On the Robustness of Collecting Data from Distributed Locations

Ciprian Jichici

ciprian.jichici@genisoft.eu

Teaching Assistant, Faculty of Mathematics and Computer Science

General Manager, Genisoft

Microsoft Regional Director, Romania

Workshop of e-Infrastructure for Public Sector

Timisoara, 22nd of March, 2013

Agenda

The Problem

The Challenges

The Solutions

The Problem (theory)

A large organization O has N distributed, data-generating locations (DL1, DL2, .. DL_N). Data is generated continuously, at highly variable rates, by different software and/or hardware components.

The organization has M data processing and analysis locations (CL1, CL2, .. CL_N) which are used to transform raw data into information. The information is used to support decisions that have significant impact on the evolution of O.

It is safe to assume that M << N.

The Problem (theory) - continued

There are communication channels between DLi and CLj. Data transfer speeds are variable and availability is less than 100%. For a given timespan of 24 hours, it is safe to assume for each DLi that it will have at least a total of 1 hour of communication channel availability with at least one CLj (not necessarily the same).

BASIC GOAL

Provide a data transfer platform that enables O to consolidate data in each CLj with a latency as close as possible to zero (near real-time).

Provide a data transfer platform that enables O to distribute changes to each DLi with a latency as close as possible to zero (near real-time).

The Problem (theory) - continued

There are pieces of data labeled as “critical”. Once a piece of “critical” data is generated at DLi, it needs to get as quickly as possible to all CLj’s as well as to all DLk’s (where $k \neq i$).

ADVANCED GOAL

Provide a data transfer platform that enables O to handle “critical” data with a latency as close as possible to zero (near real-time).

The Problem (real life)

An easy one:

A public sector agency has offices in 200 locations across the country. Each location generates new data on a daily basis. The agency needs to have all the new data collected in its central location no later than 7:00 AM the next day.

There are pieces of data that need to reach the central location and all offices in a maximum of 2 hours.

The Problem (real life)

A slightly more complex one:

A public sector agency has 15 subordinated agencies, each of them having offices in 50 locations across the country. Each location generates new data on a daily basis. The agency needs to have all the new data from all its subordinated agencies collected in its central location no later than 7:00 AM the next day.

There are pieces of data that need to reach the central location and all offices in a maximum of 2 hours.

The Problem (real life)

A complex one:

A retail organization has 1500 sales locations distributed across an entire continent. It also has 500 mobile sales agents which perform daily routes and change data from mobile devices with unpredictable connectivity.

The organization has 5 data processing locations, out of which 3 are located in public clouds.

The organization operates with a margin of roughly 15% which creates a critical dependency between profit and accurate data.

The Challenges

Architectural patterns

Delta detection

Communication patterns

EOIO delivery (exactly once, in order delivery)

Channel disruption handling

Instrumentation

Security

Latency minimization

robust

Definition

ro·bust [ro búst]

ADJECTIVE

1. strong and healthy: strong, healthy, and hardy in constitution

2. strongly constructed: built, constructed, or designed to be sturdy, durable, or hard-wearing

3. needing physical strength: involving or requiring great physical strength and stamina

4. determined: characterized by firmness and determination and a refusal to make concessions

5. straightforward: showing clear thought and common sense

6. blunt or crude: rough and direct or crude

7. full-flavored: rich, strong-tasting, and full-bodied

8. comput **capable of recovery:** describes a computer program or system that is able to recover from unexpected conditions during operation

robust

Definition

ro·bust [ro búst]

ADJECTIVE

1. strong and healthy: strong, healthy, and hardy in constitution

2. strongly constructed: built, constructed, or designed to be sturdy, durable, or hard-wearing

3. needing physical strength: involving or requiring great physical strength and stamina

4. determined: characterized by firmness and determination and a refusal to make concessions

5. straightforward: showing clear thought and common sense

6. blunt or crude: rough and direct or crude

7. full-flavored: rich, strong-tasting, and full-bodied

8. comput **capable of recovery:** describes a computer program or system that is able to recover from unexpected conditions during operation

