Strider^{R}: Massive and distributed RDF Graph Stream Reasoning

Xiangnan REN, Olivier CURÉ, Hubert Naacke, **Jérémy LHEZ**, Ke Li

LIGM - LIP6 CNRS, FRANCE

OUTLINE

Agenda of the Presentation

- CONTEXT
- ARCHITECTURE OVERVIEW
- ENCODING - REWRITING
- CONCLUSION

CONTEXT

Importance of stream reasoning

DATA STREAMS EVERYWHERE

WAVES PROJECT

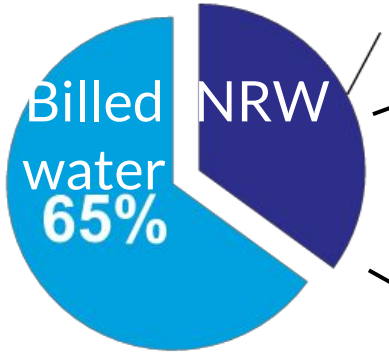
- **Smart water network management**
  - Data streams from sensors
  - Filtering errors in measures
  - Identify sources in external events

- **Main partner: Suez**
  - 650 collaborators in Europe
  - 73 billion $US in R&D

- **French project**
  - http://www.waves-rsp.org/

Why water management ?

Water SUPPLIED to the network – **Water BILLED to *customers*** = **NON-REVENUE WATER (NRW)**



48.6 billion m^3/year

Loss of US$**14 billion**/year

=

x2 the annual domestic water consumption of the USA

WAVES PROJECT

- **Objectives:**
  - Robust real time engine, modular, flexible, intelligent
  - Distribution

- **RDF representation**
  - Integration of data/knowledge from different sources
  - Reasoning capabilities

- **Other applications:**
  - Banking/payments, climate, energy, power consumption, etc

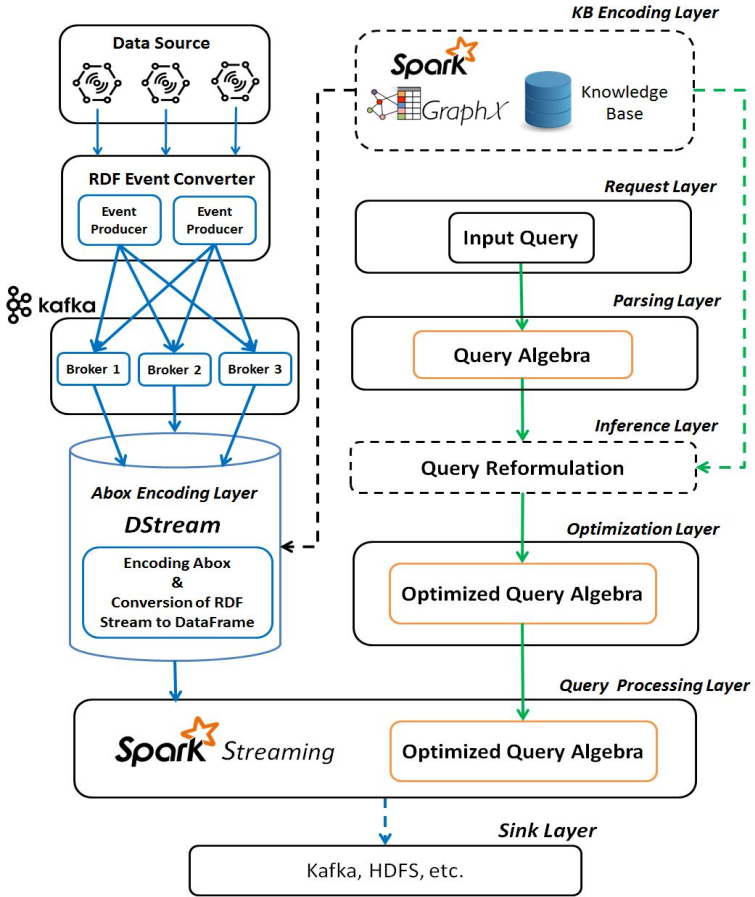
ISSUES OF STREAM PROCESSING

- **Solutions specific to the reasoning tasks**
  - **Materialization:** huge amounts of data
  - **Query rewriting:** execution time
- **Lack of performance for heavy data load**
- **Compression efficiency**
  - No distribution
  - Decompression process

