RC25537 (IRE1509-005) September 16, 2015
Computer Science

IBM Research Report

Trust the Raw Data? The Importance of Applying Data Integrity Intelligence to Building Energy Management Systems

Niall Brady, Raymond Lloyd
IBM Research
Smarter Cities Technology Centre
Mulhuddart
Dublin 15, Ireland

Research Division
Almaden – Austin – Beijing – Cambridge – Dublin - Haifa – India – Melbourne - T.J. Watson – Tokyo - Zurich

Trust the Raw Data? The Importance of Applying Data Integrity Intelligence to Building Energy Management Systems

Niall Brady bradynl@ie.ibm.com
Raymond Lloyd raylloyd@ie.ibm.com

Abstract — *As the smart building solution providers industry continues to mature, the benefits of deploying Building Energy Management Systems (BEMS) to achieve and sustain energy efficiencies goals within the real estate portfolio, becomes ever clearer. However as with any high volume data acquisition environment, BEMS are susceptible to a myriad of ongoing data integrity operational issues that beyond the obvious data outage reporting difficulties, can ultimately lead to a mistrust in the underlying data sets, and therefore a degradation in the value of any applied analytics that relies on the provisioning of accurate sensor or meter data. For example how does the smart building practitioner maintain confidence in a BEMS system accuracy and effectiveness, when there are unknown and undetected data outages, intermittent sensor or meter raw data reporting, or abnormal or corrupted sensor data events? Or how does the user become aware of a problem of sensor drift that may impact the BEMS reporting accuracy, and given the growing complexity of the sensor and metering environments, how does the user quickly fault diagnose the problem once it has been identified? These questions are critical to the notion of ongoing effective use of BEMS in an organisation, and so this paper attempts to address these known industry concerns, by setting out the background of the most important of these practical data integrity issues, by defining a data integrity validation methodology, and finally through a series of use cases, demonstrate how such an approach could effectively be applied, to build trust levels, by introducing detection and timely fault diagnosis in underlying BEMS raw data sources.*

Keywords : Smart, Building, Sensor, Meter, Raw Data, Trust, Integrity, Detection Algorithms, Fault Diagnosis, Geospatial Visualisation

1. Introduction

With the proliferation of BEMS system managing energy alerting processes across the Smart Buildings community, which have the capability to detect energy wastage scenarios, more companies are building a reliance on such systems to achieve their ongoing energy savings objectives [1]. As this dependency increases, so too does the need for trust in the underlying building environmental data to ensure proper alert reporting, and an avoidance of false positives which only serve to reduce the confidence in the use and effectiveness of such systems over time. Equally maintaining the integrity and subsequent ongoing performance is paramount, particularly with weak or no direct linkage between system BEMS practitioners and system maintainers which more than likely reside with an organisation's IT department [2].

1.1 Understanding the Raw Data Integrity Landscape

In order to better understand the data integrity problem a background study of historical data integrity performance, looking at the various data quality indicators was performed. The initial characterisation exercise was carried out on a number critical data points, taken from IBM Dublin Research's Living Lab comprehensive dataset, a data warehouse which has been acquiring deep dive building parametric data for over 3 years now as part of its ongoing Smart Building's research efforts.

A data quality indicator summary from a sample interval time period from a single meter 1194 are presented in Figure 1 below to highlight the unexpected levels of variability in the data reporting.

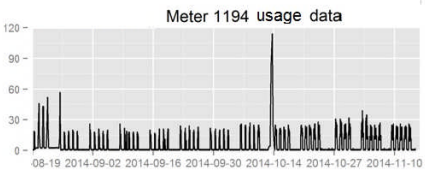

Fig 1(a) : Sample of Meter Point 1194 Usage

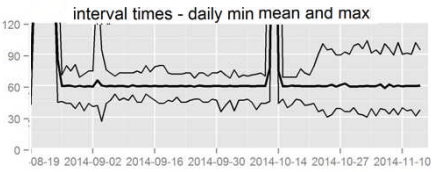

Fig 1(b) : Meter Point 1194 Reporting Interval

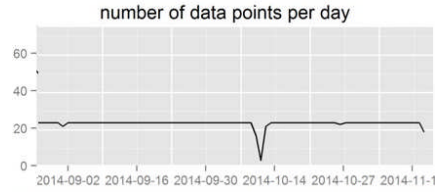

Fig 1(c) : Meter Point 1194 Daily Interval Count

Several observations are clearly identifiable from the plots in Figure 1, namely

- The interval times (the timestamps of data points as they arrive into the warehouse) vary significantly
- Over time there was a widening of the maximum and minimum reporting intervals
- There are missing days in evidence

