Scalability in RDF Stream Processing Systems

Alejandro Llaves and
Oscar Corcho

Ontology Engineering Group
Universidad Politécnica de Madrid
Madrid, Spain
allaves@fi.upm.es

May 31 2015

What we do to improve **Scalability in *our* RDF Stream Processing System**

<u>Alejandro Llaves</u> and Oscar Corcho

Ontology Engineering Group
Universidad Politécnica de Madrid
Madrid, Spain
allaves@fi.upm.es

May 31 2015

Outline

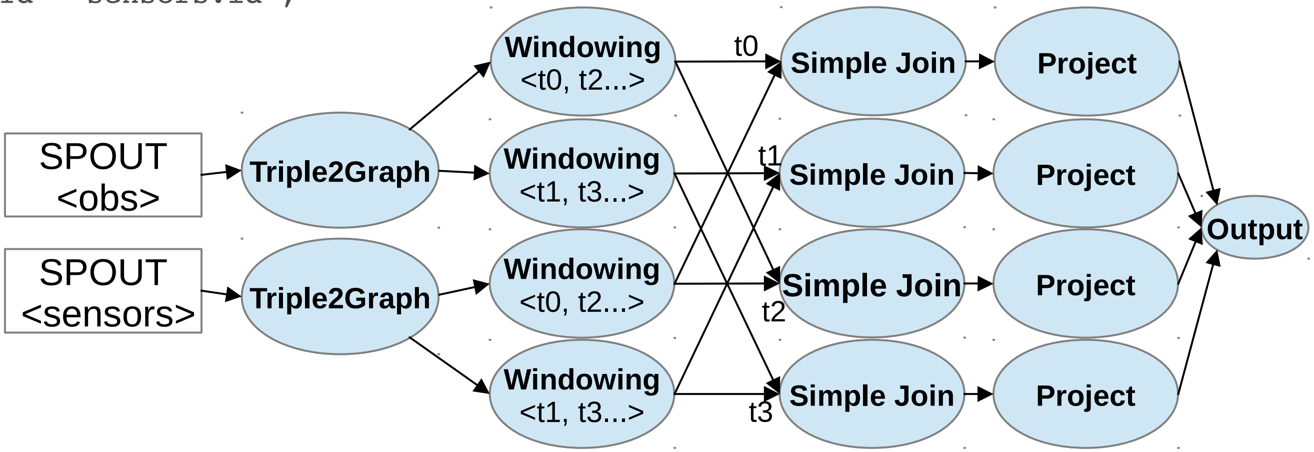
- Towards efficient processing of RDF data streams
- Architecture overview
- Parallelizing the pre-processing of sensor data streams
- Example of use: CSIRO's Sensor Cloud
- Discussion and future work

