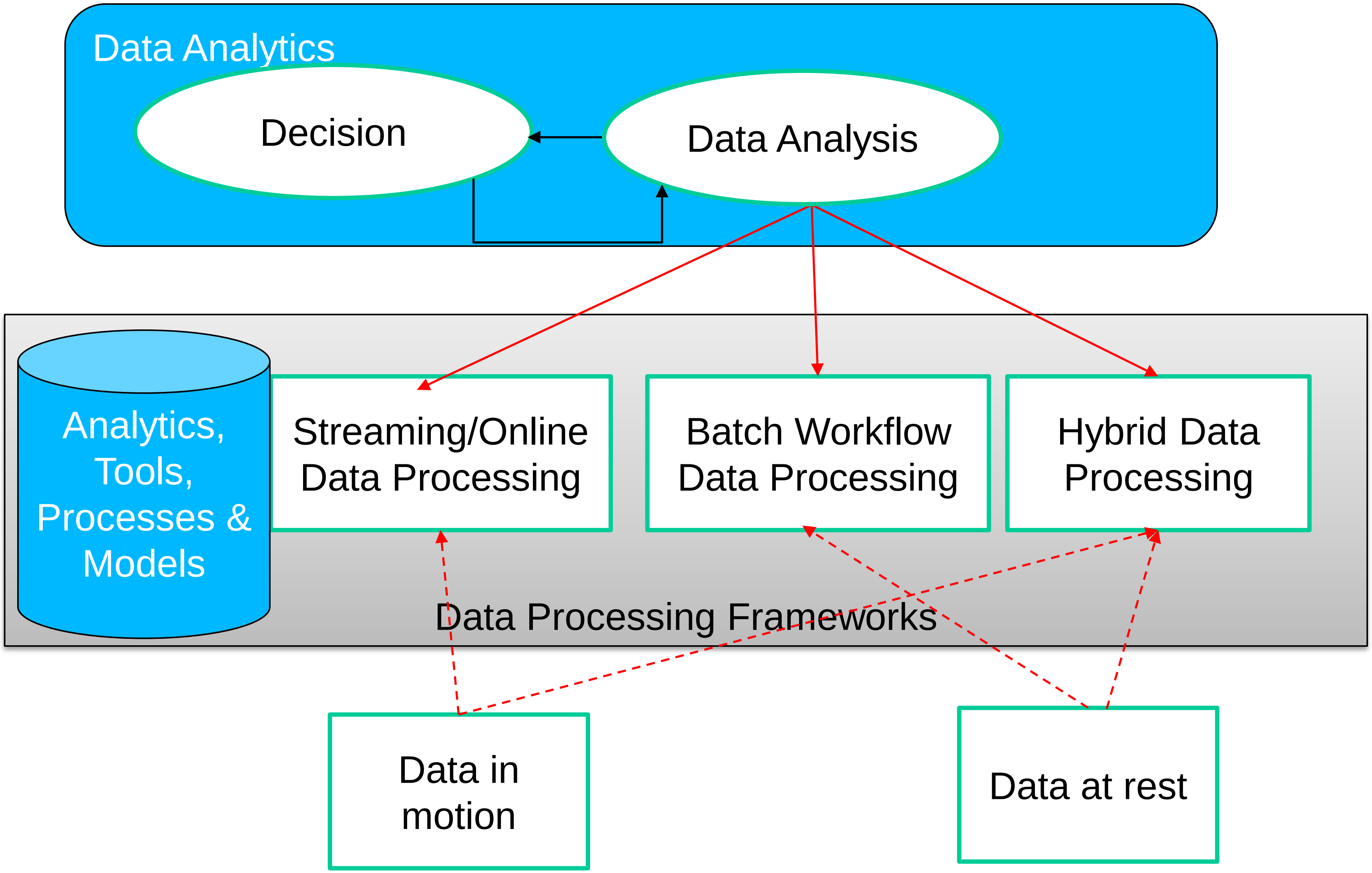
- Static data – data at rest
  - Hadoop file systems
  - Large scale storage data systems
    - iRODS, BigQuery, and other NoSQL
  - Web services for Data-as-a-Service (e.g., GIS)
- Real time data – data in motion
  - Cloud data platforms
  - Several MOM (Message-oriented Middleware)
    - E.g., Apache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

Data Processing Framework

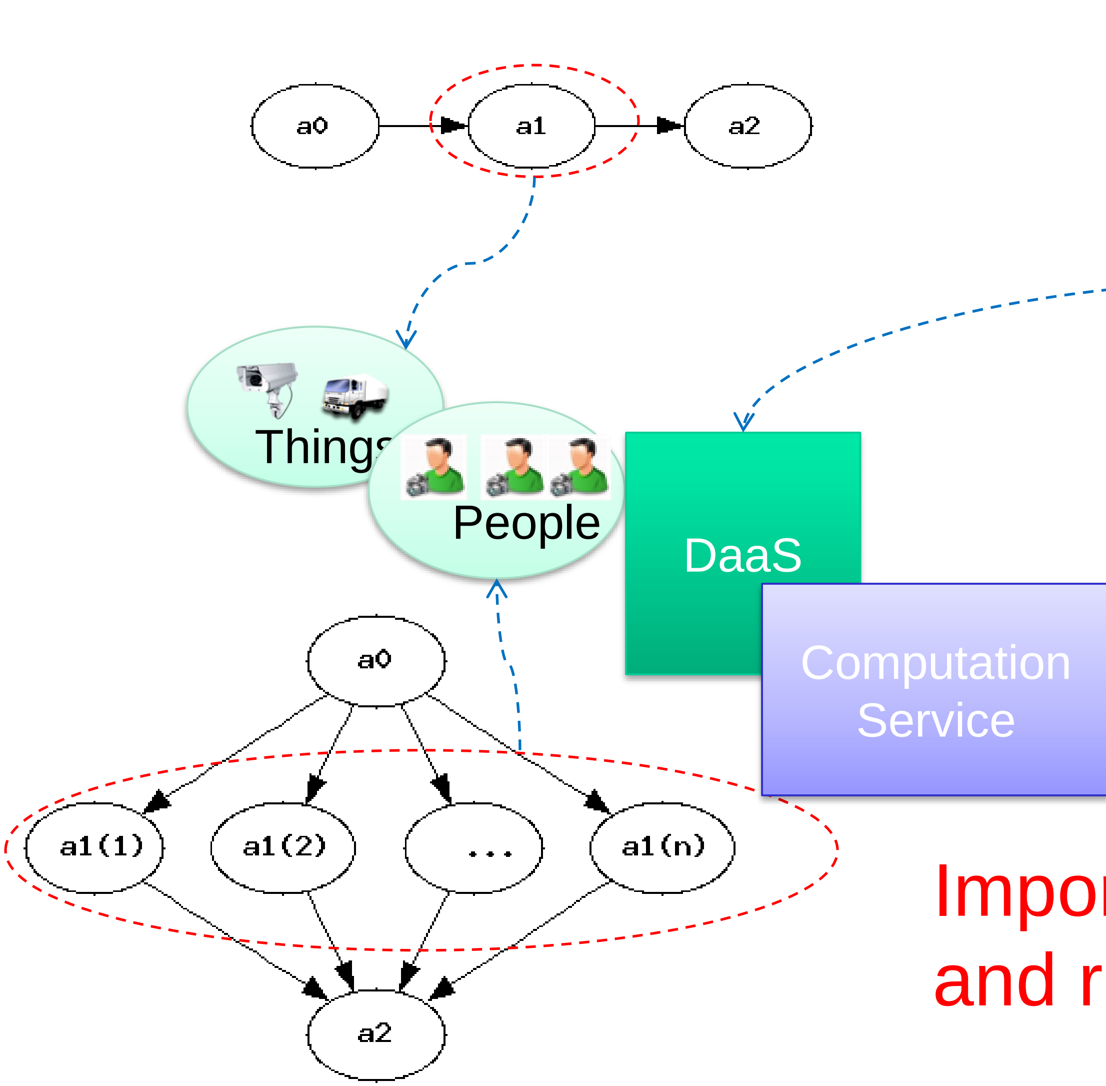
- Batch processing
  - Mapreduce/Hadoop
  - Data pipelines/Data flows
  - Scientific workflows

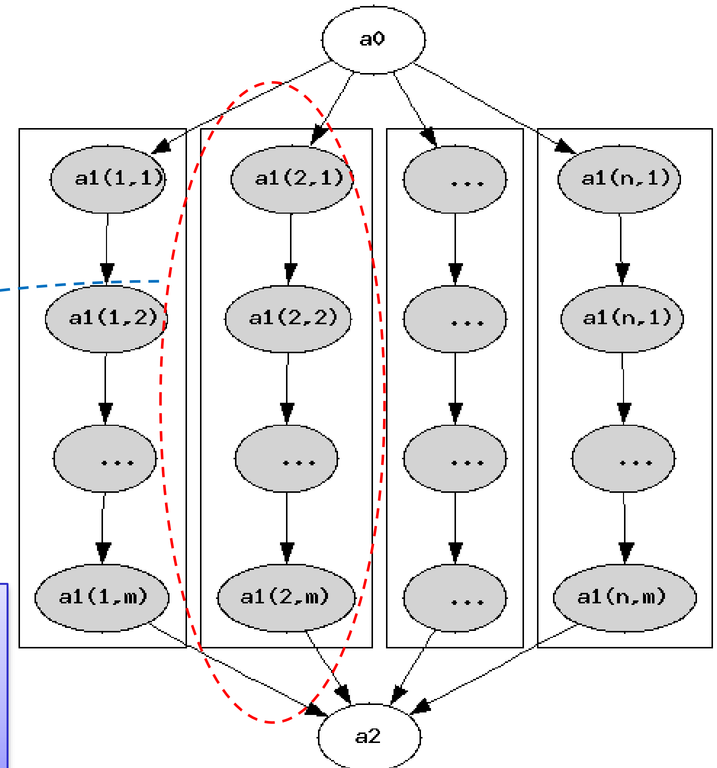
- (Near) realtime streaming processing
  - Apache Flink, Apache Kafka Streaming, Apache Apex, Apache Spark

