Toward Semantic Data Stream Technologies and Applications

Raja CHIKY
raja.chiky@isep.fr
2013-2014
About me

- Associate professor in Computer Science – LISITE-RDI
- Research interest: Data stream mining, scalability and resource optimization in cloud computing, recommender systems
- Research field: **Large scale data management**

1. Real-time and distributed processing of various data sources
2. Use semantic technologies to add a semantic layer
3. Distributive and collaborative data mining
4. Optimizing resources in large scale systems
5. Modeling and validation of complex systems
Outline

- What is a data stream?
- Data stream management systems
- New applications
- Semantic Data Stream
- Conclusion
What is a data stream?

- Golab & Oszu (2003): “A data stream is a real-time, continuous, ordered (implicitly by arrival time or explicitly by timestamp) sequence of items. It is impossible to control the order in which items arrive, nor is it feasible to locally store a stream in its entirety.”
- Massive volumes of structured data, items arrive at a high rate.
Data Stream Management Systems

<table>
<thead>
<tr>
<th></th>
<th>DBMS</th>
<th>DSMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data model</td>
<td>Permanent updatable relations</td>
<td>Streams and permanent updatable relations</td>
</tr>
<tr>
<td>Storage</td>
<td>Data is stored on disk</td>
<td>Permanent relations are stored on disk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Streams are processed on the fly</td>
</tr>
<tr>
<td>Query</td>
<td>SQL language</td>
<td>SQL-like query language</td>
</tr>
<tr>
<td></td>
<td>Creating structures</td>
<td>Standard SQL on permanent relations</td>
</tr>
<tr>
<td></td>
<td>Inserting/updating/deleting data</td>
<td>Extended SQL on streams with windowing</td>
</tr>
<tr>
<td></td>
<td>Retrieving data (one-time query)</td>
<td>Continuous queries</td>
</tr>
<tr>
<td>Performance</td>
<td>Large volumes of data</td>
<td>Optimization of computer resources to deal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>with Several streams</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Several queries</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ability to face variations in arrival rates</td>
</tr>
<tr>
<td></td>
<td></td>
<td>without crash</td>
</tr>
</tbody>
</table>
## World of data streams

<table>
<thead>
<tr>
<th>Website logs</th>
<th>Network monitoring</th>
<th>Financial services</th>
<th>Weather forecasting</th>
</tr>
</thead>
</table>

Data may come from humans, sensors or machines
New applications: 360-degree view of the customer

Why?

What?

How?

When/Where?

Who?

Operational data

Interaction data

Contextual data

Descriptive data

Behavioral data
New applications: 360-degree view of the customer (Contd.)

<table>
<thead>
<tr>
<th>Descriptive data</th>
<th>Interaction data</th>
<th>Operational data</th>
<th>Behavioral data</th>
<th>Contextual data</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Profile</td>
<td>• Multi-channel</td>
<td>• Orders</td>
<td>• Web clicks</td>
<td>• Geographical data</td>
</tr>
<tr>
<td>• Contractual data</td>
<td>• Email/chat/call center</td>
<td>• Transactions</td>
<td>• Usage history</td>
<td>• Devices</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Payment history</td>
<td>• Social media</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Opinions</td>
<td></td>
</tr>
</tbody>
</table>

Internal and External sources
New applications: Social networks

- **Twitter**: A massive Data stream
  - Free social networking and micro-blogging service
  - 3 billion requests per day

- Sentiment Analysis
- Tracking trending hashtags
- Detect breaking news
- Analyze Marketing campaigns
More applications on Twitter

- E.g. http://twitter.com/#search?q=just%20landed%20in
Combination of stream and static data

- E.g., Twitter + MetaCarta

What have you seen?

* BOTTARI: An augmented reality application for personalized **recommendation of restaurants** in Seoul
Too much data streams but not enough knowledge
Semantic Web, Linked Data

- **Semantic Web**
  - Collaborative movement to promote common data formats on the World Wide Web
  - Inclusion of semantic content in web pages
  - From unstructured and semi-structured documents to a « Web of data »

- **Linked Data**
  - Best practices to represent, publish, link data on the Semantic Web
  - Linked Data Cloud: collection of datasets that have been published in Linked data format
Linked data
Semantic Web technologies for data stream

- Annotate stream data with semantic metadata
- Apply **Linked Data principles** to publish streaming data
- Interlink streaming data with **existing** datasets
- Integrate data stream processing + **reasoning**

**Objectives**: interoperability, automation, enrichment
Existing prototypes
Introduction to RDF

Statement: "The author of http://www.w3schools.com/rdf is Jan Egil Refsnes"

