Chapter 1
Data Engineering for the Factory of the Future:
From Factory Floor to the Cloud – Part 1: Performance Evaluation of State-of-the-Art Data Formats for Time Series Applications

Emmanuel Oyekanlu
Corning Incorporated, New York, USA

David Kuhn
Corning Incorporated, New York, USA

Grethel Mulroy
Corning Incorporated, New York, USA

ABSTRACT

In this chapter, the benefits that can be derived by using different existing data formats for industrial IoT (IIoT) and factory of the future (FoF) applications are analyzed. For factory floor automation, in-depth performance evaluation in terms of storage memory footprint and usage advantages and disadvantages are provided for various traditional and state-of-the-art data formats including: YAML, Feather, JSON, XML, Parquet, CSV, TXT, and Msgpack. Benefits or otherwise of using these data formats for cloud based FoF applications including for setting up robust Delta Lakes having very reactive bronze, silver, and gold data tables are also discussed. Based on extensive literature survey, this chapter provides the most comprehensive data storage performance evaluation of different data formats when IIoT and FoF applications are considered. The companion chapter, Part II, provides an extensive Pythonlibraries and examples that are useful for converting data from one format to another.

DOI: 10.4018/978-1-7998-7852-0.ch001

INTRODUCTION

In many vertical industries all over the world, numerous legacy machines generate huge amount of operational data that the legacy machines have no means of directly funneling into Industry 4.0 applications. These trapped operational data would have otherwise been valuable for generating insights that may have served to improve manufacturing processes (Accenture, n.d.; Fogg, 2020; Oyekanlu, 2018a). Also, across many industries, data scientists spend over 80 percent of their time just scrubbing data to make it fit for analytical purposes (Snyder, 2019); and by extension, to make it fit for inclusion in Gold Tables in the cloud. In some other cases, companies often find themselves unable to control and manage data at scale. This implies that due to excessive data volume, speed, and lack of transparent methods of ensuring data veracity, companies key decision makers may not be able to control, derive insight from, or operationalize generated dataset. These issues always prevent key decision makers from leveraging data for strategic smart manufacturing initiatives. In addition, most factory floors are still populated with legacy machines; and this often hamper efforts to capture, and process varied data types and deliver analytics insights with high agility.

Being able to easily integrate legacy machines into industrial IoT (IIoT) initiatives will enable manufacturers to begin to have deeper insight on factory shop floor performance from newly generated data as soon as the dataset are generated from those legacy machines. Connecting equipment and systems to FoF and IIoT systems will allow for greater visibility, remote machine monitoring, and accurate performance reporting (Fogg, 2020). In many instances, due to difficulties resulting from lack of easy means of integrating legacy machine situated at factory floors with data analytics repositories at the factory's edge computing platform, or in the cloud; analytics data lakes are often built without sufficiently factoring-in the factory floors' entire analytics needs. Such integration issue oftentimes leads to compromising the factory floor's data preparation, structure, veracity, and lineage. Sometimes, reactive data analytics and useful insights will need data inputs from both the internal factory floor and from external factory customers. However, data from customers may not be available. At some other times however, the data may be obtainable, but may not be readily available in formats that are compatible with the data formats being used on factory floor analytics system.

Naturally, people lacked trust in the data (Accenture, n.d.). This issue makes the challenges of data security and verification difficult. It also makes obtaining strategic values from data almost impossible (Accenture, n.d.). Additionally, employees, due to lack of required training skills, are often limited to using data in some specific legacy formats such as comma separated values (CSV), text (TXT) and JavaScript Object Notation (JSON). For such employees, working with data that are generated in recently available Big Data formats such as Feather, Avro, Optimized Row Columnar (ORC) and Parquet formats is a difficult proposition. Employees' lack of comfortability with these new data formats may sometimes make generating insights and deriving values from data, right from the factory shop, and floor across several FoF tools, such as the factory edge computing platform, Hadoop Big Data platform, and the company's cloud system to be quite burdensome.

In the IIoT ecosystem, enormous amounts of industrial data are generated through several different devices and systems throughout the supply chain. This includes machines, assembly lines, mobile devices, utility meters, smart sensors, automated appliances, routers, robots, and others (Databerg, 2019). The data generated from these industrial devices can be in structured, semi-structured and unstructured formats. As shown in Figure 1, generating data in structured formats allows for greater efficiency in terms of storage and data usage performance, especially during computation. For trustful analytics and