WAVES

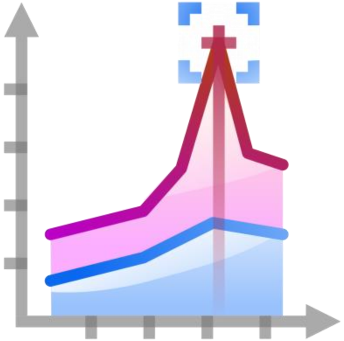
Big Data Platform for Real-time Semantic Stream Management

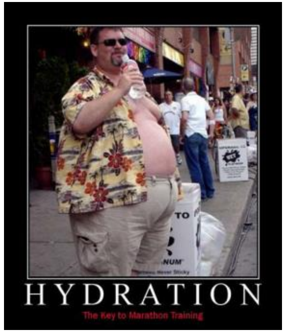
Badre BELABBESS

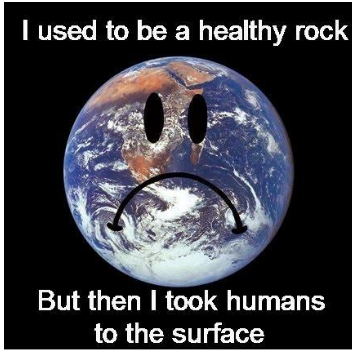
LIGM, Marne-La-Vallée, FRANCE
ATOS Bezons, FRANCE

A GOOD OLD STORY

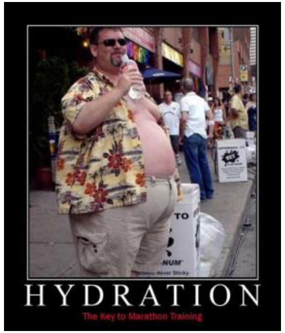
WAVES
ATOS SE

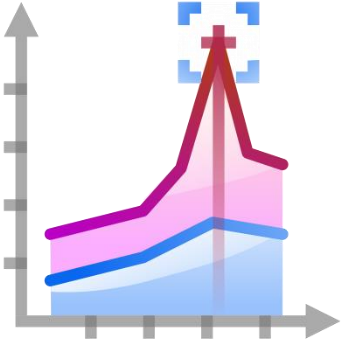


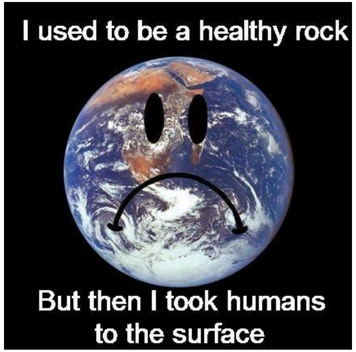




A GOOD OLD STORY







INTRODUCTION

- **Presenter:** Badre BELABBESS, PHD candidate
- **Research sites:**
  - *Atos SE:* Large european IT Company, Bezons, France
  - *LIGM*: Ponts ParisTech, UPEM, CNRS (UMR 8049), ESIEE Paris
- **Main research topics:** Big Data frameworks, real-time stream processing, system architecture
- **WAVES project:**
  - 3 year research project funded by the French government
  - Several partners: Industrial & Academic
  - *Distributed Open source platform intended for the new forms of massive semantic data streams processing.*

OUTLINE

Agenda of WAVES Presentation

ABOUT WAVES

Massive Semantic Streams empowering Innovative Big Data Platform

WAVES IN A NUTSHELL

- **Main aspects:**
  - Real-Time processing
  - RDF data streams/Sparql queries
  - Reasoning Capabilities/Inferences
- **Objectives:**
  - Robust RSP engine
  - Modularity and flexibility
  - Distribution – Industrial Cluster
- **Applications:**
  - **Potentially:** Banking/payments, climate, energy, power consumption, etc
  - **Currently:** Water Network Management

WHY WAVES ?

- **C-SPARQL**
  - high input loads → Precision/Recall decrease
  - Not designed to be distributed

M. Kolchin, P. Wetz, E. Kiesling, and A. M. Tjoa. **Yabench: A comprehensive framework for RDF stream processor correctness and performance assessment.** In Web Engineering - 16th International Conference, ICWE, Lugano, Switzerland, June 6-9, 2016.

