Privacy Preserving, Peta-scale Stream Analytics for Domain-Experts

NRP 75 Big Data

Michael H. Böhlen
Department of Computer Science (IfI)
University of Zürich

Swiss SKA Days
23.5.2017
NRP 75 “Big Data”

- NFP 75 funding: CHF 25 million.
- The research phase will last 5 years.
- 172 pre-proposals
- 55 full proposals
- 36 approved projects
  - Module 1: Computing and information technology (13 projects)
  - Module 2: Societal, regulatory, and educational challenges (8 projects)
  - Module 3: Applications (15 projects)
Outset and Goal

Outset:
- Production of Big Data will soon outpace the availability of both data storage and data scientists who know how to handle such data.

- Society is increasingly concerned about data protection.

Goal:
- We need stream-processing systems that continuously analyse incoming data (rather than store it) and allow non-computer scientists to specify its analysis in a privacy-preserving manner.
Privacy Preserving, Peta-scale Stream Analytics for Domain-Experts

We develop a petabyte-scale, privacy-preserving processing system for commodity hardware.

1. We provide a declarative stream query language with extensions for statistical operations and capacity for real-time operations.
2. The language permits users to specify the desired privacy level.
3. The system compiler translates the statistical functions and privacy specifications into executable computations.
4. The runtime environment selects the approach for optimising exec usrng (e.g. existing systems Apache Flink, Spark Streaming or Storm)
Project Team

Michael Böhlen
Avi Bernstein
Daniele Dell’Aglio
Muhammad Saad
Yasett Acurana
Pengcheng Duan
Use Case 1: ASKAP

- Real-time processing of high-volume ASKAP (Australia Square Kilometer Array Pathfinder) astrophysics data.
- ASKAP is a high-volume astronomy survey instrument that includes 36 12-meter antennas, each of collecting 36 independent beams of about 1 degree each.
- Every dish generates 18 TB of samples per second, which gets reduced to 0.6 TB/s by beamformers and then a correlator computes correlations between the samples reducing it to a total of 2.5 GB/s of raw data.
Expected Outcomes from ASKAP

- ASKAP data will allow us to
  - test the **efficiency** of the PB-scale data processing capability (addressing the big data *volume* and *velocity* challenges)
  - test the **auto-tuning** when adding/removing heterogeneous computational and network capacity during runtime
  - evaluate the **abstractions** offered to non-data-science trained scientists.

- We will test these goals by providing astronomers with our interface and evaluate if the stream specifications they come up with can suitably process real-world ASKAP data within the required time frame.

- Open questions: privacy, data combination/merging from multiple streams, meta-data
Use Case 2: IPTV

- Switching behavior of 3M TV viewers, socio-demographic data, video feeds of 200 channels, electronic program guide (EPG).

- **Differential privacy** describes a promise made by a data holder to a data provider: “You will not be affected by allowing your data to be used, no matter what other information.”

\[
Pr[A(D) \in S] \leq \epsilon * Pr[A(D + x) \in S]
\]

- Differentially private mechanisms can make confidential data available for data analysis, without resorting to non-disclosure agreements, data anonymization, or restricted access.
Expected Outcome from IPTV

- Test the **effectiveness** of the **expressivity** of the proposed query language:
  - mathematical and statistical operators (such as the imputation)
  - privacy extensions
- Explore how far into the query processing chain we will be able to embed privacy ensuring operators.
- Complex distribution constraints:
  - Most IPTV providers do not want to continuously stream all channels to a particular consumer.
  - System may have to run some operations in the IPTV operators facilities.
  - Interesting cross-domain optimization constraints.
Project Schedule

<table>
<thead>
<tr>
<th>Project Quarters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16</td>
</tr>
</tbody>
</table>

WP1: Use cases, requirements

WP2: Language development

WP3: Compiler development

WP4: Runtime environment

WP5: Optimization, tuning

WP6: Evaluations

**start-to-start**
WP2: Language development

- The goal of WP2 is to develop a **declarative query language** with **functional extensions** for linear algebra and privacy, which can later be compiled into a stream graph.
- The operators shall include
  - relational operators (join, aggregation, etc)
  - mathematical and linear algebra operators (FFT, SVD, correlation, etc), and
  - privacy-ensuring operators.
- The linear algebra operators and privacy extensions that are needed to implement our use cases will be worked out in close interaction with the domain experts.
  - CSIRO Wallaby (Bärbel Koribalski, Keith Bannister)
  - Zattoo, Genistat
WP2: Language development

