CIRUS : A Cloud Infrastructure for Real-time Ubilytics (aka ubiquitous big data analytics)

Didier Donsez
Université de Grenoble
LIG / ERODS
P.N@imag.fr
Thanks to Manfred for the introduction of my talk

Cyber-Physical Systems - A Road to Connected Assistance and Global Services

Status: Accepted

Cyber-Physical Systems (CPS) are integrations of computation, networking, and physical processes. CPS are affecting all domains of digital technology, and the resulting innovation will create business models that change the rules.

In his keynote, Professor Broy will show how we can reap the benefits of CPS in global services and assistance networks, while minimizing the inherent risks of disruptive technology.

Session details

Speaker(s):
Manfred Broy

Session Type:
Keynote

Experience level:
Beginner

Track:
Keynote

Schedule info

Session Time Slot(s):
Cassiopée - Thursday, June 19, 2014 - 09:00 to 09:45
Web X.0 (X > 3)
Emerging ICT domains

- Cloud Computing
- Big Data Analytics
- Internet of (Every)Things
- Social Networks
- Mobile computing
- Crowd sourcing
- Open data
- ...

Internet of Things (IoT)

- Industrial IoT (IIoT)
- RFID / NFC
- phones
- sensor nodes
- robot

Instrumentation
Communication
Mediation
Decision
Action
Mining
Big Picture of Internet of (Every)Things, Data and Services

Network cell size

WAN
MAN
LAN
WLAN
WSN
PAN
BAN

Home Automation SOHO
Smart Public Space Smart Building Industry 4.0
Smart Cities Urban Spaces

Geographic Scale
Big Picture of Internet of (Every)Things, data and services
Internet(s) of Data, Things and Services

- Internet of (Chatty) Things
- Internet of Everything
- Internet of People
- Internet of My Things
- Industrial Internet of Things (IIoT) : *Industry 4.0*
- Fog Computing
- Cyber-Physical Systems
- ...
What is Cloud Computing?

- On-demand computing
  - IaaS, PaaS, SaaS
  - Public, Private, Hybrid, Community, User-Centric, Souverain

- Advantages
  - Virtualization, TCO, Resilience, Elasticity, Energy efficiency, Big Data Analytics …

- Drawbacks
  - Confidentiality (Privacy, Industrial properties, …)
  - Souverainety
Cloud Services Models

- **Software as a Service**
  - Salesforce, Steam, ...

- **Platform as a Service**
  - Google App Engine, Amazon Hadoop ...

- **Infrastructure as a Service**
  - Amazon EC2, ...

- **Virtual/Physical Infrastructure (FaaS)**

- **IT cooling**

- **Smart Green Grid**
  - H₂
Cloud Computing:
UbiCloud, Cloud of Things, ...

• UbiCloud
  – Clouds with/for Ubi-terminals (smartphones, tablets, cars, IDS, ...)

• Cloud of Things (CoT)
  – Cloud for Things (data collection and long-term storage ...)
  – *Things are facilities in the FaaS*
Cloud of Things

Software as a Service (SaaS)
(on-demand access to any applications)

Platform as a Service (PaaS)
(on-demand platform of delivering your own application)

Infrastructure as a Service (IaaS)
(on-demand cpu/storage/nw infrastructures)

Virtual/Physical Infrastructure (FaaS)

IT cooling

Xively, Axeda, Eurotech

Deltadrone, Cloud Robotics

Smart Green Grid

D. Donsez, Intergiciels IoT
What is Big Data (Analytics)?
Big data is like teenage sex

• “Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it…”

• Dan Ariely, Professor at Duke University, TED speaker
The Data Deluge and the next IoT Data Deluge

1ZB = 10^{12} GB

IN 2010 THE DIGITAL UNIVERSE WAS
1.2 ZETTABYTES

IN A DECADE THE DIGITAL UNIVERSE WILL BE
35 ZETTABYTES

90% OF THE DIGITAL UNIVERSE IS UNSTRUCTURED

IN 2011 THE DIGITAL UNIVERSE IS
300 QUADRILLION FILES
The 4+1V of Big Data

**Volume**

- 40 Zettabytes (40 trillion gigabytes) of data will be created by 2020, an increase of 300 times from 2005.
- 2.5 Quintillion Bytes (2.3 trillion gigabytes) of data are created each day.
- 6 billion people have cell phones.
- World population: 7 billion.