Data Analytics: Analysis + Decision



Analysis: workflow models

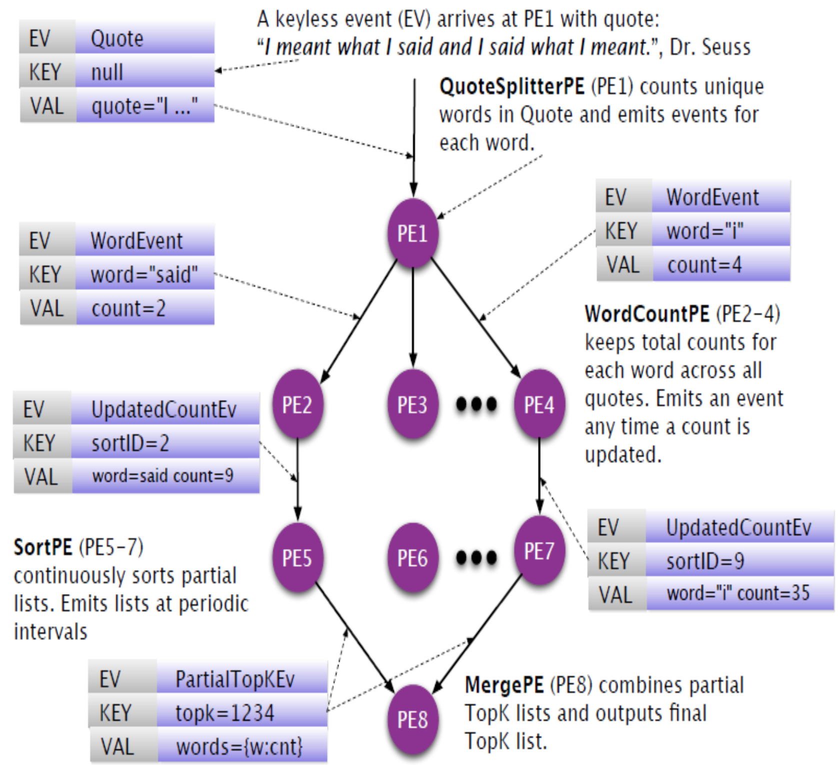




Important notes: Structures and resources

Analysis: 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

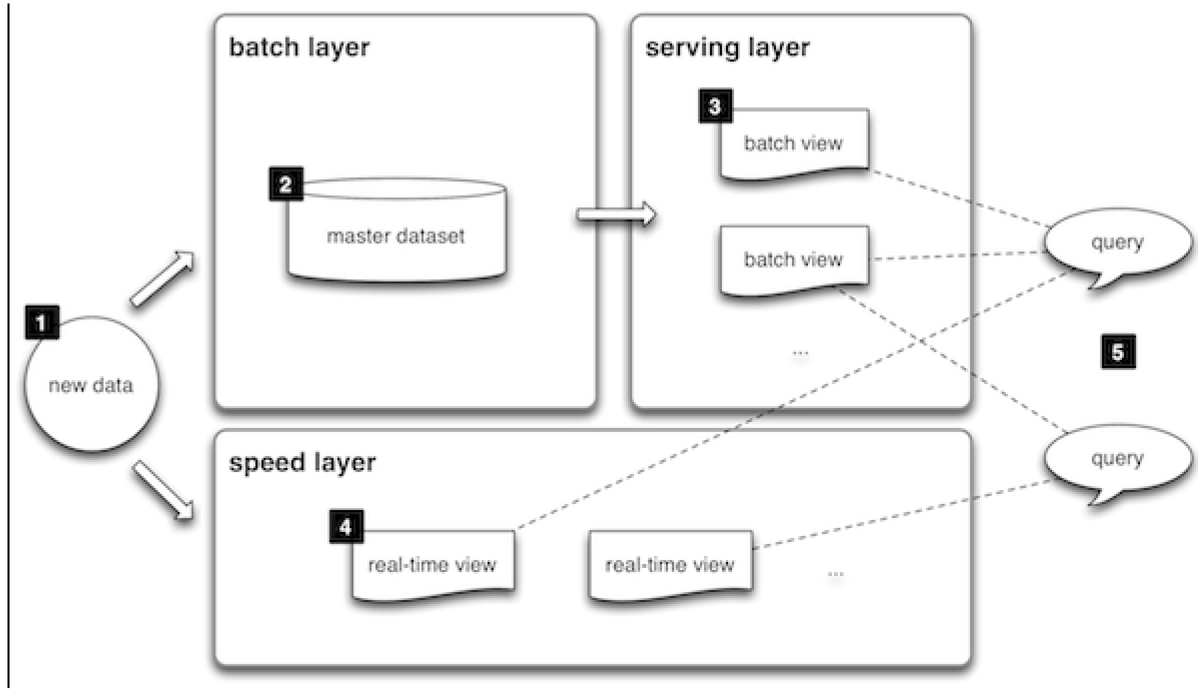
Figure 1. Word Count Example

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/staff/truong/dst/pdfs/truong-dst2018-lecture5.pdf

Analysis: hybrid data processing

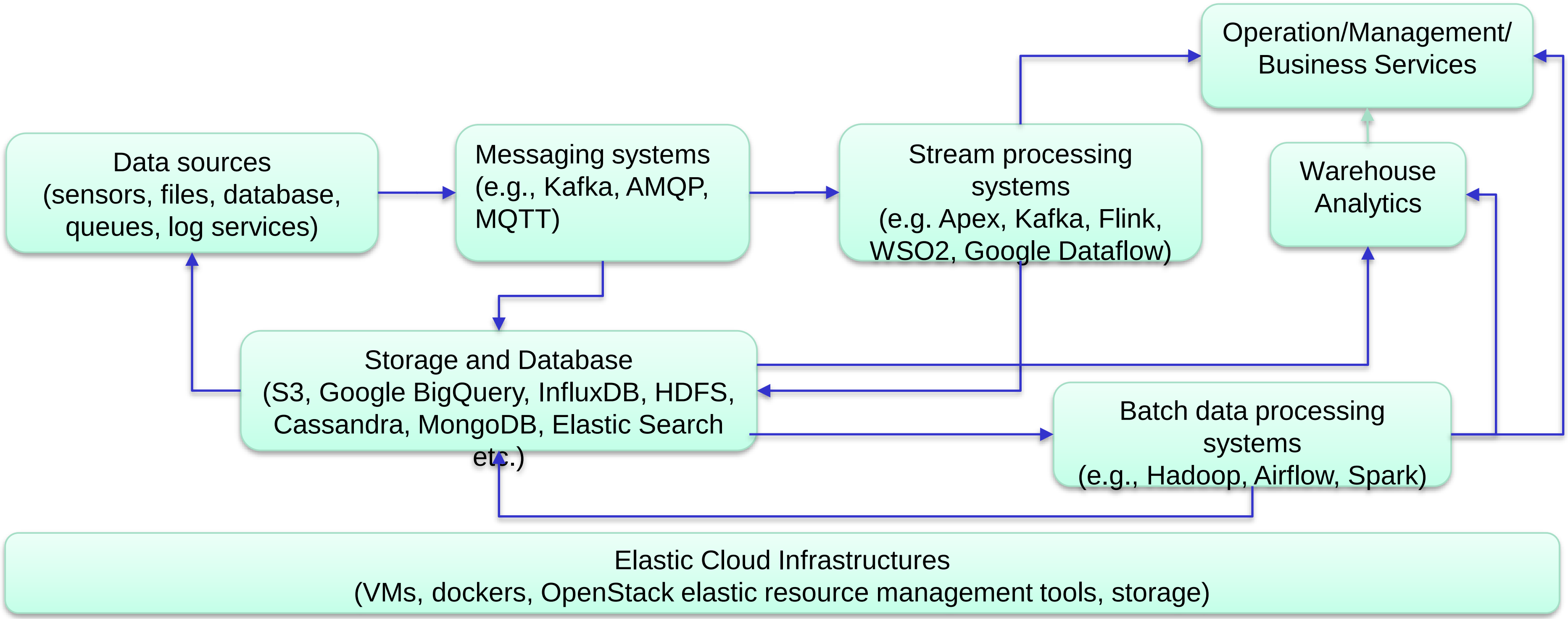
Combine batch processing and streaming processing
e.g., https://spark.apache.org/



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

Which scenarios should we use a combination?

Cloud services and big data analytics



What do we mean by quality-aware data analytics:

Able to determine quality and incidents, establish their relationships and optimize the system accordingly based on constraints on quality and incidents

Incidents

[1]The IT Infrastructure Library (ITIL) defines an incident *"as an unplanned interruption to an IT service or reduction in the quality of an IT service or a failure of a Configuration Item that has not yet impacted an IT service"*.

Check: https://en.wikipedia.org/wiki/ITIL

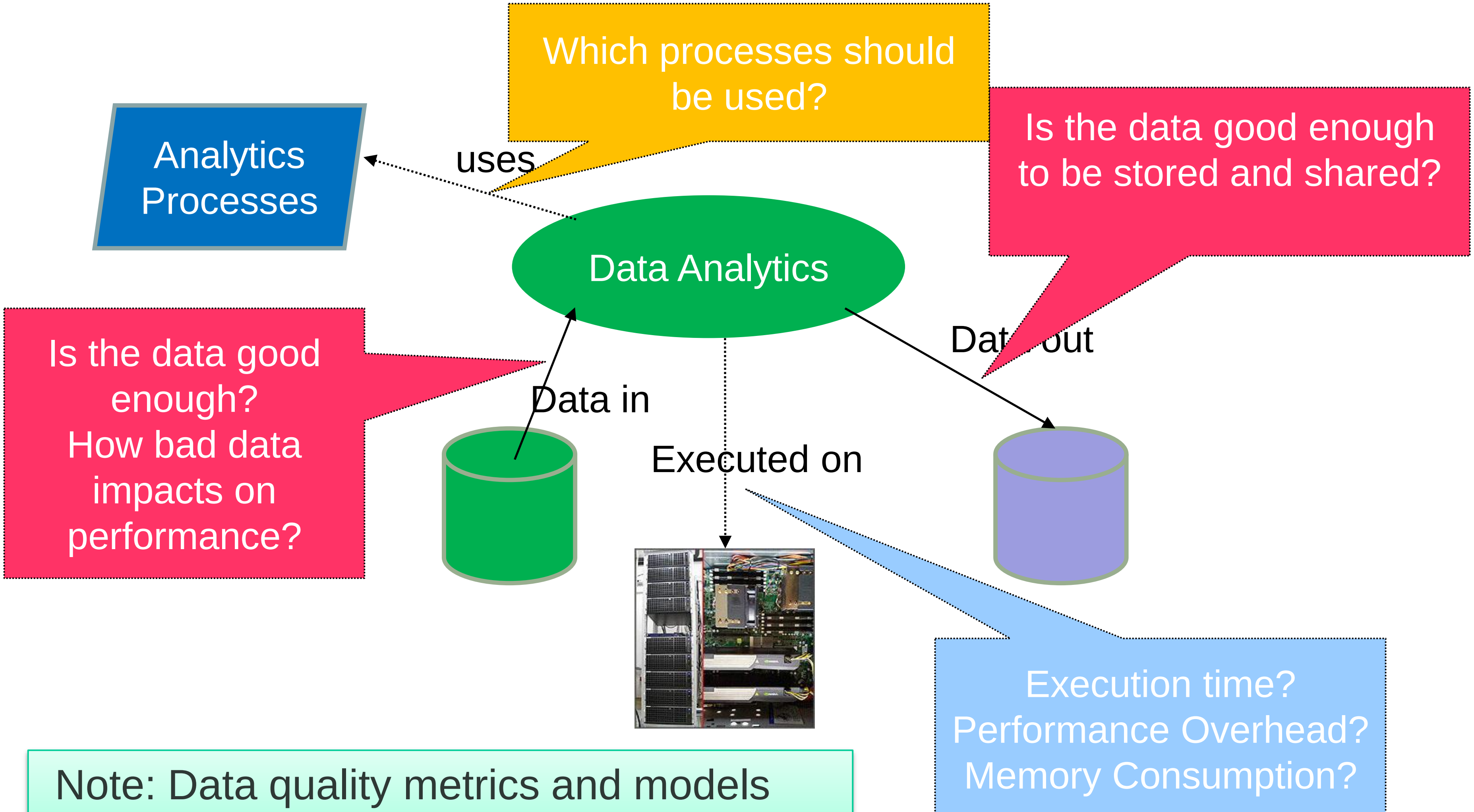
- System incidents
- Data incidents
- Processing incidents
- Cross systems and cross layers

ITIL: https://www.axelos.com/best-practice-solutions/itil

Quality of Analytics (QoA)

- Characterize the results of analytics processes
- Different elements of QoA
  - Performance (e.g. Execution time)
  - Quality of data/data quality
  - Cost
  - Data format of output results
  - Etc.
- Customer: expects QoA
- Provider: offers QoA and enforces QoA

