Emerging Distributed Computing and Challenges for Services Engineering

Hong-Linh Truong

Faculty of Informatics, TU Wien

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

- See emerging trends in distributed systems and computing
- Have a critical look at use cases and analyze use cases
- See the service engineering technologies needed for such use cases
Outline

- Some emerging models
  - IoT resources
  - (Big/realtime) data provisioning models
  - Computational infrastructures/frameworks provisioning
  - Human computation provisioning
  - Machine Learning as a service
  - Blockchain

- Use cases

- Advanced services engineering
  - Single service/platform engineering
  - Multi-platform services engineering
WHICH ARE EMERGING FORMS OF DISTRIBUTED COMPUTING MODELS, SYSTEMS AND APPLICATIONS THAT YOU SEE?
Emerging data provisioning models

- Large (near-) realtime data
  - Satellites and environmental/city sensor networks (e.g., from specific orgs/countries)
  - Machine-to-machine (e.g., from companies)
  - IoT/Sensor data
  - Manufacturing data
  - Social media (e.g., from people + platform providers)

- Open (business, science and government) data
  - Open science and engineering data sets
  - Open government data
  - Open business data

- Marketable data
  - Statistics and business data
  - Commercial data in general

Data are assets
Satellite data

Source: https://sentinel.esa.int/web/sentinel/sentinel-data-access
Large-scale (near-)realtime data: properties and issues

Some properties
- Having massive data
- Requiring large-scale, big (near-) real time processing and storing capabilities
- Enabling predictive and realtime data analytics

Some issues
- Timely analytics
  - Performance and scalability
- Quality of data control
- Handle of unknown data patterns
- Benefit/cost versus quality tradeoffs
Example of open data

a comprehensive list of open data catalogs curated by experts from around the world.

268 registered data catalogs available.
Static + Realtime Open Data

The next stage in open data.
Plenario is a centralized hub for open datasets from around the world, ready to search and download.

One database. One map.
All data in Plenario exists on a single map and a single timeline, making it incredibly easy to access multiple datasets at once—especially those originally housed at different data portals.

Here's an example: A collection of data pulled from Southside Chicago during summer 2013.

<table>
<thead>
<tr>
<th>Source</th>
<th>Dataset</th>
<th>Count</th>
<th>Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chicago Police Department</td>
<td>Crimes - 2001 to present</td>
<td>38,766</td>
<td>600</td>
</tr>
<tr>
<td>City of Chicago</td>
<td>Business Licenses</td>
<td>2,492</td>
<td>125</td>
</tr>
<tr>
<td>City of Chicago</td>
<td>Food Inspections</td>
<td>1,226</td>
<td>60</td>
</tr>
</tbody>
</table>
Telecommunication

https://dandelion.eu/datamine/open-big-data/
Open data: properties and issues

Some properties

- Having large, multiple data sources but mainly static data
  - Real-time, open data is growing
- Having good quality control in many cases
- Usually providing the data as a whole set

Some issues

- Fine-grained content search
- Balance between processing cost and performance
- Correlation/combination with real-time/private data
Marketable data examples

Real-Time Targeting

Data Buying and Targeting through the BDEX suite of tools truly takes the industry to the next level. Through BDEX's unique tool set Advertisers, Publishers and Retailers alike have the ability to target with a level of granularity that was never before possible.

- Buy Data That is Only Seconds Old
- Filter Based on Data Quality (conversions)
- Create Custom Audience Groups
- Combine an Unlimited Number of Data Points
- Set Budgets by Data Point
- Manage Campaigns in Real-Time

View data available on BDEX with this Data Visualization Tool
Marketable data examples
Marketable data examples
And trend in monetizing data

Source: https://databrokerdao.com/
Marketable data: properties and issues

Some properties
- Can be large, multiple data sources but mainly static data
- Having good quality control
- Have strong data contract terms
- Some do not offer the whole dataset

Some issues
- Multiple levels of service/data contracts
- Compatible with other data sources w.r.t. contract
- Cost w.r.t. up-to-date data
- Near-realtime data marketplaces
in EU call for proposals

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Publication date:</td>
<td>27 October 2017</td>
</tr>
</tbody>
</table>

**Types of action:**

<table>
<thead>
<tr>
<th>IA Innovation action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deadline:</td>
</tr>
<tr>
<td>single-stage</td>
</tr>
<tr>
<td>16 October 2018</td>
</tr>
</tbody>
</table>

**Deadline: 28 March 2019 17:00:00**

<table>
<thead>
<tr>
<th>RIA Research and Innovation action</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSA Coordination and support action</td>
</tr>
<tr>
<td>Deadline:</td>
</tr>
<tr>
<td>single-stage</td>
</tr>
<tr>
<td>31 October 2017</td>
</tr>
</tbody>
</table>

**Deadline: 17 April 2018 17:00:00**

Time Zone: (Brussels time)

Emerging computing infrastructure and platform provisioning models

- **Infrastructure-as-a-Service**
  - Machine as a service
  - Storage as a Service
  - Database as a Service
  - Network as a Service (think about Network Function Virtualization with 5G)