Architectural Patterns

The data collection/transfer process is a complex task -> needs to be locally coordinated by an agent

The agent must talk to the outer world -> best approach is to model it as a service

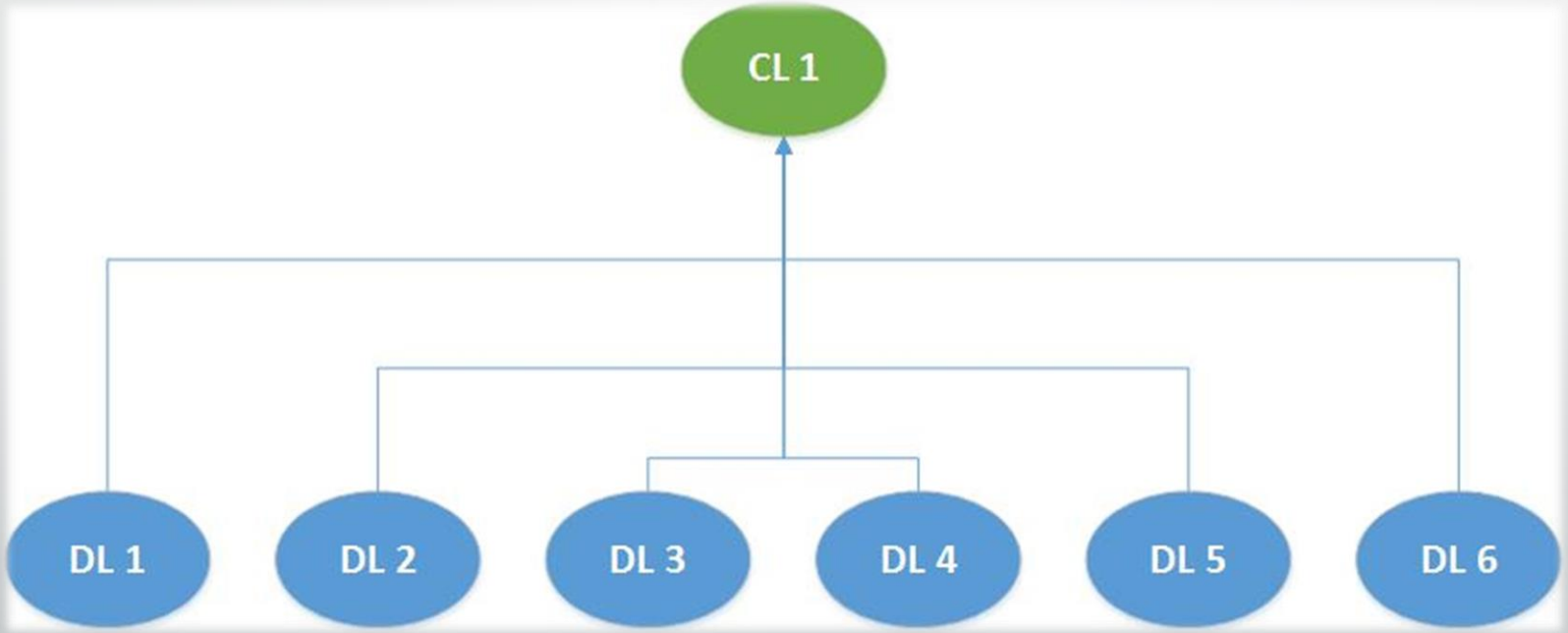
Open discussion:

What is the degree of autonomy for each agent?

