Advanced Services Engineering, Summer 2016, Lecture 6

Advanced service-based data analytics: Models, Elasticity and APIs

Hong-Linh Truong
Distributed Systems Group, TU Wien

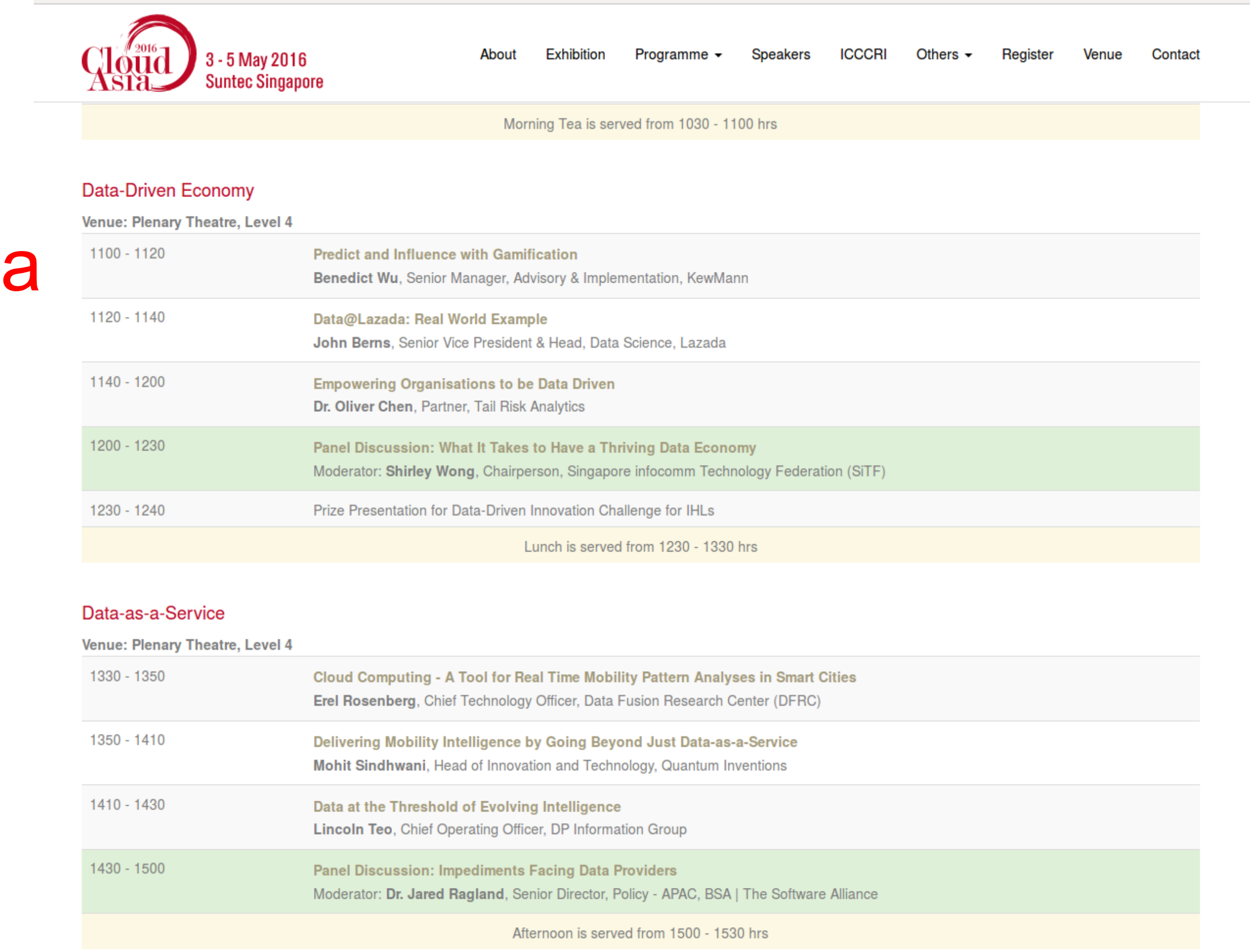
truong@dsg.tuwien.ac.at
dsg.tuwien.ac.at/staff/truong
@linhsolar

Outline

- Principles of elasticity for advanced service-based data analytics
- Data analytics within a single system
- Data analytics across multiple systems
- APIs management

Recall

data-as-a-service and data marketplaces are key elements for data-driven economy



Morning Tea is served from 1030 - 1100 hrs

Data-Driven Economy

Venue: Plenary Theatre, Level 4

1100 - 1120	**Predict and Influence with Gamification** **Benedict Wu**, Senior Manager, Advisory & Implementation, KewMann
1120 - 1140	**Data@Lazada: Real World Example** **John Berns**, Senior Vice President & Head, Data Science, Lazada
1140 - 1200	**Empowering Organisations to be Data Driven** **Dr. Oliver Chen**, Partner, Tail Risk Analytics
1200 - 1230	**Panel Discussion: What It Takes to Have a Thriving Data Economy** Moderator: **Shirley Wong**, Chairperson, Singapore infocomm Technology Federation (SiTF)
1230 - 1240	Prize Presentation for Data-Driven Innovation Challenge for IHLs

Lunch is served from 1230 - 1330 hrs

Data-as-a-Service

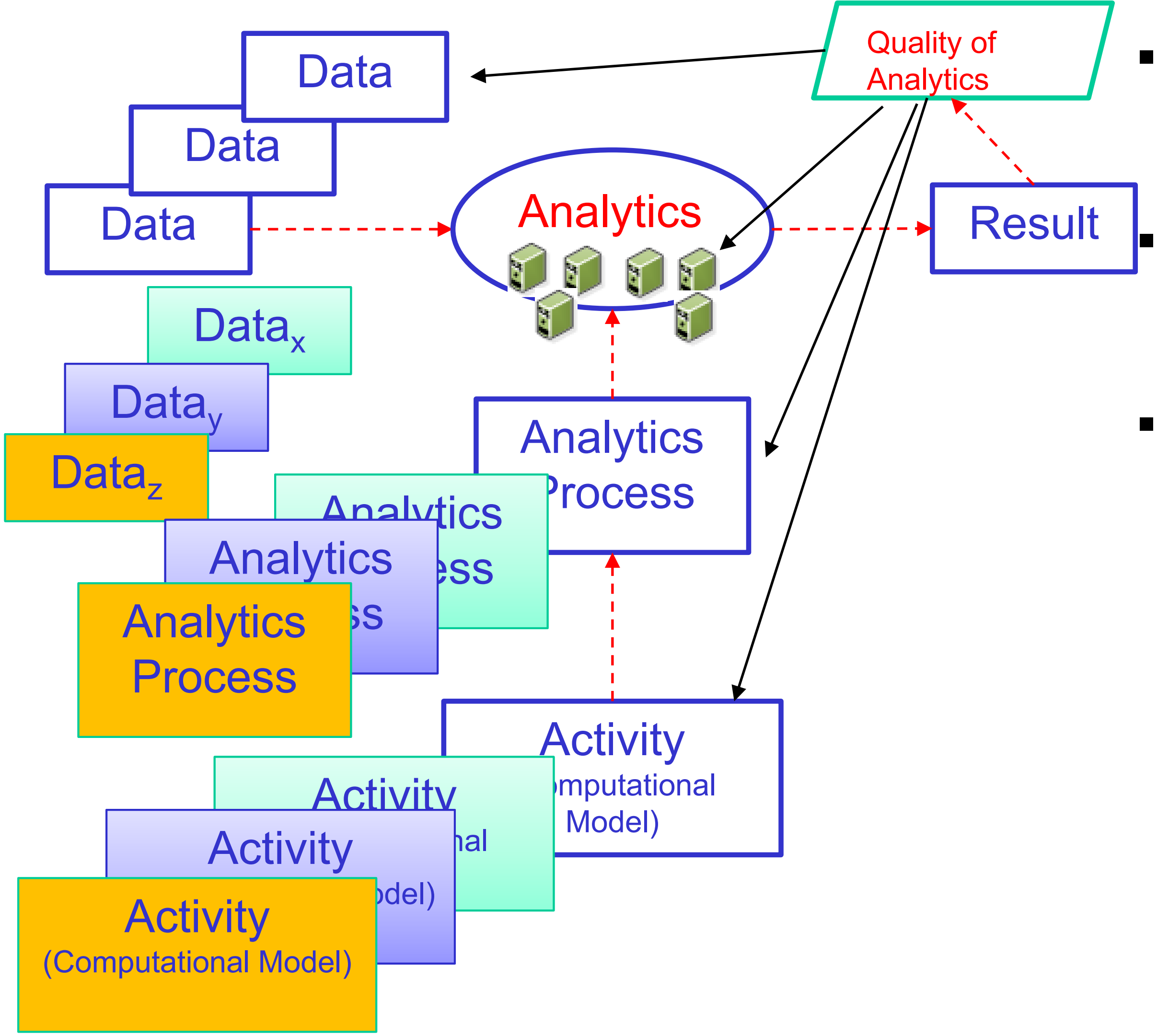
Venue: Plenary Theatre, Level 4

1330 - 1350	**Cloud Computing - A Tool for Real Time Mobility Pattern Analyses in Smart Cities** **Erel Rosenberg**, Chief Technology Officer, Data Fusion Research Center (DFRC)
1350 - 1410	**Delivering Mobility Intelligence by Going Beyond Just Data-as-a-Service** **Mohit Sindhwani**, Head of Innovation and Technology, Quantum Inventions
1410 - 1430	**Data at the Threshold of Evolving Intelligence** **Lincoln Teo**, Chief Operating Officer, DP Information Group
1430 - 1500	**Panel Discussion: Impediments Facing Data Providers** Moderator: **Dr. Jared Ragland**, Senior Director, Policy - APAC, BSA | The Software Alliance

