Architecting a Universal Sensor Analytics Platform for Utilities

Sean Murphy, Michael Andersen

September 28th, 2018 – CURENT
Presentation Outline

1. Introduction and Motivation
   Sean Murphy (~10 minutes)

2. Time Series Storage and Analysis
   Michael Andersen (~30 minutes)

3. Q&A (~10 minutes)
The Analytics Pipeline

0. Capture
1. Access
2. Acquire
3. Wrangle and Clean
4. Analyze and Model
5. Visualize, Report, and Propagate
6. Move Results to Production
The Analytics Pipeline

0. Capture
1. Access
2. Acquire
3. Wrangle and Clean
4. Analyze and Model
5. Visualize, Report, and Propagate
6. Move Results to Production

80-98% of time/resources

The Fun Part!!!! What people care about.

What makes money

0 1 2 3 4 5 6

The Fun Part!!!! What people care about.

What makes money
Annual Data Volumes

<table>
<thead>
<tr>
<th># of Streams</th>
<th>SCADA</th>
<th>PMU</th>
<th>DFR</th>
<th>Sampling Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000</td>
<td>1000</td>
<td>100</td>
<td>10</td>
<td>0.1 Hz</td>
</tr>
<tr>
<td>10000</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>1 Hz</td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>10 Hz</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>100 Hz</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>1 Khz</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>10 Khz</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100 Khz</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1 MHz</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10 MHz</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100 MHz</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1 GHz</td>
</tr>
</tbody>
</table>

- Megabytes
- Gigabytes
- Terabytes
- Petabytes
- Exabytes
- Zettabytes
Two Options
Really Only One Option
The Three Generations of Big Data Systems

(1) Hadoop
- Workload: Batch processing
- Dominant Analytics Paradigm: Classic Business Analytics
- Features: Map reduce, Disk oriented Generic, Disk bandwidth
- Limiting Reagent: Disk bandwidth

(2) Spark
- Workload: Iterative processing
- Dominant Analytics Paradigm: Machine Learning
- Features: In memory, Better tooling
- Limiting Reagent: Memory capacity

(3) Custom
- Workload: Continuous processing
- Dominant Analytics Paradigm: ML and Deep Learning
- Features: Data type specific, Industry and application focused
- Limiting Reagent: Compute
Jevons Paradox

In 1865, a twenty-nine-year-old Englishman named William Stanley Jevons published a book, “The Coal Question,” in which he argued that the bonanza couldn’t last. Britain’s affluence, he wrote, depended on its endowment of coal, which the country was rapidly depleting. He added that such an outcome could not be delayed through increased “economy” in the use of coal—what we refer to today as energy efficiency. He concluded, in italics, “It is wholly a confusion of ideas to suppose that the economical use of fuel is equivalent to a diminished consumption. The very contrary is the truth.”*

“occurs when technological progress increases the efficiency with which a resource is used (reducing the amount necessary for any one use), but the rate of consumption of that resource rises because of increasing demand”

https://www.newyorker.com/magazine/2010/12/20/the-efficiency-dilemma
Thank You!

Any questions?

Sean Murphy
Co-Chief Executive Officer
PingThings, Inc.
sean@pingthings.io
http://www.pingthings.io/