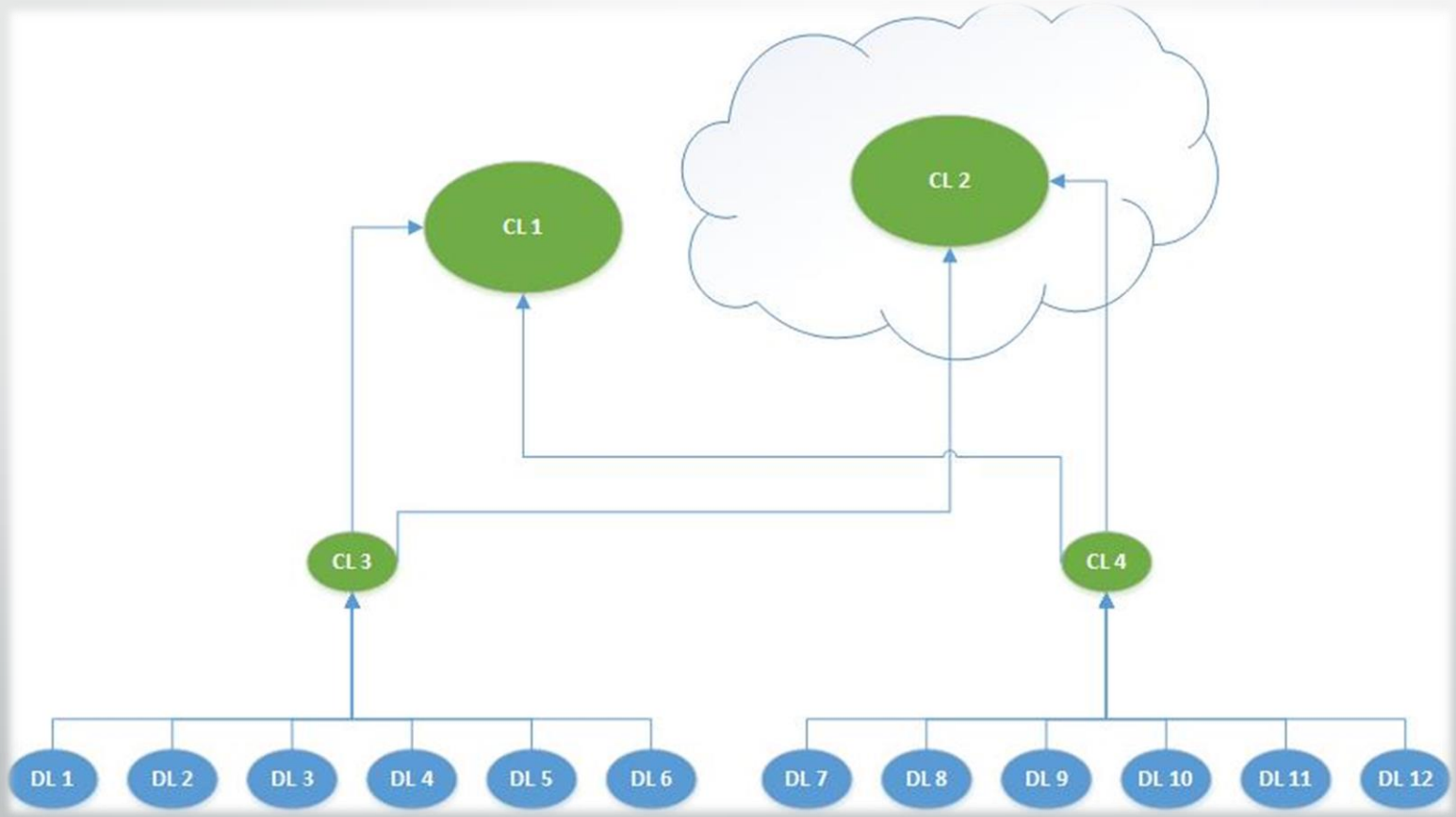
Should each agent “talk” to its closest neighbors only? Should there be a hierarchy between agents?

Should there be a “central authority” for agents?

