Searching time series with Hadoop in an electric power company
A. Bérard (TELECOM PARISTECH), G. Hébrail (ELECTRICITE DE FRANCE - R&D), FRANCE

Problem definition
• Long univariate time series
• Finding predefined pattern(s) with sliding/jumping windows

Applications & Datasets
• Electric power consumption
  • 35M customers every 30’ for 1 month
  • 3.3TB uncompressed
    → Finding specific consumption shapes
• Electric power network frequency
  • 1” sampling rate for 1 month
  • 145MB uncompressed
    → Finding specific drops in frequency

Goal of the work
• Assess the strengths/weaknesses of brutal sequential search in Hadoop
• Define adequate data structures and search functions
Searching time series with Hadoop in an electric power company
A.Bérard (TELECOM PARISTECH), G.Hébrail (ELECTRICITE DE FRANCE - R&D), FRANCE

Generic data structures and search functions for long time series in Hadoop

Multi-distance search (TopK)

- The right distance to use is unknown

• Computing TopK with several distances
• Inter-ranking results

CREATE TEMPORARY FUNCTION TopK as 'com.UDAF.TopK';
SELECT TopK(id, values) FROM table WHERE ...

• Various optimizations

Experimental results

- Scalability on electric power consumption data
- Precision on network frequency data

<table>
<thead>
<tr>
<th></th>
<th>Whole Matching</th>
<th>Subsequence Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>l = 5</td>
<td>l = 20</td>
</tr>
<tr>
<td>Min</td>
<td>3.00</td>
<td>5.62</td>
</tr>
<tr>
<td>Mean</td>
<td>3.20</td>
<td>5.85</td>
</tr>
<tr>
<td>Max</td>
<td>3.32</td>
<td>6.15</td>
</tr>
</tbody>
</table>