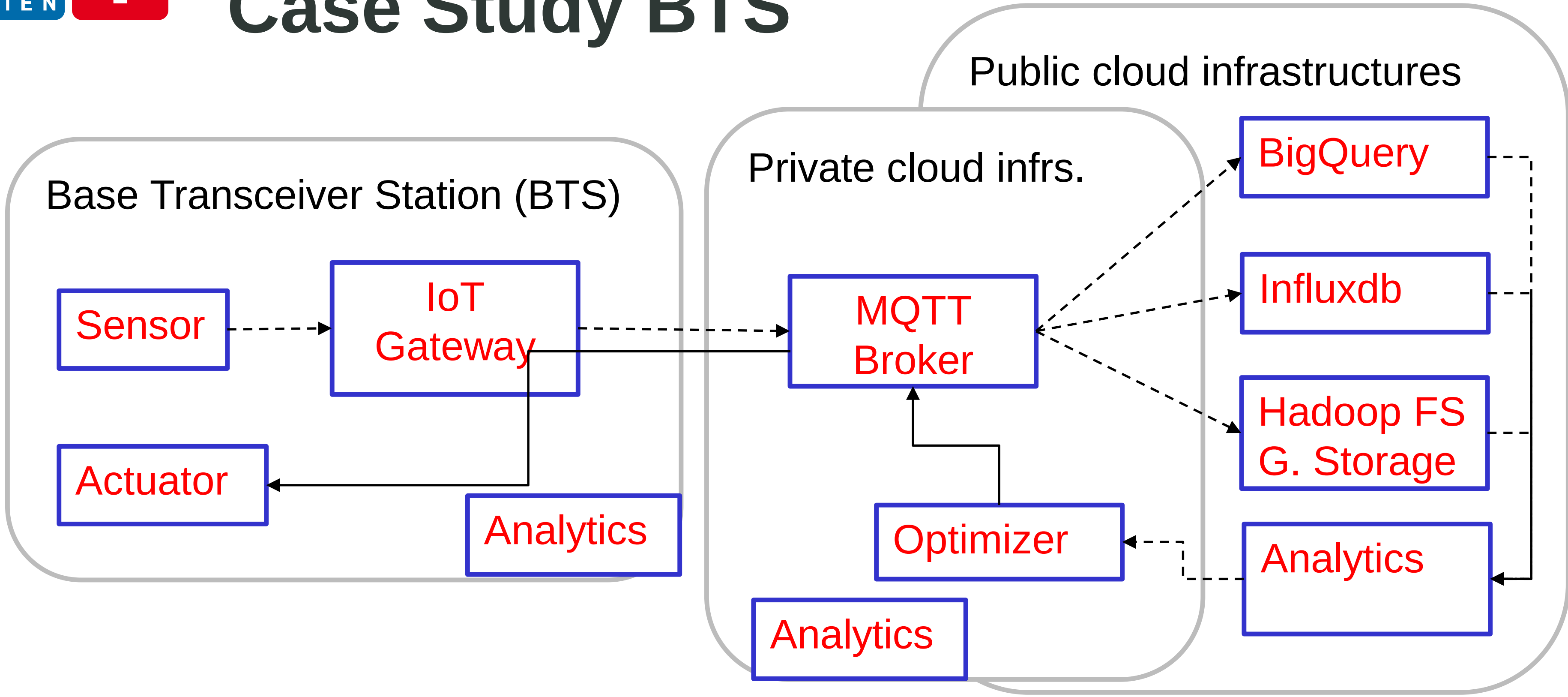
A simple QoA view



Note: Data quality metrics and models are strongly domain-specific

INCIDENTS IN CLOUD-BASED BIG DATA

Case Study BTS



- Large-scale systems (1K+ BTS)
- Flexible back-end clouds
  - Generic enough for other applications (e.g., in smart agriculture)
- With bad infrastructures for IoT and connectivity

If you monitor alarms in BTSs and see this

What could be happened?

Challenges

The ultimate goal of the (domain) data scientist is to meet

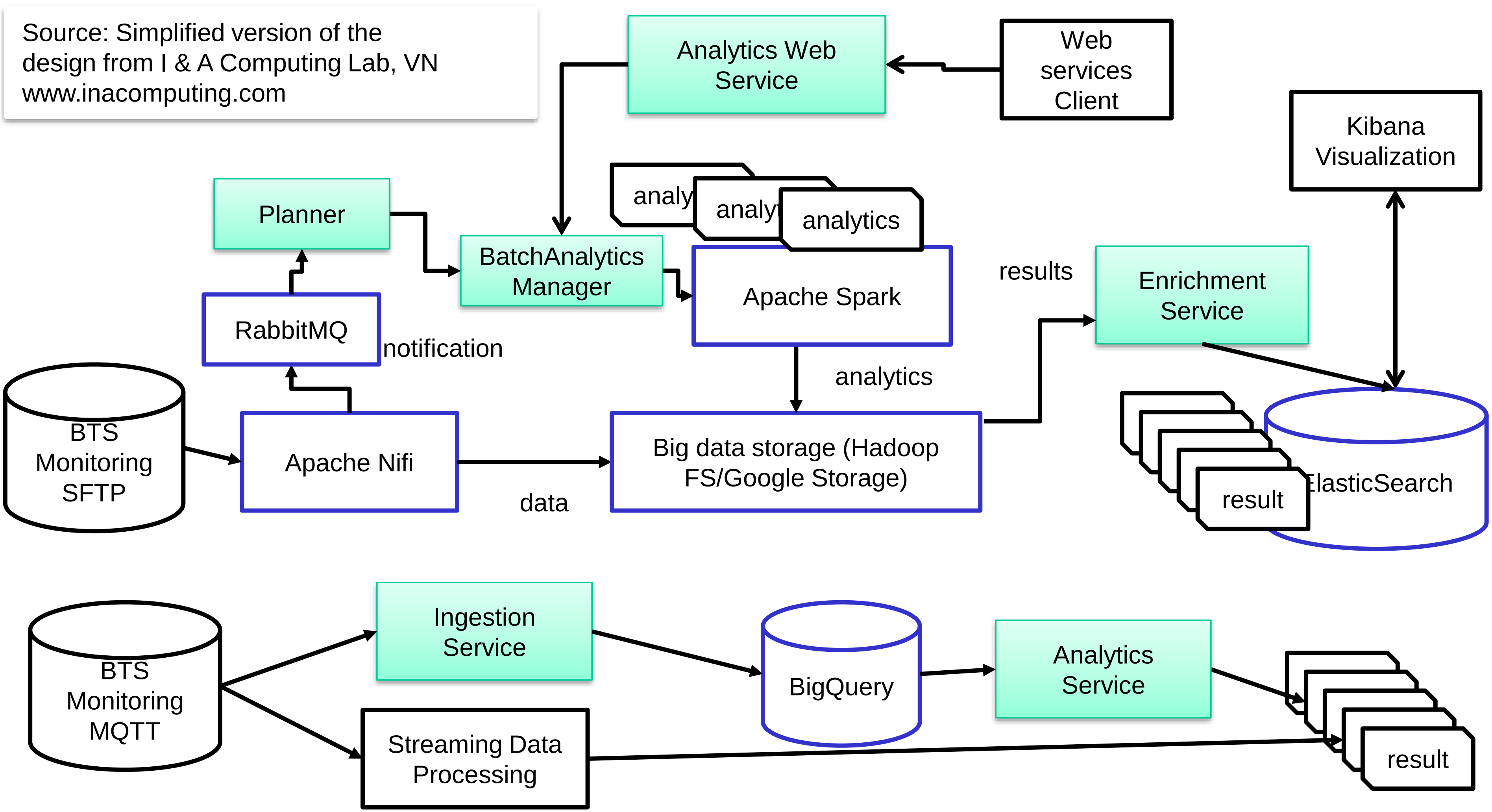
Quality of Analytics (QoA)

QoA: cost, performance (response time), quality of data (up-to-date ness, accuracy)

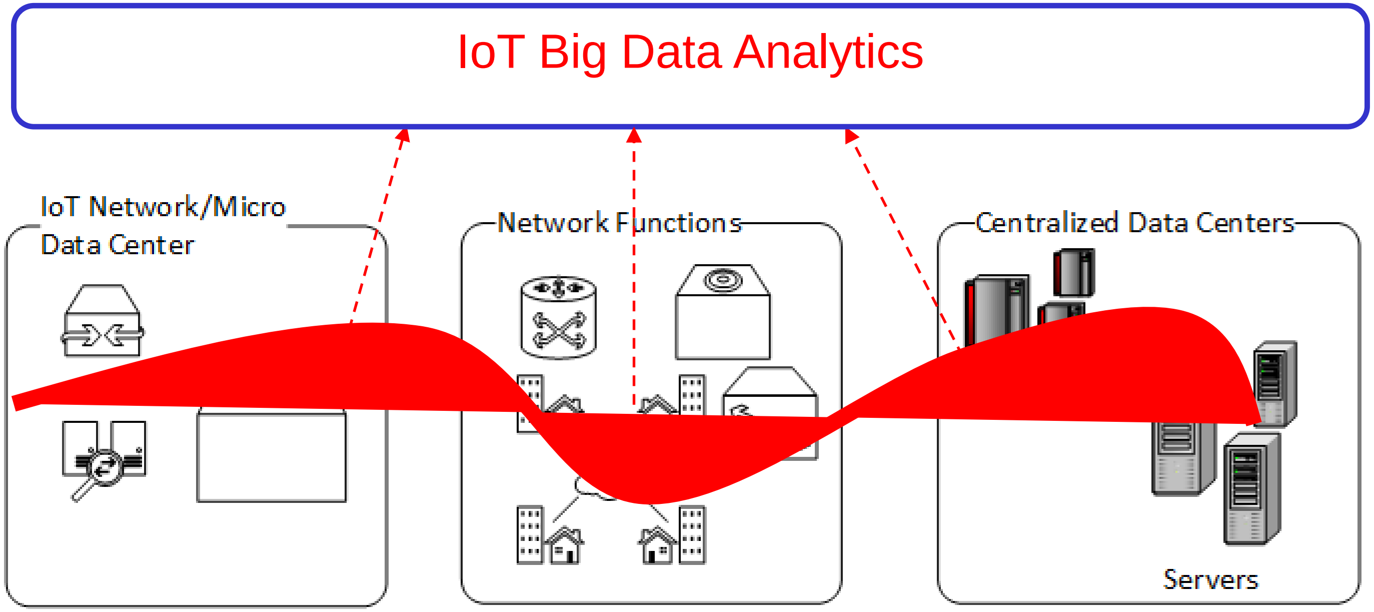
But there are many interactions that might cause incidents that lead to unexpected QoA

Hong-Linh Truong , Aitor Murguzur, Erica Yang, Challenges in Enabling Quality of Analytics in the Cloud, ACM JDIQ Challenge paper, 2017.

Problem 1: the complexity of software stacks and subsystems



Porblem 2: Complexity of the underlying virtual computing and network infrastructures



- Heavily based on virtual resources
  - IoT, Network functions and Clouds

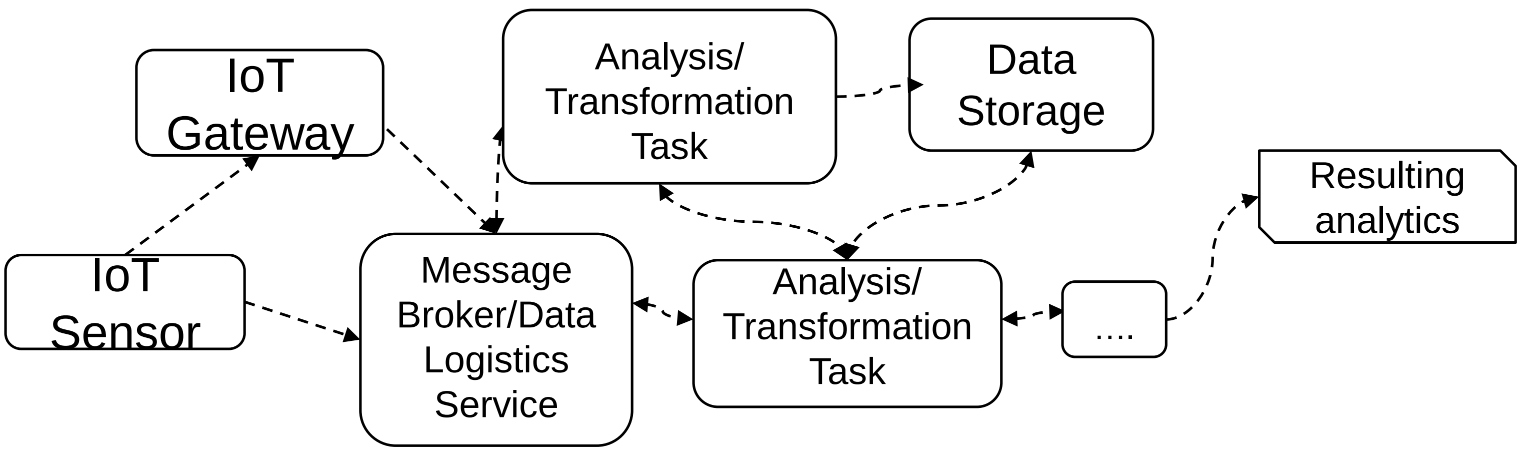
The SINC Concept: http://sincconcept.github.io

Incident monitoring and analytics

- Classification of incidents:
  - to quantify incidents and identify possible data sources, monitoring techniques and analytics.
- Measurement/Instrumentation:
  - to provide mechanisms for measurement and data collection for incidents.
- Incident analytics:
  - to find out the root cause and dependencies of incidents.