Afternoon is served from 1500 - 1530 hrs

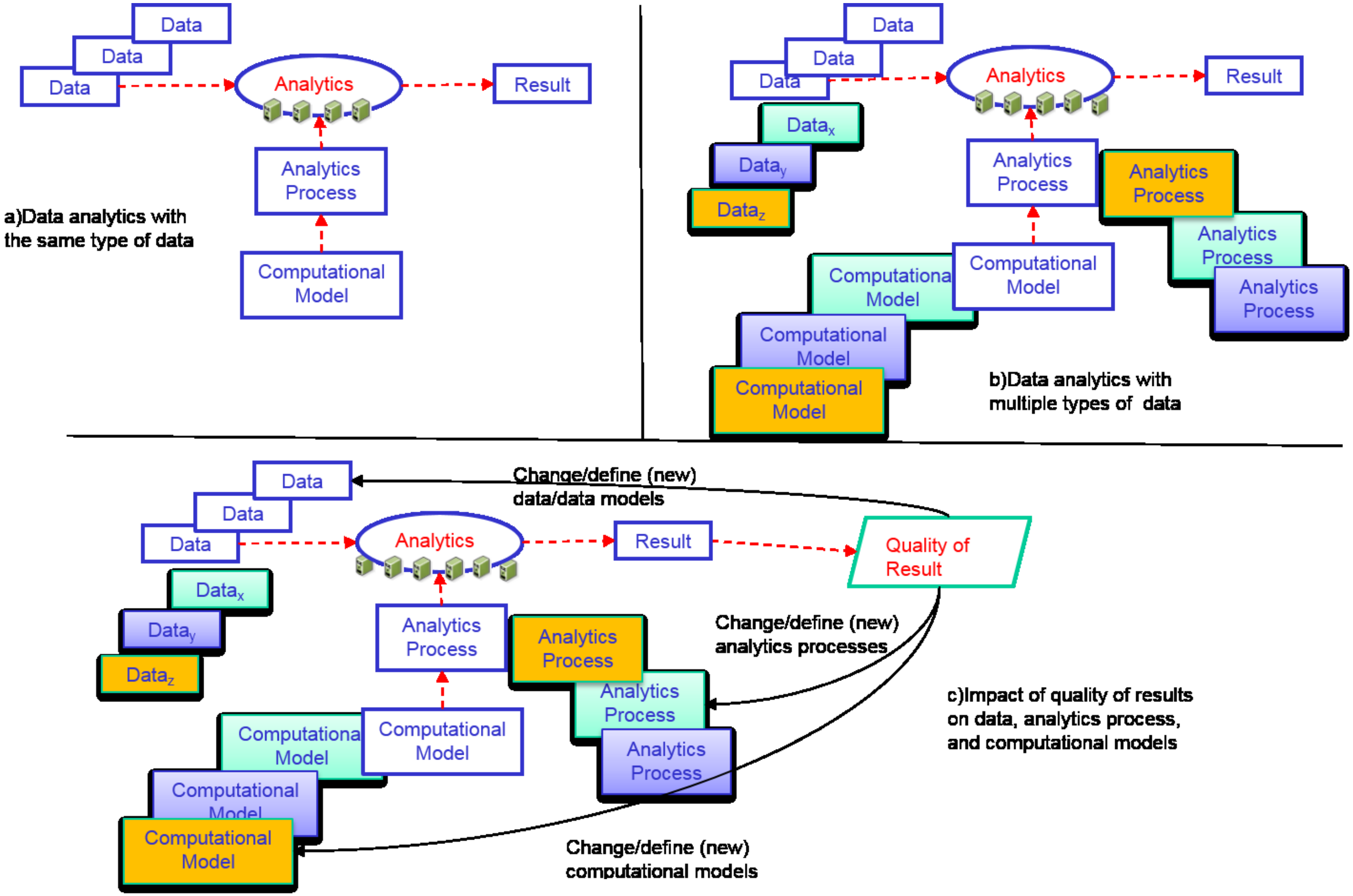
PRINCIPLES OF ELASTICITY FOR DATA ANALYTICS

Complex dependencies in (big) data analytics



- **More data** → more computational resources (e.g. more VMs)
- **More types of data** → more computational models → more analytics processes
- Change **quality of analytics**
  - Change quality of data
  - Change response time
  - Change cost
  - Change types of result (form of the data output, e.g. tree, visual, story, etc.)

Complex dependencies in (big) data analytics



Hong-Linh Truong, Schahram Dustdar, "Principles of Software-defined Elastic Systems for Big Data Analytics", (c) IEEE Computer Society, IEEE International Workshop on Software Defined Systems, 2014 IEEE International Conference on Cloud Engineering (IC2E 2014), Boston, Massachusetts, USA, 10-14 March 2014

Elasticity principles can be used to support dynamic quality of analytics

Elasticity Principles: Elasticity of data and computational models

- Multiple types of objects from different sources with complex dependencies, relevancies, and quality
- Different data and computational models for the same analytics subject
- New analytics subjects can be defined and analytics goals can be changed
- Decide/select/define/compose not only computational models for analytics subjects but also data models based on existing ones

Management and modeling of elasticity of data and computational model during the analytics

Elasticity Principles: Elasticity of data resources

- Data provided, managed and shared by different providers
- Data associated with different concerns (cost, quality of data, privacy, contract, etc.
- Static data, open data, data-as-a-service, opportunistic data (from sensors and human sensing)
- Not just centralized big data and total data ownership

Data resources can be taken into account in an elastic manger: similar to VMs, based on their quality, relevancy, pricing, etc.

Elasticity Principles: Elasticity of humans and software as computing units

- Human in the loop to solve analytics tasks that software cannot solve
- Human-based compute units can be scaled up/down with different cost, availability, and performance models
- Human-based compute units + software-based compute units for executing computational models
- Elasticity controls can be also done by humans

Provisioning hybrid compute units in an elastic way for computational/data/network tasks as well as for monitoring/control tasks in the analytics process

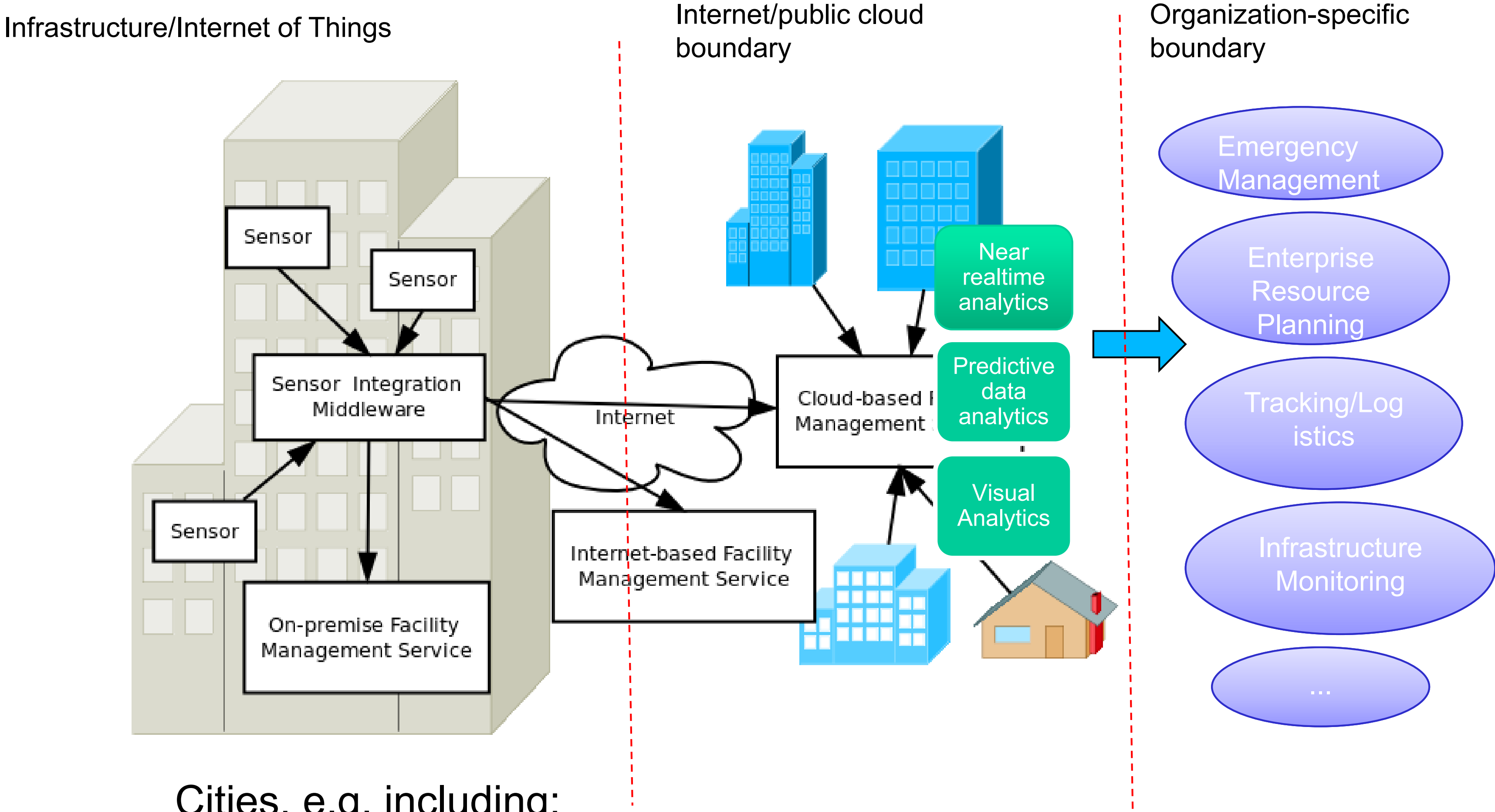
Elasticity Principles: Elasticity of quality of analytics

- Definition of quality of analytics
  - Trade-offs of time, cost, quality of data, forms of output
- Using quality of analytics to select suitable computational models, data resources, computing units
- Multi-level control for the elasticity based on quality of analytics

Able to cope with changes in quality of data, performance, cost and types of results at runtime

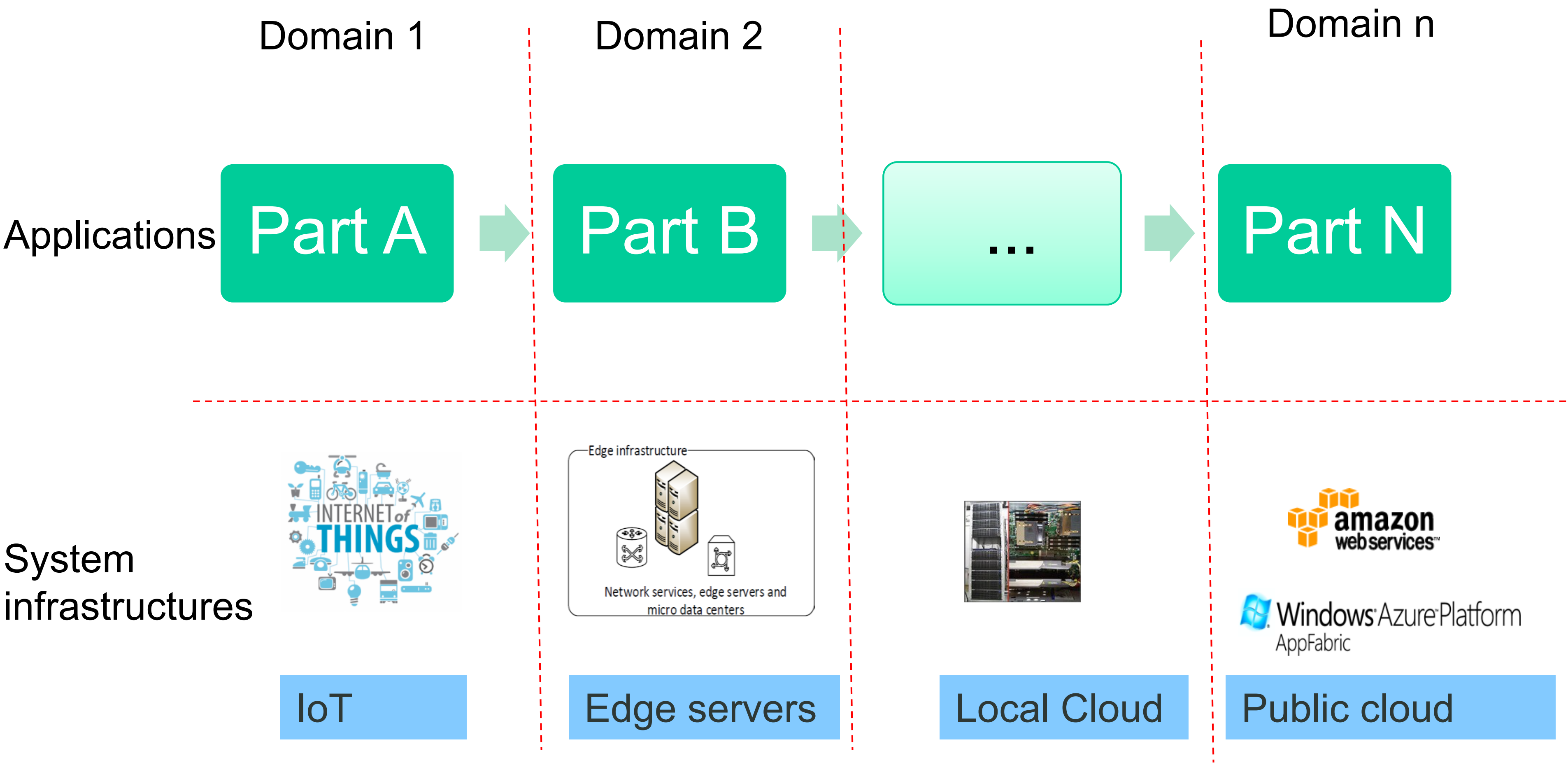
Advanced service-based analytics – which are fundamental engineering questions?