Towards efficient processing of RDF data streams

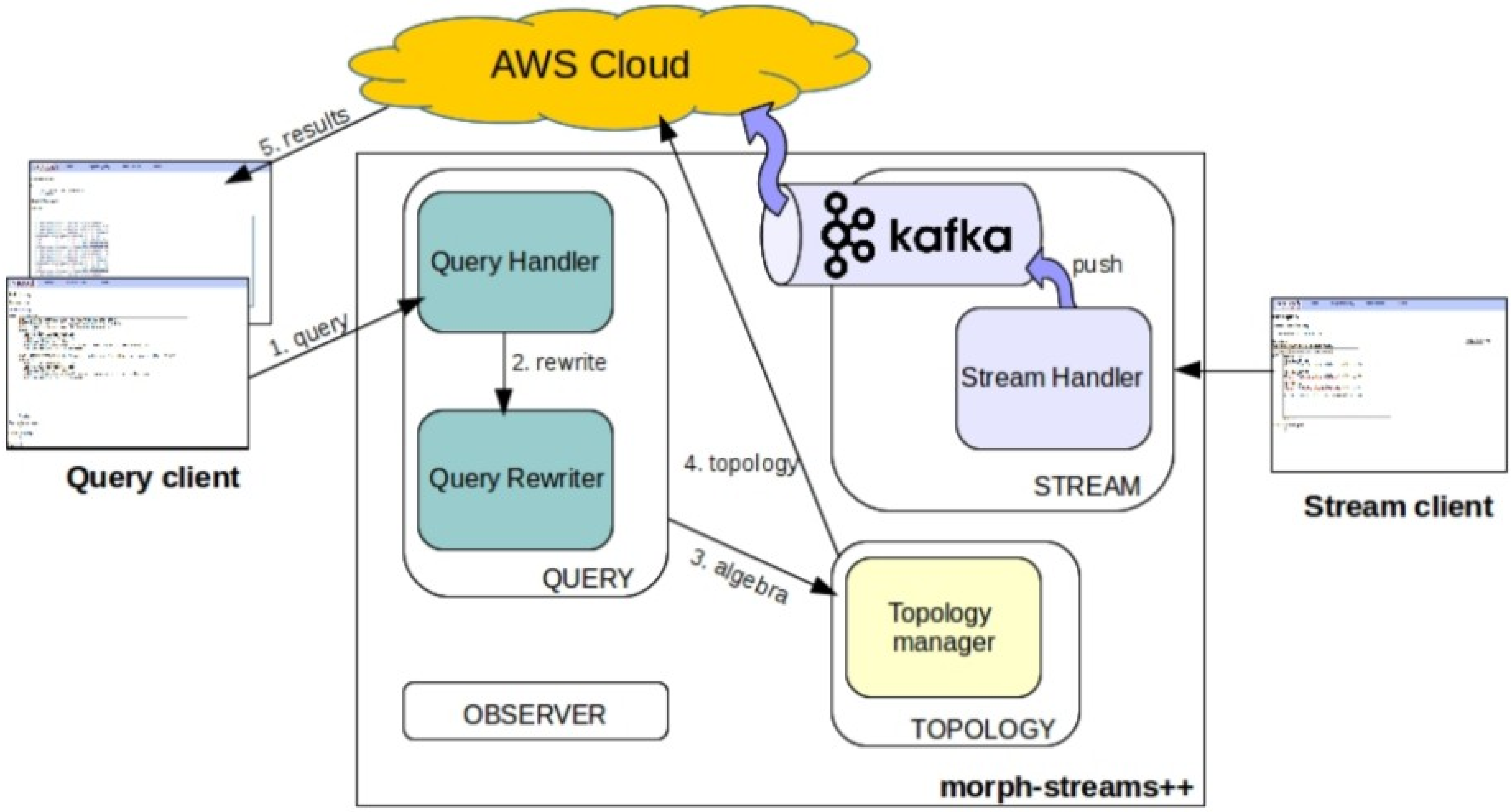
- **Goal:** to develop a stream processing engine capable of adapting to variable conditions, such as changing rates of input data, failure of processing nodes, or distribution of workload, while serving complex continuous queries.
- Example of query execution parallelization (OrdRing 2014)

Storm topology example (4 nodes)

```
SELECT ?obs.value ?sensors.location
FROM NAMED STREAM <obs> [60 SEC TO NOW]
FROM NAMED STREAM <sensors> [60 SEC TO NOW]
WHERE obs.sensorId = sensors.id ;
```



morph-streams++ architecture



Parallelizing the pre-processing of sensor data streams

Methodology

1. Transform data input into field-named tuples
2. Add semantic annotations (if needed)
3. Publish tuples to multiple channels
4. Convert tuples to RDF on (query) demand

