CELEBRATING 25 YEARS OF IES

Tech Mahindra

IoT Analytics: Data Quality challenges

Connected World.
Connected Experiences.

Abstract

Sensor data quality plays an important role in IoT (Internet of Things) Analytics. Often the organizations fail to assess and preempt this cost of quality, leading to failed IoT deployment. Organizations must ensure the completeness, consistence and correctness of the IoT data streams to enhance the quality of insights, enabling deployment of effective IoT projects and realize optimal RoI. Errors arising in sensor data are primarily due to the inherent randomness of the physical system which cannot be eliminated completely, but can be controlled and mitigated. This article focuses on discussing the impact of data quality (DQ) issues in the context of IoT Analytics and presents a high-level view of the potential sensor data errors impacting the data accuracy and the generated insights.

Introduction

In an advanced engineering & manufacturing Industry or a sensor enriched asset like gas turbine, automobile or a machine, Internet of things (IoT) is the forerunner in data generation. However, when it comes to the end-users, it is not the raw data that they find value in; rather the ease of interpretation of the information, the ability to generate meaningful insights and its representation in a schematic layout or graphical visualization for easy interpretation; makes IoT solution desirable to consumers. This is exactly why the IoT Analytics footprint is exponentially increasing, enabling a data driven ecosystem which will expose the user to real-time insights, quick & informed decisions through various descriptive, predictive and prescriptive use cases like condition monitoring, predictive maintenance, anomaly detection, process optimization etc. The real-time monitoring ability of the IoT system and the relative ease of use has opened up a whole new range of opportunities spawning complete product lifecycle - engineering, testing, manufacturing, quality assurance and aftermarket.

The sensors being the fundamental components in generating the raw data, quality and post processing of this data plays a primary role in the performance of IoT systems and its downstream applications. According to The Data Warehouse Institute (TDWI), poor data costs U.S. companies around $600 billion (€545 billion) every year. The sensors could be deployed in harsh and unpredictable environments, subjected to high temperature, humidity, dust, vibration, remote locations etc. Inevitably, this implies that the sensors are prone to failures, packet loss, rapid attrition, malfunction, malicious attacks, ware & tare , theft and tampering; causing them to produce unusual and erroneous readings. According to the report published by Hubspot , it is apparent that around 40% of IoT stakeholders face difficulty in capturing useful and reliable data. Besides, Cisco released a survey that highlights a perception gap between technology (IT) and the value delivered (Business) - 35% of surveyed IT executives perceived their IoT project as successful, but only 15% of business executives believe that the initiative was a success; primarily attributing to the capturing of invaluable data. These findings reflect the need for organizations to invest considerable efforts to ensure Sensor data quality for a successful IoT implementation.

DATA QUALITY & SENSING

One of the key steps in Data Analytics is the data preparation, which is a combination of data quality, profiling, standardization and transformations. This is an important step which constitutes up to 70% of the overall efforts, to ensure that the Statistical and ML models provide meaningful insights, enabling delivery of the expected business value and RoI. There are multiple dimensions of data quality, however here we will focus on the key quality aspects impacting the IoT Analytics.

Accuracy

This can be defined as the closeness of agreement between a measurement and reference standard. In case of a sensor network, this is the measure of whether all the devices are reporting the same data or data within an acceptable deviation.

Consistency

This parameter tells whether the data is consistent with the context in which they were generated. Environmental factors could influence the sensor data consistency.