Advanced service-based data analytics (1)


Infrastructure/Internet of Things
Internet/public cloud boundary
Organization-specific boundary
Sensor
Sensor
Sensor Integration Middleware
Sensor
On-premise Facility Management Service
Internet
Internet-based Facility Management Service
Cloud-based F Management
Near realtime analytics
Predictive data analytics
Visual Analytics
Emergency Management
Enterprise Resource Planning
Tracking/Logistics
Infrastructure Monitoring
...

Cities, e.g. including:
10000+ buildings
1000000+ sensors

Advanced service-based data analytics -- fundamental concepts



Domain 1

Domain 2

Domain n

Applications

Part A

Part B

...

Part N

System infrastructures

IoT

Edge servers

Local Cloud

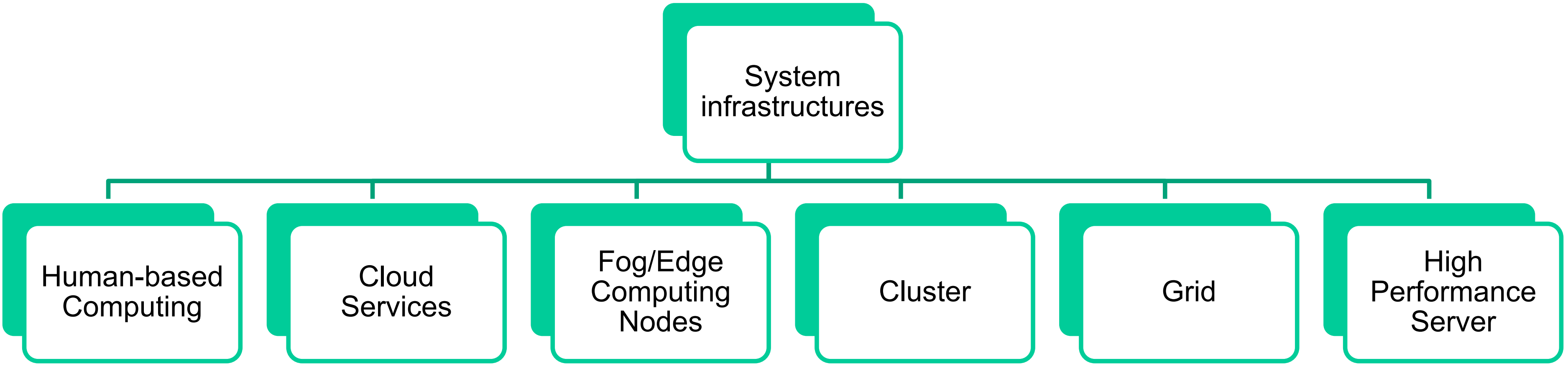
Public cloud

Design questions

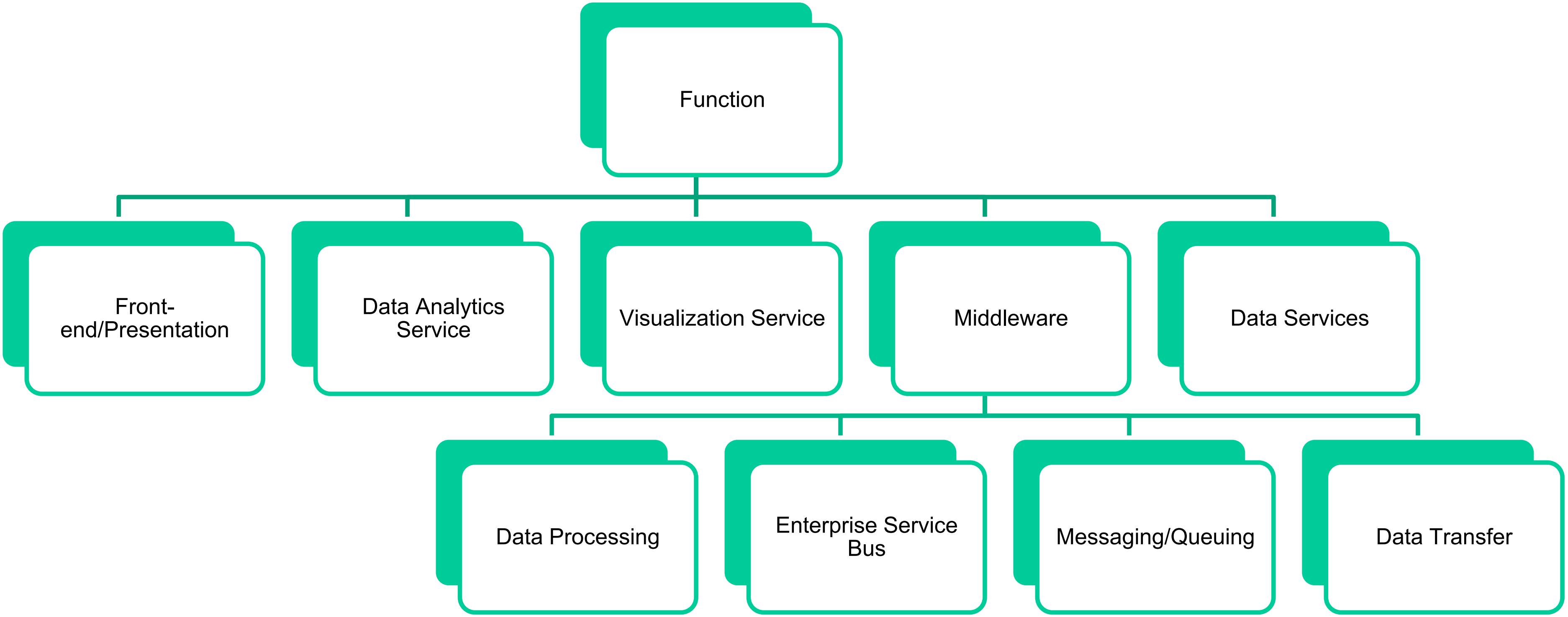
Part = a (composite) service unit

- Which system infrastructures are used?
- Which interfaces are suitable for units?
- Which programming models are used within units?
- Which are fundamental units to be used?
- How do different units interact?
- Which non-functional parameters are important and how to measure them?

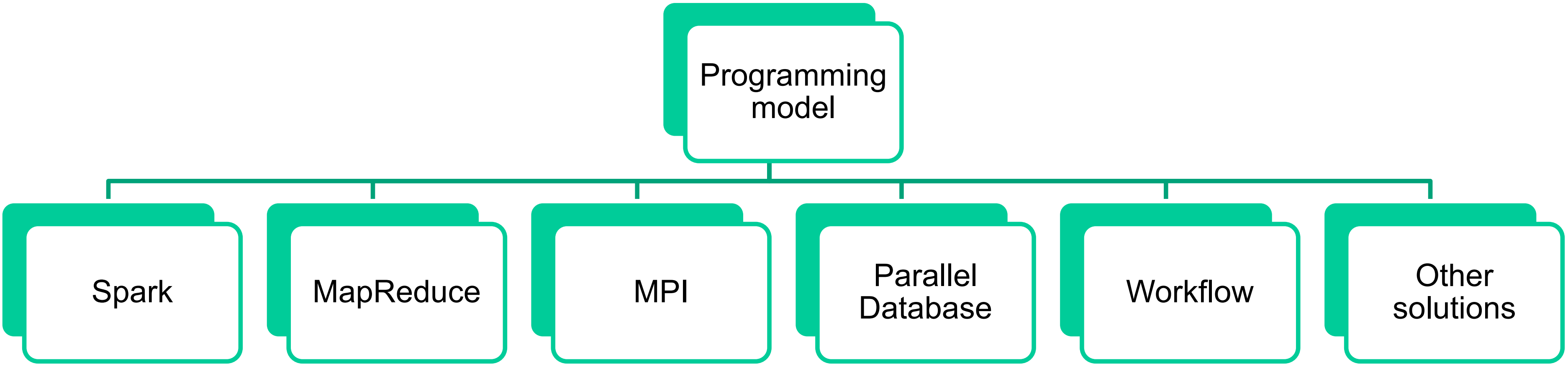
Fundamental concepts – system infrastructure unit



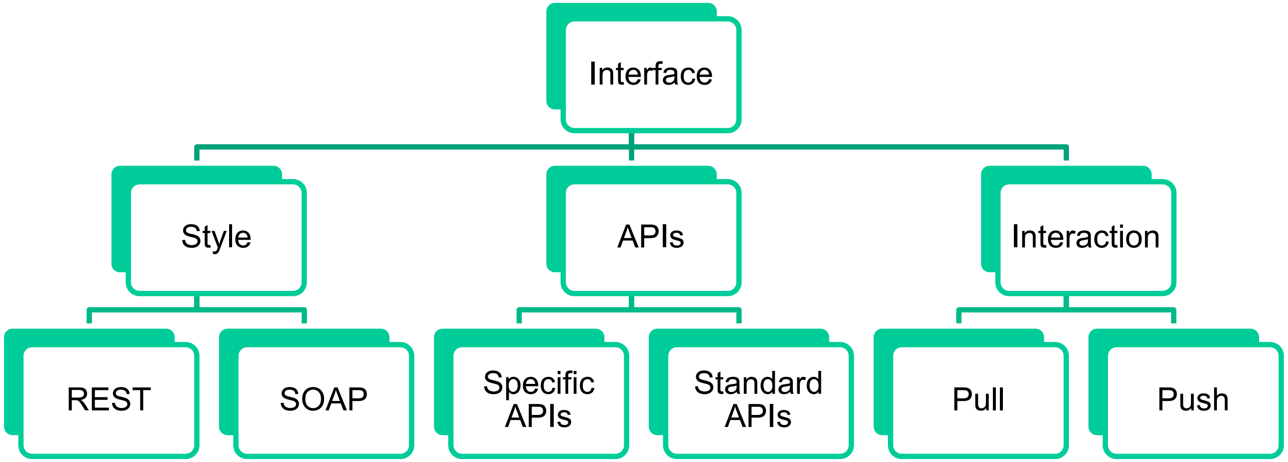
Fundamental concepts – unit functions



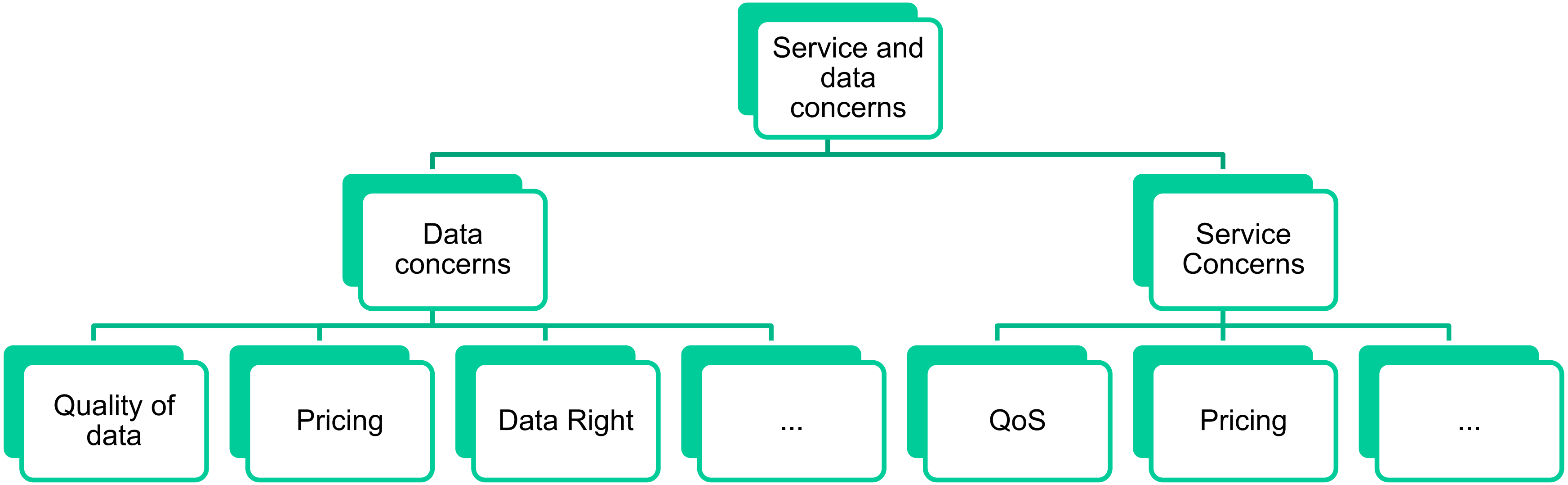
Fundamental concepts – programming model within units



Fundamental concepts – interfaces between units



Fundamental concepts – services and data concerns


Service and data concerns

Data concerns

Service Concerns

Quality of data

Pricing

Data Right

...