From the study it became clear that data quality, and resultant data quality indicators are more variable than expected and therefore worthy of ongoing monitoring, and correction, and without which will inevitably have downstream impacts on output reporting and resultant analytics effectiveness and credibility.

1.2 Data Integrity Acquisition Issues

Enterprise data sources by their nature are inevitably quite dirty [3][4]. Table 1 presents a non-exhaustive summary of the typical data integrity failure modes that such smart building energy systems encounter during their normal day to day operation, which includes those experienced in the study detailed above.

Failure Mode	Operational Causes
Data Outage	Hardware failure at sensor or network level
	Battery failure in wireless networks
	Data Integrator or Warehouse Database outage
Data Intermitency	Network Traffic Delays
	Data Compression Invocation*
	Server response Time
	Random reporting delays
Data Drift	Sensor hardware failure
	Scheduled Maintenance delinquency
	Calibration failure
Data Abnormailties	Noisy infrastructure generating random data values
	Corrupted database or interface
	Sensor failure
	Sensor replacement
	Sensor reboot

** handlling sensor data compression particularly in wireless networks where battery saving algorithms are being invoked causing the sensor node only to report intermittently is a problem that needs special attention from a data integrity validation perspective*

Table 1 : Data Integrity Failure Mode Summary

And so with the many sensor and meter data failure modes, coupled with large volume sensor network estates, and the gap between system practitioners and system support teams, one can see that the management and maintenance of data quality becomes a challenging task. This paper presents a data integrity management methodology that will help the user to help manage this environment effectively, by proposing data integrity model, defining failure modes, and discussing possible failure mode detection capabilities.

Further in the paper several use cases are presented that are aimed at helping the reader to see the practical implications and value of deployment of such a data integrity modelling and tooling within their environments to help

confidently address the maintenance of a healthy raw data sensor and meter estates.

2 Defining a Data Integrity Model

It is clear that there is need to develop management solutions to the various data integrity and flow issues that have been outlined in the previous section. In the following section a data integrity model schema is proposed, and follow on detail of failure mode definitions and various failure mode detection scenarios are presented.

2.1 Data Integrity Model Schema

As outlined in data model schema presented in Figure 2 below, current BEMS positioning is more than likely taking the multiple direct building environment meter and sensor feeds or through intermediate warehousing with only rudimentary raw data error checking.

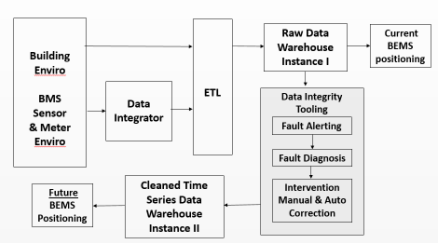

Fig 2 : Data Integrity Model Schema

In the new environment, data integrity tooling (identified as the shading area in Figure 2) would be proposed. Raw data integrity testing, including detection, alerting, diagnosis and appropriate intervention (manual and automatic) that would then deliver a proper synchronised clean and continuous time series database to future BEMS instantiations.

2.2 Data Integrity Failure Mode Definitions

Within the proposed data integrity tool provisioning outlined in data model – the first step in the process was to define appropriately categorised data integrity failure modes experienced in the field relating to raw data sensor and meter data acquisition issues. A first pass non exhaustive list of these possible categories are defined in Table 2 below

Category	Failure Mode Definition	Alert Detection Approach	Data Point Type Applicability
1	Data point outage	Detect a data point outage over a defined time interval	Meter/Sensor
2	Data point intermittent	Detect a change in frequency in data point	Meter/Sensor
3	Data point base drift	Detect a significant drift (>10%) in data point reference over a defined time interval	Sensor
4	Data point abnormal reporting	Detect abnormal sensor behaviour over a defined time interval	Meter/Sensor

Table 2 : Data Integrity Failure Mode Definition Summary

Once the definitions were clarified, putting in place the appropriate detection techniques and follow on alerting capability as to possible data integrity issues, is required.