<table>
<thead>
<tr>
<th></th>
<th>Month</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_1$</td>
<td>2012/3</td>
<td>180</td>
</tr>
<tr>
<td>$p_2$</td>
<td>2012/4</td>
<td>240</td>
</tr>
<tr>
<td>$p_3$</td>
<td>2012/5</td>
<td>250</td>
</tr>
<tr>
<td>$p_4$</td>
<td>2012/6</td>
<td>150</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Month</th>
<th>PV</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q_1$</td>
<td>2012/3</td>
<td>350</td>
</tr>
<tr>
<td>$q_2$</td>
<td>2012/4</td>
<td>340</td>
</tr>
<tr>
<td>$q_3$</td>
<td>2012/5</td>
<td>330</td>
</tr>
<tr>
<td>$q_4$</td>
<td>2012/6</td>
<td>550</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Month</th>
<th>Day</th>
<th>Su</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>2012/3</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>$t_2$</td>
<td>2012/3</td>
<td>2</td>
<td>40</td>
</tr>
<tr>
<td>$t_i$</td>
<td>2012/6</td>
<td>30</td>
<td>70</td>
</tr>
</tbody>
</table>

Databases systems: join and aggregate

<table>
<thead>
<tr>
<th></th>
<th>Month</th>
<th>CV</th>
<th>PV</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012/5</td>
<td>250</td>
<td>330</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>2012/6</td>
<td>150</td>
<td>550</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>2012/4</td>
<td>240</td>
<td>340</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>2012/3</td>
<td>180</td>
<td>350</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>

Database systems: standard deviations

12/17
WP2: Language development

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>CV</th>
<th>PV</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012/3</td>
<td>-25</td>
<td>-42.5</td>
<td>-0.5</td>
<td></td>
</tr>
<tr>
<td>2012/4</td>
<td>35</td>
<td>-52.5</td>
<td>-5.5</td>
<td></td>
</tr>
<tr>
<td>2012/5</td>
<td>45</td>
<td>-62.5</td>
<td>-3.5</td>
<td></td>
</tr>
<tr>
<td>2012/6</td>
<td>-55</td>
<td>157.5</td>
<td>9.5</td>
<td></td>
</tr>
</tbody>
</table>

Linear algebra: matrix transposition

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>C</th>
<th>2012/3</th>
<th>2012/4</th>
<th>2012/5</th>
<th>2012/6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>w1</td>
<td></td>
<td>'CV'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-25</td>
<td>35</td>
<td>45</td>
<td>-55</td>
<td></td>
</tr>
<tr>
<td></td>
<td>w2</td>
<td>'PV'</td>
<td>-42.5</td>
<td>-52.5</td>
<td>-62.5</td>
<td>157.5</td>
</tr>
<tr>
<td></td>
<td>w3</td>
<td>'SD'</td>
<td>-0.5</td>
<td>-5.5</td>
<td>-3.5</td>
<td>9.5</td>
</tr>
</tbody>
</table>

Linear algebra: matrix multiplication
**WP2: Language development**

$$x$$

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>CV</th>
<th>PV</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>'CV'</td>
<td>6900</td>
<td>-12250</td>
<td>-860</td>
</tr>
<tr>
<td>2</td>
<td>'PV'</td>
<td>-12250</td>
<td>33275</td>
<td>2025</td>
</tr>
<tr>
<td>3</td>
<td>'SD'</td>
<td>-860</td>
<td>2025</td>
<td>133</td>
</tr>
</tbody>
</table>

Database systems: aggregation, Cartesian product, projection

$$y$$

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>CV</th>
<th>PV</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_1$</td>
<td>'CV'</td>
<td>2300</td>
<td>-4083</td>
<td>287</td>
</tr>
<tr>
<td>$y_2$</td>
<td>'PV'</td>
<td>-4083</td>
<td>11092</td>
<td>675</td>
</tr>
<tr>
<td>$y_3$</td>
<td>'SD'</td>
<td>-287</td>
<td>675</td>
<td>44</td>
</tr>
</tbody>
</table>
WP3: Process specification compilation

- Seamless integration of
  - operators over relations,
  - operators over streams, and
  - operators over multi-dimensional arrays.

- We start out with CQL’s stream algebra SRA and extend it with linear algebra and privacy-ensuring operators. A possible algebraic representation of the IPTV example is as follows:

```
istream(
  privatized[diff_priv, eps = 0.1](
    \pi[StartTime, Duration, IP](
      correlateANDimpute[...](Range(1Day)(IPTV1),
                                Range(1Day)(IPTV2),
                                EPG))))
```
WP4: Runtime Environment, Optimization and Tuning

- Build a distributed runtime environment for stream processing.
- Based on Apache Flink, Spark Streaming or Apache Storm.
- The seamless integration of operators over relations, operators over streams, and operators over multi-dimensional arrays.
- Physical algebra that supports streams, relations and arrays and precisely defines the semantics of operations.
- Use the physical algebra as a basis for optimization.
Summary

- ASKAP data processing system in collaboration with CSIRO Wallaby (Bärbel Koribalski, Keith Bannister).
- Leverage
  - stream processing,
  - database querying, and
  - mathematical, statistical and linear algebra functionality
  to build a system for **declarative stream analytics for domain experts**.
- Allow data providers to specify the desired level of privacy of their data