ARCHITECTURE

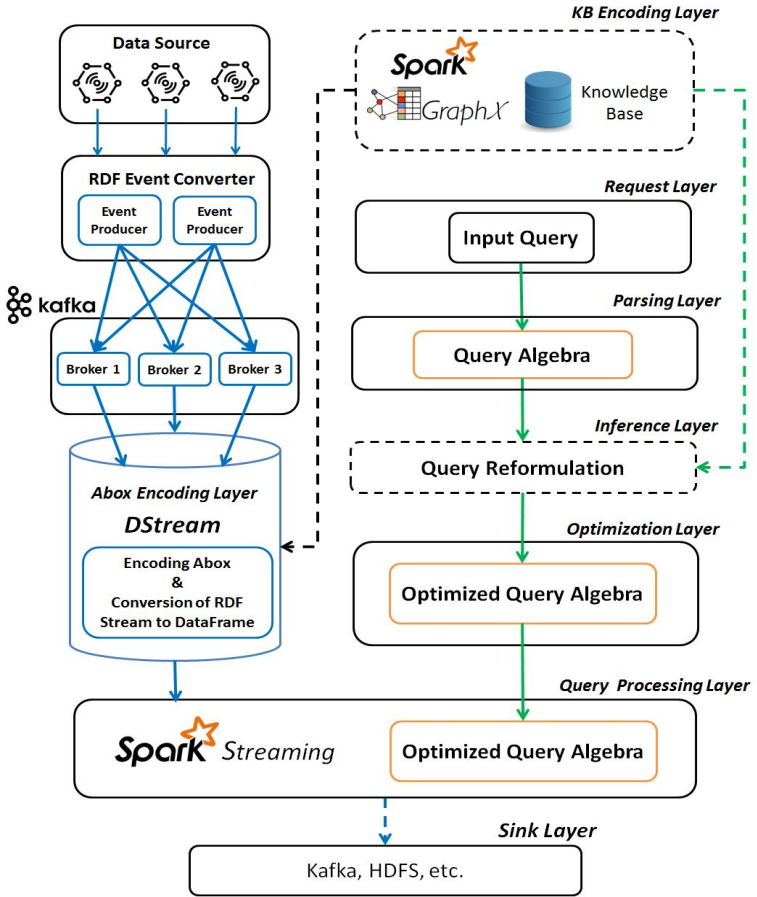
Strider organization

STRIDER ARCHITECTURE



ISWC 2017, p.559-576
VLDB 2017, vol.10, nb.12

STRIDER ARCHITECTURE



- **Data flow management:**
  - Apache Kafka
  - Data streams partitioned
- **Computing core**
  - Query processing
  - Parallel query execution
- **Encoding of the data**
  - Static knowledge base: offline encoding (Abox + Tbox)
  - Dynamic data: encoding on the fly (Abox)

ISWC 2017, p.559-576
VLDB 2017, vol.10, nb.12

ENCODING - REWRITING

An encoding form conserving the hierarchy
used to rewrite the queries

LITEMAT

➢ **Principle:**

- Binary structure conserving the semantics of the ontology
- Each identifier is prefixed by its parent's
- Conversion as integer identifiers
- Supports RDFS

➢ **Advantages:**

- Compression maintaining hierarchy
  - ✓ No need for ontology at query runtime
  - ✓ Execution performance
- Easy query rewriting

IEEE Big Data 215, p.1823-1830
ESWC 2017, p.79-93

LITEMAT EXAMPLE

Concepts	Compression
owl:Thing → DUL:PhysicalObject → ssn:Sensor → ssn:SensingDevice → dbo:Engine → DUL:SocialObject	1

LITEMAT EXAMPLE

Concepts	Compression
owl:Thing	1
DUL:PhysicalObject	101
ssn:Sensor	
ssn:SensingDevice	
dbo:Engine	
DUL:SocialObject	110

LITEMAT EXAMPLE

Concepts	Compression
owl:Thing	1
→ DUL:PhysicalObject	101
→ → ssn:Sensor	10101
→ → → ssn:SensingDevice	
→ → dbo:Engine	10110
→ DUL:SocialObject	110