2.3 Failure Mode Detection Algorithms

Properly crafted detection algorithms are critical for intelligently detecting and differentiating between the various failure modes, and where appropriate machine learning techniques can further be applied [5]. But given that this is a positional paper, details of more indepth statistical algorithms are not presented here, but are planned in future publications, with only some simple baseline aggregation test algorithms provided within the different failure mode categories being presented for demonstration purposes.

However, as the following use cases show, even such simple algorithms can be effective in detecting many of the data integrity issues likely to be experienced in normal day to day operations. For example, while category 1, that of data outage is relatively straightforward to detect, additional consideration is needed to

be given to areas like outage tolerances, to allow for normal occasional temporary outage windows that will occur, to avoid false alerts scenarios. Also detection frequencies need to be considered, which are dependent on the criticality of the data points in question, and upfront assessments of the potential loss and impact on BEMS analytics engines if such outages are not detected and remediated quickly.

2.4 Failure Mode Alert Reporting and Diagnosis

While there are several existing effective methods to output the subsequent data integrity alerts, the proposed method presented in the data integrity tool instantiation is that of RSS feed subscription, where alerts are delivered through the user's RSS feed reader. The feed reader could be standalone, or integrated into the users email or browser.

RSS is an existing web based technology that checks for feed updates automatically at specified time intervals. Updates are provided by subscribing to an RSS feed URL, where the updates are delivered directly to the feed reader. In this case, the URL is a web service that provides and analytical checks required e.g. checking if a specific sensor has reported in the last number of hours. Therefore RSS was considered a pragmatic solution to data integrity alert reporting and given that it can be embedded in email applications, like Lotus Notes or Outlook, it allows users to have almost continuous visibility of data integrity performance through their email application during their working day and where they can see almost immediately on their workspace when data integrity alerts are forthcoming. Setting up the RSS subscription takes only a few seconds making it another useful reporting advantage.

It is envisioned that a summary set of data integrity RSS feeds by the failure mode categories defined in Table 2 would always be present on the users email application, as an effective way to continue to monitor the smart building data flow quality and consistency.

Furthermore as an enhancement to data integrity management capabilities, an additional feature relating to the use of simple visualisation techniques that make use of geospatial data relating to the sensor and meter estate and help in quick and efficient fault diagnosis, an example of which is also presented in the next section.

Finally as per data integrity model schema presented in Figure 2 above, the final stage of the proposed data integrity toolset, that of intervention, data cleansing or data correction algorithms, manual and automated solutions for populating a clean continuous time series database for future BEMS application, and appropriate use case demonstration is to be considered for future publications. Some Research work has already been published in this space. [6]

3.0 Data Integrity Failure Mode Detection Use Case Outline

In order to demonstrate the data integrity model from concept to actual use, the following section outlines a practical set of actual deployed and operational use cases for the given defined model.

As mentioned in the previous section while there are several methods to output the subsequent data integrity alerts, the method deployed in the use case presentation, relate to relevant screen shots of RSS feed output within a Lotus Notes email environment, but equally accessible directly through a browser.

3.1 Category 1 Usage Case Example

A screenshot of a defined category 1 data integrity alert, and corresponding alert message detail relating to a data outage problem, is presented in Figure x below as it would be presented in a Lotus Notes email RSS event notification window.