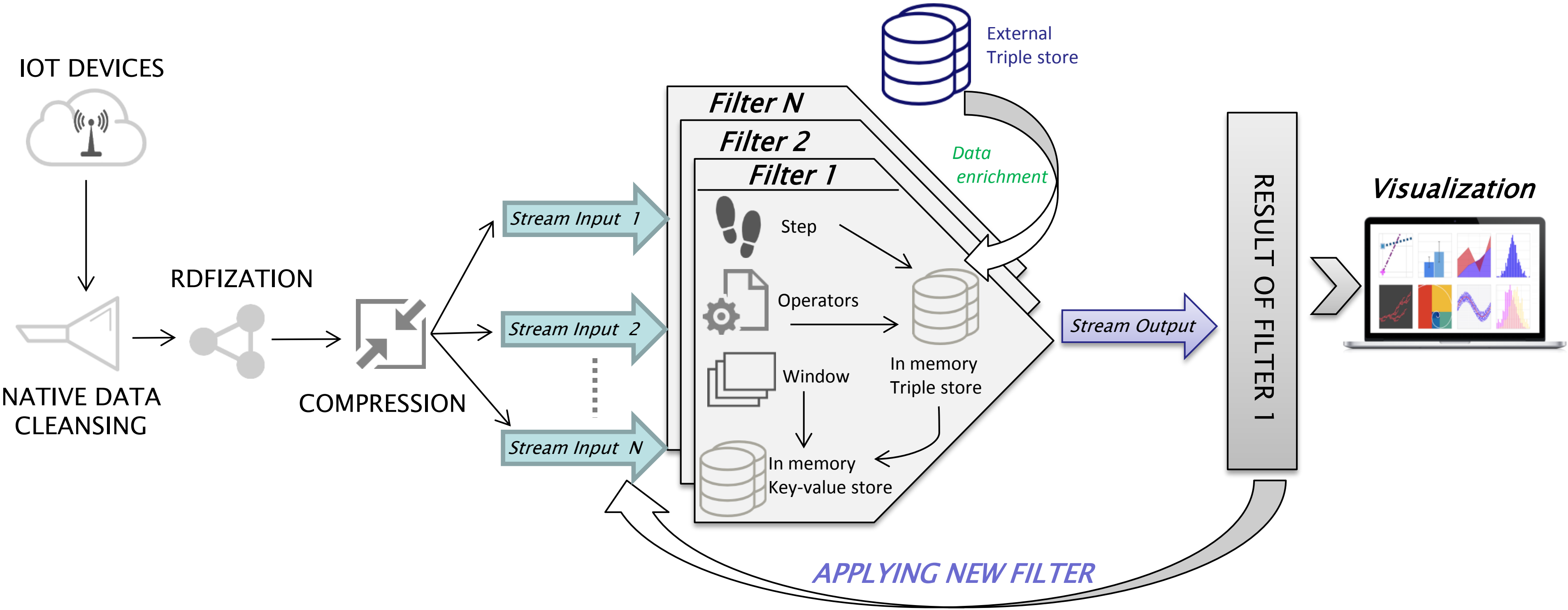
- **CQELS-Cloud**
  - Early stage/not open-source
  - Impossible to define specific queries/input data/parameters

→ Need to create a new RSP engine **industry ready** with **high precision/recall**, ability to **parallelize processing** and **open-source.**