- **Edge/Fog computing**
  - Distributed edge/fog systems
    - analytics at the edge
    - Network functions and other system operations at the edge/fog systems
Emerging computing infrastructure and platform provisioning models

- Platform-as-a-Service
  - Application middleware
  - Computational frameworks
  - Data processing frameworks
  - Management middleware (e.g., monitoring, control, deployment)

- Technologies
  - Virtualization
  - Microservice architectures
  - Serverless computing
  - Machine learning/deep learning
  - Blockchain
  - Etc.
Examples

Data Processing Framework
- Amazon Elastic MapReduce
- Apache Flink
- Apache Apex
- Apache Spark
- Google Cloud Dataflow
- Kafka SQL
- Azure Stream Analytics

Data Transfer/Messaging Middleware
- StormMQ
- Apache Nifi
- Kafka
- MQTT
- Amazon SQS
- Cloud AMQP

Data Storages
- MongoLab
- Amazon S3
- Cassandra
- Google BigQuery
- Elastic Search
- Influx DB
- Built around Mapreduce programming models and Hadoop software ecosystems
- From “The Forrester Wave™: Big Data Hadoop Distributions, Q1 2016”: Top Hadoop solution providers are Cloudera, Hortonworks, IBM, MapR Technologies, and Pivotal Software

Spark ecosystem

Programming with Java, Scala, Python, R
We can have a separate modules

Spark SQL + DataFrames
Streaming
MLlib Machine Learning
GraphX Graph Computation

Spark Core API
R SQL Python Scala Java

Figure source: https://databricks.com/spark/about
ELK Stack

Building using elastic components: Elasticsearch, Elasticsearch Hadoop, Kibana, and Logstash

https://www.elastic.co/

TICK Stack

Main from services of Influx

https://www.influxdata.com

Focus on time series data

- Collect
- Storage
- Visualize
- ETL

TICK

Telegraf
- Time-Series Data Collector

InfluxDB
- Time-Series Data Storage

Chronograf
- Time-Series Data Visualization

Kapacitor
- Time-Series Data Processing
Machine Learning Stack
Cloud-based Analytics

A lot

Things

Social Platforms

Environments

Critical Infrastructures

A few

Data/Service Platforms

Data Profiling and Enrichment

Data Storage

Data Processing

Data Query

A lot

Applications

Applications

A lot

Data Analytics Algorithms/Processes

...
Emerging computing infrastructure/platform provisioning models—properties and issues

Some properties
- Rich types of services from multiple providers
  - Better choices in terms of functions and costs
- Concepts are similar but diverse APIs
- Strong dependencies/tight ecosystems

Some issues
- On-demand information management from multiple sources
- APIs complexity and API management
- Cross-vendor integration
- Execution in Multi-cloud environments
- Data locality
- Service mess/discovery
Emerging human computation models

- Crowdsourcing platforms
  - (Anonymous) people computing capabilities exploited via task bids
- Expert as Individual Compute Unit
  - An individual is treated like „a processor“ or “functional unit“. A service can wrap human capabilities to support the communication and coordination of tasks
- A set of individuals as collectives
  - A set of people and software that are initiated and provisioned as a service for solving tasks

The main point: humans are a computing element
Examples of human computation (2)

```java
import edu.umass.cs.automan.adapters.MTurk_

object SimpleProgram extends App {
  val a' = MTurkAdapter { mt =>
    mt.access_key_id = "XXXX"
    mt.secret_access_key = "XXXX"
  }

  def which_one() = a'.RadioButtonQuestion { q =>
    q.budget = 8.00
    q.text = "Which one of these does not belong?"
    q.options = List(
      a'.Option('oscar, "Oscar the Grouch"),
      a'.Option('kermit, "Kermit the Frog"),
      a'.Option('spongebob, "Spongebob Squarepants"),
      a'.Option('cookie, "Cookie Monster"),
      a'.Option('count, "The Count"
    )
  }

  println("The answer is " + which_one())
}
```

Human computation models – properties and issues

Some properties
- Huge number of people
- Capabilities might not know in advance
- Unpredictable behavior
- Simple coordination models

Some issues
- Reliability
- Quality control
- Reliability assurance
- Proactive, on-demand acquisition
- Incentive strategies
- Collectives
Discussion time:

DO I NEED TO STUDY THEM ALL? WHY?
USE CASES/SCENARIOS
Critical infrastructures/services for citizens and business

Figure source: http://uidai.gov.in/images/AadhaarTechnologyArchitecture_March2014.pdf
Cities, e.g. including:
10000+ buildings
1000000+ sensors

Can we combine open government data with building monitoring data?
Earth Observation

Figure source:
https://eodatacube.eu/

ASE Summer 2018 35
Video analytics + business applications/public security

Use Case 3: Video Analytics

Figure 4: Example of video analytics

Figure source: https://portal.etsi.org/portals/0/tbpages/mec/docs/mobile-edge_computing__introductory_technical_white_paper_v1%2018-09-14.pdf
Edge/Cloud ML-based Video analytics

Chinese police are using smart glasses to identify potential suspects

Figure source: https://techcrunch.com/2018/02/08/chinese-police-are-getting-smart-glasses/

Figure caption 2: Video data include cabin and roadway views. This view illustrates possible relationships between observed driver behavior and the roadway context outside the vehicle, which includes other vehicles, cyclists, pedestrians, and lane markings.

