Big Data Scale and Energy Insight

Josh Gray
Verlitics
905 NW 12th Avenue
Portland, Oregon 97209
joshg@verlitics.com

ABSTRACT

We consider challenges associated with building a big data system specifically to support data streams typical of more advanced Non-Intrusive Load Monitoring (NILM) approaches. Aspects of *data velocity* and *data volume* are addressed as they relate to time-sequenced electrical sensor data at scale. An approach to efficiently transform time-sequenced data so it can be mined for insights using common tools and approaches is described.

Keywords

Energy Analytics, Energy Data Mining, Big Data, Time Sequence Data Store, Map/Reduce, Energy Data

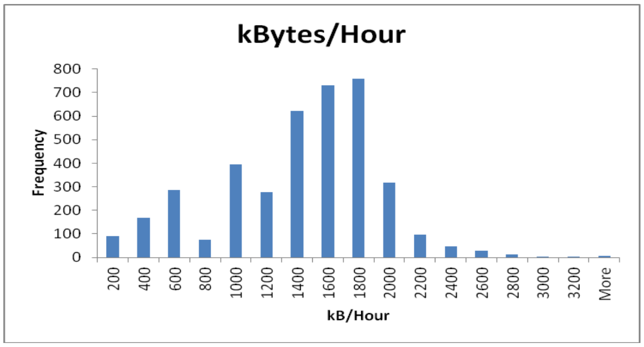
1. INTRODUCTION

Utility advanced metering infrastructure (AMI) is often held up as an example of a *big data* opportunity. At a typical data rate of a few measurements *per minute* multiplied by more than 20,000,000[1] deployed devices, the volume and velocity of information is significant. With many interesting NILM techniques requiring data rates far exceeding AMI capabilities, designing a system to handle this while still enabling common data mining tools and techniques requires special care.

2. ENERGY DATA

Target data rates certainly vary by NILM approach—with some techniques utilizing peak sampling rates in excess of 20 kHz.

Typical data streams used by Verlitics field and algorithm development activities were used to evaluate communication and data infrastructure. A random sample of 4,000 hours of stored energy analytics data yielded an average size of 1.35 megabytes/hour of recorded energy data—highly compressed from the measured data streams.



2.1 Local Capture and Storage

Continuous, unbroken communications with offsite servers are not practical in most environments where long-term deployment is intended. Data must be buffered during lapses in data connectivity that will occur. Any lapse in communications that exceeds the capacity of local data buffering will necessarily result in data loss.

The intended output of many NILM strategies is to inform when events of interest have occurred—when a particular load began consuming energy or changed its operational characteristics. Missing critical data when a notable event occurred can make estimation of the time of the event imprecise, and depending upon the approach may make the event undetectable completely.

Very low cost sensing hardware motivates limiting local buffer storage capacity for measured energy data. This limited capacity can significantly curtail retry capabilities in communicating information to connected storage servers—particularly when wireless communications are involved.

Buffer Capacity	Occurrences of Gaps
2 seconds	461
5 seconds	67
10 seconds	6
30 seconds	3
2 minutes	2
10 minutes	2

Representative 7-day Period

The table above illustrates the number of occurrences where gaps in transmitted energy data would have occurred, based on various limitations in local device buffering capacity over a week of typical field deployment. Objectives for data continuity, data capture requirements and measurement device cost targets need to be carefully considered against anticipated NILM strategies for broad deployments.

2.2 Time Precision

Mining information from time-sequenced event databases—such as that produced by common NILM techniques—often relies on correlating timestamps between data streams. Coincidence in time is a key factor in determining event correlation, and indeterminate clock drift can significantly complicate determination of correlation.

There are several key considerations in delivering an accurately-timed stream of energy data for purposes of mining event correlations originating from NILM-identified observations.

[1] U.S. Energy Information Administration (eia.gov) sited 20,334,525 deployed "smart" electricity meters in 2010.

- **Commodity UTC Clocks Drift**—Common integrated circuit clock chips drift significantly over time. Aggressive periodic correction is required from a connected, accurate time source.
- **Incomplete Embedded NTP Protocols**—Embedded devices with clock chips and Network Time Protocol support may have incomplete implementations that do not adequately compensate for network delay per the specification.
- **Environmental Exacerbates Drift**—Electrical panel installation is outdoors in many areas and sensing hardware may need to be as well. Wider temperature swings can affect clock drift velocity.