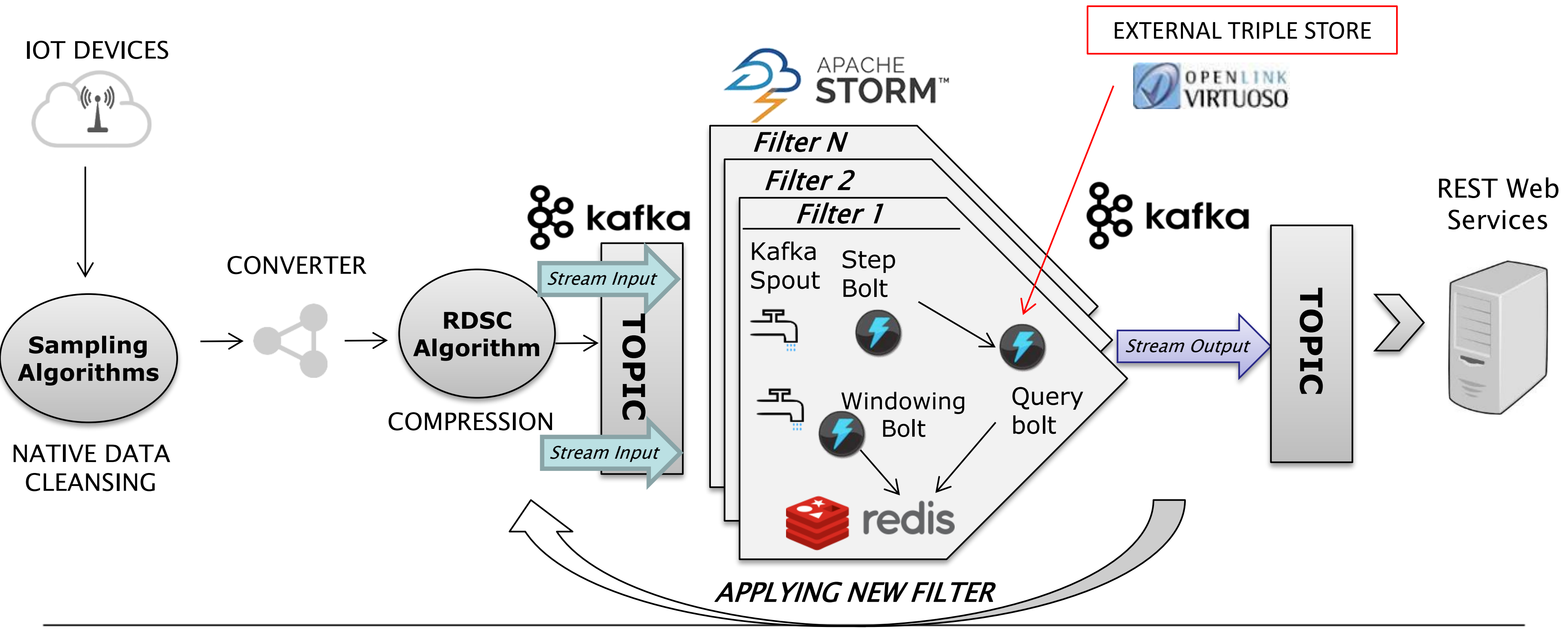
OVERALL ARCHITECTURE

Combining Big Data and Semantic Web technologies

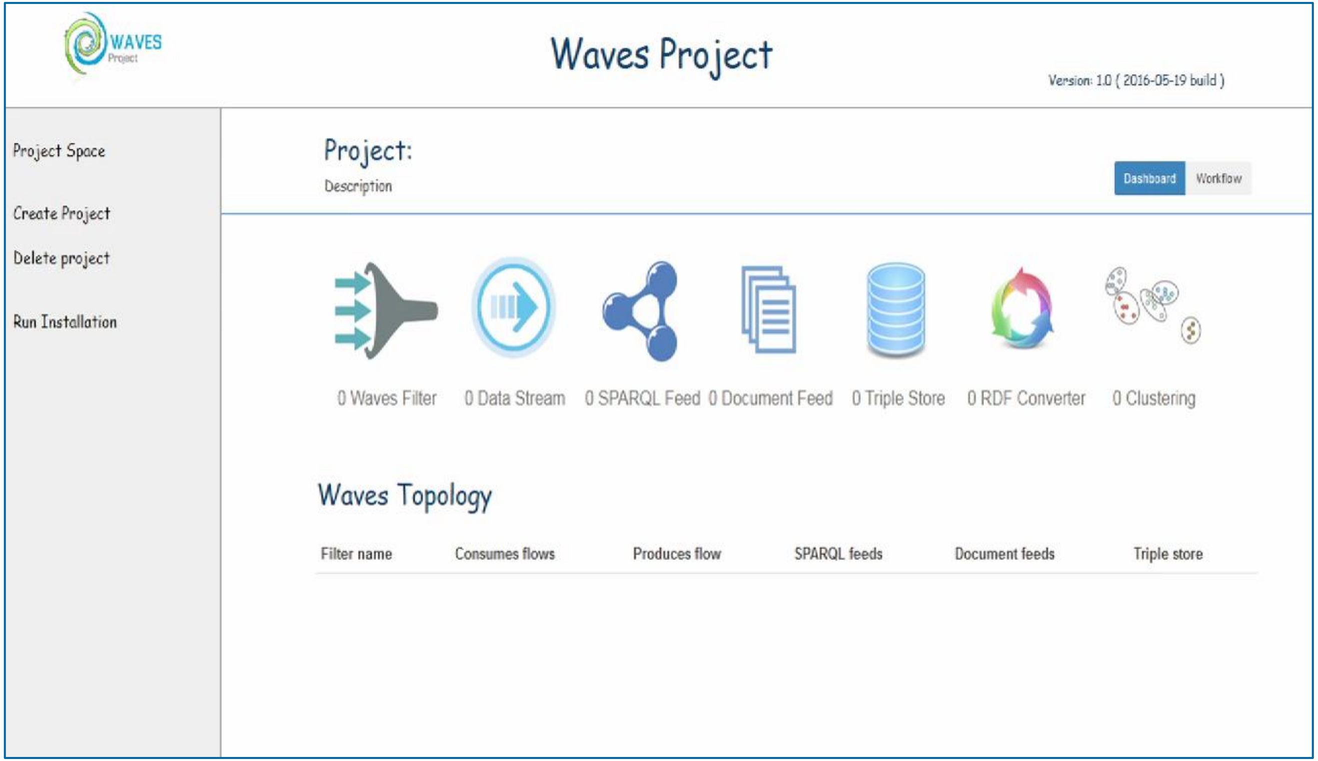
LOGICAL ARCHITECTURE



IMPLEMENTATION ARCHITECTURE



How to configure WAVES ?