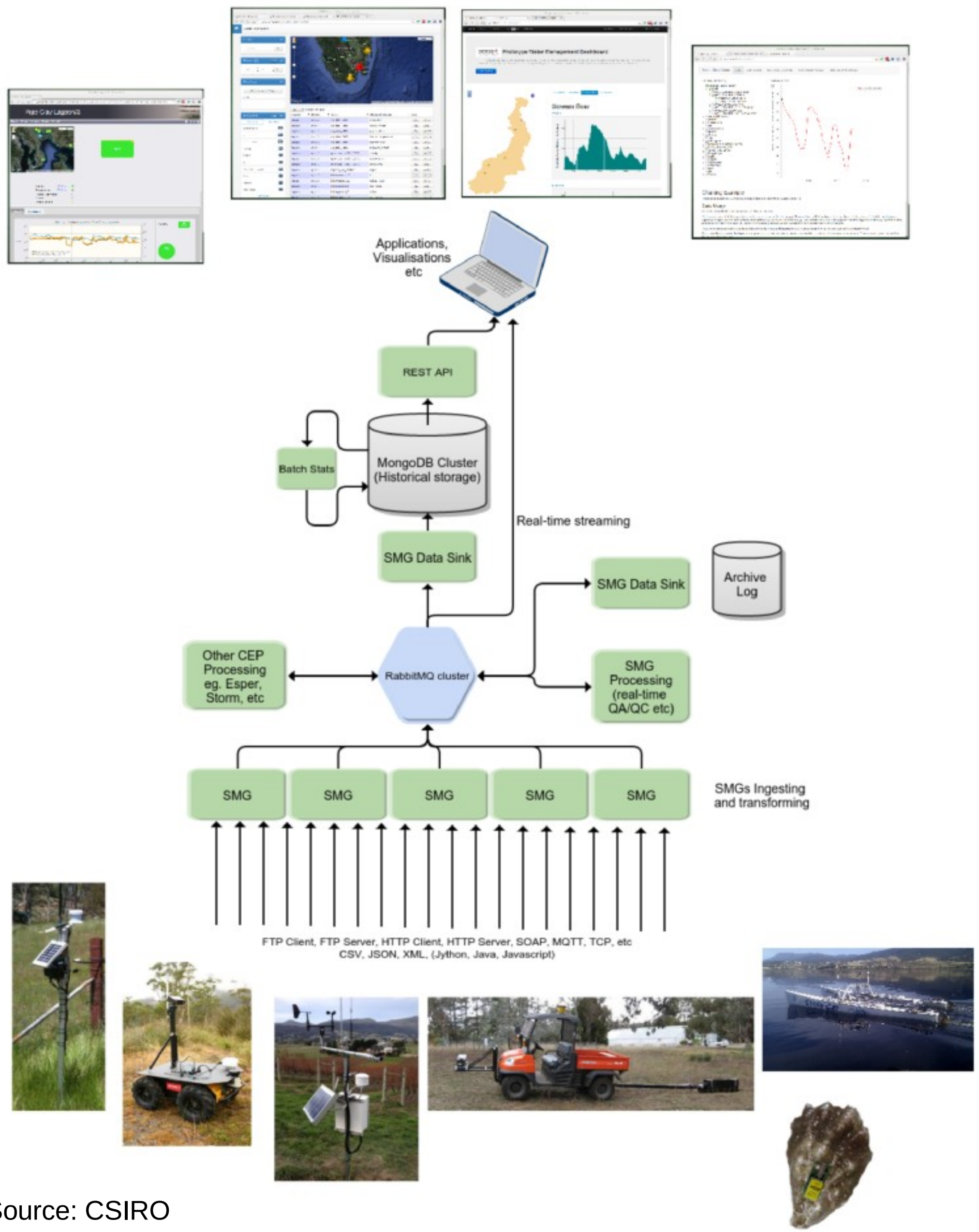
Focus

- Storm topologies
- Environmental sensor observations
- Using Semantic Sensor Network (SSN) ontology

Example of use: CSIRO's Sensor Cloud (1/3)

Sensor Cloud

- Viticulture, water management, weather monitoring, oyster farming...
- RESTful API – JSON
- Network → Platform → Sensor → Phenomenon → Observation
- Lack of semantic descriptions, e.g. *rain_trace* vs *Rain.*
- Multiple HTTP requests to query various streams.



Source: CSIRO

Example of use: CSIRO's Sensor Cloud (2/3)

1. Sensor Cloud messages to field-named tuples

phenomenon

```
<sample time="2015-05-28T16:30" value="48" sensor="bom_gov_au.94961.air.rel_hum"/>
```

sampling time

network platform sensor

```
["2015-05-28T16:32", "2015-05-28T16:30", "48", "bom_gov_au", "94961", "air", "rel_hum",
"-43.3167", "147.0075"]
```