2.3 Server Storage

Energy data streaming from sensing devices is naturally suited for time-series storage—organized by device identifier. Any NILM technique applied that will utilize historical context needs to consume the data sequence from a particular measurement device in chronological order.

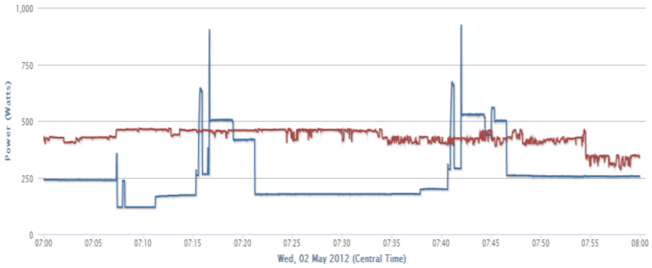
Groups of energy data readings grouped by device and adjacency in time are collected together and stored as convenient units of information. In this work, the underlying mechanism of storage is a distributed key-value data store with records containing an hour of energy data from a particular energy sensing device.

2.4 Time Sequence NILM Analysis

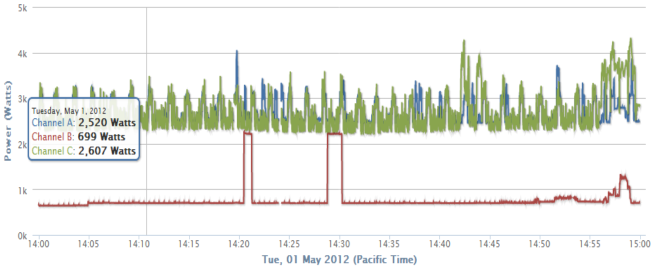
Primary to enabling data mining from NILM analysis is to produce informational events from raw energy data streams. Many NILM strategies appropriately utilize facility-specific historical context to shape probabilistic expectations in interpreting future data. This requires an energy data sequence to be analyzed in chronological order.

Depending upon the type of analysis, the results are deposited into a separate data store. Correlations can be queried or found experimentally. The work described here uses both indexed queries on key data metrics such as timestamps, as well as map/reduce compatible stores allowing efficient experimentation and discovery of less-structured patterns in the events produced by NILM identification.

This separation of processing steps allows appropriate, adaptive NILM to compensate for sometimes dramatically different individual characteristics of each facility. It benefits from the facility's historical context, while enabling effective mining of patterns in produced events and correlations across facilities and against alternate data sources.



A 'typical' residential NILM application, Split-Phase A & B Legs



A 'typical' small commercial NILM application, 3-Phase A, B & C

The information generated by NILM analysis and deposited into data stores is generally free of context required for interpretation. It consists of 'event' information (e.g., load on, load change, diagnostic symptom detected, etc.).

It is advantageous to locate time-adjacent data from the same energy sensing device on the same computational node in the analytics cluster. This allows optimization of the NILM analysis as the code can more generally move to the data rather than the other way around.

Finally, an appropriate amount of energy information is chosen for purposes of dispatching NILM processing for a particular meter and time duration. Selection of this involves trading off granularity of job analysis dispatch and latency of results.

3. RESULTS

We built and have been validating this infrastructure, having accumulated and stored over 120,000 meter-hours of measured energy information. Scale is currently being validated to the first tier of 50,000 simultaneous data streams by scaling out cloud-based servers.

Flexibility of the job processor has facilitated continued NILM strategy development and implementation validation against stored data sets with cloud scale velocity. Alternate analytics implementations can easily be run against substantially all available energy data iteratively, speeding development and catching changes in results over a broader data population more quickly.

The structured NILM output has enabled active development in mining valuable trends, correlations and observations across the energy data set and in the context of other time-sequence data sources.

About the speaker:

Josh Gray is co-founder and leads the product development team at Verlitics, an energy data analytics company. Previously he co-founded Emme, a company focused on the design and manufacture of award-winning energy and environmental control systems.