COMPRESSION TECHNIQUE

Reducing the data size and exposing the results.

RDSZ ALGORITHM

- **Research paper:**

Fernandez Arias J., Sanchez L. ,Fuentes-Lorenzo D., Corcho, O.**RDSZ: An Approach for Lossless RDF Stream Compression**. In he Seman tic Web: Trends and Challenges, LNCS, vol. 8465, pp. 52–67. Springer **(2014)**

- **General approach:**
  - Decomposition of an item into a triple pattern and a set of variable bindings
  - Ordering the triples in the RDF graph of the item.
  - Iterating over ordered list and replacing the subject + object by variables.
  - Comparaison and new representation based on N-1 item

RDSZ: PATTERN & BINDINGS

➢ WAVES EVENT

```
@prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .
@prefix qudt:  <http://data.nasa.gov/qudt/owl/qudt#> .
@prefix ssn:   <http://purl.oclc.org/NET/ssnx/ssn#> .
@prefix waves: <http://waves.org/resource#> .

waves:event_1j_sh   ssn:hasValue        waves:obs_1j_sh ;
                    ssn:isProducedBy    waves:Q_DT01 ;
                    ssn:startTime   "2015-01-01T01:15:00"^^xsd:dateTime
                    rdf:type            ssn:SensorOutput .

waves:Obs_1j_sh    qudt:numericValue   1.3E-1 ;
                   rdf:type    ssn:ObservationValue .
```

Pattern
?x0 <http://purl.oclc.org/NET/ssnx/ssn#hasValue> ?x1 .
?x0 <http://purl.oclc.org/NET/ssnx/ssn#isProducedBy> ?x2 .
?x0 <http://purl.oclc.org/NET/ssnx/ssn#startTime> ?x3 .
?x0 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ?x4 .
?x1 <http://data.nasa.gov/qudt/owl/qudt#numericValue> ?x5 .
?x1 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ?x6 .

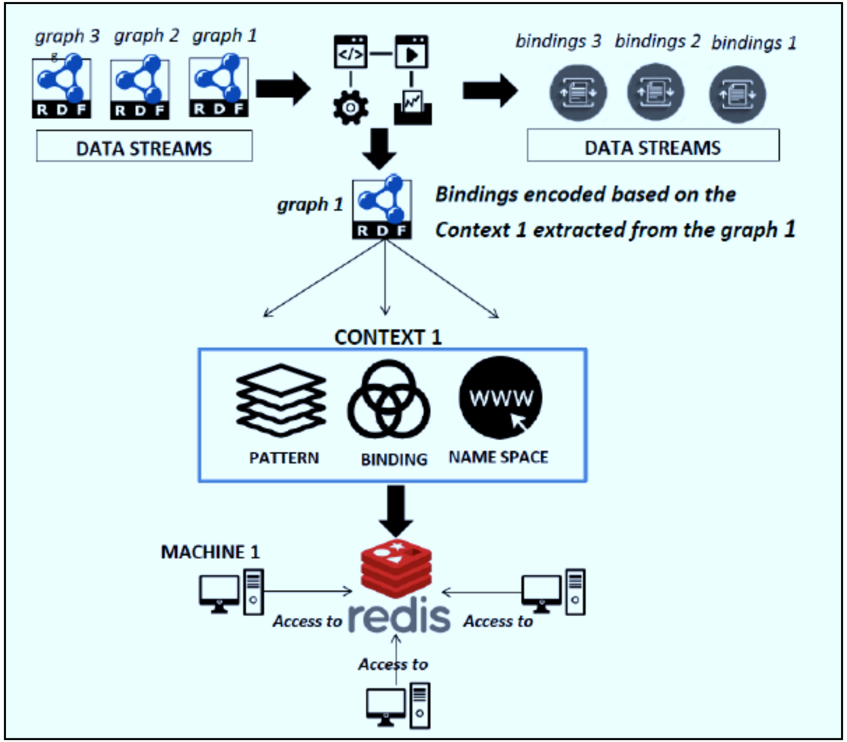
Bindings

Variable	Value
?x0	<http://waves.org/resource#event_1j_sh>
?x1	<http://waves.org/resource#obs_1j_sh >
?x2	<http://waves.org/resource#Q_DT01>
?x3	"2015-01-01T01:15:00 "^^xsd:dateTime
?x4	<http://purl.oclc.org/NET/ssnx/ssn#SensorOutput>
?x5	"1.3E-1"^^xsd:double
?x6	<http://purl.oclc.org/NET/ssnx/ssn#ObservationValue>

