Farming Big Data Must Allow for Hills and Valleys that Affect Quality and Reliability

Ronald Hiller
BLX.io
ron@blx.io

Mahmond Daneshmand
Stevens Institute of Technology
mdaneshm@stevens.edu

Abstract - Processing and analyzing Big Data from uncontrolled environments, such as a large agricultural setting, introduces many variables that can affect the integrity and usability of the data that's collected. This paper discusses some design challenges to consider when building such a system and their effects on the reliability of the resulting information. It addresses certain important aspects of quality and reliability of Big Data, with emphasis on applications in farming and agriculture.

Keywords - Internet of Things, data acquisition, data integrity, agricultural data analytics, quality and reliability by design

Processing and analyzing Big Data from uncontrolled environments, such as a large agricultural setting, introduces many variables that can affect the integrity and usability of the data that's collected.

Garbage in, garbage out. The well-used expression is frequently applicable to the Big Data world, and it's a problem that can be compounded by the Internet of Things (IoT) when data acquisition, collection and processing are spread out geographically and face hostile environmental conditions.

American engineer and statistician W. Edwards Deming believed that if you improved quality, you automatically improved productivity, and that unclean, data raw data leads to the wrong decisions, and ultimately increases expenses and reduces revenue. If we apply this to Big Data, integrity and usability can be hampered by the quality and the reliability of the technology used to gather data from the site, as well as the infrastructure that transports it to the analysis system. If you are monitoring a large agricultural plot of land to maintain a vast irrigation system, for example, or monitoring usage and quality in a big city, any equipment regardless of size, including sensors and RFID devices, will need to be hardy enough to weather rain, wind, dirt and vibration.

But the overall quality and reliability of a Big Data system is about more than just making sure a sensor doesn't go down. It actually starts in the design phase, the so-called "quality and reliability by design." It's critical to begin by understanding what exactly needs to be measured, how often, and how accurate the measurement must be, while factoring in the reliability of the system that is collecting and processing the measurements. For example, if you are measuring the electrical current to a pump, the goal may be simply to know if it's off, running or overloaded, and may only require infrequent measurements within 30 per cent accuracy to achieve this goal. It may even be acceptable if a small number of measurements never even make it to the processing engine.

Sometimes, imprecise measurements are sufficient, but other times, more accuracy is required. When seeding a crop as part of a well-controlled agricultural field that will be monitored over the course of the growing season, the number of planted seeds per square meter may need to fall within a percentage point.

Understanding the particular goals and the information sought ultimately determines the equipment to be deployed and the corresponding cost. In many applications, IoT allows for the use of very low-cost sensors. So while it is certainly possible to measure a pump's input current within one per cent, it would require more expensive equipment. On the other hand, cheap sensors may break down easily or not reliably gather all of the Big Data required to build a complete picture of the subject.

In complex distributed data acquisition systems such as these, data quality has two dimensions: the accuracy of the underlying measurements, and the timely delivery of those

measurements to the processing system. If there's a need to process data in real-time, and the data arrives delayed, out of order, or both, it will affect the complexity of the processing, and the ultimate value that can be extracted from the data.

Sometimes data is better never than late, and sometimes better late than never. If processing needs to be done in real-time, data that arrives late is useless (and may be harmful to the analysis); but if delayed processing is acceptable, the analysis can be done across the whole (resorted into time order) dataset later without negative effect. A combination of approaches is usually required.

The mechanics of getting the measurements back for processing can be a challenge. All types of communications equipment, including Wi-Fi, cellular and Bluetooth, are feasible depending on the situation, but building and maintaining that infrastructure is a project in itself, as large farms may not be well covered with communications technology. The scale and topography make this an especially difficult challenge, although the authors have developed special protocols and techniques for handling these scenarios so that reliability and quality of data are not adversely affected.

Gathering Big Data through distributed IoT devices has changed the nature of research data collection, particularly in agricultural settings. From a utility perspective, having a lot of imperfect data can be a worthy supplement to a small number of carefully controlled measurements. For example, there can be tremendous value in a continuous collection of geo-tagged measurements across multiple farms over a wide geographic region. Traditional agricultural research stations have more controlled measurements but at a smaller scale and over a small area.

Generally, the data will be coming back in real-time or close, so researchers can get an immediate look at it, and possibly modify the experiment or data collection to explore in more detail a particular facet that has been uncovered, as well as identify any imperfections in the data collection process that might affect reliability and integrity.

The good news is we're not starting from scratch when it comes to designing and implementing systems that use IoT to gather Big Data for challenging environments such as agricultural properties because there are other industries we can learn from. Supervisory, Control and Data Acquisition (SCADA) systems used for industrial control in areas such as oil and gas, hydroelectric, chemical plants, and even spacecraft, can inspire how Big Data collection can be done reliably and accurately using IoT devices across a widespread area with topographic challenges.

REFERENCES

[1] Ron Hiller, Mahmoud Daneshmand, "Big Data, IoT: Solving the world's water woes", http://www.smartgridnews.com/story/big-data-iot-solving-worlds-water-woes/2016-02-18

[2] Mahmoud Daneshmand, Catherine Sawolaine, "Network Reliability", IEEE Communications Magazine June 1993

[3] Deming, W. Edwards (1993). The New Economics for Industry, Government, and Education. Boston, Ma: MIT Press. p. 132. ISBN 0262541

AUTHOR BIOGRAPHY

Ronald Hiller is the founder of BLX.io, which was established to advance agricultural automation. Previously, he founded Quantiva to bring statistical machine learning techniques to web performance management and has led a variety of innovative software development, networking, and performance analysis projects at Bell Labs and other companies. An IEEE veteran of 35 years, he is a member of the Big Data Community as well as a long time member of IEEE Computer, Cloud Computing and Communications Societies.

Dr. Mahmoud Daneshmand is professor of business intelligence analytics at Stevens Institute of Technology. He is an expert in big data analytics, Internet of Things/sensor and RFID data streams analytics, data mining, machine learning, probability and stochastic processes, and statistics. He has more than 35 years of teaching, research and publications, consultation, and management experience in academia and industry, including Bell Laboratories, University of Texas, University of Tehran and New York University. Daneshmand holds key leadership roles with IEEE Journals Publications as well as IEEE Major Conferences. He is a member of the steering committee for IEEE BDI; leader of IEEE Big Data Standardization; and, chair of the steering committee of the IEEE IoT Journal, among others.