PREPARATION OF INDUSTRIAL DATA FOR IMPLEMENTATION IN BIG DATA

Angel Dimitrov, Nikolina Ivanova, Rosita Nesheva, Nikolay Yanev

University of Mining and Geology “St. Ivan Rilski”, 1700 Sofia; E-mail: nikolay.yanev@mgu.bg

ABSTRACT. In recent years, Big Data has become the main tool for companies to improve their efficiency. Modern corporations are forced to reorganise their businesses due to the incredible amount of new information. The mining industry is not an exception. A serious challenge in integrating Big Data is the preparation of the data itself. In this report, we will examine the basic stages of preparing industrial data for its implementation in Big Data. These stages are related to studying the data sources, data integrity assessment and the data formats. We will also review the implementation of Big Data in mining industry.

Key words: Big Data, mining industry, industrial data

ПОДГОТОВКА НА ИНДУСТРИАЛНИ ДАННИ С ЦЕЛ ИЗПОЛЗВАНЕТО ИМ В BIG DATA

Ангел Димитров, Николина Иванова, Росита Нешева, Николай Янев

[1]Минно-геоложки университет „Св. Иван Рилски“, 1700 София

РЕЗЮМЕ. През последните години Big Data се превърна в основно средство за подобряване на ефективността на компаниите. Съвременните компании са принудени да преобразуват бизнеса си заради неимоверно нарастващия обем на информацията. Минната индустрия не прави изключение от тази тенденция. Сериозно предизвикателство пред внедряването на Big Data е предварителната обработка на използваните данни. В доклада ще бъдат разгледани основните етапи при подготовката на индустриални данни с цел приложените им в Big Data. Тези етапи са свързани с изследването на източниците на данни, оценка на качеството на данните и форматите за описание на данните. Ще се обърне внимание и на приложението на Big Data в минната индустрия.

Ключови думи: Big Data, минна индустрия, индустриални данни.

INTRODUCTION

The latest industry paradigm requires real-time data collection (and storage), processing, and analysis. Smart factories must be able to gather as much data as possible in order to make the right decisions and actions possible, as well as to analyse potential faults and alarms in order to avoid production loss.

Machines must be able to assign functions to other machines or seek instructions from a higher level or layer. Aside from the theory concept, real-time capability denotes high-quality technological solutions.

Industrial data has evolved into industrial big data, according to a recent General Electric whitepaper (Shefali, 2017), with the volume of data produced every minute. According to the survey, 152K data samples per second, 9M data per minute, 545M per hour, 4B samples per turn, 13B samples per day, and 4T samples per year are produced in a single machine in the manufacturing of baby products.

This data only applies to one machine, additional information such as geolocation, faults, alerts, and process logs may be even some. There is a huge need to process this data, turn it to information, and use it in a constructive and predictive way, generating knowledge value and building digital trust, as it comes from various sources and in various formats.

Big data is viewed as a collection of approaches, methods, and tools for working with huge volumes of data collected form different sources which sometimes are generated in real-time. The data is “big” not only for its volume, but also for its diversity and complexity in its core. Data management and processing is a difficult task when using traditional methods for manipulation and storage (multidimensional arrays, text files and/or relational databases).

According to an IDC research, 44 zettabytes of data were generated last year and the expectations are that the number will grow to 163 zettabytes by the year of 2025 (SAP, 2017).

In a research done by several global mining companies, it becomes clear that Big Data analyses rapidly develop the efficiency of ore mining, analysis, transport, and handling. Using Big Data makes procedures faster and more efficient on every level.

Big Data analysis mainly stimulates better asset utilisation and higher productivity and deals with material flow delays. This is done by sensors embedded in mining operations. These sensors generate huge amounts of geoscience, status, and operational data in real-time which is due to the improvement of WiFi, 3G and 4G speed. Real-time data is collected from the moment of the preparation for mining until the final ore transportation. This way, the data is processed and distributed much faster.

Data preparation

When preparing industrial data for implementation in Big Data, several key components should be considered – source, model, format, and quality of the data.

Data sources

Contemporary platforms for Big Data assimilate huge amounts of different inputs from different sources in real-time. These sources can be divided into three basic categories: (Violino, 2018)

- Data generated from social media – comments, blogs, videos, pictures etc. Social media users grow exponentially; therefore, data generated from social media also grows each year.
- Data generated from business information systems – some of the bigger organisations process thousands of transactions per day including market transactions, e-commerce, client information, credit transactions, etc. The data generated from business information systems more and more often includes semi-structured data like pictures and comments, which makes them even more difficult to manage.
- Data generated from machines/sensors – IoT machines and devices have embedded sensors and detectors which can send and receive digital data. IoT sensors help organisations to collect and process machine data from devices, vehicles, and equipment. The number of the devices that generate data also grows – meteorological, road sensors, surveillance cameras, satellite pictures etc.

Data sources in mining industry can be classified as direct and indirect. Direct data sources, like global positioning system (GPS), conventional geodetic measurements and prices monitoring, represent data received from devices especially designed for data collection. Indirect data sources are data generated as a side product from mining operations, like drilling and explosives work, control systems and railway tracks.