- The subject of the statement above is: http://www.w3schools.com/rdf
- The predicate is: author
- The object is: Jan Egil Refsnes
RDF Stream

- RDF Stream is defined as an ordered sequence of pair.
- Each pair consists of a RDF triplet and a timestamp:
  \[
  \langle \text{subj}_i, \text{pred}_i, \text{obj}_i \rangle, T_i
  \]
  \[
  \langle \text{subj}_{i+1}, \text{pred}_{i+1}, \text{obj}_{i+1} \rangle, T_{i+1} \quad (T_i \leq T_{i+1})
  \]

- Example:
  \[
  \langle :\text{traveller1} :\text{justLanded} :\text{placeA} \rangle, T_1
  \]
  \[
  \langle :\text{traveller2} :\text{justLanded} :\text{placeB} \rangle, T_1
  \]
  \[
  \langle :\text{traveller3} :\text{justLanded} :\text{placeA} \rangle, T_2
  \]
  \[
  \langle :\text{traveller1} :\text{justLanded} :\text{placeC} \rangle, T_3
  \]
RDF Stream – Several definitions

- Several possibilities:
  - An RDF stream is an infinite sequence of timestamped events (triples or graphs), ...

\[
<\langle s, p, o \rangle_i, t_i > \\
<\langle s, p, o \rangle_{i+1}, t_{i+1} > \\
<\langle s, p, o \rangle_{i+2}, t_{i+2} > \\
... \\
\]

- An RDF stream is an infinite sequence of triple occurrences \(<\langle s, p, o \rangle, t_\alpha, t_\omega >\) where \(<s, p, o >\) is an RDF triple and \(t_\alpha\) and \(t_\omega\) are the start and end of the interval
Approaches

In RDF Stream models
(timestamps, events, time intervals, triple-based, graph-based …)

In RDF Stream
query languages
(windows, stream selection, CEP-based operators, …)

In implementations
(RDF native, query rewriting, continuous query registration, scalability, static vs streaming data…)

Heterogeneity
Example: SRBench

Q: Detect if a hurricane has been observed

“A hurricane has a sustained wind (for more than 3 hours) of at least 33 metres per second or 74 miles per hour (119 km/h).”

```
ASK
WHERE {
STREAM <http://www.cwi.nl/SRBench/observations> [RANGE 10800s SLIDE 600s]
  {?observation om-owl:procedure ?sensor ;
    om-owl:observedProperty weather:WindSpeed ;
    om-owl:result [ om-owl:floatValue ?value ] .}
}
GROUP BY ?sensor
HAVING ( AVG(?value) >= "74"^^xsd:float )
```

```
ASK
FROM STREAM <http://www.cwi.nl/SRBench/observations> [RANGE 3h STEP 10m]
WHERE {
  ?observation om-owl:procedure ?sensor ;
  om-owl:observedProperty weather:WindSpeed ;
  om-owl:result [ om-owl:floatValue ?value ] .}
GROUP BY ?sensor
HAVING ( AVG(?value) >= "74"^^xsd:float )
```
W3C RDF Stream Processing Community

http://www.w3.org/community/rsp/

RDF Stream Processing Community Group

The mission of the RDF Stream Processing Community Group (RSP) is to define a common model for producing, transmitting and continuously querying RDF Streams. This includes extensions to both RDF and SPARQL for representing streaming data, as well as their semantics. Moreover this work envisions an ecosystem of streaming and static RDF data sources whose data can be combined through standard models, languages and protocols. Complementary to related work in the area of databases, this Community Group looks at the dynamic properties of graph-based data, i.e., graphs that are produced over time and which may change their shape and data over time.

Get involved!

Anyone may join this Community Group. All participants in this group have signed the W3C Community Contributor License Agreement (CLA).

JOIN THIS GROUP

or learn how to join or request an account.

Note: Community Groups are proposed and run by the community. Although W3C hosts these conversations, the groups do not necessarily represent the views of the W3C Membership or staff.
W3C RSP Community Group mission

“The mission of the RDF Stream Processing Community Group (RSP) is to define a common model for producing, transmitting and continuously querying RDF Streams. This includes extensions to both RDF and SPARQL for representing streaming data, as well as their semantics. Moreover this work envisions an ecosystem of streaming and static RDF data sources whose data can be combined through standard models, languages and protocols (...)”
Conclusion: Research Challenges

- Benchmarking: LSBench, SRBench
- Data Quality: Dealing with incomplete and noisy data
- Reasoning on streams: formal semantics, efficiency, scalability
- Distributed and parallel processing: data stream sources are distributed in nature
- Scalability: approximation, load shedding, etc.
Thanks for your attention

raja.chiky@isep.fr