QoS

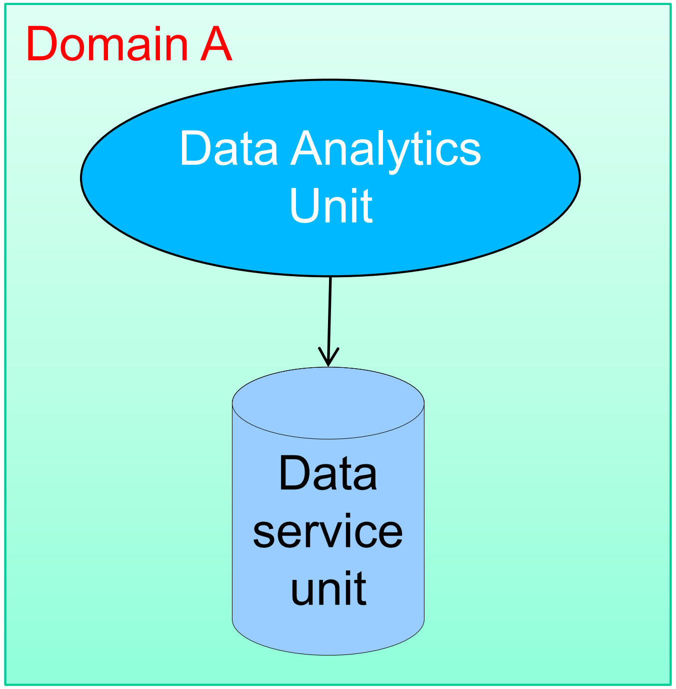
Pricing

...

You see we need to deal with many techniques and frameworks

WE NEED TO START FROM DATA ANALYTICS WITHIN A SINGLE SYSTEM

Data analytics within a single system



- They are complex enough but do not meet all requirements
- In a single domain
  - Tightly coupled computing infrastructures
    - E.g., in the same cloud
  - Computation and data are close
  - Several concerns can be by-passed

Not always provisioned under the „Service Unit“ model

Data analytics within a single system

Some papers

1. Andrew Pavlo, Erik Paulson, Alexander Rasin, Daniel J. Abadi, David J. DeWitt, Samuel Madden, and Michael Stonebraker. 2009. A comparison of approaches to large-scale data analysis. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of data (SIGMOD '09), Carsten Binnig and Benoit Dageville (Eds.). ACM, New York, NY, USA, 165-178. DOI=10.1145/1559845.1559865 http://doi.acm.org/10.1145/1559845.1559865

2. Leonardo Neumeyer, Bruce Robbins, Anish Nair, Anand Kesari: S4: Distributed Stream Computing Platform. ICDM Workshops 2010: 170-177

3. Jerry Chou, Mark Howison, Brian Austin, Kesheng Wu, Ji Qiang, E. Wes Bethel, Arie Shoshani, Oliver Rübel, Prabhat, and Rob D. Ryne. 2011. Parallel index and query for large scale data analysis. In Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC '11). ACM, New York, NY, USA, , Article 30 , 11 pages. DOI=10.1145/2063384.2063424 http://doi.acm.org/10.1145/2063384.2063424

4. Boduo Li, Edward Mazur, Yanlei Diao, Andrew McGregor, Prashant J. Shenoy: A platform for scalable one-pass analytics using MapReduce. SIGMOD Conference 2011: 985-996

5. Fabrizio Marozzo, Domenico Talia, Paolo Trunfio: A Cloud Framework for Parameter Sweeping Data Mining Applications. CloudCom 2011: 367-374

6. Yingyi Bu, Bill Howe, Magdalena Balazinska, Michael D. Ernst: HaLoop: Efficient Iterative Data Processing on Large Clusters. PVLDB 3(1): 285-296 (2010)

Data analytics within a single system – some examples

Message Passing Interface (MPI) + Cluster-based File system

Parallel Database (SQL/NonSQL)

MapReduce + Google File System

Yahoo S4

Hadoop + HDFS

Spark

Dryad+LINQ

Scientific/Business Workflow

A short, good overview in Chapter 6: Cloud Programming and Software Environments, Book: Distributed and Cloud Computing – from Parallel Processing to the Internet of Things, Kai Hwang, Geoffrey C. Fox and Jack J Dongarra, Morgan Kaufmann, 2012

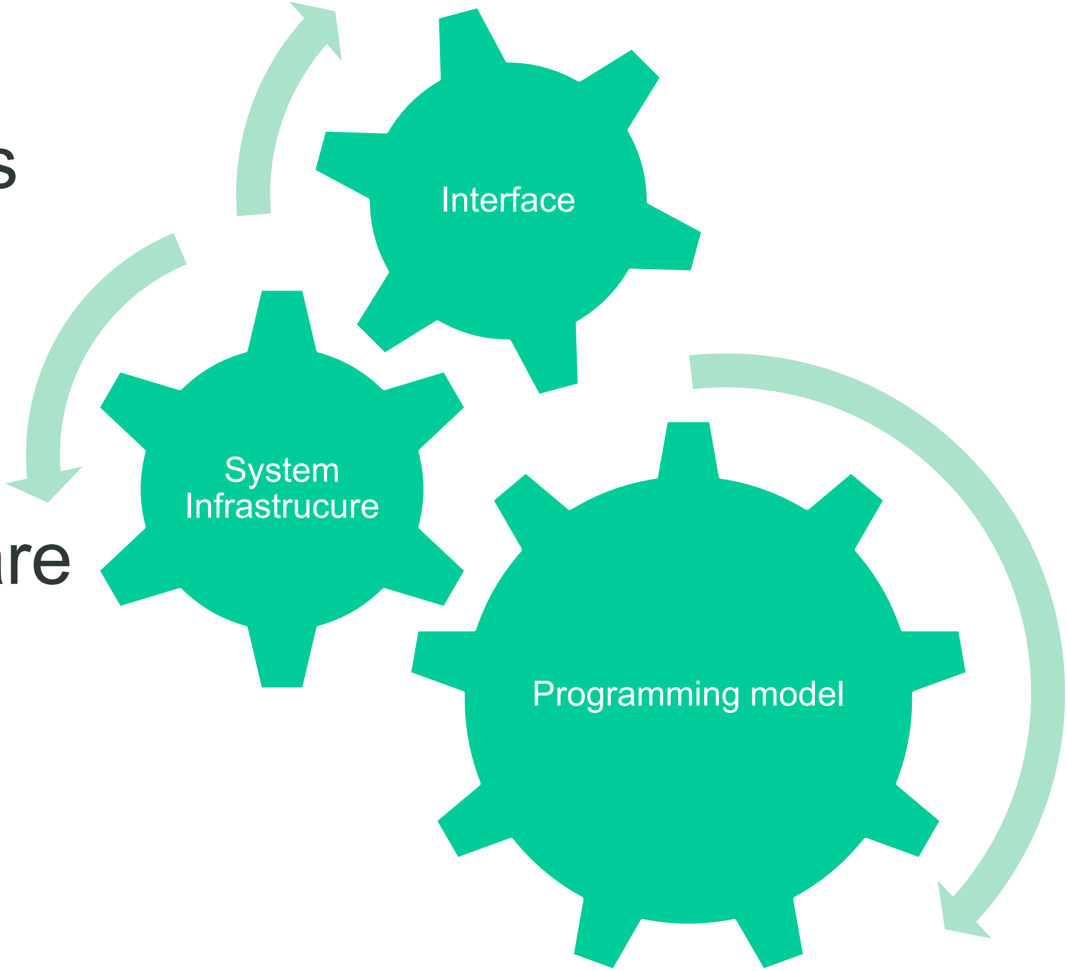
WHY SHOULD ANALYTICS UNITS BE „CLOSED“ TO DATA UNITS?

WHICH CONCERNS COULD BE IGNORED IN SINGLE SYSTEM DATA ANALYTICS?

WHICH ARE THE ISSUES THAT WE NEED TO CONSIDER WHEN OUR DATA UNITS ARE IN DIFFERENT SYSTEMS?

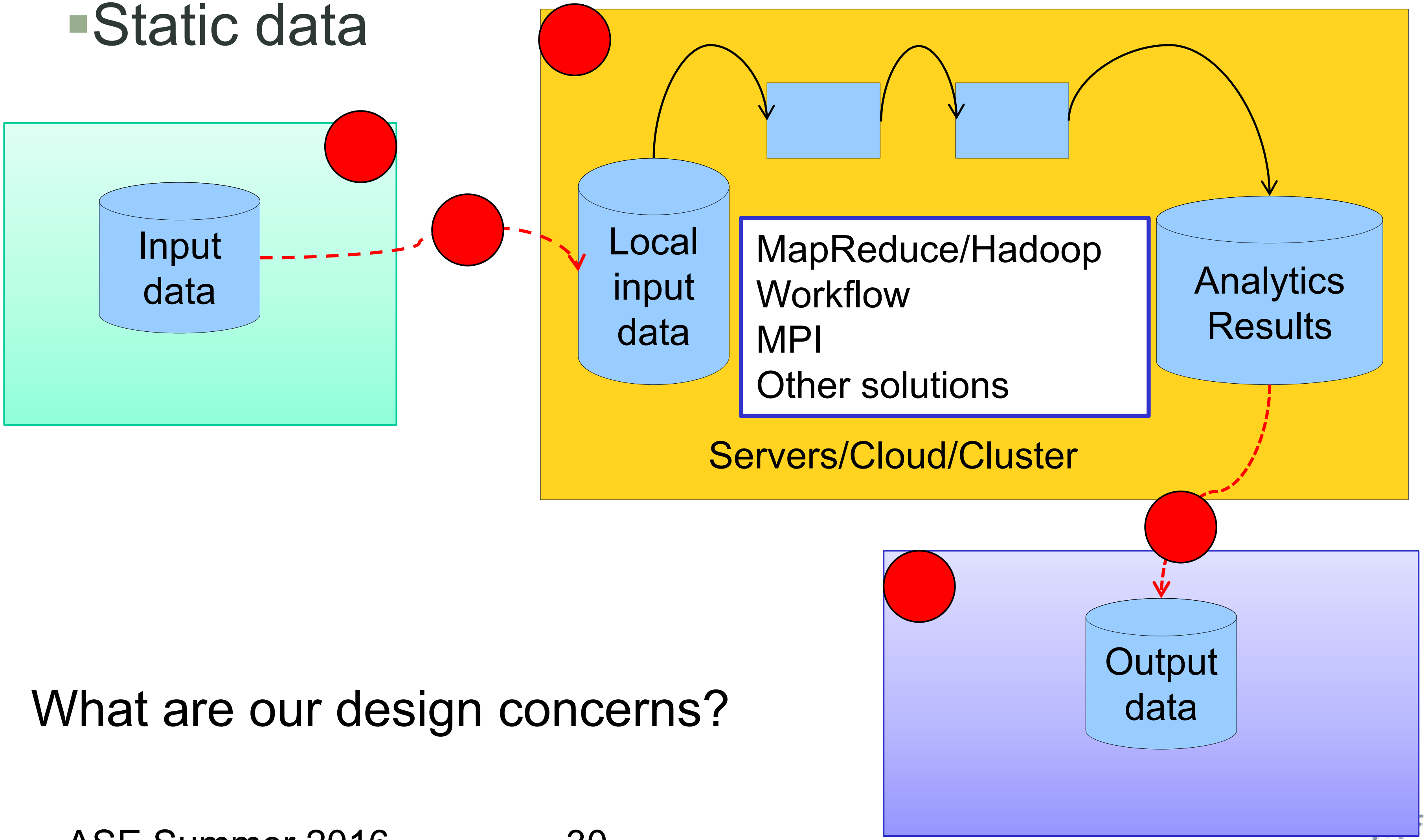
Data analytics across multiple systems – design choice

