Mining Big Data in Real Time

Albert Bifet

Turing/SLAIS 2012 Conference
BIG DATA

Measure and React
Motivation

Source: IDC’s Digital Universe Study (EMC), June 2011

Data is growing
## Motivation

<table>
<thead>
<tr>
<th>Memory unit</th>
<th>Size</th>
<th>Binary size</th>
</tr>
</thead>
<tbody>
<tr>
<td>kilobyte (kB/KB)</td>
<td>$10^3$</td>
<td>$2^{10}$</td>
</tr>
<tr>
<td>megabyte (MB)</td>
<td>$10^6$</td>
<td>$2^{20}$</td>
</tr>
<tr>
<td>gigabyte (GB)</td>
<td>$10^9$</td>
<td>$2^{30}$</td>
</tr>
<tr>
<td>terabyte (TB)</td>
<td>$10^{12}$</td>
<td>$2^{40}$</td>
</tr>
<tr>
<td>petabyte (PB)</td>
<td>$10^{15}$</td>
<td>$2^{50}$</td>
</tr>
<tr>
<td>exabyte (EB)</td>
<td>$10^{18}$</td>
<td>$2^{60}$</td>
</tr>
<tr>
<td>zettabyte (ZB)</td>
<td>$10^{21}$</td>
<td>$2^{70}$</td>
</tr>
<tr>
<td>yottabyte (YB)</td>
<td>$10^{24}$</td>
<td>$2^{80}$</td>
</tr>
</tbody>
</table>

Data is growing
Motivation

Source: IDC’s Digital Universe Study (EMC), June 2011

Data is growing
Motivation

Source: IDC’s Digital Universe Study (EMC), June 2011

Data is growing
Motivation

Meanwhile, the number of IT professionals in the world will grow by less than 1.5 times.

Source: IDC’s Digital Universe Study (EMC), June 2011

Data is growing
Streaming Data

Big Data & Real Time
**Big Data**

*Big data—a growing torrent*

-$600$ to buy a disk drive that can store all of the world’s music

5 billion mobile phones in use in 2010


---

**Big data** refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.
Big Data


**Big data** refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.
BIG Data

- Volume
- Variety
- Velocity

3 Vs
Methodology

Sampling and distributed systems
Big Data does not need big machines, it needs big intelligence
Real time analytics

We want to analyze what is happening now.
We want to analyze what is happening **now**.
Time and Memory

Number 8 Wire Mentality

Time and memory are the resource dimensions of the process.
Time and memory are the resource dimensions of the process.
Algorithms

Classification, Regression, Clustering, Frequent Pattern Mining.
Applications

- sensor data: industry, cities
- telecomm data
- social networks: twitter, facebook, yahoo
- marketing: sales business

Data may come from: humans, sensors, or machines.
New applications: social networks

Twitter: A Massive Data Stream

- Micro-blogging service
- Built to discover what is happening at any moment in time, anywhere in the world.
- 3 billion requests a day via its API.

MOA-TweetReader: a real-time system to

- read tweets in real time
- detect changes
- find the terms whose frequency changed
Sentiment Analysis on Twitter

Sentiment analysis
Classifying messages into two categories depending on whether they convey positive or negative feelings

*Emoticons* are visual cues associated with emotional states, which can be used to define class labels for sentiment classification.

<table>
<thead>
<tr>
<th>Positive Emoticons</th>
<th>Negative Emoticons</th>
</tr>
</thead>
<tbody>
<tr>
<td>:)</td>
<td>:(</td>
</tr>
<tr>
<td>:-()</td>
<td>:-()</td>
</tr>
<tr>
<td>: )</td>
<td>: (</td>
</tr>
<tr>
<td>:D</td>
<td></td>
</tr>
<tr>
<td>=)</td>
<td></td>
</tr>
</tbody>
</table>

*Table*: List of positive and negative emoticons.
New problem: structured classification

New methods for structured classification

- sequences, trees, graphs

Example: Lord of the Rings

Action, Adventure, Fantasy
New problem: structured classification

New methods for structured classification

- sequences, trees, graphs
- frequent pattern mining techniques
New problem: structured classification

New methods for structured classification

- sequences, trees, graphs
- frequent pattern mining techniques
- multi-label data mining
  - Example: Lord of the Rings → Action, Adventure, Fantasy

a, b → class1, class2
New Techniques: Distributed Systems

Hadoop, S4 and Storm
Hadoop
Hadoop architecture
Apache Mahout

Mahout: open source framework
Pig: Similar to SQL
A = LOAD 'data' USING PigStorage() AS (f1:int, f2:int, f3:int);
B = GROUP A BY f1;
C = FOREACH B GENERATE COUNT ($0);
DUMP C;

Pig: Similar to SQL
Apache S4
A keyless event (EV) arrives at PE1 with quote: "I meant what I said and I said what I meant.", Dr. Seuss

**QuoteSplitterPE (PE1)** counts unique words in Quote and emits events for each word.

**WordCountPE (PE2–4)** keeps total counts for each word across all quotes. Emits an event any time a count is updated.

**SortPE (PE5–7)** continuously sorts partial lists. Emits lists at periodic intervals

**MergePE (PE8)** combines partial TopK lists and outputs final TopK list.

---

<table>
<thead>
<tr>
<th>PE ID</th>
<th>PE Name</th>
<th>Key Tuple</th>
</tr>
</thead>
<tbody>
<tr>
<td>PE1</td>
<td>QuoteSplitterPE</td>
<td>null</td>
</tr>
<tr>
<td>PE2</td>
<td>WordCountPE</td>
<td>word=&quot;said&quot;</td>
</tr>
<tr>
<td>PE4</td>
<td>WordCountPE</td>
<td>word=&quot;i&quot;</td>
</tr>
<tr>
<td>PE5</td>
<td>SortPE</td>
<td>sortID=2</td>
</tr>
<tr>
<td>PE7</td>
<td>SortPE</td>
<td>sortID=9</td>
</tr>
<tr>
<td>PE8</td>
<td>MergePE</td>
<td>topK=1234</td>
</tr>
</tbody>
</table>
Storm

Storm from Twitter
Storm

Stream, Spout, Bolt, Topology
Runaway complexity in Big Data
Nathan Marz, 2012
Data Streams

Big Data & Real Time
Data Streams

Thanks!