Completeness

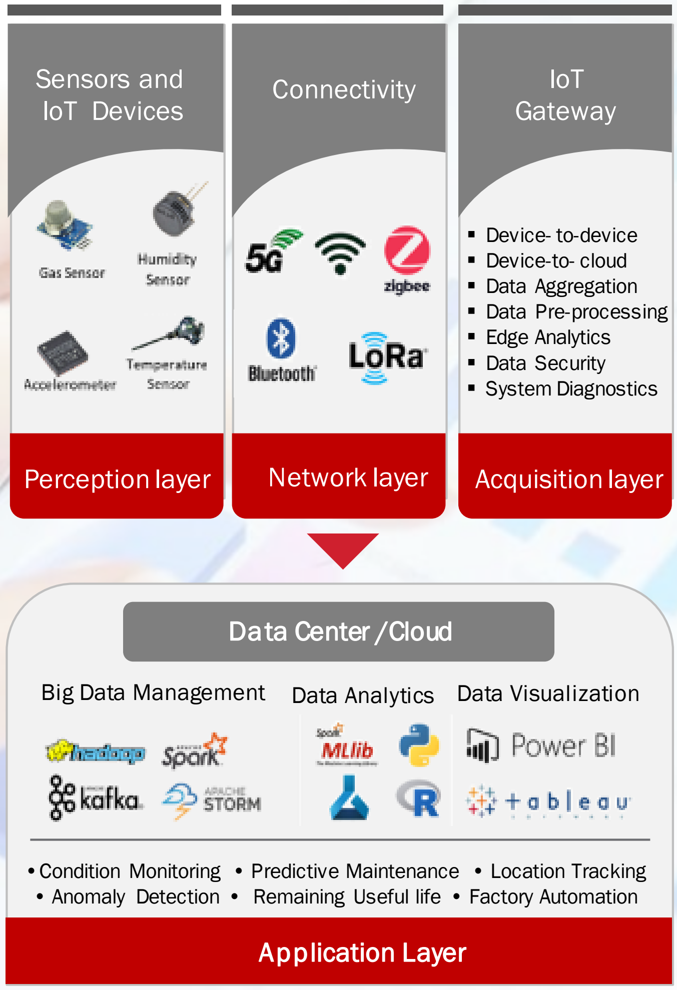
This parameter tells whether the system has accumulated all the data values or are there any gaps or missing values in a series of reported events / sensor values that should have been captured.

Time-Frame

Since the sensor data is a time-series, it is imperative to ensure that the data is captured within a reasonable time frame. If the data is coming from a wide variety of IoT devices, we need to ensure that the collective data set is synchronized

Typically, Sensors will produce Time Series data and it is imperative to consider the end objective or use case to decide on a hybrid, time or frequency domain based transformation during the data preprocessing. If the problem revolves around how a signal changes with time, then we work in time domain; where as if problem is more around how much of the signal lies within each given frequency band over a range of frequencies then we work in frequency domain. Sensors in an IoT network can be classified into ceand expects them to answer it by sending back the results to a sink node. (II) Event detection - a sensing report is triggered not by a query, but by the occurrence of an event (a fire in a forest, a gas leakage in a coal mine, or a flood in an agriculture field). (III) The focus of this article revolves around this third category called Data collection - sensor data is collected continuously over a long time period and is stored in a centralized database. The data collected & transmitted by sensors are pre-processed to be transformed into information, which is further processed through Statistical algorithms, Machine Learning (ML) models and visualization applications.

An abstraction of a typical IoT architecture is shown in Figure 1



High-quality data input is imperative to train and tune the machine learning (ML) models. The training is done on the historical data, which is deployed on the near-real-time data stream to perform the inference. For example, the unexpected failure of a mining excavator can create a considerable disruption, damage, and economic loss to both the mining company and the downstream applications like the ore refinery and the metal market. Predictive models trained on high-quality data sets would enable the reliability of the equipment by detecting potential failures before significant problems arise.

TYPES OF ERRORS IN SENSOR DATA

International Standardization Organization (ISO) defines the ‘measurement error’ as the result of a measurement minus the true value of the measure. Neither the true value of the measure nor the result of the measurement can ever be known to its exact value due to the uncertainty in the data. There are several types of measurement errors related to sensor data and some of the major ones are as below.

Outliers

Outliers are unexpected observations, which deviate from the majority of observations and are also known as anomalies and spikes. They are values that exceed the thresholds or considerably deviate from the normal behavior.