W3H: what, when, where and how for incidents

Too complex with many types of software. Can we have a simplified taxonomy for mapping incidents?



Large number of data sources (e.g., IoT devices)

Large-scale brokers & data transfer/logistics services

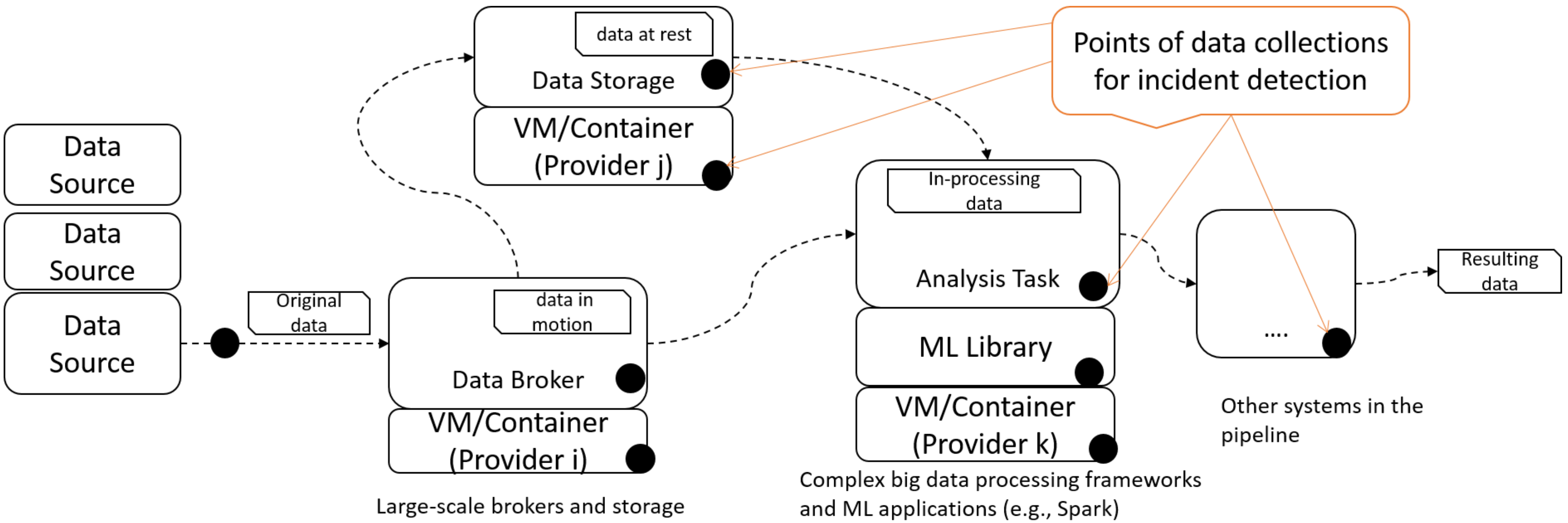
Complex big data processing frameworks

Other systems in the pipeline

Hong-Linh Truong, Manfred Halper, **Characterizing Incidents in Cloud-based IoT Data Analytics**,, The 42nd IEEE International Conference on Computers, Software & Applications Tokyo, Japan, July 23-27, 2018.

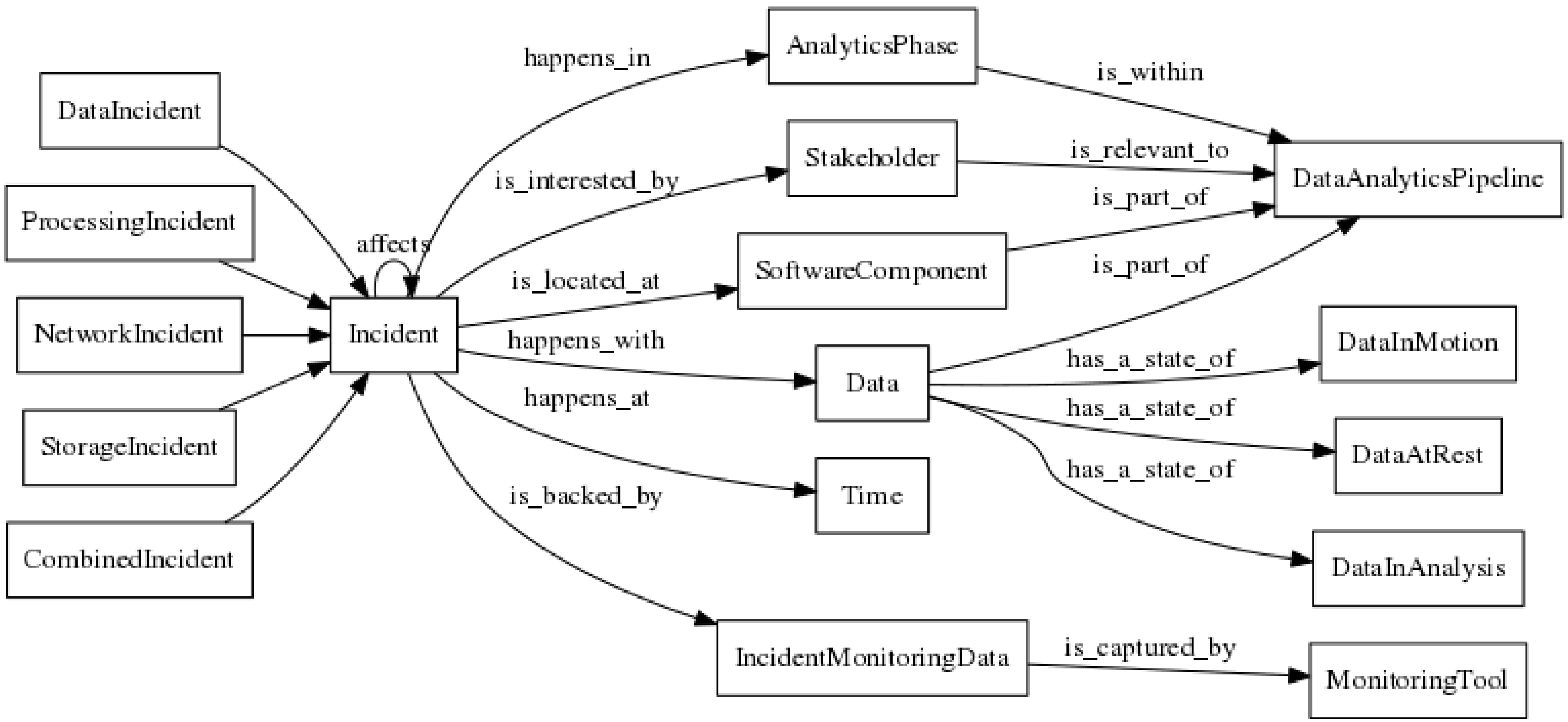
Points of instrumentation for gathering data for incident analytics

Capture monitoring data to analyze and solve incidents, especially incidents related to data quality, across subsystems in ensembles to achieve quality of results



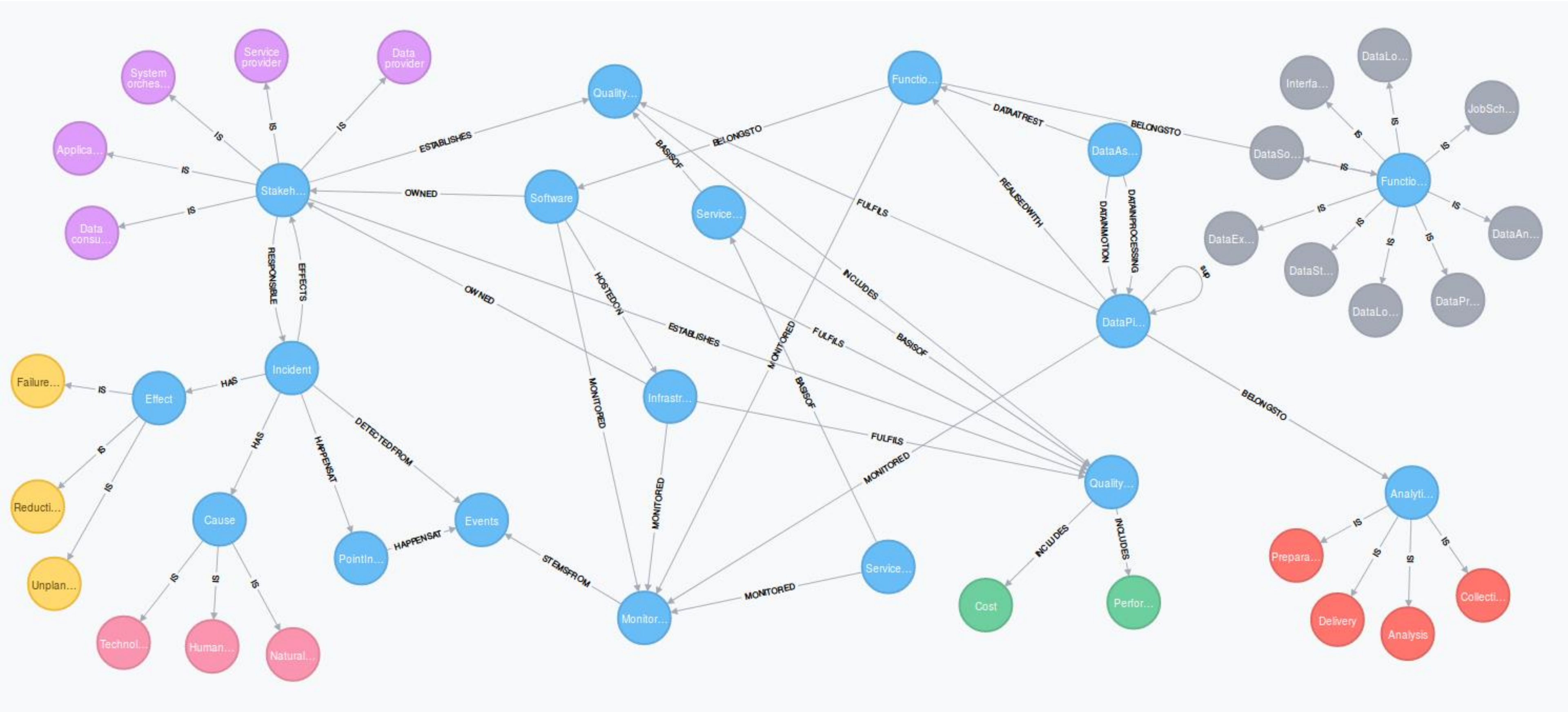
Hong-Linh Truong, Manfred Halper, **Characterizing Incidents in Cloud-based IoT Data Analytics**,, The 42nd IEEE International Conference on Computers, Software & Applications Tokyo, Japan, July 23-27, 2018.