RDSC: WAVES CONTRIBUTION

- **Context ID & comparaison fashion**

  Comparison with first item ⇨
  - $?O_0 <P_O> ?O_2$
  - $?O_2 <P_2> ?O_3$
- **Initial Binding in context (shared and immutable)**
  - Replacing redudant values
- **Prefixes:** Encoded URI table → smaller Patterns
- **Name spaces**
  - Reducing URLs length by using association between each pattern and the list of prefixes with their namespaces extracted.
- **GZIP compression (activated if needed)**
  - Replacing Zlib by a more adapted algorithm



RDSC: COMPRESSION RESULTS

WAVES
ATOS SE

- A set of 1000 Files(1.4 Gb) each one containing a waves event

  - RDSC deactivating GZIP and Namespaces: **62,4%**
  - RDSC activating GZIP: **78.3%**
  - RDSC activating GZIP and Namespaces: **84.6%**

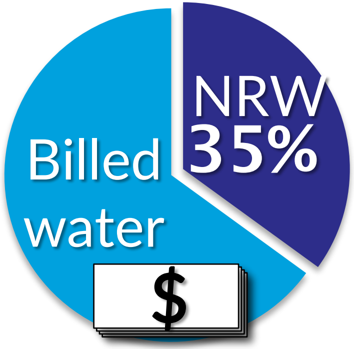
File number	Size	RDSZ	RDSZ+gzip	RDSZ+Ns	D-RDSZ (RDSZ+gzip+Ns)
1	1.3 Kb	84.3%	89.2%	92.6%	93.5%
1000	1.4 Gb	62.4%	78.3%	72.4%	84.6%

USE CASE

Smart Water Management Network.

Why water management ?