Missing Values

The transmitted sensor data are lost due to reasons such as unstable wireless connection, sensor device outages also due to its limited battery life, environmental interferences, random occurrences of local interferences, malicious attacks etc. leading to missing data.

Bias

There could be a small offset in the average signal output, which is result of manufacturing imperfections, temperature differences, electronic noise, external interference, amongst other things. This type of error would usually require sensor calibration to derive the true value.

Drift

These are sensor readings that deviate from its true value over a time period, due to the degradation of sensing material. It is often associated with electronic aging of components or reference standards in the sensor.

Noise

This is unknown and unwanted modifications a signal may suffer during capture, storage, transmission, processing, or conversion. Noise correction techniques mostly includes signal processing solutions

Constant Value

These are usually caused by a faulty sensor or transmission problems, leading to sensor readings with a constant value over time; though the readings might belong to a normal range.

Uncertainty

This reflects the lack of exact knowledge of the value of the measure and is often the result of either inherent limitation in the accuracy with which the sensed data is acquired or limitations imposed by efficiency, battery life etc. In most cases the values are small enough not to significantly affect the results.

Stuck at Fault

If the sensor reading remains zero or provides maximal value over an extended period of time, then it is called stuck-at-zero or stuck-at-one respectively, which is primarily due to the malfunctioning of the sensor

Sensor data is like any other data coming from different sources that needs cleansing, analysis and governance. It can be considered as a stream of information or values with some distinct properties, which are juxtaposed against time. We cannot simply discard information siting DQ issues as all data is destined to have some value and meaning even if it is not known at time of collection. In most scenarios, these DQ issues are attributed to external factors which are beyond human control and needs to be addressed through data science techniques. Based on various studies, practical experiences and publications related to sensor data quality; the most significant & common sensor DQ issues impacting the IoT Analytics has been identified as **outliers and missing values**. Several statistical methods & machine learning techniques have been suggested to detect, quantify and rectify these DQ issues.

Outlier Detection : Fundamental approaches to address them

Unsupervised - Determine the outliers with no prior knowledge of the data and processes the data as a static distribution, pinpoints the most remote points, and flags them as potential outliers. *Ex: Clustering, PCA*

Supervised - Model both normality and abnormality and requires pre-labeled data, tagged as normal or abnormal. *Ex: Neural Network (CNN Classification)*

Semi-Supervised - Model only the normality (or in a few cases model abnormality). It is considered semi-supervised as the algorithm is trained on normal class but recognizes abnormality. *Ex: Once Class SVM*

Missing Data: Categorization of imputation techniques

Intransitive - The variable with missing values depends on itself. Statistical techniques such as mean imputation, List wise deletion, Interpolation and extrapolation, Hot deck imputation etc. are some of the approaches.

Transitive - The variable with missing values has dependency on other variables in the data. Data Science techniques such as Regression, Associate Rule Mining, Clustering, KNN and Artificial Neural Network are some of the approaches.

However, in the IoT context, the applicability of Intransitive techniques has limitations since the IoT data has latent (hidden) dynamics and interdependency, unlike a customer data from a retail store.

It can be noticed that the larger percentage of the publications which proposed sensor data error detection and quantification methods have recommended data science techniques as the most suitable solutions for addressing sensor DQ. There are also hybrid approaches, which incorporates more than one type of method in detecting sensor data errors. We should deploy the right & lean approach to detect, quantify and rectify DQ issues, ensuring retention of maximum information, as well as considering whether the end-objective is to deploy an exceedance based condition monitoring solution or to build more complex predictive algorithms & Machine Learning models to generate deeper insights from the data.