Classification of incidents



Hong-Linh Truong, Manfred Halper, **Characterizing Incidents in Cloud-based IoT Data Analytics**,, The 42nd IEEE International Conference on Computers, Software & Applications Tokyo, Japan, July 23-27, 2018.

Example of incident classification



See https://www.researchgate.net/publication/324170664_Characterizing_Incidents_in_Cloud-based_IoT_Data_Analytics

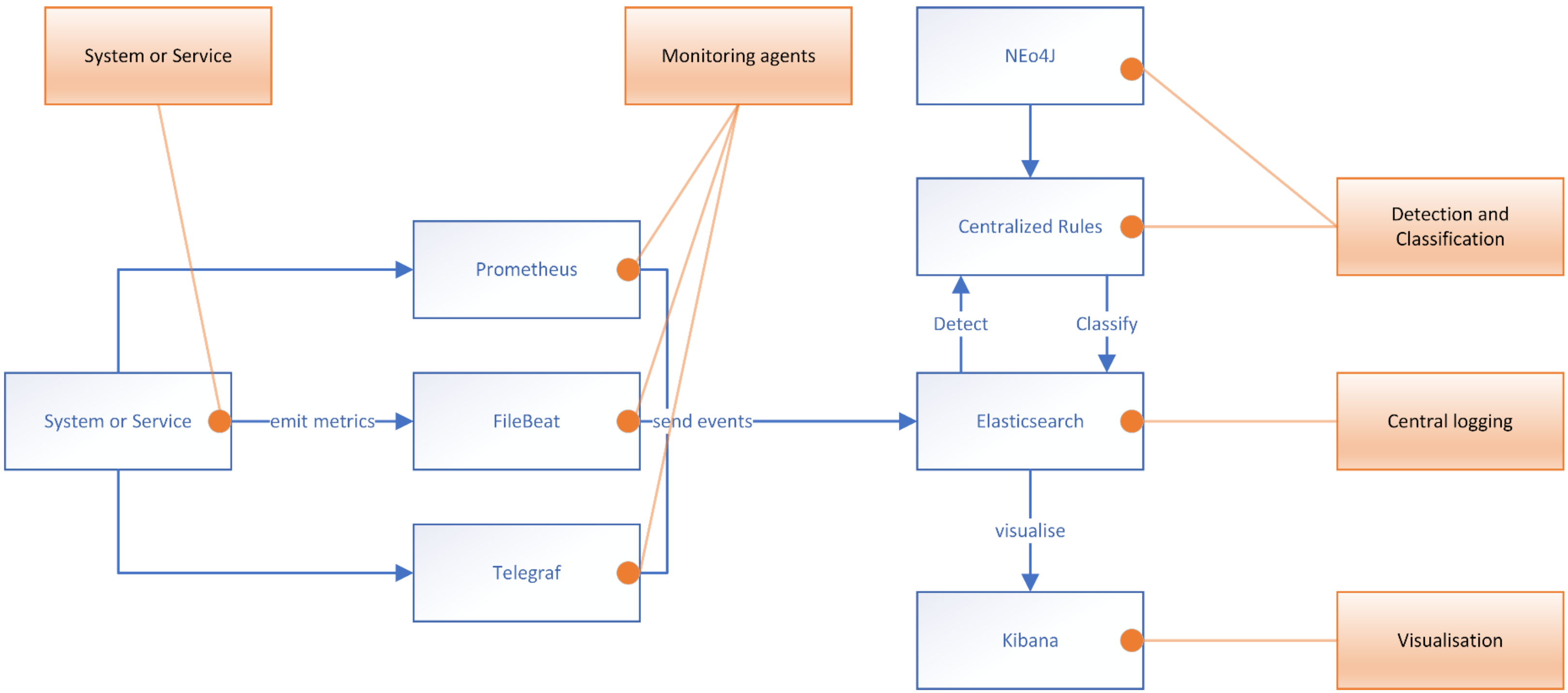
Monitoring and Analytics

Not just fast, distributed and cross layer monitoring

→ Hard to collect some incident related data for quality of data

→ Analytics: will be based on big data principles with ML but dependency analysis is not trivial

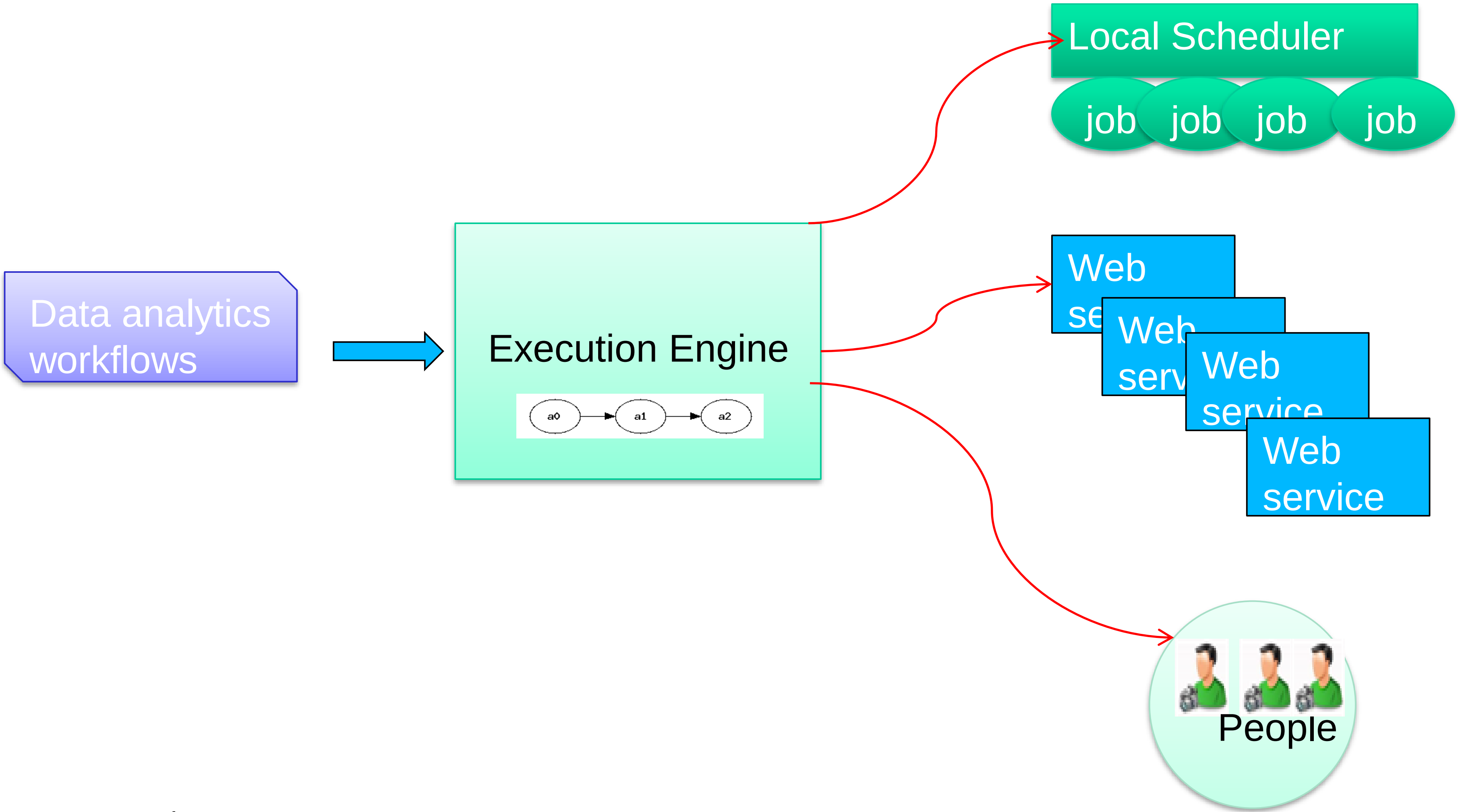
One example of tools for monitoring



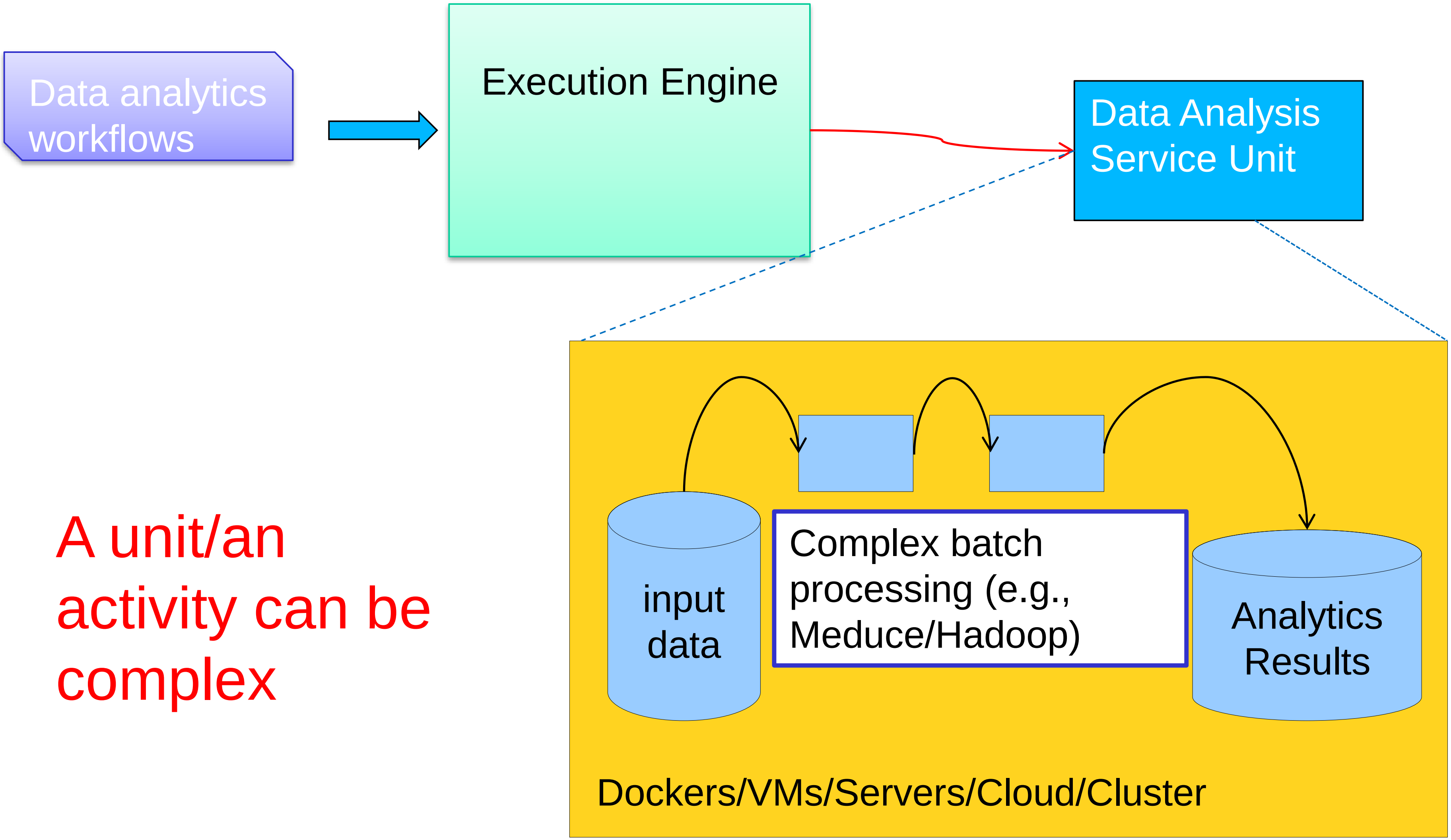
Check: https://github.com/rdsea/bigdataincidentanalytics