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

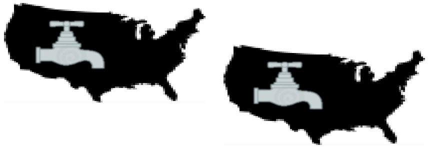


48.6 billion m^3/year

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

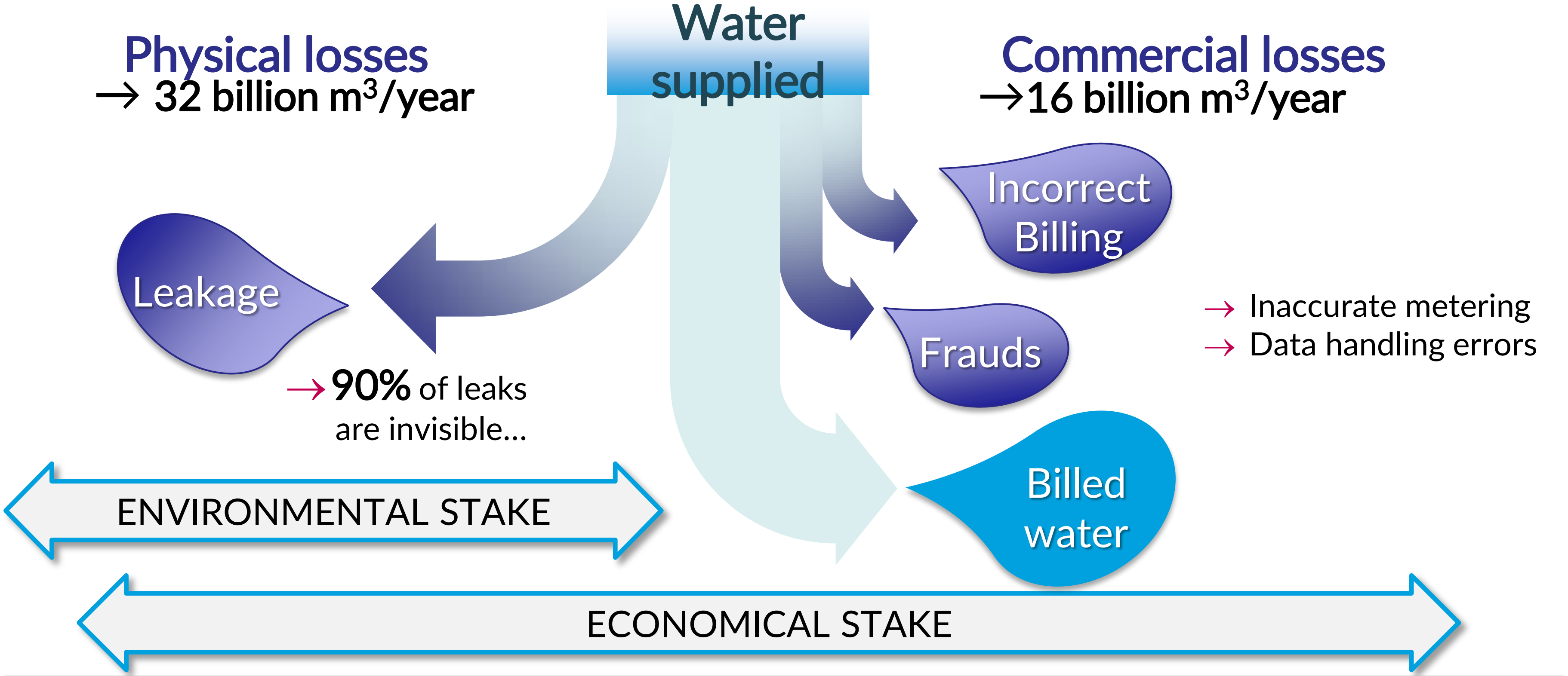
=

x2 the annual domestic water consumption of the USA



Why water management ?

WAVES
ATOS SE



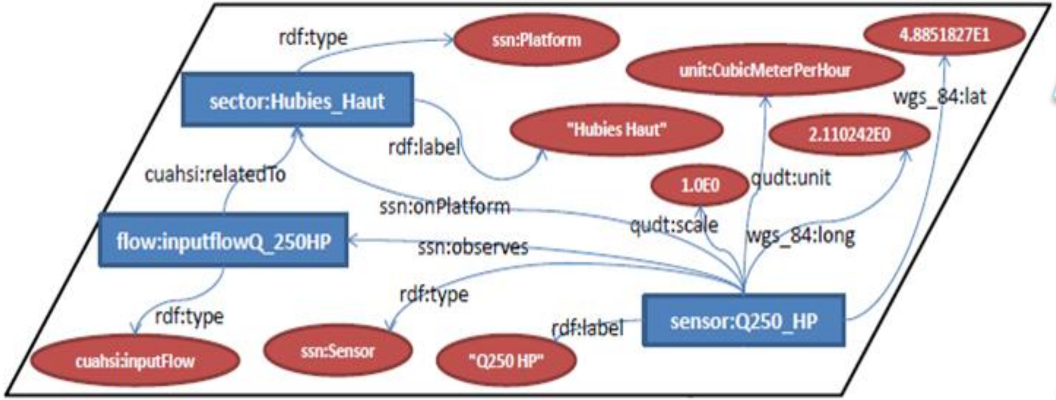
Data Modeling

- **Ontologies:**
  - SSN: *Semantic Sensor Network*
  - CUAHSI: *Consortium of Universities for the Advancement of Hydrologic Science Inc*
  - QUDT3: *Quantities, Units, Dimensions and Data Types Ontologies*
  - WGS84: *World Geodetic System 1984*

- Event:
  - Time stamp
  - Observation ID
  - Numeric value of the observations

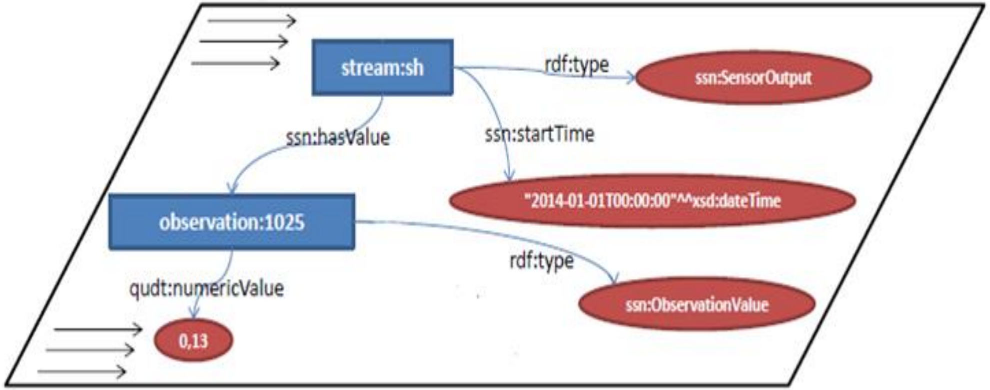
- **Data:**
  - Static
  - Dynamic

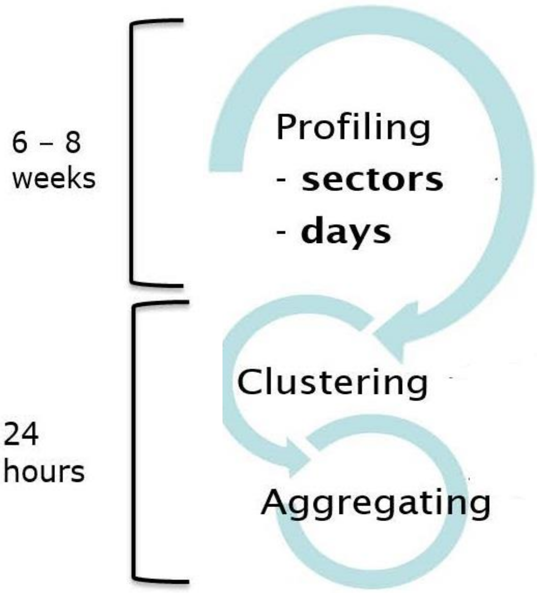
Data Modeling



- Static Data

- Dynamic Data





Use **Pearson's correlation** to evaluate similarities between sector's time series consumption