LITEMAT EXAMPLE

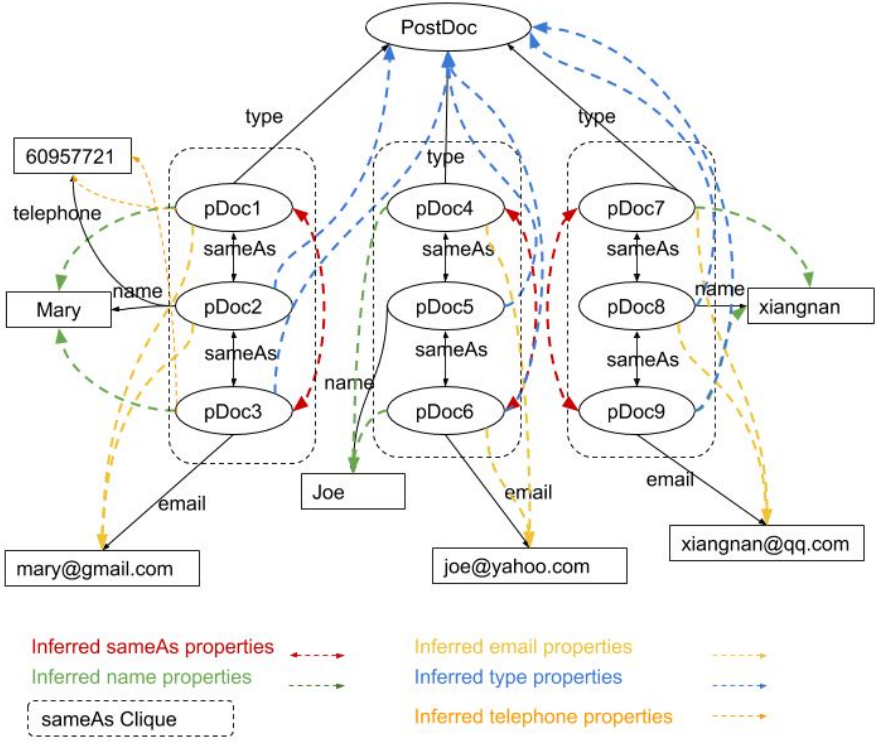
Concepts	Compression
owl:Thing	1
DUL:PhysicalObject	101
ssn:Sensor	10101
ssn:SensingDevice	101011
dbo:Engine	10110
DUL:SocialObject	110

LITEMAT EXAMPLE

Concepts	Compression
owl:Thing	100000 = 32
DUL:PhysicalObject	101000 = 40
ssn:Sensor	101010 = 42
ssn:SensingDevice	101011 = 43
dbo:Engine	101100 = 44
DUL:SocialObject	110000 = 48

- Identify the subclasses specific to a concept in a specific interval
  - e.g. [40, 48[= all physical objects

SAMEAS REPRESENTATION



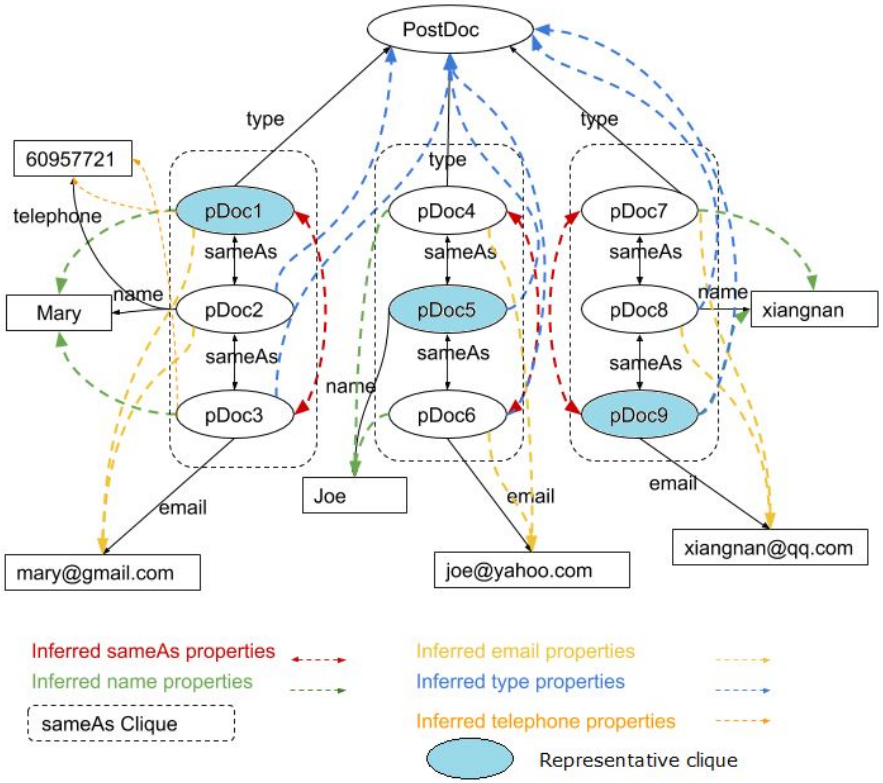
Limitation of queries

```
SELECT ?e ?n
WHERE {
        ?x rdf:type pDoc1
        ?x email ?e          ← pDoc3
        ?x name ?n           ← pDoc2
}
```