QOA IN DATA ANALYTICS WORKFLOWS

Data analytics workflow execution models



Data analytics workflow execution models



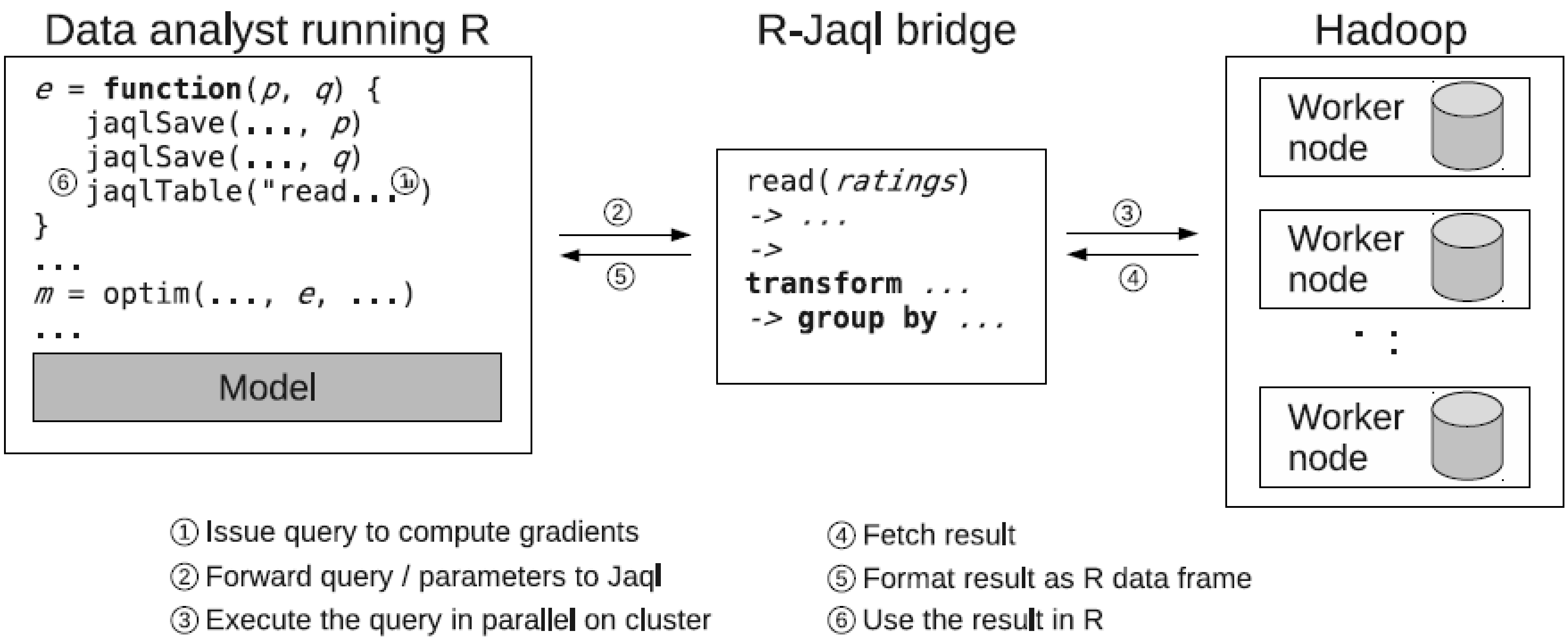
A unit/an activity can be complex

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

Check also: http://www.infosys.tuwien.ac.at/staff/truong/dst/pdfs/truong-dst2018-lecture5.pdf

Some examples (3)



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

Some examples (4): Airflow from Airbnb

- Workflow is a DAG (Direct Acyclic Graph)
  - http://airbnb.io/projects/airflow/

- Task/Operator:
  - BashOperator, PythonOperator, EmailOperator, HTTPOperator, SqlOperator, Sensor,
  - DockerOperator, HiveOperator, S3FileTransferOperator, PrestoToMysqlOperator, SlackOperator

Example for processing signal file

```
DAG_NAME = 'signal_upload_file'

default_args = {
    'owner': 'hong-linh-truong',
    'depends_on_past': False,
    'start_date': datetime.now(),
}

dag = DAG(DAG_NAME, schedule_interval=None, default_args=default_args)


stations=["station1", "station2"]

def checkSituation(**kwargs):
    f = 'f'
    t = 't'
    return t

downloadlogscript="curl  file:///home/truong/myprojects/mygit/rdsea-mobifone-training/data/opensignal/sample-Oct182016.csv -o /opt/data/air

t_downloadlogtocloud=  BashOperator(
    task_id="download_signal_file",
    bash_command=downloadlogscript,
    dag = dag
    )


t_analytics=  BashOperator(
    task_id="analyticsinternetusage",
    bash_command="/usr/bin/python /home/truong/myprojects/mygit/rdsea-mobifone-training/examples/databases/elasticsearch/uploader/src/uploa
    dag = dag
    )
t_sendresult =SimpleHttpOperator(
    task_id='sendresults',
    method='POST',
    http_conn_id='station1',
    endpoint='api/update/credit',
    data=json.dumps({"userphone": "066412345","credit":10}),
    headers={"Content-Type": "application/json"},
    dag = dag
    )

t_analytics.set_upstream(t_downloadlogtocloud)
t_sendresult.set_upstream(t_analytics)
```