Clustering pair sector's event between comparable sectors and days to detect singularity by using different models and further optimizations

Aggregating anomalies based on time interval using agglomerative clustering

Anomaly Detection

WAVES
ATOS SE

Sector 1	Sector 2	Correlation
Brezin	Garches	0.985
Brezin	Gobert	0.979
Brezin	Haut-Clagny	0.988
Brezin	Hubies-Detendu	0.991
Garches	Gobert	0.985
Garches	Haut-Clagny	0.988
Garches	Hubies-Detendu	0.989
Gobert	Guyancourt-Detendu	0.969

Models	KMeans	Agglomerative Clustering	OPTICS	DBSCAN	ROCK
Precision	0.90	0.87	0.97	0.84	0.78
Recall	0.65	0.70	0.85	0.69	0.57
α-threshold	0.60	0.80	0.75	0.80	0.78
β-threshold	0.45	0.54	0.32	0.44	0.78
distance used	Euclidean	Cityblocks	Euclidean	Manhattan	-

Day 1	Day 2	Correlation
Monday	Tuesday	0.958
Monday	Thursday	0.955
Monday	Friday	0.943
Saturday	Sunday	0.933
Tuesday	Wednesday	0.778
Monday	Wednesday	0.745

EVALUATION

Exposing Results and Advancement

GLOBAL SET-UP

List of sensors that have measures between 5 and 12

vs

Overall consumption represented by the sum of input fow grouped by the observation timestamp.

Query	Type	Filter	OPTIONAL	GROUP BY
Q1	Simple	✓	X	X
Q2	Complex	X	✓	✓

- Three load scenarios:
  - Scenario1: small for 1,500 triples/sec
  - Scenario2: medium for 8,000 triples/sec
  - Scenario3: high for more than 20,000 triples/sec

- Scenarios run on real cluster using Amazon virtual Machines:
  - 5 nodes/10 nodes/20 nodes
  - C-SPARQL (1 node only)

QUERY EXAMPLES

- **Range**: 2 sec / 4sec
- **Step:** 4 sec / 1 sec

➢ Simple Query

```
PREFIX ssn:<http://purl.oclc.org/NET/ssnx/ssn#>
PREFIX qudt:<http://data.nasa.gov/qudt/owl/qudt#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX cuahsi: <http://his.cuahsi.org/ontology/cuahsi#>
CONSTRUCT{
    waves:event1 ssn:isProducedBy ?sensor;
    ssn:startTime ?time ;
    qudt:numericValue ?value.
    ?sensor ssn:observes ?flow.
}
WHERE
{
    ?event ssn:isProducedBy ?sensor;
    ssn:hasValue ?observation;
    ssn:startTime ?time;
    ?observation qudt:numericValue ?value.
    ?sensor ssn:observes ?flow.
    FILTER( ?value > "100"^^xsd:double || ?value < "1"^^xsd:double )
}
```

➢ Complex Query

```
PREFIX ssn:<http://purl.oclc.org/NET/ssnx/ssn#>
PREFIX qudt:<http://data.nasa.gov/qudt/owl/qudt#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX cuahsi: <http://his.cuahsi.org/ontology/cuahsi#>

CONSTRUCT {
             ?event ssn:startTime ?time .
             ?event ssn:isProducedBy  _:o .
             _:o ssn:onPlatform ?sector .
             _:o qudt:numericValue ?totalSum .
}
WHERE {
    SELECT DISTINCT ?sector ?event ((?inputSum-?outputSum) as ?totalSum) ?time  \
    WHERE {
        SELECT ?sector ?event ?time (SUM(?input_value) AS ?inputSum) (SUM(?output_value) AS ?outputSum)
        WHERE {
            ?observation qudt:numericValue ?input_value.
            ?event ssn:isProducedBy ?sensor;
                   ssn:hasValue ?observation;
                   ssn:startTime ?time.
            ?sensor ssn:observes ?flow;
                    ssn:onPlatform ?sector.
            ?flow a cuahsi:InputFlow;
                  cuahsi:relatedTo ?sector.
            OPTIONAL
            {
                ?observation qudt:numericValue ?output_value.
                ?event ssn:isProducedBy ?sensor;
                       ssn:hasValue ?observation;
                       ssn:startTime ?time.
                ?sensor ssn:observes ?flow;
                        ssn:onPlatform ?sector.
                ?flow a cuahsi:OutputFlow;
                      cuahsi:relatedTo ?sector.
            }
        }
        GROUP BY ?sector
    }
}
```