- Programming models for data analytics service
- Data service units
- Supporting middleware units



Data analytics across multiple systems – programming models (1)

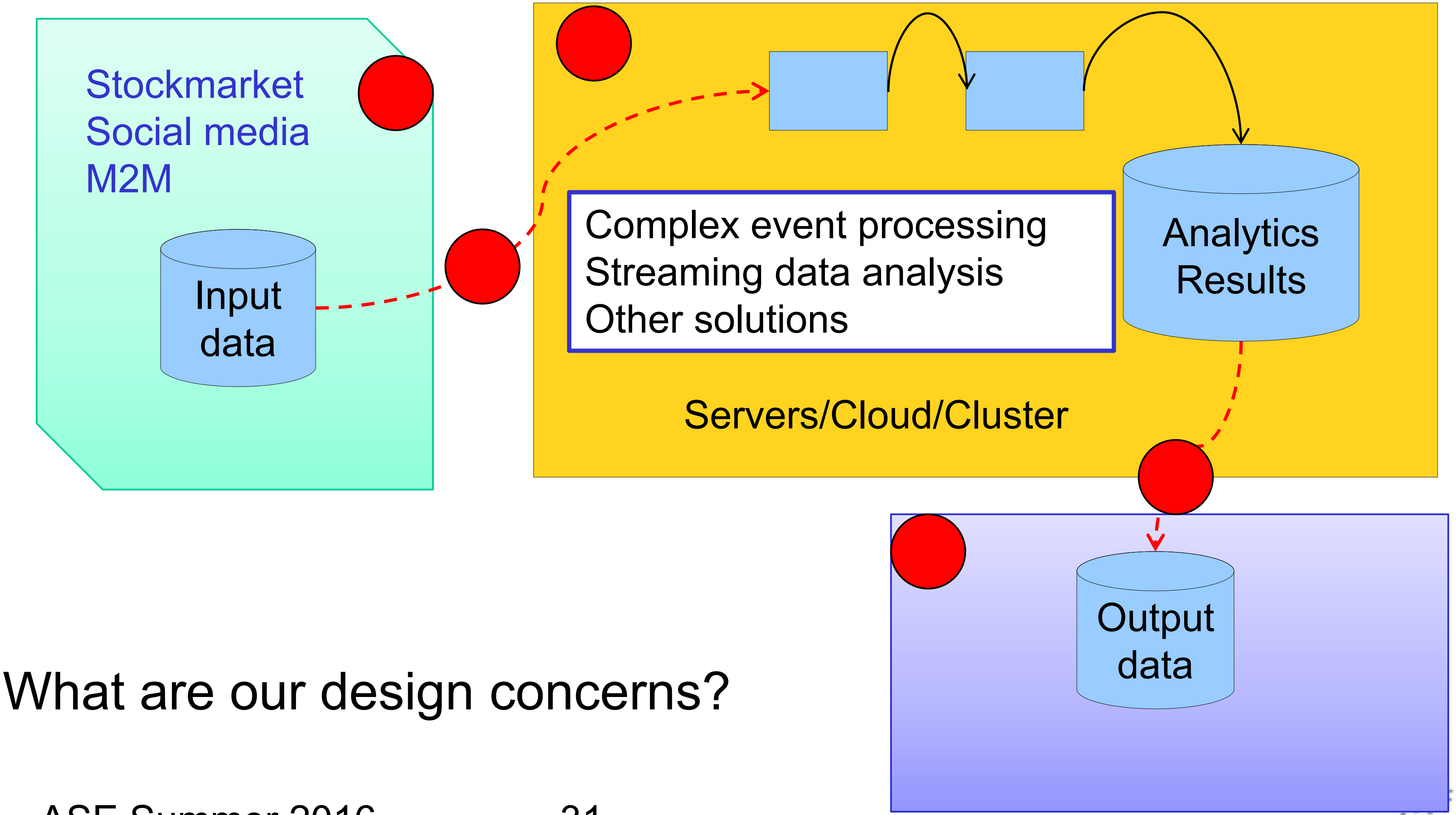
- Static data



What are our design concerns?

Data analytics across multiple systems – programming models (2)

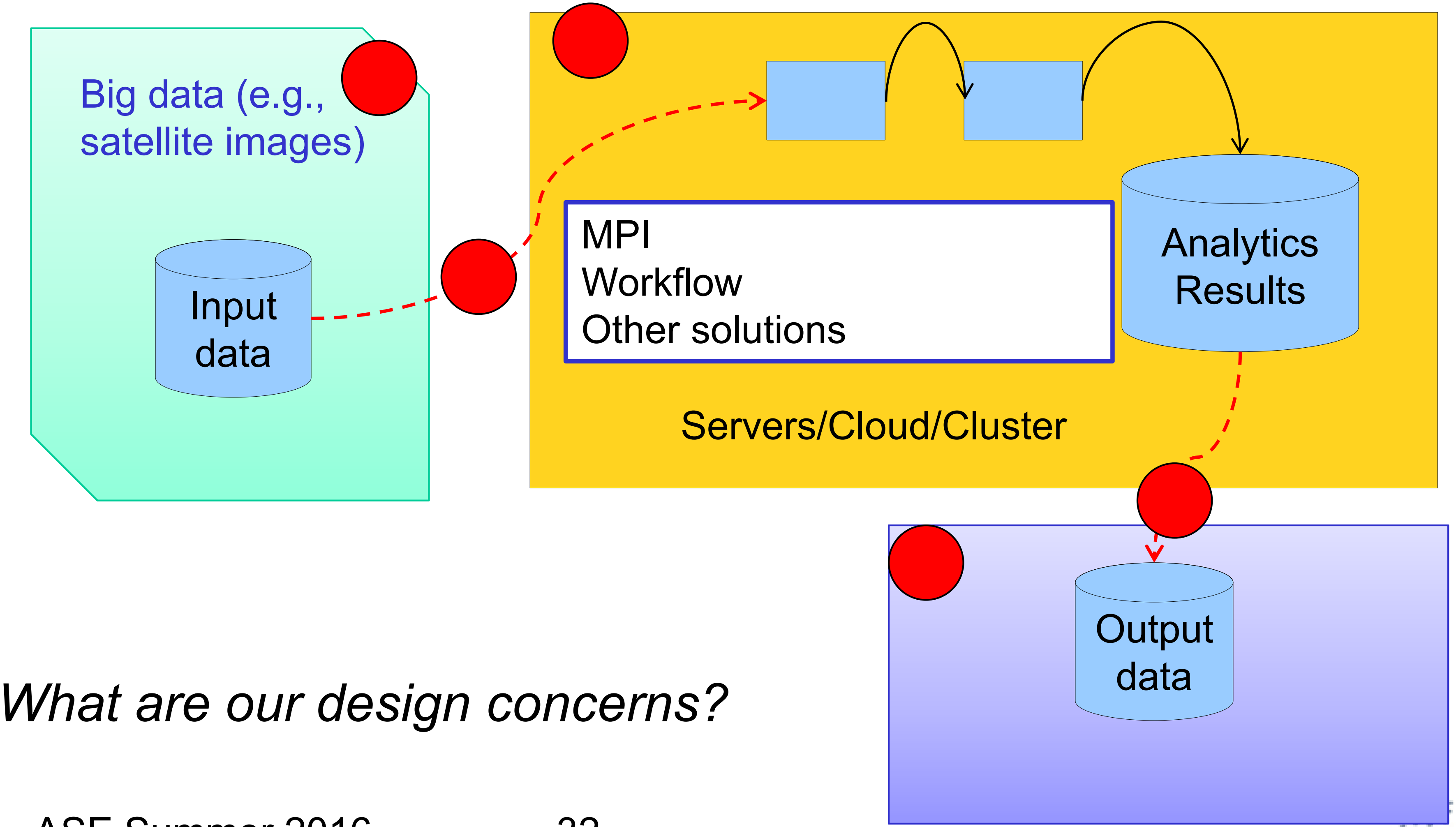
- Near-realtime data



What are our design concerns?

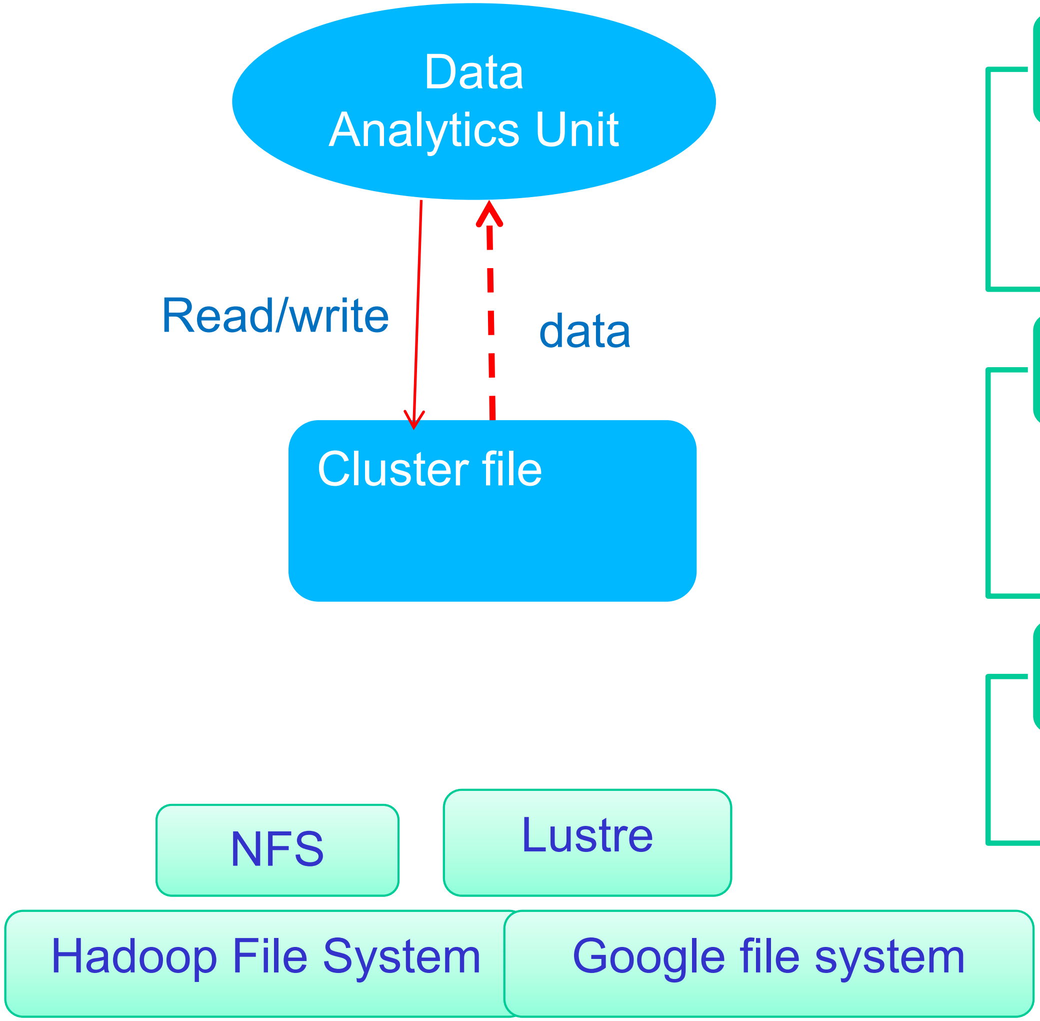
Data analytics across multiple systems – programming models (3)

- Near-realtime data



What are our design concerns?