Some examples (5): 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));
```

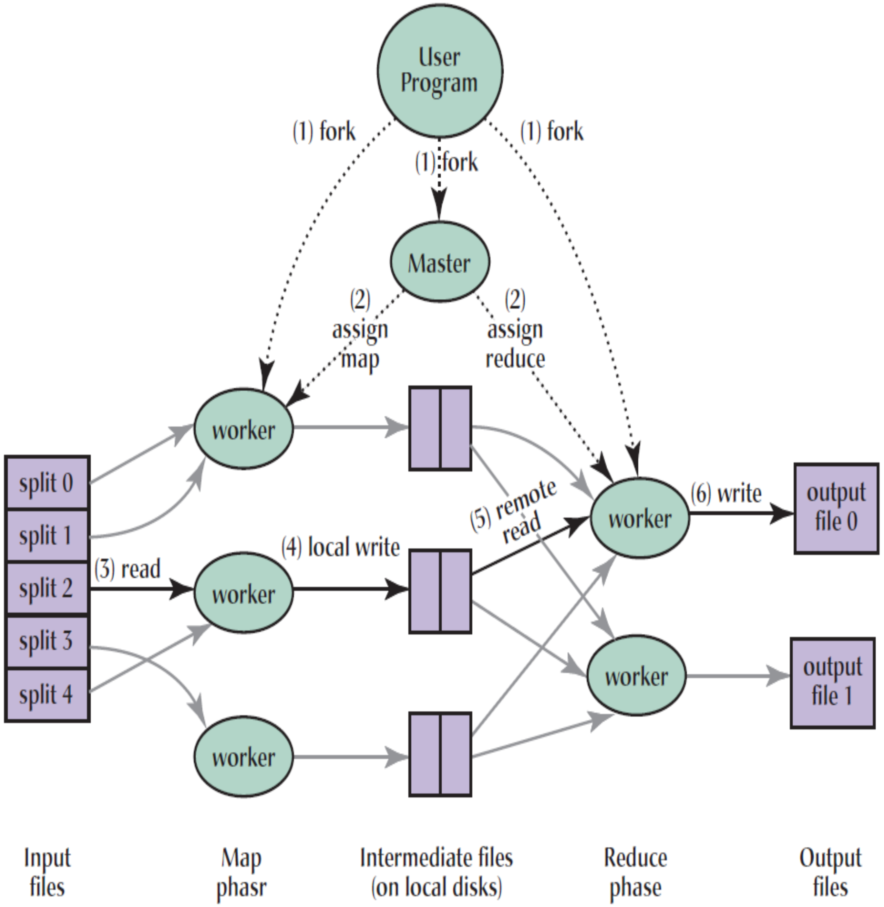


Fig. 1. Execution overview.

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

Apache Beam

- Goal: separate from pipelines from backend engines

Read data analytics → Post-processing result → Store analysis result

So how do we enable QoA-aware analytics?

Solutions

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

Which tools do you need for such solutions?

We will focus on quality of data as it has not been studied well

Mostly performance but not data quality

Logged in as: dr.who

All Applications

- Cluster
  - About
  - Nodes
  - Node Labels
  - Applications
    - NEW
    - NEW_SAVING
    - SUBMITTED
    - ACCEPTED
    - RUNNING
    - FINISHED
    - FAILED
    - KILLED
  - Scheduler
- Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
299	0	0	299	0	0 B	12 GB	0 B	0	4	0	4	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[MEMORY]	<memory:256, vCores:1>	<memory:3072, vCores:1>

Show 20 entries   Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI	Blacklisted Nodes
application_1494674366445_0299	liep	InA-BachPhu-BTS-Data-Analytics	SPARK	default	Fri Jun 2 04:31:00 +0200 2017	Fri Jun 2 04:32:35 +0200 2017	FINISHED	SUCCEEDED		History	N/A
application_1494674366445_0298	liep	InA-BachPhu-BTS-Data-Analytics	SPARK	default	Fri Jun 2 03:54:18 +0200 2017	Fri Jun 2 03:55:39 +0200 2017	FINISHED	SUCCEEDED		History	N/A
application_1494674366445_0297	liep	InA-BachPhu-BTS-Data-Analytics	SPARK	default	Fri Jun 2 03:43:46 +0200 2017	Fri Jun 2 03:44:32 +0200 2017	FINISHED	SUCCEEDED		History	N/A
application_1494674366445_0296	liep	InA-BachPhu-BTS-Data-Analytics	SPARK	default	Thu Jun 1 18:42:49 +0200 2017	Thu Jun 1 18:44:13 +0200 2017	FINISHED	SUCCEEDED		History	N/A
application_1494674366445_0295	liep	InA-BachPhu-BTS-Data-Analytics	SPARK	default	Thu Jun 1 18:39:11 +0200 2017	Thu Jun 1 18:40:03 +0200 2017	FINISHED	SUCCEEDED		History	N/A
application_1494674366445_0294	liep	InA-BachPhu-	SPARK	default	Thu Jun 1 18:28:15 +0200	Thu Jun 1 18:29:00 +0200 2017	FINISHED	SUCCEEDED		History	N/A

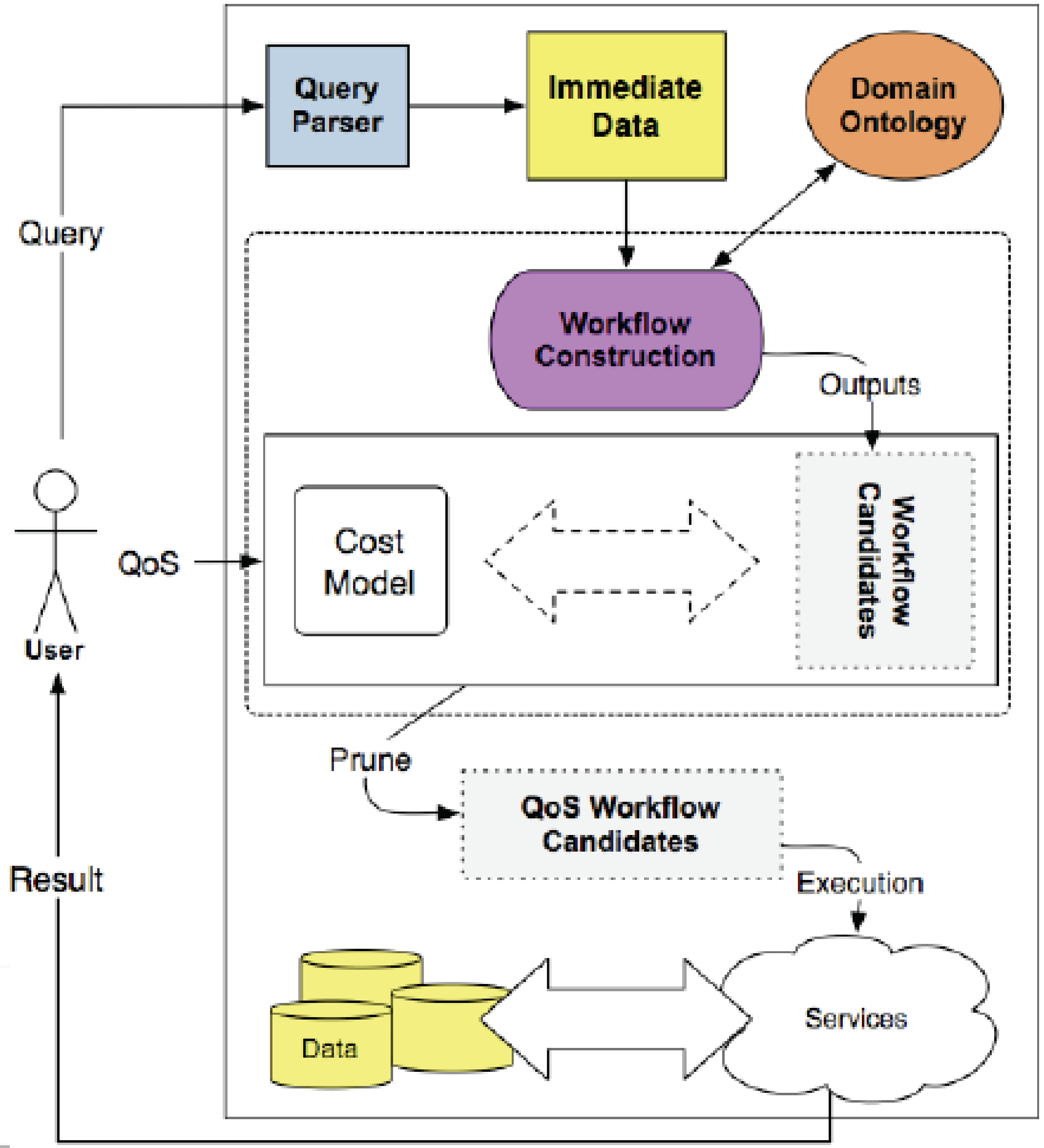
Executors

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write
Active(8)	0	0.0 B / 3.7 GB	0.0 B	7	0	0	550	550	2.8 m (5.6 s)	29.0 MB	270.3 KB	690.4 KB
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0 ms (0 ms)	0.0 B	0.0 B	0.0 B
Total(8)	0	0.0 B / 3.7 GB	0.0 B	7	0	0	550	550	2.8 m (5.6 s)	29.0 MB	270.3 KB	690.4 KB

If a job is failed due to the quality of data, how do you know?

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

Data Operations and cost with BigQuery

US (multi-region) Monthly

Operation	Pricing	Details
Active storage	$0.02 per GB	The first 10 GB is free each month. See Storage pricing for details.
Long-term storage	$0.01 per GB	The first 10 GB is free each month. See Storage pricing for details.
Streaming Inserts	$0.01 per 200 MB	You are charged for rows that are successfully inserted. Individual rows are calculated using a 1 KB minimum size. See Storage pricing for details.
Queries (analysis)	$5 per TB	First 1 TB per month is free, see On-demand pricing for details. Flat-rate pricing is also available for high-volume customers.

Source: https://cloud.google.com/bigquery/pricing

Just think about a simple example:

If you want to implement cost together data size and performance, what would be your way?

Provenance info

NiFi Data Provenance

Displaying 1,000 of 1,000
Oldest event available: 06/08/2017 04:27:03 UTC

Showing the most recent 1,000 of 1,000+ events, please refine the search.

Filter | by component name

Date/Time	Type	FlowFile Uuid	Size	Component Name	Component Type
06/09/2017 04:26:33.202 UTC	DROP	5f5e74f6-f28e-4cb8-b70e-07c5f8407bc4	8.33 MB	PutBachPhuHDFS-DYNAMIC-DATA	PutHDFS
06/09/2017 04:26:33.202 UTC	ATTRIBUTES_MODIFIED	5f5e74f6-f28e-4cb8-b70e-07c5f8407bc4	8.33 MB	PutBachPhuHDFS-DYNAMIC-DATA	PutHDFS
06/09/2017 04:26:33.202 UTC	SEND	5f5e74f6-f28e-4cb8-b70e-07c5f8407bc4	8.33 MB	PutBachPhuHDFS-DYNAMIC-DATA	PutHDFS
06/09/2017 04:26:32.703 UTC	RECEIVE	5f5e74f6-f28e-4cb8-b70e-07c5f8407bc4	8.33 MB	GetBachPhuSFTP-DYNAMIC-DATA	GetSFTP
06/09/2017 04:26:32.200 UTC	RECEIVE	348c8722-7d2b-44d6-9103-d7e699ee19f0	1.79 KB	Get-INA-OPYSPARK-HDFS	GetHDFS
