Enabling the Rapid Development of New Grid Analytics from Sensor Data at Scale

August 4th, 2018 – Portland, OR

Sean Patrick Murphy
Michael Andersen
Kevin Jones
The Core Problem

Hypotheses
Algorithms
Prototypes

Operational Deployment
The Analytics Pipeline

0. Capture
1. Access
2. Acquisition
3. Wrangling and Cleaning
4. Analysis and Modeling
5. Visualization, Reporting, and Propagation
6. Moving Results to Production

- Data Exploration
- Rapid Prototyping
- Test and Evaluation
The Analytics Pipeline is Broken – Part 1

- Sensor and Comm Errors
- NDA, DSA, and Red Tape
- Ancient Software and Network
- Wasted Effort on each Project
- Dated Tooling (EXCEL)
- Traditional Channels

0. Capture
1. Access
2. Acquisition
3. Wrangling and Cleaning
4. Analysis and Modeling
5. Visualization, Reporting, and Propagation
6. Moving Results to Production

- Data Exploration
- Rapid Prototyping
- Test and Evaluation

PingThings
The Analytics Pipeline is Broken – Part 2

Sensor and Comm Errors
NDA, DSA, and Red Tape
Ancient Software and Network
Wasted Effort on each Project
Dated Tooling (EXCEL)
Traditional Channels

Capture
Access
Acquisition
Wrangling and Cleaning
Analysis and Modeling
Visualization, Reporting, and Propagation
Moving Results to Production

Data Exploration
Test and Evaluation
Rapid Prototyping

PingThings
Annual Data Volumes

![Data Volume Heatmap](image)

**SCADA**
- Number of Streams: 100,000
  - Sensor Measurement Frequency: 0.1 Hz
  - Annual Data Volume: 4.6 TB

- Number of Streams: 10,000
  - Sensor Measurement Frequency: 1 Hz
  - Annual Data Volume: 470.2 GB

- Number of Streams: 1,000
  - Sensor Measurement Frequency: 10 Hz
  - Annual Data Volume: 47.0 GB

- Number of Streams: 100
  - Sensor Measurement Frequency: 100 Hz
  - Annual Data Volume: 4.7 GB

- Number of Streams: 10
  - Sensor Measurement Frequency: 1 KHz
  - Annual Data Volume: 481.5 MB

- Number of Streams: 1
  - Sensor Measurement Frequency: 10 KHz
  - Annual Data Volume: 48.1 MB

**PMU**
- Number of Streams: 100,000
  - Sensor Measurement Frequency: 0.1 Hz
  - Annual Data Volume: 459.2 TB

- Number of Streams: 10,000
  - Sensor Measurement Frequency: 1 Hz
  - Annual Data Volume: 45.9 TB

- Number of Streams: 1,000
  - Sensor Measurement Frequency: 10 Hz
  - Annual Data Volume: 459.2 TB

- Number of Streams: 100
  - Sensor Measurement Frequency: 100 Hz
  - Annual Data Volume: 4.5 PB

- Number of Streams: 10
  - Sensor Measurement Frequency: 1 KHz
  - Annual Data Volume: 44.9 PB

- Number of Streams: 1
  - Sensor Measurement Frequency: 10 KHz
  - Annual Data Volume: 448.5 PB

**DFR**
- Number of Streams: 100,000
  - Sensor Measurement Frequency: 0.1 Hz
  - Annual Data Volume: 459.2 TB

- Number of Streams: 10,000
  - Sensor Measurement Frequency: 1 Hz
  - Annual Data Volume: 448.5 PB

- Number of Streams: 1,000
  - Sensor Measurement Frequency: 10 Hz
  - Annual Data Volume: 44.9 PB

- Number of Streams: 100
  - Sensor Measurement Frequency: 100 Hz
  - Annual Data Volume: 4.4 EB

- Number of Streams: 10
  - Sensor Measurement Frequency: 1 KHz
  - Annual Data Volume: 448.5 PB

- Number of Streams: 1
  - Sensor Measurement Frequency: 10 KHz
  - Annual Data Volume: 448.5 PB

---

**PingThings**

Q3 2018 | CONFIDENTIAL | 6
# The Three Generations of Big Data Systems

<table>
<thead>
<tr>
<th>Workload</th>
<th>(1) Hadoop</th>
<th>(2) Spark</th>
<th>(3) Custom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dominant Analytics Paradigm</td>
<td>Batch processing</td>
<td>Iterative processing</td>
<td>Continuous processing</td>
</tr>
<tr>
<td>Features</td>
<td>Classic Business Analytics</td>
<td>Machine Learning</td>
<td>ML and Deep Learning</td>
</tr>
<tr>
<td>Limiting Reagent</td>
<td>Map reduce</td>
<td>In memory</td>
<td>Data type specific</td>
</tr>
<tr>
<td></td>
<td>Disk oriented Generic</td>
<td>Better tooling</td>
<td>Industry and application</td>
</tr>
<tr>
<td></td>
<td>Disk bandwidth</td>
<td>Memory capacity</td>
<td>focused</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Compute</td>
</tr>
</tbody>
</table>
Universal Sensor Analytics Platform