Data analytics across multiple systems – data service units



Data Analytics Unit

Read/write

data

Cluster file

NFS

Lustre

Hadoop File System

Google file system

Interface

- Read/write data via direct , low-level read/write via IO

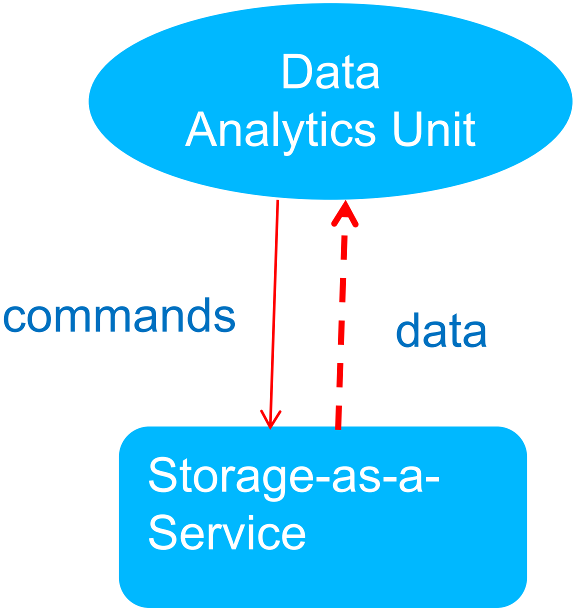
System

- Cluster or cluster of clusters
- Can be very large

Programming model

- Usually parallel processing

Data analytics across multiple systems – data service units



Interface

- Direct data transfer via REST/SOAP APIs

System

- Decouple between analytics and storage

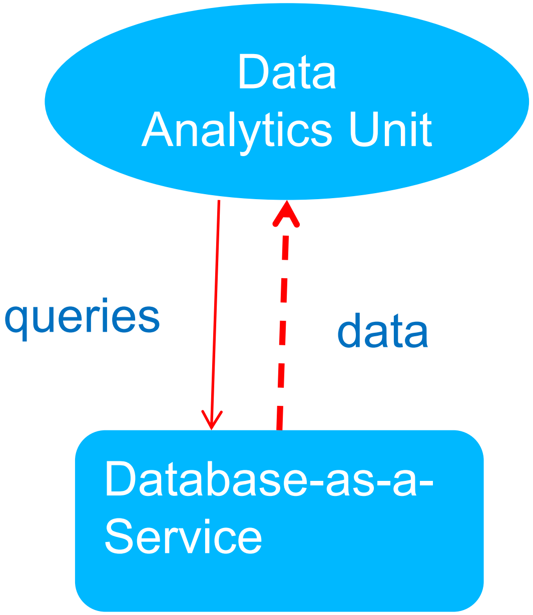
Programming model

- May require middleware for data transfer
  - Request via SOAP/REST
  - Real data transfer done by external middleware
- A rich set of programming models can be used

Amazon S3 (SOAP/REST API)

Google Storage Service (REST API)

Data analytics across multiple systems – data service units



Interface

- REST/SOAP APIs
- Mainly for commands and results

System

- Decouple between analytics unit and database
- Database as a sevice can be very large

Programming model

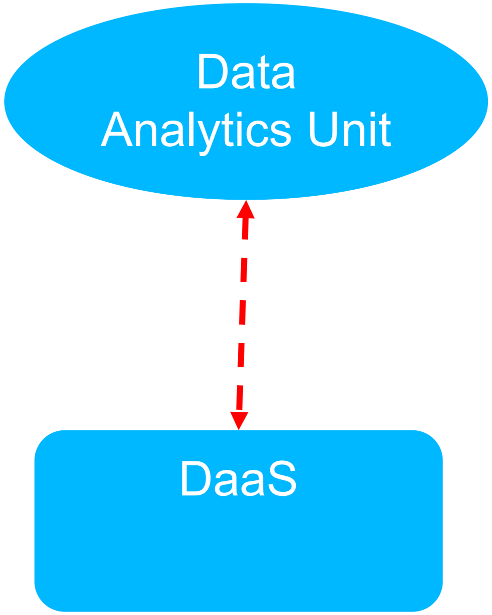
- Analytics can be done at both sides
- Analytic units can use any programming models
- Database-as-a-service can perform a lot of analytics
  - Parallel database operations

Technology

MongoDB/MongoLab
Amazon DynamoDB
Amazon SimpleDB
Cloudant Data

SkySQL
Amazon RDS
Microsoft SQL Azure
Clustrix DBaaS

Data analytics across multiple systems – data service units



Technology

Infochimps
Microsoft Azure
Xively
GNIP

Interface

- Data transfer can be uni or bi-direction
- REST/SOAP APIs

System

- Both systems for DaaS and for analytics units can be very large

Programming model

- Can be any

Middleware service unit for transfering large data -- GlobusOnline

Figure 1. Globus Online architecture.

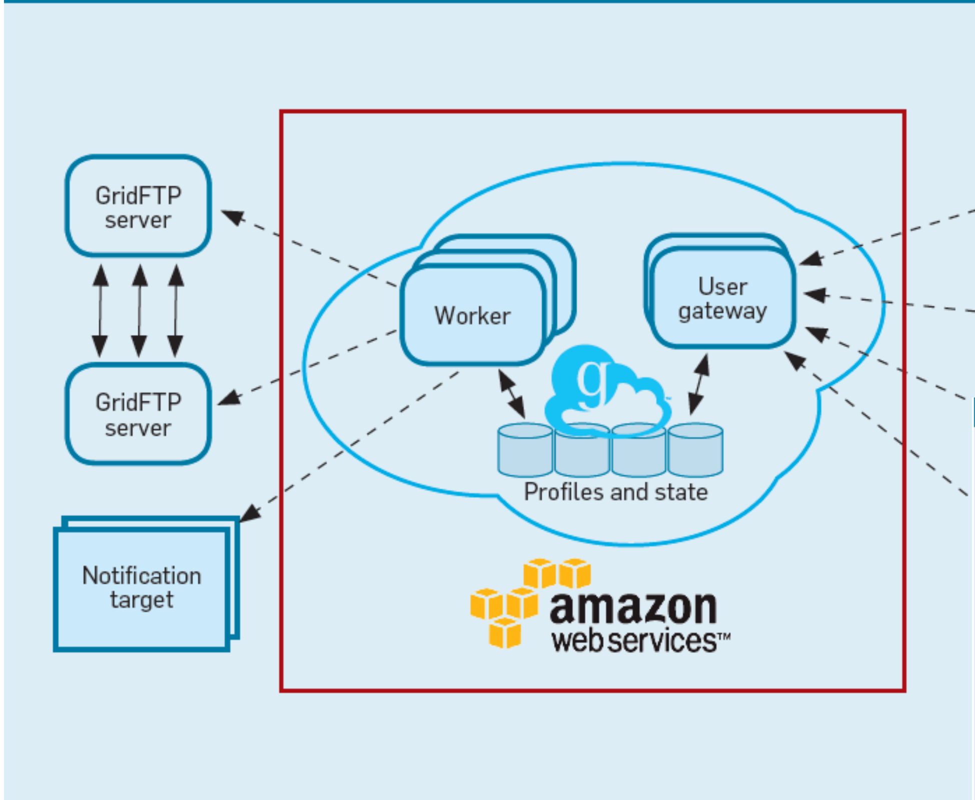
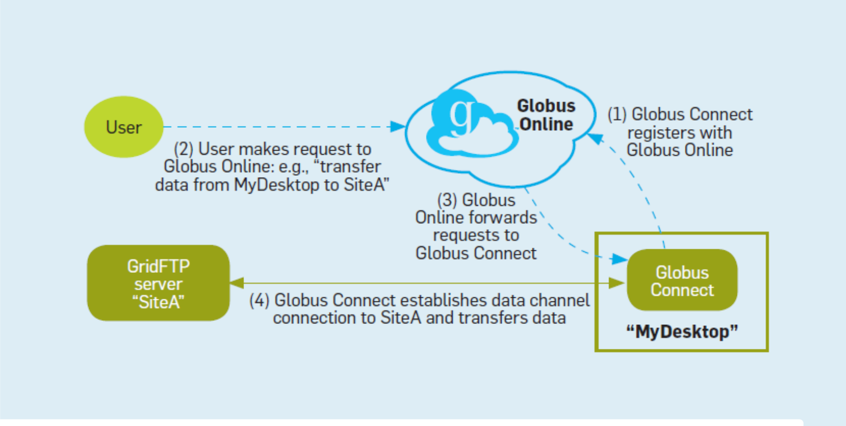


Figure 2. Globus Connect architecture.



User
Globus Online
(1) Globus Connect registers with Globus Online
(2) User makes request to Globus Online: e.g., "transfer data from MyDesktop to SiteA"
(3) Globus Online forwards requests to Globus Connect
GridFTP server "SiteA"
(4) Globus Connect establishes data channel connection to SiteA and transfers data
Globus Connect
"MyDesktop"

Source: Bryce Allen, John Bresnahan, Lisa Childers, Ian Foster, Gopi Kandaswamy, Raj Kettimuthu, Jack Kordas, Mike Link, Stuart Martin, Karl Pickett, and Steven Tuecke. 2012. Software as a service for data scientists. Commun. ACM 55, 2 (February 2012), 81-88. DOI=10.1145/2076450.2076468 http://doi.acm.org/10.1145/2076450.2076468

Middleware service units for messages/queuing

- Advanced Message Queuing Protocol (AMQP)
- Simple (or Streaming) Text Orientated Messaging Protocol (STOMP)
- Specific protocols/APIs

StormMQ

RabbitMQ

Amazon SQS

So many types of services from different providers. Anyway to simplify the management of service units for the developer/operator?

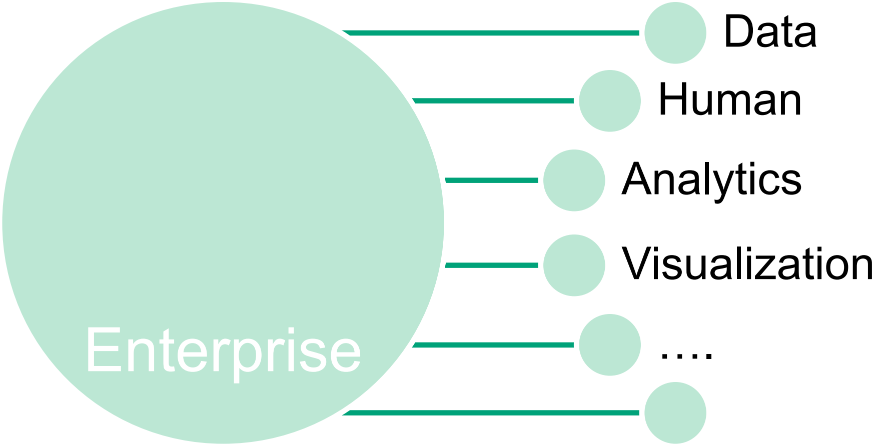
API MANAGEMENT

Ecosystem view for advanced service engineering

- Complex data analytics applications → need to understand potential service units from an ecosystem perspective
  - Interdependent systems: Social computing, mobile computing, cloud computing, data management, etc.
  - Different types of information are linked
  - Blending vertical and horizontal analytics
  - Different functions (analytics, visualization, communications, etc.)
  - Too many different types of customers (and their interactions)

APIs

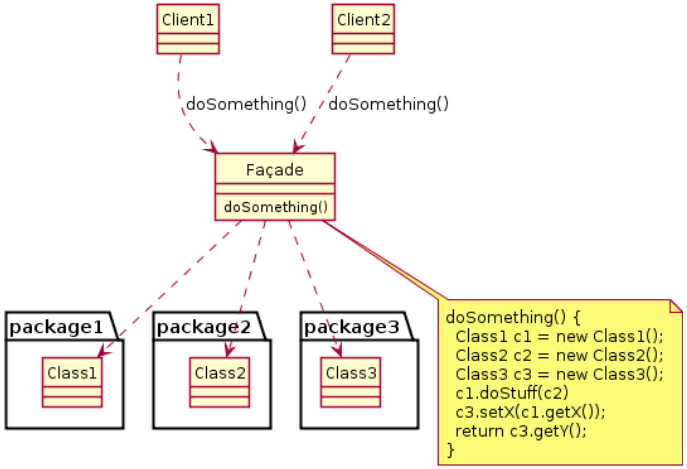
- APIs are key! Why?
  - Enable access to data and function from entities in your ecosystem
  - Virtualization



- An API is an asset
  - We need to have lifecycle, pricing, management, etc.