**Velocity**

- Most companies in the U.S. have at least 100 Terabytes (100 billion gigabytes) of data stored.
- The New York Stock Exchange captures 1 TB of trade information during each trading session.
- Modern cars have close to 100 sensors that monitor items such as fuel level and tire pressure.

**Variety**

- As of 2011, the global size of data in healthcare was estimated to be 150 Exabytes (181 billion gigabytes).
- 4 billion+ hours of video are watched on YouTube each month.
- 30 billion pieces of content are shared on Facebook every month.
- 400 million tweets are sent per day by about 200 million monthly active users.

**Veracity**

- By 2015, 4.4 million IT jobs will be created globally to support big data, with 1.9 million in the United States.
- In one survey were unsure of how much of their data was inaccurate.

**Value**

- Poor data quality costs the US economy around $3.1 trillion a year.
- 27% of respondents don’t trust the information they use to make decisions.
- 1 in 3 business leaders don’t trust the information they use to make decisions.
- By 2020, it’s anticipated there will be 420 million wearable, wireless health monitors.
- Different forms of data:
  - 4 billion+ hours of video watched on YouTube each month.
  - 30 billion pieces of content shared on Facebook every month.
  - 400 million tweets sent per day by about 200 million monthly active users.

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, WEPEEC, QAS.
IoT Big Data 5V

Volume
Velocity
Variety
Veracity
Value
Big Data and IoT

• Ubilitycs : Ubiquitous big data analytics
  • Realtime prediction on the sensor data flows
  • For realtime decision (ie action)

• Mixin with other data sources
  • Corporate data
  • Open Data (gov, ...)
  • Crowd-sourced data
  • Social networks posts/tweets
  • ...

How analyze the Iot Data Deluge?

- Fastly (hours) and Very Fastly (milliseconds)
- For speeding and improving decision supports

→ Business Intelligence tools (OLTP, OLAP, …) can't!

Now computing models are available
- Massively distributed, on-demand, fault tolerant
- But
  - All old-fashioned statistical/prediction methods must be rethink
Computing Models for Big Data

- **Post-Processing Batch**
  - TBs / PBs of stored data

- **High-latency Decision Support**
  → Map Reduce
  - Hadoop, SciDB, Spark, Giraph, ...

- **Continuous Event Streaming**
  → Event Stream Processing
  - 100 MB/s of live data
  - Storm, S4, Samza, Millwheel, ...

  - **Discretized Stream Processing**
    - Map-Update
    - MUD8P

  - **Low-latency Decision Support**
    - Sensor, S4, Samza, Millwheel, ...

- **Continuous Event Streaming**
  - Storm, S4, Samza, Millwheel, ...

- **Discretized Stream Processing**
  - Map-Update
  - MUD8P
Event Stream Processing

- Massively distributed processing of continuous flows of events (sensors data, ...)
  - Low-latency (few millisecond after)
Example: Event Stream Processing

Trending Topics

- Apache Kafka
  - Tweets
    - Parser
    - Storage
    - Total Ranking
    - Intermediate Ranking
    - Rolling Count
  - group by word
  - group by ranking

Twitter Sentiment Analysis (SA)

- Apache Kafka
  - Tweets
    - Parser
    - Calculate Sentiment
    - Storage
    - Location Data
      - group by state

VolPSTREAM (VS)

- Apache Kafka
  - CDR
    - Parser
    - Variation Detector
      - ENCR
      - ECR
      - RCR
      - ECR24
      - CT24
      - Global ACD
      - URL
      - FoFiR
      - ACD
        - Scorer
        - Storage
  - group by caller, callee
  - group by caller
  - group by caller
  - group by caller

From Maycon Bordin's ms thesis
Lambda Architecture

Nathan Marz (Twitter, Backtype)