- **PMU, uPMU, DFR, Power Quality, PQube, FNET, UGA, SEL Devices, AMI, Smart Meters, ...**
- **Hi-speed Historical Data Ingest**
- **Multi-format Data Export**
- **Fusion of Simulation and Sensor Data**
- **Modeling and Simulation Connectivity**
- **Ingestor\textsubscript{N}**
- **CSV, .d, .d2**
- **GEP/STTP**
- **COMTRADE**
- **C37.118**
- **Ingest Engine**
- **Time Series Database**
- **API Layer**
- **Analytic Pipeline**
  - Human-scale Data Engagement
  - Customizable Real-Time Dashboards
  - Spreadsheets
  - Interactive Data Exploration and Analysis
  - Rapid Deployment of Production Analytics
  - Bespoke Web and Mobile Applications

**Real-Time & Historical Analytics**

**Deep Learning and Machine Learning at Scale**

**TensorFlow**
Why You Should Care

- Built for heterogeneous time series data
- Open source
- Open data formats
- Designed to make the humans better
- 3rd generation platform
- Horizontally scalable and distributed
- Analytics as a first class citizen
- State of the art lossless compression (10:1)
(timestamp, value)

64-bit integer

$[-(2^{63}), 2^{63} - 1]$

Or

Write Patterns
Write Patterns
Read Patterns

1. Human interaction and exploration of the data
2. Analytics
3. Training ML/DL Patterns
Read Patterns

“Overview first, zoom and filter, then details-on-demand.”

The Visual Information-Seeking Mantra [Shneiderman, 1996] summarizes many visual design guidelines and provides an excellent framework for designing information visualization applications.
Random, Multi-Resolution Read Patterns

About 4 billion data points
Random, Multi-Resolution Read Patterns
Random, Multi-Resolution Read Patterns
Random, Multi-Resolution Read Patterns
Random, Multi-Resolution Read Patterns
Random, Multi-Resolution Read Patterns
Random, Multi-Resolution Read Patterns

50 datapoints (8 orders of magnitude)
Berkeley Tree Data Structure

Copy on write K-ary Tree
Partitioning static time (1933 to 2079)

Leaf nodes
- Time, value pairs + length

Internal nodes
- Pointers to children
- **Version annotations** for children
- **Aggregates** for children
  - Min, Mean, Max, Count
  - Any associative operator
Did it Work?
### Core Benchmarks

<table>
<thead>
<tr>
<th>#BTrDB</th>
<th>Streams</th>
<th>Total points</th>
<th>#Conn</th>
<th>Insert [mil/s]</th>
<th>Cold Query [mil/s]</th>
<th>Warm Query [mil/s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>500 mil</td>
<td>30</td>
<td>16.77</td>
<td>9.79</td>
<td>33.54</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>1000 mil</td>
<td>60</td>
<td>28.13</td>
<td>17.23</td>
<td>61.44</td>
</tr>
<tr>
<td>3</td>
<td>150</td>
<td>1500 mil</td>
<td>90</td>
<td>36.68</td>
<td>22.05</td>
<td>78.47</td>
</tr>
<tr>
<td>4</td>
<td>200</td>
<td>2000 mil</td>
<td>120</td>
<td><strong>53.35</strong></td>
<td>33.67</td>
<td><strong>119.87</strong></td>
</tr>
</tbody>
</table>

10x Faster than the Closest Platform
Common Analytics Patterns
Common Analytics Patterns

Clean
Common Analytics Patterns

Clean

Filter
Common Analytics Patterns

Clean

Filter

Compute Power
Directed Acyclic Graph of Calculations
## Analytics Benchmarks

<table>
<thead>
<tr>
<th></th>
<th>Distributed</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Identity</td>
<td>Phase Difference</td>
<td>Reactive/Fundamental Pwr</td>
</tr>
<tr>
<td>Input/Output streams</td>
<td>1/1</td>
<td>2/1</td>
<td>4/2</td>
</tr>
<tr>
<td>Compute changeset</td>
<td>972 μs</td>
<td>1659μs</td>
<td>1180μs</td>
</tr>
<tr>
<td>Query data [s]</td>
<td>69.8</td>
<td>104.4</td>
<td>196.9</td>
</tr>
<tr>
<td>Kernel calculation [s]</td>
<td>10.8</td>
<td>22.7</td>
<td>245.5</td>
</tr>
<tr>
<td>Delete old data [s]</td>
<td>6.7</td>
<td>6.9</td>
<td>15.8</td>
</tr>
<tr>
<td>Insert new data [s]</td>
<td>40.7</td>
<td>39.8</td>
<td>66.5</td>
</tr>
<tr>
<td>Changeset / compute time</td>
<td>1064 x</td>
<td>773 x</td>
<td>259 x</td>
</tr>
</tbody>
</table>
Example Use Case – Voltage Sag Detection
Example Use Case – Voltage Sag Detection
Example Use Case – Voltage Sag Detection

From Prototype to Production in a Week
Building a True Platform

“A platform is a business based on enabling value-creating interactions between external producers and consumers. The platform provides an open, participative infrastructure for these interactions and sets governance conditions for them. The platform’s overarching purpose: to consummate matches among users and facilitate the exchange of goods, services, or social currency, thereby enabling value creation for all participants.

*Strategy has moved from controlling unique internal resources and erecting competitive barriers to orchestrating external resources and engaging vibrant communities. And innovation is no longer the province of in-house experts and research and development labs, but is produced through crowdsourcing and the contribution of ideas by independent participants in the platform. External resources don’t completely replace internal resources—more often they serve as a complement. But platform firms emphasize ecosystem governance more than product optimization, and persuasion of outside partners more than control of internal employees.”*
The Punchline

• Free private beta for universities for the upcoming Academic Year
• Platform is pre-populated with TB of real synchrophasor data
• New sensor data is streaming into the platform
Sean Patrick Murphy
sean@pingthings.ai