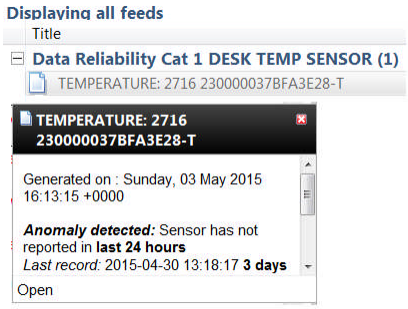

Fig 3 : Category 1 RSS Feed Alert Screen Shot

In this example the alert relates to a desk temperature sensor data outage i.e. sensor has not reported in the last 24 hours. And while such an alert is useful for detection of an outage event, of more beneficial use from a fault diagnosis perspective is the ability to overlay the sensor's geospatial data. There has been much advancements and developments in this space over the last number of years. [7]. This approach not only physically locates the non-reporting sensor but allows the user to see where it is positioned in the building and how it relates to other sensors and assets in space. Such a visualisation snapshot is presented below in Figure 4 below, where the use of the additional sensor geospatial data allows for immediate location and diagnosis of the data integrity problem, which clearly identifies a network breakage in the daisy chain infrastructure at desk DP ID 2116, which incidentally impacts also on desk DP ID's 2117 and 2118 (although RSS feed alerts for these sensor data outages were not presented in the snapshot for clarity reasons)

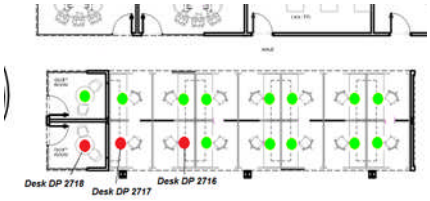

Fig 4 : Category 1 Alert Diagnosis Screenshot

3.2 Category 2 Usage Case Example

A screenshot of a defined category 2 data relating to a data intermittency problem is presented in Figure x below as it would be presented in a Lotus Notes email RSS event notification window.

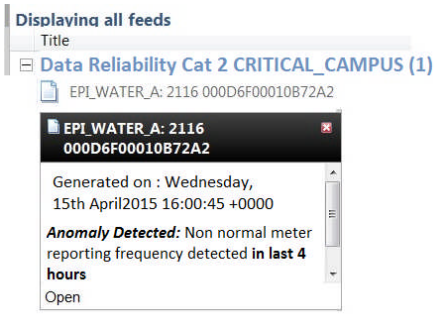

Fig 5 : Category 2 RSS Feed Alert Screen Shot

In this example the alert relates to a critical building water submeter that has resulted in intermittent meter reporting, as can be seen in Figure 6 which is quite common in wireless network environments. Algorithms to detect various forms of intermittency can be deployed but in this example for demonstration purposes a simple hourly aggregation method is deployed

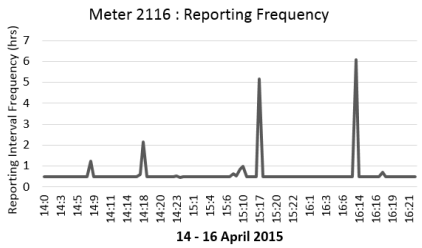

Fig 6 : Sensor DP ID 2116 Reporting Frequency

Detecting such events is important particularly for usage meter reporting – and to avoid classic meter reporting issues as highlighted in Figure 6 where such a data integrity problem is not appropriately managed and corrected.

Data outage or data intermittent outages can lead to BEMS false alerts, for example, the classic misreporting of abnormal usage behaviour levels which are due to a delay in meter reporting, is once such scenario, presented in Figure 7 below.

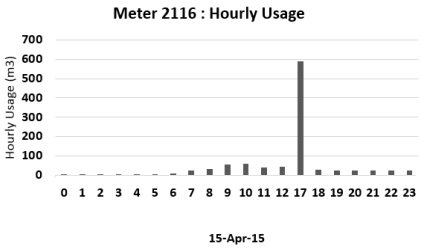

Fig 7 : Impact of Category 2 DP ID 2116 Failure

3.3 Category 3 Usage Case Example

A screenshot of a defined category 3 data integrity alert, relating to a sensor drift problem is presented in Figure x below as it would be presented in a Lotus Notes email RSS event notification window.

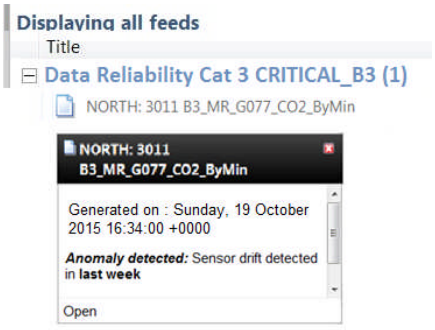

Fig 8 : Category 3 RSS Feed Alert Screen Shot

Here various algorithms can be deployed to detect drift over time – one such algorithm is that of Simple Moving Average, tested against a weekly baseline reference with appropriately applied control limits. An example of which is presented in Figure 9 below which detects an anomaly and generates an alert on the 19th October. In this specific case, Facilities Management intervention is required to replace a CO_2 sensor filter which was missed in the planned maintenance schedule.