CONCLUSION

Sensor data comes very often with a variety of quality issues, posing challenges for the data analytics and ML applications. This can occur due to diverse reasons including power outages at sensors, network issues, sensor malfunctioning or external influences. Sensor data recovery is a great challenge due to the spatio-temporal nature and their stochastic distribution. Re-transmitting the data cannot be a meaningful solution in the context of IoT Analytics as most of the IoT applications operate in real time data stream, which would mean that the delayed data is considered useless. Besides that, the computation and energy cost limit the sensor's efficiency as these devices are usually constrained in terms of battery, memory, and computational resources. High-quality data input is imperative to train and tune the machine learning (ML) models to empower IoT enabled factories & assets to make quick data-driven decisions. Poor data including inaccurate, non-consistent, missing or incomplete leads to inconvenient, time-consuming and expensive failures. Organizations must identify the sensor data quality issues and ensure the completeness, consistence and correctness of the IoT data streams to enhance the quality of insights, enabling deployment of effective IoT projects and realize optimal RoI.

ACKNOWLEDGEMENTS

You are Sensing, but are you Biased? A User Unaided Sensor Calibration Approach for Mobile Sensing
ANDREAS GRAMMENOS, CECILIA MASCOLO, JON CROWCROFT
University of Cambridge and The Alan Turing Institute

Sensor data quality: a systematic review
HUI YIE TEH, ANDREAS W. KEMPA-LIEHR & KEVIN I-KAI WANG
Journal of Big Data volume 7, Article number: 11 (2020)

Detecting Sensor Faults, Anomalies and Outliers in the Internet of Things: A Survey on the Challenges and Solutions
ANUROOP GADDAM, TIM WILKIN, MAIA ANGELOVA AND JYOTHEESH GADDAM
School of Information Technology, Deakin University, Geelong 3216, Australia;

Characteristics of IoT data quality
DAVID LOSHIN ON THE DATA ROUNDTABLE APRIL 27, 2016

A Data Imputation Model in Sensor Databases
NAN JIANG,
The University of Oklahoma, School of Computer Science

Anish Joseph

Program Manager, Engineering Analytics, Techmahindra.

Anish is an Engineering Graduate with a Post Graduate Diploma in Data Science, accreditation in International Sales & Marketing and Global Leadership. He has around 18 years of experience in various leadership roles including Analytics Consulting, Business Development & Sales Enablement, Account Management, Alliance & Strategy and Program Management; with experience ranging across several Manufacturing verticals.

Abhishek Sharma

Practice Head, Engineering Analytics, Techmahindra.

Abhishek is a Post Graduate in Computer Science, Management Consultancy and Physics. Career spawning over 20 years he has taken up various roles working on cutting edge IT Products across verticals. As part of Integrated Engineering Services, Tech Mahindra, he is responsible for Data Engineering & Science solutions for customers in Manufacturing., Healthcare, Aerospace and Automotive space.

INTEGRATED ENGINEERING SOLUTIONS (IES)

IES is a Connected Engineering Solutions business unit of Tech Mahindra. At Integrated Engineering Solutions, customers are at the core of every innovation. We align Technology, Businesses and Customers through innovative frameworks. We deliver future-ready digital convergence solutions across Aerospace and Defense, Automotive, Industrial Equipment, Transportation, Consumer Products, Energy and Utilities, Healthcare and Hi-Tech products. Our 'Connected' solutions are designed to be platform agnostic, scalable, flexible, modular and leverage emerging technologies like Networking, Mobility, Analytics, Cloud, Security, Social and Sensors, that enable launching of smart products and deliver unique connected consumer experiences, weaving a connected world. Coupled with this, our strong capabilities in Electronics, Mechatronics and Mechanical Engineering along with domain understanding and product knowledge, bring excellence to the entire lifecycle of these connected ecosystems.

Engineering AI is the competency unit of IES, focused on the descriptive, diagnostic, prescriptive, predictive and cognitive analytics of the data generated or captured from engineering & manufacturing processes and products.

Tech Mahindra

CONTACT US AT

connect@techmahindra.com
www.techmahindra.com

SOCIAL MEDIA

www.youtube.com/user/techmahindra09
www.facebook.com/techmahindra
www.twitter.com/tech_mahindra
www.linkedin.com/company/tech-mahindra