• Combine Low and High Latency BD stacks

Batch layer can compute the analytics model of the Speed Layer

http://jameskinley.tumblr.com/post/37398560534/the-lambda-architecture-principles-for-architecting
Ubilitytics
Ubiquitous Big Data Analytics

• Motivation
  – PaaS for « Ubilitytics »
    • Autonomic: scalability, fault tolerance
  – End-to-End
    • From sensors, gateways and lambda architecture (cloud)
  – « Simple as Possible »
    • → for IoT SMEs & their IoT data scientists

• Problem
  – Huge variety of needs
  – Huge variety of technologies
New trendy Job: IoT Data Scientist

Gartner says big data creates big jobs: 4.4 million IT jobs globally to support big data by 2015.
http://www.gartner.com/newsroom/id/2207915

The U.S. could face a shortage by 2018 of 140,000 to 190,000 people with "deep analytical talent" and of 1.5 million people capable of analyzing data in ways that enable business decisions. (McKinsey & Co)

Big Data industry is worth more than $100 billion growing at almost 10% a year (roughly twice as fast as the software business)

http://nirvacana.com/thoughts/becoming-a-data-scientist/

How can this guy deal with this deluge of technologies? How to make this guy productive?
Who is able to develop/ deploy Ublilytics infrastructures?

IoT Data Scientist
Towards "Ubilytics" PaaSs

Message Broker or MaaS or PSaaS
Mosquito, RabbitMQ, …
Protocols: MQTT, AMQP, STOMP, XMPP, CoAP, WebRTC, Motwin ...

In Elastic Cloud
NoSQL Store
MongoDB, Cassandra, HDFS

MapReduce
Hadoop

Realtime ESP
Storm, Samza, S4, Spark Streaming, …
topologies

Monitoring
Placement
(static,dynamic)

Mashup,
Reporting,
Dashboard, …
(history charts, …)

Storing
aggregates

Predictions
Trends, …

Deployment & (Re)Configuration (Roboconf)

Send selected sensors measurements

SmartPhone
@ Car, City, …
Embedded boards smartphones by millions

Sensors data messages:
energy Consumption, temperature, images, …

M2M Gateway
@ Home, Office, City, Warehouse
(OpenHAB, IoTSys)
Ubilitytics Example
Energy Consumption Forecast

Domain : Smart Grid
  • 2125 individual smart plugs in 40 houses
  • measuring and sending instant load (W) and cumulative load (kW)

Sensor Dataset*:
  • Events contain instant load (W) and cumulative load (kW)
  • 130 millions events/day (on one month)
  • 3 GB/day

Challenge* : Forecasts loads at 1min, 5min, 15min, 60min and 120min

Goal : anticipate electricity demand
  for adjusting the production (ie. save energy and avoid blackout)

* DEBS GC 2014 http://www.cse.iitb.ac.in/debs2014/?page_id=42#
Ubilitytics Example

Energy Consumption Forecast

MQTT Broker + Storm Topology + Cassandra DB on a Azure VM cluster

instant load (W) and cumulative load (kW)
Smart Campus (ie Small Smart City)

OpenHAB, Galileo, MQTT, Storm, Azure, ...
Smart Campus
L’École Polytechnique de l'université Grenoble-I, plus communément appelée Polytech’Grenoble, est l’école d’ingénieurs de l'Université Joseph Fourier. Elle appartient au réseau Polytech qui regroupe treize écoles d’ingénieurs universitaires dans toute la France. L’école propose différentes formations couvrant ainsi l’ensemble des domaines technologiques modernes. [Site web de Polytech’Grenoble.]
# SmartCampus

![Map of Polytech Grenoble](image)

## Description
Bureau de Didier Donsez

Capteurs:
- humidité : 18 %

## Evenement

<table>
<thead>
<tr>
<th>Avis</th>
</tr>
</thead>
</table>

---

D. Donsez, CIRUS, EclipseCon 2014
Smart Campus

**Description**

Horaires théoriques :

- **Direction Seyssins Le Prisme** :
  - 1 min - 6 min - 11 min - 16 min - 21 min -
- **Direction Gières** :
  - 2 min - 7 min - 12 min - 17 min - 22 min -

**Horaire**

**Avis**
Q & A