Figure source: https://www.fhwa.dot.gov/research/resources/computervision_breakthrough.cfm
Shop scale

Welcome to Amazon Go and the world's most advanced shopping technology. No lines, no checkout—just grab and go!

Now open in Seattle!
Monday to Friday 7AM–9PM

https://www.amazon.com/b?ie=UTF8&node=16008589011
IoT data in City-scale

https://arrayofthings.github.io/node.html
Smart Farming

Geo Sports in the stadium scale

Geo Sports: Picture courtesy
Future Position X, Sweden

sensing  →  analytics  →  Decision
Drones for logistics

Source: DHL Trend Report “Unmanned Aerial Vehicles”
http://www.dhl.com/content/dam/downloads/g0/about_us/logistics_insights/dhl_trend_report_uav.pdf
Blockchain and IoT for Disaster Management?

Predictive Maintenance in Smart Buildings

IoT & Cloud Platforms

Things/Sensors

Data Analytics Platforms (Clouds, Edges)

Human/Expert Provisioning Platform

CRITICAL CLEAN CHILLER EVAPORATOR

<<send data>>

<<analyze data>>

<<monitor>>

<<invoke experts to predict and solve problems>>

<<notify possible problem>>

<<maintain evaporator>>

Predictive Maintenance in Smart Buildings

ASE Summer 2018
Predictive Maintenance in Telcos

- Offers cloud services for handling IoT data
- Offers cloud services for big data analytics
- Offers human-based services for complex problem solving

IoT Data Hub → Cloud Data Analytics Platform → Support Team

- <<analyze data>>
- <<notify problem>>
- <<control sensors>>
- <<monitor>>

Base Transceivers Station

ASE Summer 2018
CONVERGENCE OF MULTIPLE COMPUTING MODELS
Today’s Computing Models

- Internet infrastructure and software connect *contents, things, and people*, each has different roles (*computation, sensing, analytics, etc.*)
Emerging forms of computing models, systems and applications

- Big and high performance centralized data analytics
- IoT data streaming analytics
- Large-scale applications spanning data centers and edge servers/gateways
- Adaptive collective systems of humans and machines
Summary of emerging models wrt advanced service-based systems

Engineering advanced service-based systems

utilize/consist of

- Emerging data provisioning models
- Emerging computational infrastructure/platform provisioning models
- Emerging human computation models
- Emerging data provisioning models

Challenges in Virtualization, Programming, Communication, and Coordination, etc.
ADVANCED SERVICES ENGINEERING’S FOCUS
The service model can be applied to things, people and software

Consumption, ownership, provisioning, price, etc.

„basic component“/“basic function“ modeling and description / microservices
Single service/platform engineering – service unit provisioning

- Provisioning software, things and human capabilities under services
- E.g.g., video analytics, machine learning service, Spark cluster, BigQuery, HDFS, Ethereum, etc.

Internet-scale multi-platform services engineering – required technologies

Multi-platform Services Engineering for Software, Things and People

- IoT
- Middleware (e.g., StormMQ)
- Workflows (e.g., Airflow)
- Crowd platforms, human-based service platforms (e.g., Mturks)
- Data services (e.g., Azure, S3)
- Billing/Monitoring (e.g., thecurrencycloud)
- Blockchain
- Data analysis/Computation services in cluster (e.g., Hadoop)

ASE Summer 2018
Service engineering – the elasticity

- 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.)

Service engineering -- big/near-real time data impact

- Which are data concerns that impact the data processing?
- How to use data concerns to optimize data analytics and service provisioning?
- How to use available data assets for advanced services in an elastic manner?
- What are the role of human-based services in dealing with complex data analytics?
## Advanced service engineering -- Steps

### Single service/platform engineering

<table>
<thead>
<tr>
<th>Service units for representing fundamental things, people and software</th>
<th>Provisioning of fundamental service units</th>
<th>Engineering with single service units</th>
</tr>
</thead>
</table>

### Understanding Properties/Concerns

<table>
<thead>
<tr>
<th>Data /Service/Application concerns; their dependencies</th>
<th>Monitoring, evaluation and provisioning of concerns</th>
<th>Utilization of data/service concerns</th>
</tr>
</thead>
</table>

### Large-scale, multi-platform services engineering

<table>
<thead>
<tr>
<th>Identify platform/application problems</th>
<th>Identify the scale, complexity and *city</th>
<th>Design units, selection of existing service units;</th>
<th>Development and integration, optimization</th>
</tr>
</thead>
</table>
Exercises

- Read papers mentioned in slides
  - Get their main ideas
- Check services mentioned in examples
  - Examine capabilities of the mentioned services
    - Including price models and underlying technologies
  - Examine their size and scale
  - Examine their ecosystems and dependencies
- Work on possible categories of single service units that are useful for your work
  - Some common service units with capabilities and providers
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