Big Data Analysis

This article is part of the Digitalization Applications 101 Technology Training Course, which provides a comprehensive understanding on the basic concepts of digitalization terminologies, technologies and its applications in the steel industry. The course was developed by the Digitalization Applications Technology Committee as an introductory course to educate industry personnel in digitalization.

Author

Ed LaBruna
Partner, Janus Automation,
Bridgeville, Pa., USA
ed.labruna@janusautomation.com

Definition

Big data refers to the handling of extremely large data sets, which require a scalable architecture to make the storage, manipulation and analysis of this information efficient.

Analytics is the systematic computational analysis of data or statistics. Organizations may apply analytics to business data to describe, predict and improve business performance. Specifically, areas within analytics include predictive analytics, prescriptive analytics, enterprise decision management, descriptive analytics, cognitive analytics and big data analytics.

These data come from countless sources: smartphones and social media posts, sensors, point of sale terminals, cameras, computers and programmable logic controllers (PLCs), among others.

In these data, great potential and opportunities for different industrial sectors are hidden. An in-depth analysis of this data can provide companies with a large amount of information that gives them a competitive advantage and improves their decision-making. In order to access these benefits, companies need qualified professionals with the necessary skills to extract valuable information, in the process known as data mining.

Big data is characterized by what is known as the 3 Vs:

1. Volume. With the current growth in data generation, it is estimated that by 2025 the digital universe will reach 175 zettabytes; that is, 175 followed by 21 zeros. The main challenge with data volume is not so much storage, but how to identify relevant data within giant data sets and make good use of them.
2. Velocity. Data is being generated at an increasingly rapid rate. The challenge for data scientists is to find ways to collect, process and use large amounts of data.
3. Variety. The data comes in different forms, mainly classified as structured and unstructured. Structured data is data that can be arranged in an orderly manner within the columns of a database. This type of data is relatively easy to enter, store, consult and analyze. Unstructured data is more difficult to sort and extract value from. Examples of unstructured data include emails, social media posts, word processing documents, audio, video and photo files, webpages, etc.

History

The term “big data” first appeared around 2005, when it was coined by O’Reilly Media. However, the use of big data and the need to understand all available data have been around for much longer.

In 2006, Hadoop was created by Yahoo!, which was built on Google MapReduce. Its goal was to index the entire World Wide Web. Today open-source Hadoop is used by many organizations to process large amounts of data.

Technology

Hadoop is an open-source framework for storing data and running applications on basic hardware

clusters. It provides massive storage for any type of data, enormous processing power, and the ability to handle virtually unlimited tasks or jobs. Among its characteristics are:

- The ability to store and process large amounts of any type of data quickly. With constantly increasing volumes and variety of data, especially when it comes to social media and the Internet of Things, this is a key consideration.
- Processing power. Distributed computing model quickly processes big data. The more computing nodes are used, the more processing power.
- Fault tolerance. Data and application processing are protected against hardware failure. If a node goes down, jobs are automatically redirected to other nodes to make sure that distributed computing doesn't fail. Multiple copies of all data are automatically stored.
- Flexibility. Unlike traditional relational databases, there is no need to pre-process the data before storing it. The user can store as much data as they like and decide how to use it later. This includes unstructured data such as text, images and video.
- Different open-source computing frameworks can be used. Its developers present it as "a general and fast engine for large-scale data processing.
- Low cost. The open-source framework is free and uses basic hardware to store large amounts of data.
- Scalability. The user can easily grow the system to handle more data simply by adding nodes. Little administration is required.

Application Example

Industrial sensors allow large amounts of information to be obtained from industrial processes. The following case is a development for a steel company where the creation of machine-learning models was sought for the prediction of defects in the product at the exit of a certain line. For this project, 18-month historical information was required from different sources of information, including information collected by the plant supervisory control and data acquisition (SCADA) system, process logs, level 2, personal digital assistants and information from other systems. The total of number of variables considered was in excess of 700, so a big data work scheme was proposed as the best one to do the analysis.

The architecture used is shown in Fig. 1. A cluster of three machines was used in a parallel processing environment. Apache Spark was used for the analysis that led to the selection of variables and information necessary for the training of machine learning models.

Figure 1

Architecture of machine-learning module.

Figure 2

Big data analysis methodology.

The selected variables and the corresponding data were used to train a machine-learning system that finally predicts defects on the line. Variables used on the analysis correspond to a different area before the actual line and variables of that line.

The majority (70%) of the data was used to train the AI system leaving a portion (30%) to be used to test the model with data that was not part of the learning and tuning of the system.

A basic conceptual schema of this process in shown in Fig. 2.

One of the key aspects is to combine business or process experts with data scientists. That combination improves the chance to reach better results.

References

1. B. Purcell, "The Emergence of 'Big Data' Technology and Analytics," *Journal of Technology Research.*
2. B. Thakur and M. Mann, "Data Mining for Big Data: A Review," *International Journal of Advanced Research in Computer Science and Software Engineering.*
3. Apache Hadoop, http://hadoop.apache.org.
4. Wikipedia Analytics. ✦