system time

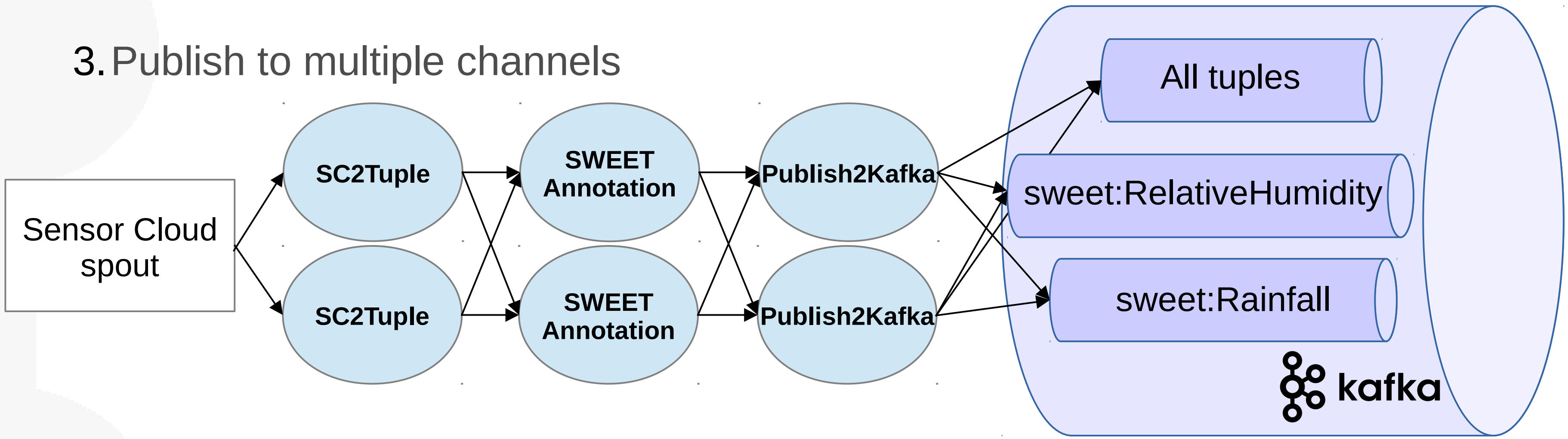
latitude longitude

2. SWEET annotations for phenomena

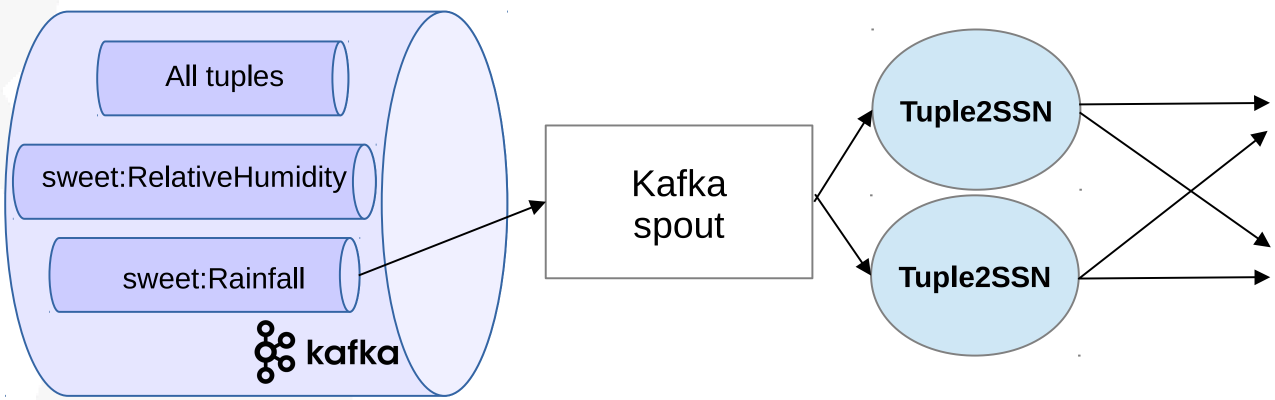
Sensor Cloud phenomena	SWEET annotations
rain_trace, Rain, rainfall-per-hour	Rainfall
air_temp, temperature_deg_c, temperature, average-air-temperature	Temperature
wind_dir, average-wind-direction	Direction
wind_spd_kmh, average-wind-speed	WindSpeed
rel_hum, average-relative-humidity	RelativeHumidity
dewpt	DewPoint
Evap	Evaporation

Example of use: CSIRO's Sensor Cloud (3/3)

3. Publish to multiple channels



4. Convert tuples to SSN model on (query) demand



to be continued...

```
SELECT SUM(?obsValue) ?location
FROM NAMED STREAM <http://sensorcloud.linkeddata.es/obs/> [NOW-24 H]
WHERE {
  ?obs a ssn:Observation.
  ?obs ssn:observedBy ?sensor.
  ?obs ssn:observedProperty sweet:Rainfall.
  ?obs ssn:observationResult ?result.
  ?result ssn:hasValue ?obsValue.
  ?obs geosparql:hasGeometry ?geom.
  ?geom geosparql:asWKT ?location.
}
GROUP BY ?sensor
HAVING(SUM(?obsValue) >= 50)
```

Query: heavy rainfall events

Conclusion

- Division of work into simple tasks.
- Parallelize any parallelizable task.
- Delay RDF generation and convert on demand.

Future work

- Evaluation and benchmarking.
- SSN mapping interface.
- Topology package: executing distributed queries (Storm).
- theObserver (theO) package: monitoring scalability metrics for adaptive query processing.

Thanks!

The presented research has has been funded by Ministerio de Economía y Competitividad (Spain) under the project ”4V: Volumen, Velocidad, Variedad y Validez en la Gestión Innovadora de Datos” (TIN2013-46238-C4-2-R), by the EU Marie Curie IRSES project SemData (612551), and supported by an AWS in Education Research Grant award.

Alejandro Llaves
allaves@fi.upm.es