DATA 698 Project Proposal: Interval Data Standardization, Cleanup, and Validation

Dan Smilowitz

September 18, 2017

Background

New York Energy Manager (NYEM) is an offering of the New York Power Authority (NYPA) providing insights into energy usage. The NYEM portal delivers to customers a variety of analytical services, ranging from simple data visualization (data presentment) to building modeling. The desired impact of these services is that

> By delivering relevant and visually accessible information, NYEM helps drive insights to improve building energy performance, reduce environmental impact and lower energy bills

Data is currently processed at three levels of granularity and frequency:

1. **Monthly utility bill data** to allow evaluation of overall building performance and identification of cost-saving strategies
2. **Building-level data** recorded at 15-minute intervals to reveal strategies for cost-saving load shifting and usage insights
3. **Deep submetering** to maximize performance and minimize costs of equipment and systems within a building

NYEM is an important part of NYPA's energy services strategy, and strongly supports one of its key strategic initiatives through 2020; it is also a focal point in NYPA's efforts to become the nation's first digital utility.

Problem Description

NYPA is investigating a number of possible delivery methods for the many analytical and advisory services its customers seek to access via NYEM – some involving internal development, others involving partnerships with third-party vendors to leverage the experience of the energy management marketplace. In order to allow the possibility of multiple entities accessing and analyzing NYEM data, a common, easily-understood data structure must be established.

The data entering the NYEM system at the first two levels (monthly utility and building-level interval) is managed by a software partner that NYPA initially partnered with in 2013. Unfortunately, this legacy data has presented a number of data quality issues – there are frequently data readings missing from facilities' data sets, as well as anomalous data points that likely represent invalid data. In order for services to be delivered to customers based on this data, these issues must be addressed.

NYEM's efforts to expand the collection and analysis of deep submetering data is based upon a partnership with a separate technology vendor. As the NYEM team deploys sensors at facilities across the state and deep submetering data begins to enter the system, the complexity and volume of the data flowing into the system will increase, and a change in structure will likely be necessary.

This problem presents an interesting challenge, as it is the first customer-facing digitization effort that NYPA is undertaking, and embraces the spread of Internet of Things (IoT) devices and the volume of data they create. If executed, it will ensure the accuracy of the analyses and recommendations provided by one of NYPA's flagship customer offerings.

Previous Approaches & Hypothesis

The previously-referenced legacy data structure is currently hosted on a multi-tenant database owned by the existing software partner – as such, NYPA does not currently have direct access to raw data for the provision of analytical and advisory services. The vendor has provided extracts of the data relevant to NYPA's customers enrolled in the NYEM service, which is organized in a schema that provides a great deal of redundancy and unused database columns; this harms the efficiency of computations based on this data. Further, the vendor's platform has been unable to properly handle the quality issues of the raw data collected, harming the reliability of analytics performed on this data.

With the growth of IoT sensors and data, it is expected that an existing solution for IoT devices can be leveraged to handle the processing of data from building-level and deep submeter interval devices. One likely candidate for this solution is Project Haystack,

> an open source initiative to streamline working with data from the Internet of Things . . . [standardizing] semantic data models and web services with the goal of making it easier to unlock value from the vast quantity of data being generated by the smart devices that permeate our homes, buildings, factories, and cities.

If this, or any other data standardization method, is determined to be suitable for the data that is the key of NYEM, existing data stores may be migrated to this standard. The migration of this data will, in itself, propose quite a challenge given the inefficient data structures currently in place.

Once a common data structure is determined and implemented, a methodology must be determined for detecting and correcting incoming abberant data. The identification of these data points, as well as the proper imputation methods (possibly including normalization), will need to be configured separately but automatically for each facility. These methods will also likely need to be applied to data retroactively to ensure consistency.