Inference of properties

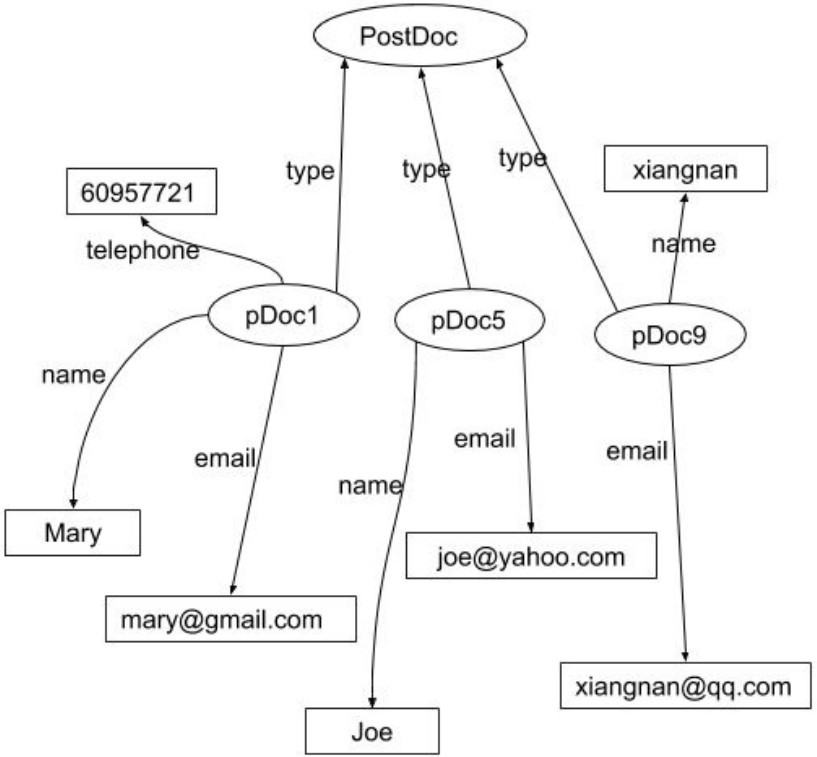
- Massive graph
- Complex manipulation (encoding, update…)

SAMEAS REPRESENTATION



- **One representative clique selected**
  - Properties share the same identifier
  - The other identifier is a reference
- **Advantages**
  - More compact graphs
  - Use of the dictionary for transformation
  - Update of sameas values has no performance impact

SAMEAS REPRESENTATION



(virtual representation)

- **No more query problems**
  - Encoding using the identifier clique's encoding

- **Dictionary holding the sameAs identifiers**

id. value	concept value	concept
31	31	pDoc1
31	32	pDoc2
31	33	pDoc3

PARTIAL ENCODING

Stream:

_:x1 id “Q250”
_:x1 date 30/03/2017
_:x1 pressureMeasure _:x2
_:x2 value 4.5

LiteMat dictionnary:

id	=	40
hLocation	=	54
pMeasure	=	56

Partial encoding:

e1 40 i13 — Encoded individual
e1 date 30/03/2017 — Encoded property
e1 56 m1 — Encoded concept
m1 value 4.5

➢ **Some of the identifiers are not present in the static knowledge base**

QUERY REWRITING

- **Query reformulation:**
  - x SELECT ?x WHERE { ?x rdf:type DUL:PhysicalObject . }
  - ✓ SELECT ?x WHERE { ?x rdf:type ?v .
    FILTER (?v >= 40 && ?v < 48) }

- **Usage in WAVES:**
  - Encoding of the static knowledge base and the queries
  - Partial encoding of streams

- **Rewriting of classes, properties and sameAs**
  - Using the identifiers from the static knowledge base

CONCLUSION

CONCLUSION

- **Waves project**
  - RDF Stream Processing engine
  - Use case: drinkable water network management

- **Strider**
  - Distributed RDF graph stream with reasoning
  - Support of RDFS and sameAs

- **LiteMat**
  - Compression with identifiers
  - Entity identifiers represent the ontology semantics

FUTURE WORKS

- **Increase the expressivity of supported ontologies**
  - Transitive properties, inverseOf

- **Improve partitioning**
  - of the dictionaries
  - Improvement of FILTER on distributed streams

THANK YOU