Check http://www.apiacademy.co for some useful tutorials

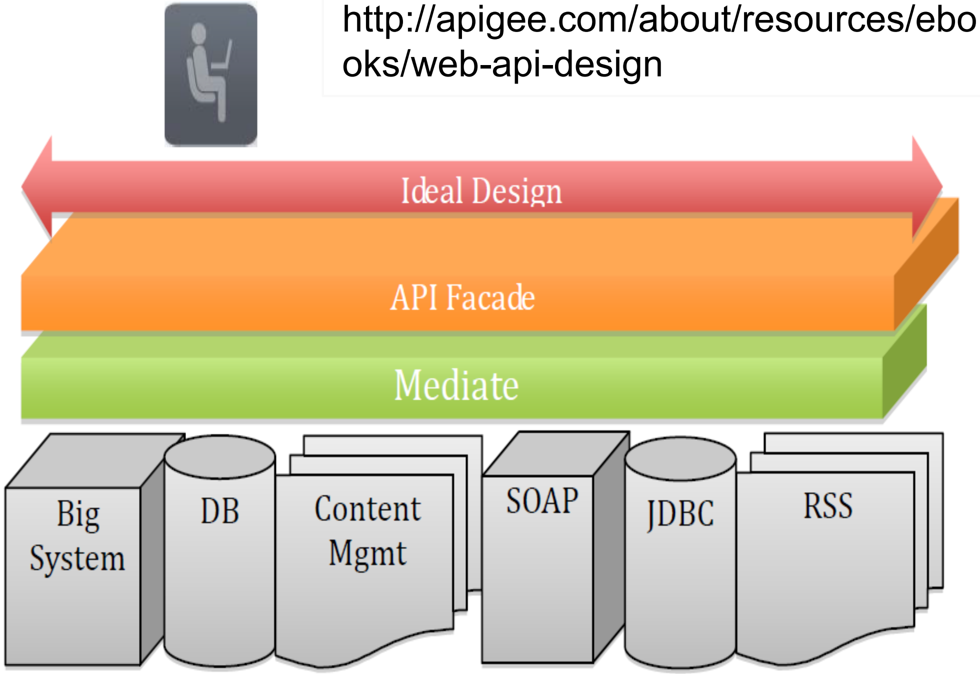
API Fasade



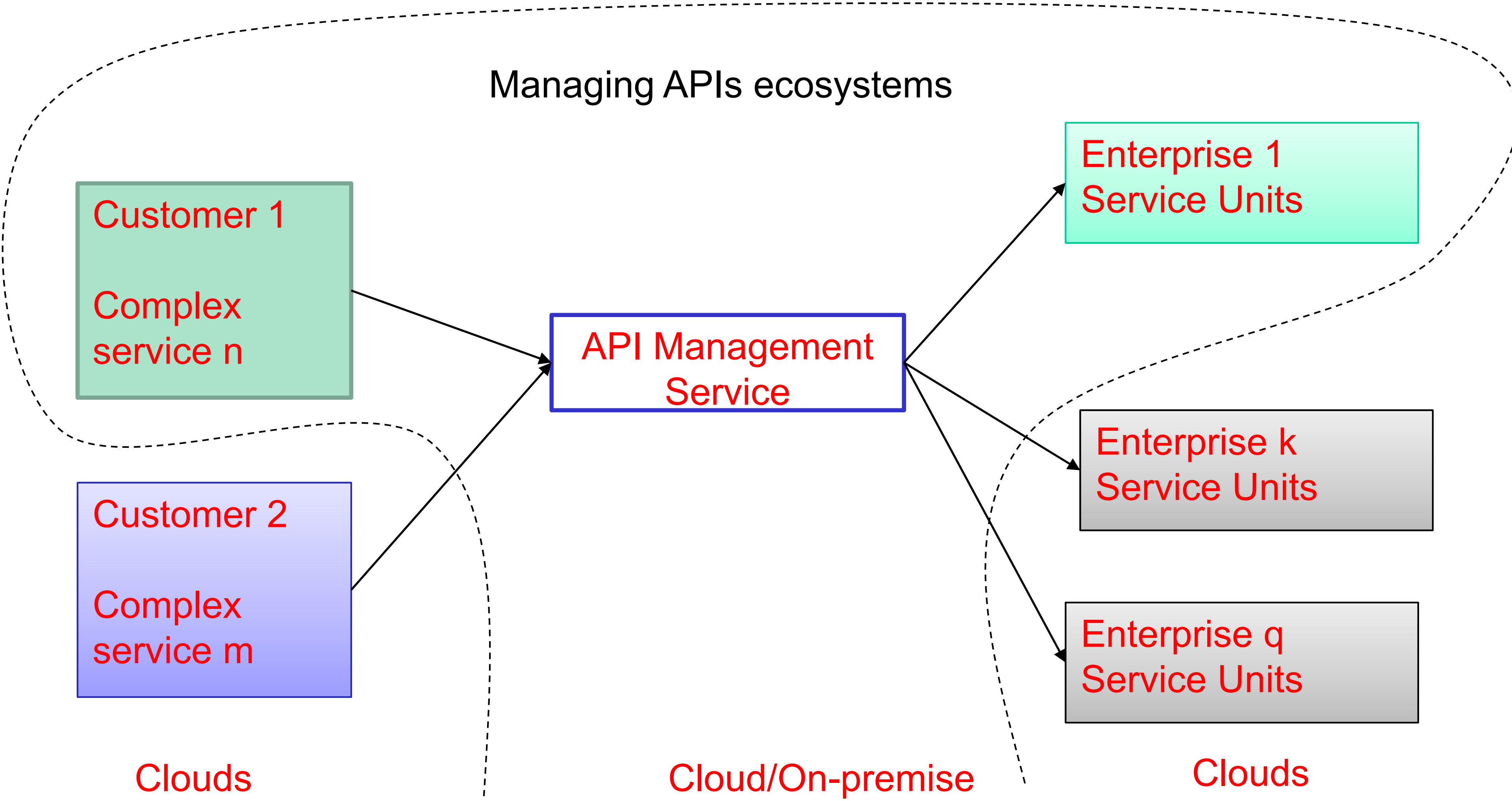
```
doSomething() {
  Class1 c1 = new Class1();
  Class2 c2 = new Class2();
  Class3 c3 = new Class3();
  c1.doStuff(c2)
  c3.setX(c1.getX());
  return c3.getY();
}
```

Sourre:
https://en.wikipedia.org/wiki/Facade_pattern

Source: Web API Design, Brian Mulloy
http://apigee.com/about/resources/ebooks/web-api-design



API management & APIs as a service


Managing APIs ecosystems

Customer 1

Complex service n

Customer 2

Complex service m

API Management Service

Enterprise 1 Service Units

Enterprise k Service Units

Enterprise q Service Units

Clouds

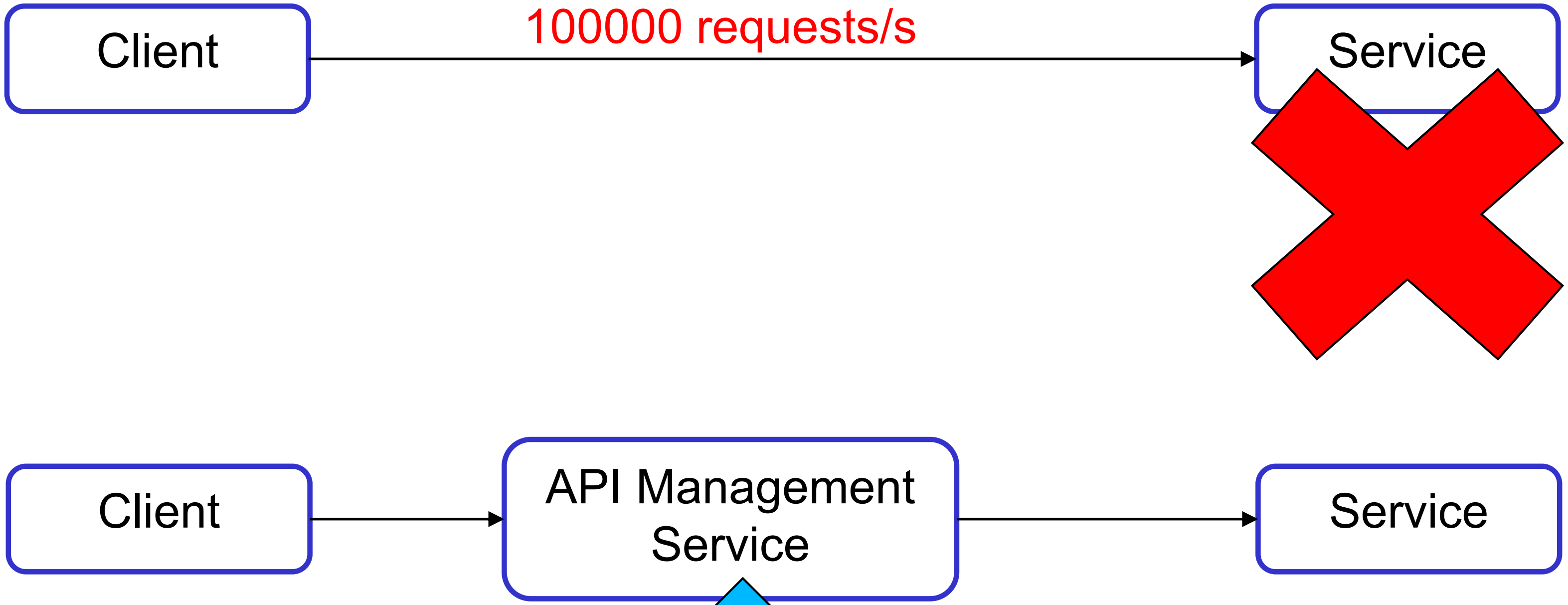
Cloud/On-premise

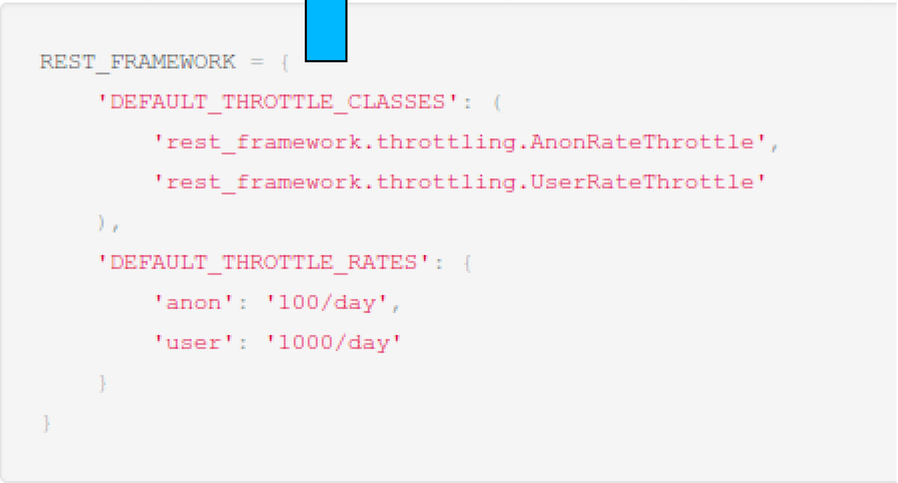
Clouds

Development of APIs

- Not just the functions behind the APIs
  - This we have learned since a long time
- Emerging (business/service) management aspects
  - Usage control and security
  - Any where from any device for any customer
    - Interfaces (communications, inputs/output formats)
  - APIs as a service:
    - Availability and reliability of APIs are important – think APIs are similar to a service that your client will consume

Prevent too many accesses?





```
REST_FRAMEWORK = {
    'DEFAULT_THROTTLE_CLASSES': (
        'rest_framework.throttling.AnonRateThrottle',
        'rest_framework.throttling.UserRateThrottle'
    ),
    'DEFAULT_THROTTLE_RATES': {
        'anon': '100/day',
        'user': '1000/day'
    }
}
```

Code: http://www.django-rest-framework.org/api-guide/throttling/#how-throttling-is-determined

How can we use API management for data/service contracts?