PRECISION & RECALL

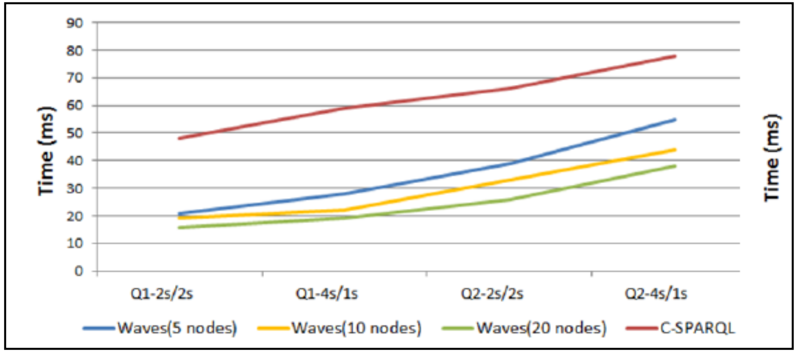
WAVES VS SPARQL

		(a) Scenario1		(b) Scenario 2		(c) Scenario 3	
		WAVES	C-SPARQL	WAVES	C-SPARQL	WAVES	C-SPARQL
Precision	Q1-2s/2s	**100%**	**100%**	**100%**	94%	**98%**	80%
	Q1-4s/1s	**100%**	**100%**	**100%**	88%	**84%**	78%
Recall	Q2-2s/2s	**100%**	93%	**97%**	95%	**79%**	56%
	Q2-4s/1s	**100%**	91%	**94%**	84%	**72%**	43%

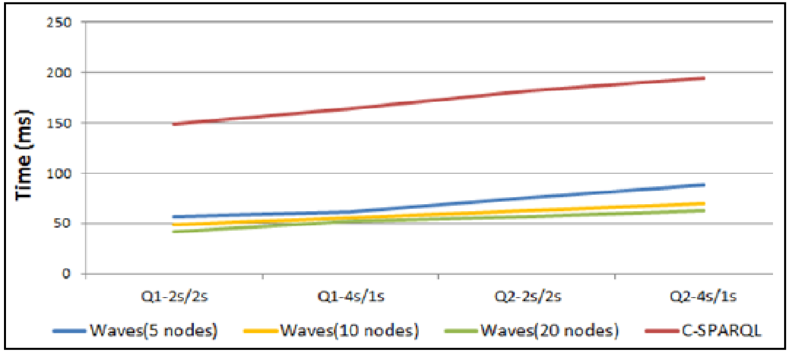
- **Simple query & low load scenario:**
  - WAVES & C-SPARQL remain performant
- **Complex query & medium load scenario**:
  - C-SPARQL shows precision and recall decrease by **3 points** on average
- **Complex query & high load scenario:**
  - C-SPARQL shows signficant precision and recall dropdown by **30 points** on average

EXECUTION TIME

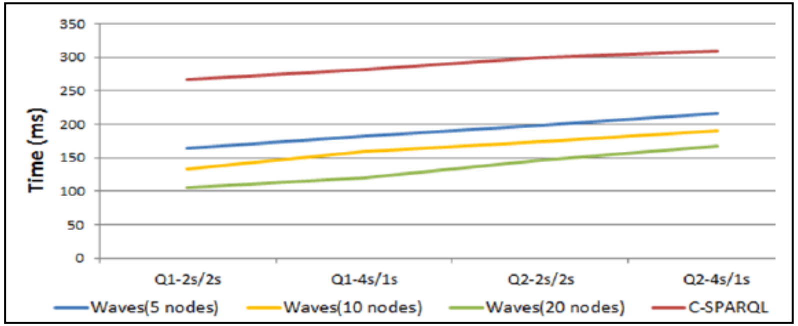
- Scenario 1: 1,500 triples/sec



- Scenario 2: 8000 triples/sec



- Scenario 3: 20000 triples/sec



CONCLUSION

Where are we heading ?

Future Perspectives

- **Enrichment with Linked Data and Social Media to determine the cause of anomalies (e.g., very high or low consumption, etc.) :**
  - Is there something happened in social networks: natural disaster, etc.
  - Special events: holidays, festivals, marathon, etc.

- **Decision making: a potential anomaly could be considered as a real anomaly or not:**
  - Invoking background context to make decision
  - Extending reasoning capabilities in WAVES

www.waves-rsp.org

THANK YOU