Big Data

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

Michael F. Peintinger
Managing Director, QuinLogic (SMS group), Cincinnati, Ohio, USA
michael.peintinger@quinlogic.us

The vision of a highly flexible smart factory producing customer-specific products at low additional cost in a short time to market is becoming a reality. Big data is the fuel of this fourth Industrial Revolution. The driving force behind this is the ever-increasing capability of analyzing data and the interaction with cyber-physical systems for commercial gains. This article defines the term in context of the steel industry and explores challenges as well as potential benefits.

What Is Big Data?

Everything we do is increasingly leaving a digital trace. When we browse the internet or partake in online shopping, our location and payment information is tracked and recorded, creating a profile of who we are and what we do. The same is true for material we produce. During production, a vast amount of data is captured from sensors generating a digital twin of the physical piece of material. Relating data from individual process steps generates even bigger data sets describing not only the current state but also the entire genealogy of the product. Considering the huge number of products that are manufactured, the amount of data aggregated over a given timeframe is larger than what can be analyzed by humans or commonly used software tools, and this is when the label "big data" is used.

Definition: Big Data — Big data describes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage and process data within a tolerable elapsed time.

Benefits of Big Data Analytics

The main benefit that can be gained from big data analysis is the detection of patterns and a better understanding of correlations and dependencies as well as the derivation of predictive models. Applications in the steel industry are, e.g., the root-cause analysis of defects detected by a surface inspection system at the hot mill and tracing it back to events at the caster. Training artificial intelligence (AI) algorithms on historical big data also enables predictive analytics. Monitoring incoming data in real time can trigger alarms and allows for corrective action once such a pattern is detected again. High-speed networks and integrated long-term data storage make plantwide big data integration feasible.

Types of Data

Structured data is located in a fixed field within a defined record, e.g., in a spreadsheet or a relational database. Order, customer and financial data are examples. As the name suggests, this kind of data is usually stored according to a predefined data model and this kind is also used in traditional data analysis. Unstructured and semi-structured data and its analysis is one of the main characteristics of the term big data. An estimated 80% of business-relevant information is unstructured. Examples are images, videos, uncategorized websites and documents. Another way of categorizing is by data that the business currently owns or generates and therefore has and controls access to, denoted as internal data, and data that is generated and exists outside of the business, denoted as external data. Sales statistics, human resources records, bank account transactions but also closed-circuit television data that is recorded on-premise are examples of internal data. External data is all data outside of the business and the amount is almost infinite. It can be either public (anyone can obtain free of charge with little effort) or private (behind a paywall/restricted access and usually must be obtained through a third party). Examples of external data are weather data, social media posts, geolocation and navigation services, as well as government census data.[1] The most common data in steel mills is internal and structured: order information, setpoints of equipment and data captured from sensors. Often unstructured data is transformed into structured data. Images (unstructured) from surface inspection systems are analyzed to detect, classify and categorize defects on coils, and stored in relational databases according to a data model (structured).

Challenges and Technology

Challenges for big data analytics are the capture, aggregation, validation, storage and provisioning of large amounts of data. The results of data mining and data analytics become better with the quality of data, but the larger the amount of data that is available, the more susceptible it is to flaws. Every day, live data is captured from credit card transactions, smartphones and fitness trackers that come equipped with location tracking, microphones that allow recording of conversations, cameras to take photos, and videos as well as gyroscopes and biometric sensors.

In manufacturing, the capturing computer system needs to be able to connect to a variety of data sources from different vendors (databases, sensors, programmable logic controllers, etc.). Data validation rules can help to keep the data free from flaws and avoid "garbage-in/garbage-out" scenarios in the analytics. Devices for data storage must be large and fast. Recently such systems became affordable and capable data warehouses can be implemented as an on-premise solution. Cloud storage or cloud-on-premise hybrid solutions are other options.[2]

Data Mining and Big Data Analytics

Traditional data analytics (data analytics not using big data) usually relies on human expert knowledge in combination with statistical methods. The four characteristics of big data make the analysis of big data vastly different:

- Variety (structured, semi-structured and unstructured).
- Velocity (batch, streaming and real time).
- Volume (terabytes to zettabytes).
- Veracity (cleanliness or messiness).

Definition: Big Data Analytics — Big data analytics is the process of analyzing larger data sets with the aim of uncovering useful information, test models and hypotheses. The results can lead to new revenue opportunities, improved operational efficiency, more efficient marketing and other business benefits.

Definition: Data Mining — Data mining is the process of analyzing data from different viewpoints and summarizing it into useful information. These include detecting abnormalities in records, cluster analysis of data files and sequential pattern mining using machine learning, statistical models A/B testing (also known as split testing), deep learning, natural language processing, and image/video analytics to uncover clandestine or hidden patterns.

Data Visualization

Visualization of the results of data mining and analytics helps in understanding the insight it creates. Reports based on traditional data analytics use (one-dimensional) line charts, pie charts, scatterplots and heat maps. To visualize results from big data, a vast variety of software tools supporting all kinds of charts were created. One now established method is the implementation of management dashboards, concise decision supporting user interfaces that display all mission-critical information.

References

1. B. Marr, *Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance,* John Wiley & Sons, 2015.
2. F. Provost and T. Fawcett, *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking, 1st Ed.,* O'Reilly Media Inc., 2013.
3. M. Kleppmann, *Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable and Maintainable Systems,* O'Reilly Media, 2017.

✦