06/09/2017 04:26:32.195 UTC	DROP	64457e3f-0699-4404-a80f-5740674eab82	1.79 KB	Put-INA-BP-SFTP	PutSFTP
06/09/2017 04:26:30.513 UTC	RECEIVE	31eb9ddc-ebb2-47cb-b09c-0ba1f7598f7a	1.79 KB	Get-INA-OPYSPARK-HDFS	GetHDFS
06/09/2017 04:26:30.505 UTC	DROP	14571cd6-e4fa-4cda-8038-7906d9263d4e	1.79 KB	Put-INA-BP-SFTP	PutSFTP
06/09/2017 04:26:28.765 UTC	RECEIVE	9030a70d-7b2f-4657-88d7-553021b072e2	1.79 KB	Get-INA-OPYSPARK-HDFS	GetHDFS
06/09/2017 04:26:28.761 UTC	DROP	eb142c05-b27e-4a43-bd7c-bc4ed83c8c46	1.79 KB	Put-INA-BP-SFTP	PutSFTP
06/09/2017 04:26:27.037 UTC	RECEIVE	f512b40e-9ba7-4f4f-aea5-171abc8cb26c	1.79 KB	Get-INA-OPYSPARK-HDFS	GetHDFS
06/09/2017 04:26:27.027 UTC	DROP	b7ac1627-7c74-48a8-98da-6ff22ec099f9	1.79 KB	Put-INA-BP-SFTP	PutSFTP
06/09/2017 04:26:25.259 UTC	RECEIVE	d1eb4033-5cc7-42ec-8268-ffaa6a70b2ce	1.79 KB	Get-INA-OPYSPARK-HDFS	GetHDFS
06/09/2017 04:26:25.253 UTC	DROP	9be59f31-03f7-4cae-8288-6487a3f40bb2	1.79 KB	Put-INA-BP-SFTP	PutSFTP
06/09/2017 04:26:23.551 UTC	RECEIVE	86ca4fa5-3b93-4546-842c-be0fc9747a3d	1.79 KB	Get-INA-OPYSPARK-HDFS	GetHDFS
06/09/2017 04:26:23.542 UTC	DROP	9fd8ee9f-1522-408c-96fb-8788021d060f	1.79 KB	Put-INA-BP-SFTP	PutSFTP
06/09/2017 04:26:21.813 UTC	RECEIVE	1cd88fd8-eb8d-4ec0-aa37-7430d5584cd9	1.79 KB	Get-INA-OPYSPARK-HDFS	GetHDFS
06/09/2017 04:26:21.802 UTC	DROP	bdc4f5e6-2bac-41d6-a6db-fc50ecd19946	1.79 KB	Put-INA-BP-SFTP	PutSFTP
06/09/2017 04:26:20.094 UTC	RECEIVE	5b8d9cee-6309-4e24-be75-6ed6e3ec1447	1.79 KB	Get-INA-OPYSPARK-HDFS	GetHDFS
06/09/2017 04:26:20.081 UTC	DROP	59b60a7f-11b5-4dc6-bfde-ea220002e536	1.79 KB	Put-INA-BP-SFTP	PutSFTP
06/09/2017 04:26:18.366 UTC	RECEIVE	ad652e87-3e34-4072-bc7a-51c5a067e39e	1.79 KB	Get-INA-OPYSPARK-HDFS	GetHDFS
06/09/2017 04:26:18.363 UTC	DROP	a08b2524-49c4-40ac-b2cd-c33499e6cb4f	1.79 KB	Put-INA-BP-SFTP	PutSFTP
06/09/2017 04:26:16.494 UTC	RECEIVE	7e108a1a-e66e-485e-9b5f-486d0cac7a55	1.79 KB	Get-INA-OPYSPARK-HDFS	GetHDFS
06/09/2017 04:26:16.490 UTC	DROP	d98e6150-3144-4039-bbd8-f3517b70be87	1.79 KB	Put-INA-BP-SFTP	PutSFTP
06/09/2017 04:26:15.351 UTC	DROP	7c172491-855f-450e-a052-d5cf49757626	301 bytes	PutBachPhuStaticData-HDFS	PutHDFS
06/09/2017 04:26:15.351 UTC	ATTRIBUTES_MODIFIED	7c172491-855f-450e-a052-d5cf49757626	301 bytes	PutBachPhuStaticData-HDFS	PutHDFS

Last updated: 04:27:10 UTC

If you are able to detect a quality problem in the analysis phase, can you trace back to the data sources?   what would be your way?

Research questions for QoD

- 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?

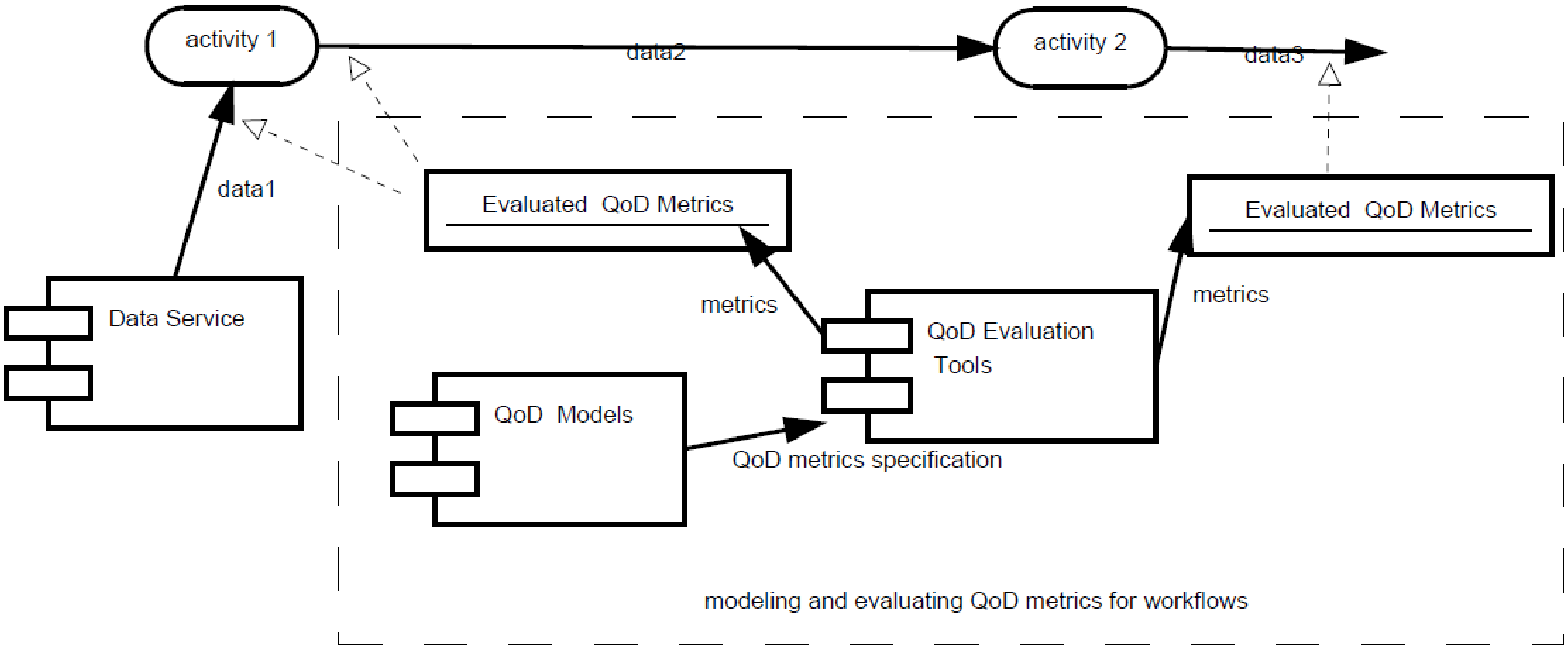
Approach

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

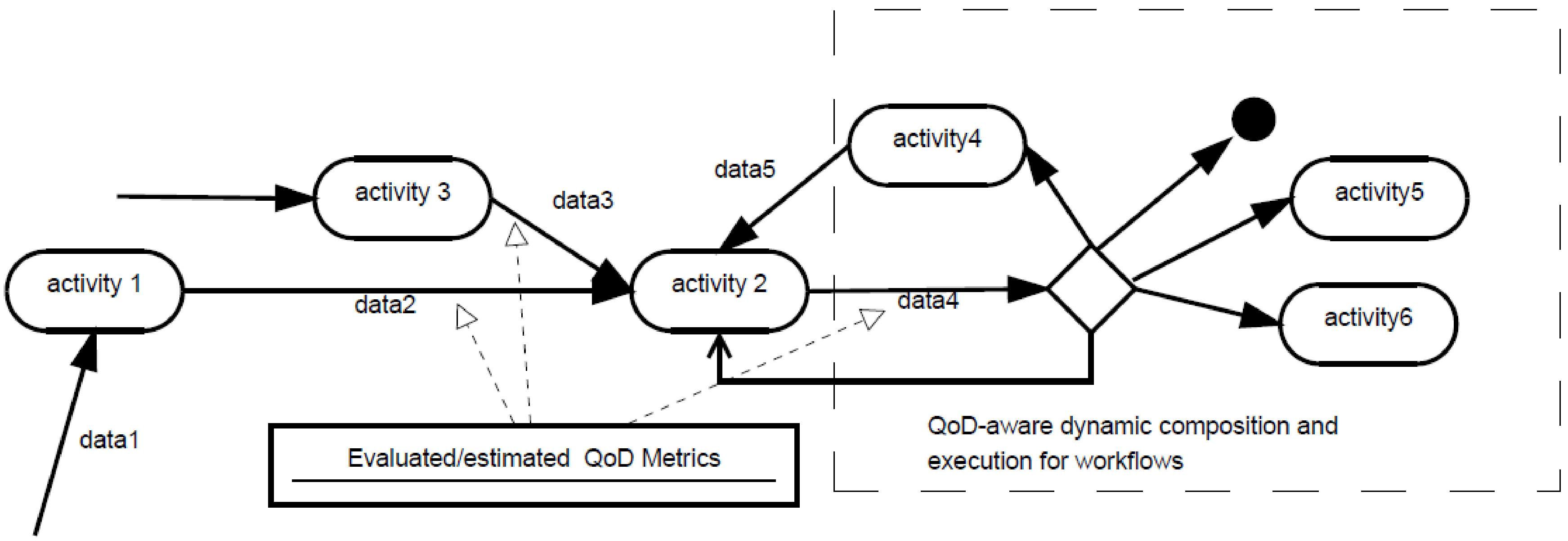
QoD-aware composition and execution

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

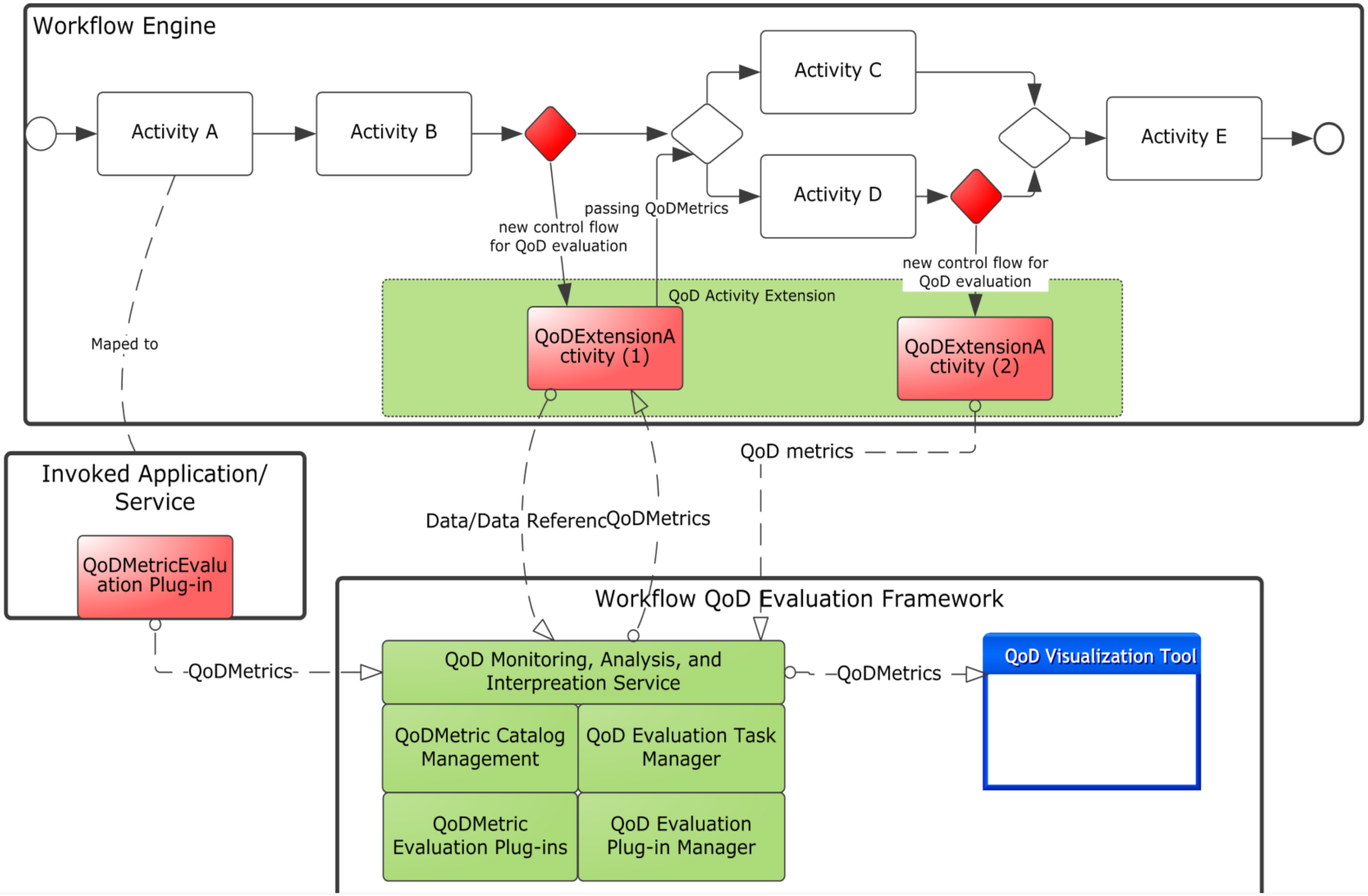


How to integrate QoD evaluators? And which concerns need to be considered?

QoD metrics evaluation

- Domain-specific metrics
  - Need specific tools and expertise for determining metrics
- Evaluation
  - Cannot done by software only: humans are required
  - Exact versus inexact evaluation due to big and streaming data
- 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

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

what kind of optimization can be done with QoD?

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.

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

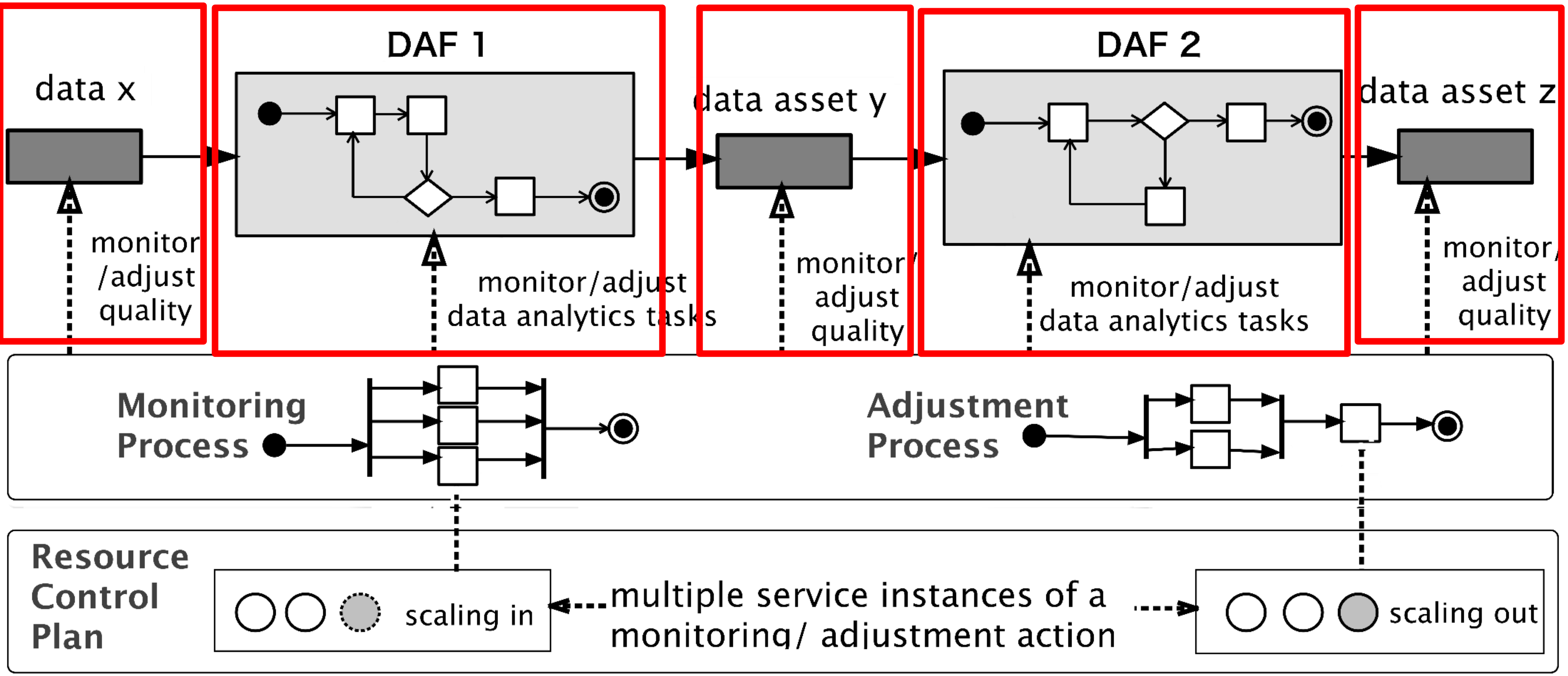
QoA: QoD, performance, cost, etc.

Quality-of-analytics driven workflows

- 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

Hong-Linh Truong, Aitor Murguzur, and Erica Yang. 2018. Challenges in Enabling Quality of Analytics in the Cloud. J. Data and Information Quality 9, 2, Article 9 (January 2018), 4 pages. DOI: https://doi.org/10.1145/3138806

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

Data elasticity

- Key techniques
  - Monitoring QoD for streaming and big data
  - Monitoring cloud resources
  - Having multiple data analysis algorithms
  - Using elasticity rules for cloud resources and analysis algorithms
    - Building your own elasticity rules/models

Exercises

- Read mentioned papers
- Examine possible incidents in your data pipelines
- 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
Faculty of Informatics, TU Wien
hong-linh.truong@tuwien.ac.at
http://www.infosys.tuwien.ac.at/staff/truong
@linhsolar