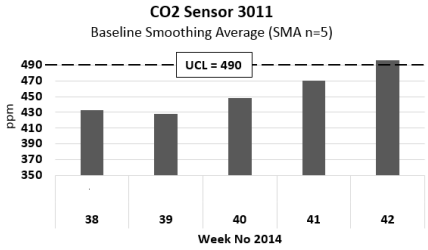

Fig 9 : Category 3 Alert Sensor Drift Detection

3.4 Category 4 Usage Case Example

A screenshot of a defined category 4 data integrity alert, relating to a possible data corruption/process shift problem is presented in Figure 10 below as it would be presented in a Lotus Notes email RSS event notification window.

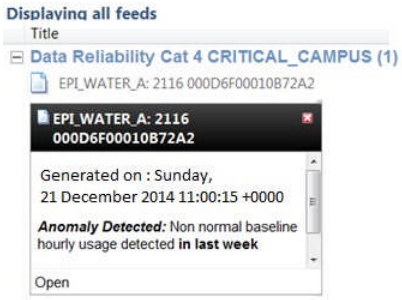

Fig 10 : Category 4 RSS Feed Alert Screen

Separating data integrity Category 4 alerts from more sophisticated multivariate regression algorithms used to detect process shifts in building performance, is difficult. So in the context of data integrity which is what is covered within the scope of this paper, the value of a Category 4 alert is seen minimally as a notification to the user to a data abnormality event, requiring investigation to see if the source of the abnormal behaviour is due to a possible data corruption issue which would need correction. Additionally however as is the case in this specific example with a building water meter, this category 4 alert, once it is confirmed that it is not a data corruption issue, is actually an indicator of a background water leak which is an additional benefit of applying such Category 4 data integrity detection approaches, one of which is presented in Figure 11 below which is based around baseline hourly average thresholds.

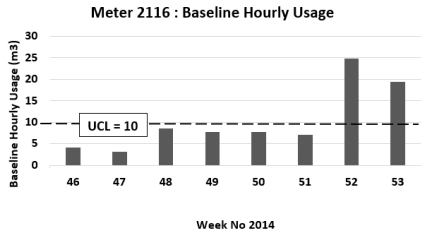

Fig 11 : Category 4 Abnormal behaviour detection

4.0 Conclusion

Given the very positive impact of BEMS system applications today, there is a growing need to put forward appropriate solutions that will properly address the myriad of data integrity issues facing the ongoing proper operation of these systems. Equally the effectiveness of the applied data analytics relies heavily on the quality and integrity of the underlying raw data from a building's sensor and meter estate. So defining appropriate data integrity failure modes and corresponding detection and correction algorithms are important first steps in helping to address this industry wide problem. Furthermore the development and testing of these first generation data integrity tools in smart building environments such as those demonstrated in the several use case scenarios presented in this paper, show the progress that is being made in addressing the raw data trust deficit that exists today. This increased confidence will in turn lead to more effective and sustainable BEMS deployments into the future.

5.0 References

1 N.Brady, Making Your Building Smarter : The Retrofit Challenge, ICEBO, Oct 2012
2 J. Sinopoli, Maintaining High Performance Control Systems,AutomatedBuildings.com White Paper, Mar 2012
3 Michael Stonebraker. Data Curation at Scale: The Data Tamer System. MIT, Oct 2014
4 Joseph M. Hellerstein, Quantitative Data Cleaning for Large Databases, UC Berkeley, Feb 2008
5 T. M. Mitchell. Machine learning. McGraw Hill series in computer science. McGraw-Hill, 1997.
6 W. E. Winkler. Overview of record linkage and current research directions. Technical report, Bureau of the census, 2006.
7 Stefan Jänicke, Comparative Visualisation of Geospatial-Temporal Data, IVAPP 2012