Architectural Patterns – A “Simple” Approach



Architectural Patterns – A Layered Approach



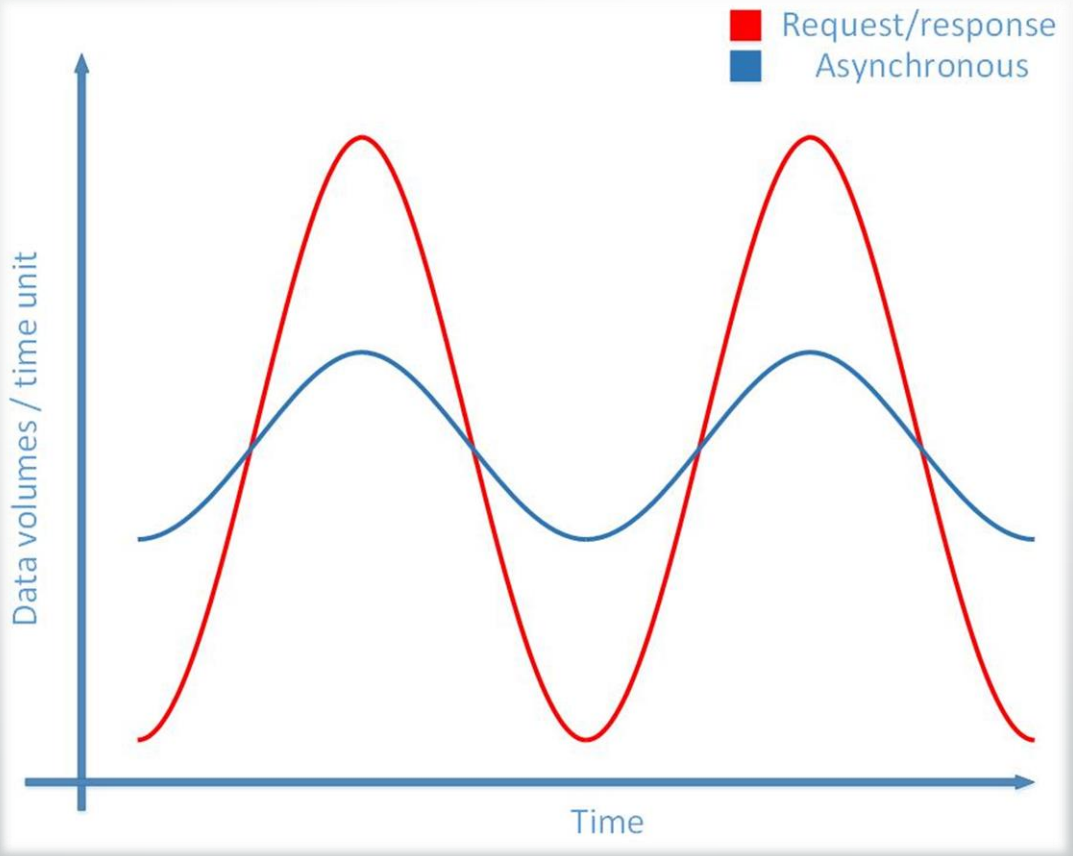
Delta Detection in Data Sources

Structured vs. Unstructured data

Platform level detection vs. Application level detection

The “quantic” effect of detection

Communication Patterns



EOIO Delivery

The fundamental pillar of robustness

Difficult to achieve

Involves several “disciplines”:

- Data packaging and sequencing
- Missing package detection
- Duplicate package detection
- Routing

Channel Disruption

Sometimes it is difficult to detect a disrupted channel

Open discussion: What is the smallest ε that defines disruption?

Data redundancy is a must to handle channel disruption?

What is the link between channel disruption and agent availability?

Instrumentation

What happens inside the agent's world?

What happens inside an agent's body?

The "quantic" effect of instrumentation

A formal definition of a "healthy" population of data collection and transfer agents

What is the best technique to detect "disease"?

What is the best technique to detect an "epidemic"?

What is the condition to activate "resurrection"?

Security

Intra-agent security

Inter-agent security

Latency Minimization

Ultimate goal: bring latency as close as possible to zero (near-real time)

In practice, a good result is to bring it down to minutes

The “cost” of latency (theory vs. practice)

The Solutions

Architecture: semi-autonomous agents, coordinated by a central authority

Communication pattern: asynchronous

Delta detection: non intrusive, synchronous or asynchronous

Logical layering:

- Physical transport
- Logical transport
- Application

Future work

Intelligent and adaptive agents

Smart routing

Dynamic agent and channel disruption detection and handling

TaaS – Transport as a Service