Issues on APIs management

- Publish
  - Business and operation planning
    - API usage schemes (e.g., pricing, data concerns)
    - API payload transform policies
    - API throttling
  - API publish and discovery (like service discovery?)
- Management
  - Management roles in enterprises, versions, etc.
- Monitoring and analytics
  - monitoring and analytics information (availability, types of customers, usage frequencies, etc.)

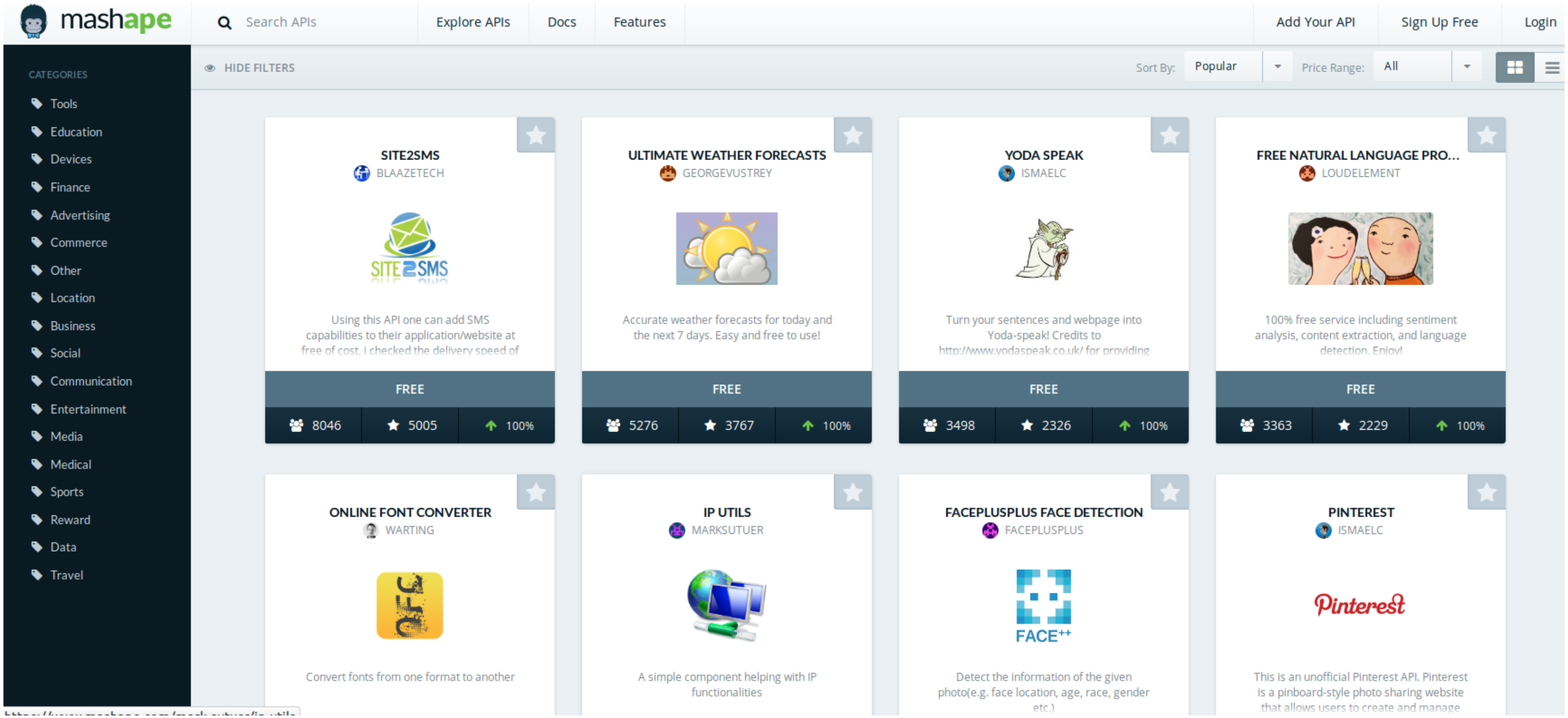
Some well-known frameworks

- http://apigee.com
- Oracle API management: http://www.oracle.com/us/products/middleware/soa/api-management/overview/index.html
- http://wso2.com/api-management/
- http://www.ca.com/us/lpg/layer-7-redirects.aspx
- https://www.mashape.com/
- http://apiaxle.com/

Build your own APIs ecosystem

- Which APIs you need? Which ones are crucial for you to build complex services?
  - Data APIs
    - Data collection
    - Visualization
    - Analytics APIs
  - Communication
  - Coordination of tasks
- API marketplaces → your APIs
- Using existing API platforms to manage your APIs

Examples of an API marketplace



SITE2SMS
BLAAZETECH
Using this API one can add SMS capabilities to their application/website at free of cost. i checked the delivery speed of
FREE
8046 ★ 5005 100%

ULTIMATE WEATHER FORECASTS
GEORGEVUSTREY
Accurate weather forecasts for today and the next 7 days. Easy and free to use!
FREE
5276 ★ 3767 100%

YODA SPEAK
ISMAELC
Turn your sentences and webpage into Yoda-speak! Credits to http://www.yodaspeak.co.uk/ for providing
FREE
3498 ★ 2326 100%

FREE NATURAL LANGUAGE PRO...
LOUDELEMENT
100% free service including sentiment analysis, content extraction, and language detection. Enjoy!
FREE
3363 ★ 2229 100%

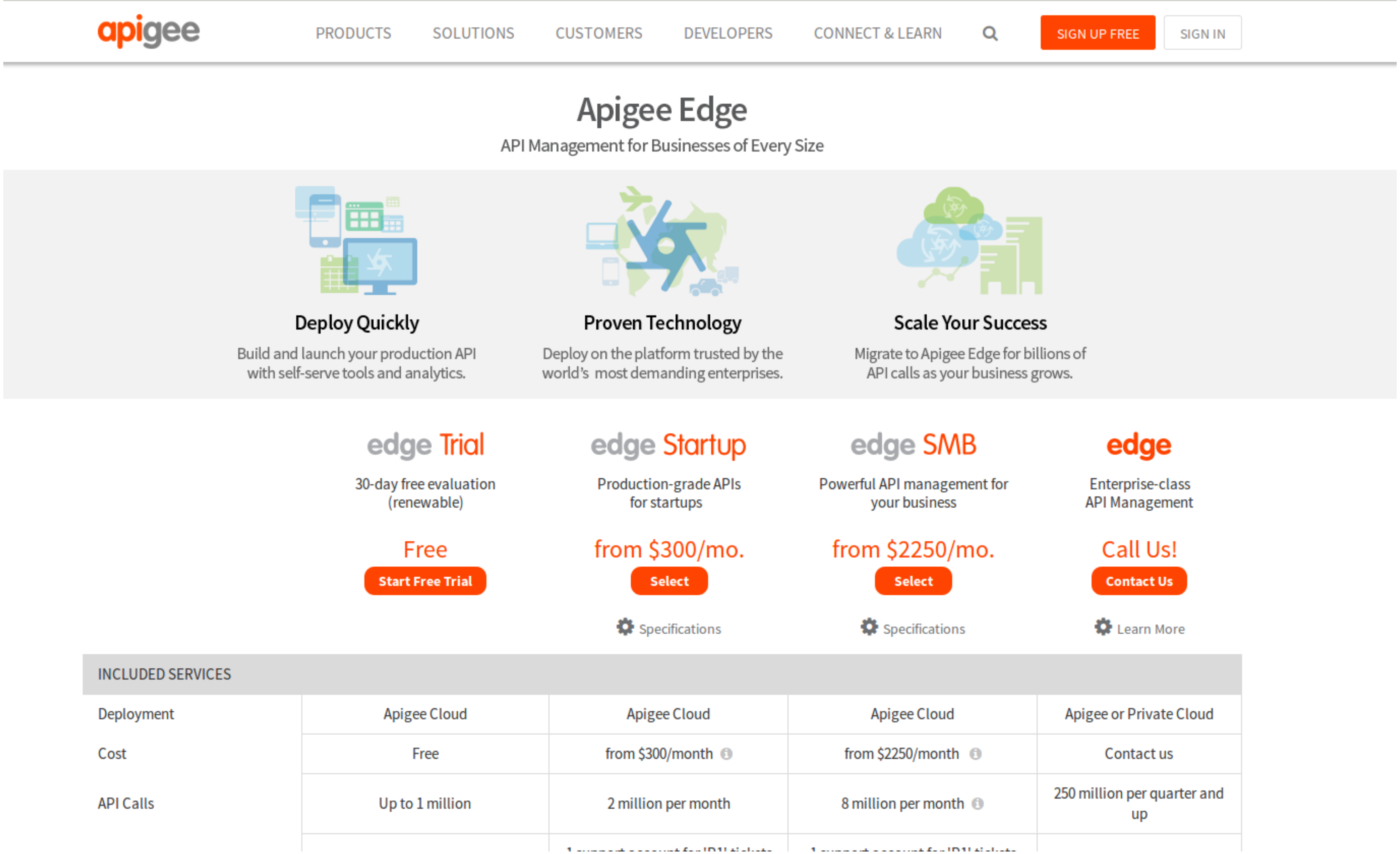
ONLINE FONT CONVERTER
WARTING
Convert fonts from one format to another

IP UTILS
MARKSUTUER
A simple component helping with IP functionalities

FACEPLUSPLUS FACE DETECTION
FACEPLUSPLUS
Detect the information of the given photo(e.g. face location, age, race, gender etc.)

PINTEREST
ISMAELC
This is an unofficial Pinterest API. Pinterest is a pinboard-style photo sharing website that allows users to create and manage

Use API Management for your mini project?



Apigee Edge

API Management for Businesses of Every Size

Deploy Quickly

Build and launch your production API with self-serve tools and analytics.

Proven Technology

Deploy on the platform trusted by the world's most demanding enterprises.

Scale Your Success

Migrate to Apigee Edge for billions of API calls as your business grows.

edge Trial

30-day free evaluation (renewable)

Free

Start Free Trial

edge Startup

Production-grade APIs for startups

from $300/mo.

Select

Specifications

edge SMB

Powerful API management for your business

from $2250/mo.

Select

Specifications

edge

Enterprise-class API Management

Call Us!

Contact Us

Learn More

INCLUDED SERVICES				
Deployment	Apigee Cloud	Apigee Cloud	Apigee Cloud	Apigee or Private Cloud
Cost	Free	from $300/month ⓘ	from $2250/month ⓘ	Contact us
API Calls	Up to 1 million	2 million per month	8 million per month ⓘ	250 million per quarter and up
		1 support account for 'P1' tickets	1 support account for 'P1' tickets	

Exercises

- Read mentioned papers
- Analyze the relationships between programming models and system infrastructures for data analytics across multiple domains
- Examine http://cloudcomputingpatterns.org and see how it supports data analytics patterns
- Develop some patterns for data analytics across multiple systems
- Setup an API management platform for your work

Thanks for your attention

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