With the development of information and communication technologies, data sources in mining industry will expand. Using sensors and intelligent microchips to measure system efficiency is rapidly increasing. It provides indication for potential accidents. Systems for unmanned aerial vehicles can be used to explore new mining areas, landscape surveillance, and measurement of tailings ponds. That is why the promotion of the Big Data concept is of great importance for the sustainable development of the mining industry.

Data structures use raw data for storage and computations; but for the needs of the data aggregation, it is vital that sets of logically similar values be used for summariing and analysis.

The different types of data that industrial processing generates and/or works with, as well as the formats in which those data are stored.

There are five types of data in general as can be seen from **Error! Reference source not found.**.

Data model

Today, the classical relational model and the currently gaining popularity NoSQL model are mainly used as data models. Traditional relational database management systems (RDBMS) are multifunctional. They provide solutions for many different types of problems. Unlike them, NoSQL and NewSQL systems are focused on solving a specific problem like operations for short term online transaction processing (OLTP) (Yanev, 2019).

Table 1. *Types of data*

Data type	Meaning	Examples
Observational	Captured in real time. Cannot be recaptured.	Sensor readings, sensory (human) observations, survey results
Experimental	Data collected under controlled conditions, in situ or laboratory based.	Gene sequences, chromatograms, spectroscopy, microscopy
Derived or compiled	Reproducible, but can be very expensive.	Text and data mining, derived variables, 3D models
Simulation	Results from using a model to study the behavior of an actual or theoretical system.	Climate models, economic models, biogeochemical models
Reference or canonical	Static or organic collection [peer-reviewed] datasets.	Gene sequence databanks, chemical structures, spatial data portals.

The biggest companies in the mining industry are still using relational databases (table 2) but NoSQL models are also becoming increasingly popular (table 3).

Table 2. *Mining companies that use NoSQL*

Company	Headquarters	Turnover/revenue
Lhoist	Belgium	2.2 bln. EUR
Barrick	Canada	9.7 bln. USD
Nevada Gold Mines	USA	4 bln. USD
Epiroc	Sweden	4.2 bln. USD
Atlas Copco	Sweden	12.2 bln. USD
Abbott	USA	34.6 bln. USD

Table 3. *Mining companies that use SQL*

Company	Headquarters	Turnover/revenue
Glencore	Switzerland	220.1 bln. USD
BHP Billiton	Australia	43.6 bln. USD
Rio Tino	Great Britain	40.7 bln. USD
China Shenhua Energy	China	37.6 bln. USD

Data format

Regarding data formats, the most popular are CSV, XML, and JSON (Fig. 1).

CSV (comma-separated values) – stores bidimensional data, list of elements separated by a comma or another delimiter. It is compact and easy to work with when used for tabular data. It is not very versatile and does not support hierarchical or object-oriented structures. CSV has been used to store data since the dawn of the first business computers.

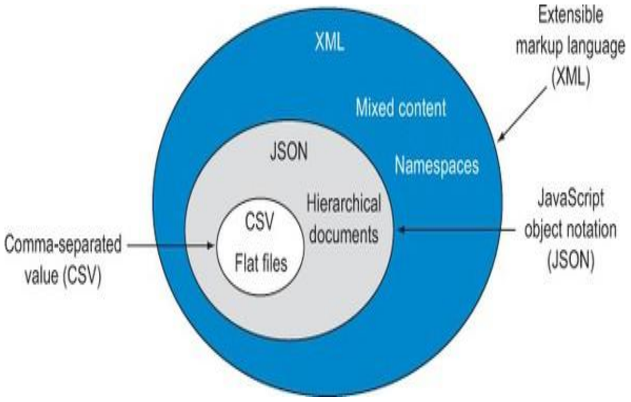
XML
Extensible markup language (XML)
Mixed content
JSON
Namespaces
Hierarchical documents
CSV
Flat files
JavaScript object notation (JSON)
Comma-separated value (CSV)

Fig. 1. Comparison between CSV, JSON and XML

JSON (JavaScript Object Notation) – an open-source file format that stores data in key-value and array formats. It is based on JavaScript and widely popular. It was developed in the early years of the 21st century by Douglas Crockford. In 2013, the language was structured under the ECMA-404 standard. It supports hierarchical data structures. It is heavier than CSV but lighter than XML. As it is based on JavaScript, it is easily interpreted and works on the client. Not all programming languages have libraries for it and, if not used correctly, JSON could pose a security threat. Extracting values is relatively easy. JSON derivatives which can be used in the mining industry are: GeoJSON and TopoJSON. Databases that use JSON are: MongoDB, CouchDB, DocumentDB, MarkLogic, OrientDB, RethinkDB, Riak, etc.

XML (Extensible Markup Language) – a meta language in which information is stored in simple text files and the structure is defined by additional (meta) data. XML’s first specification was first developed in 1998 by the World Wide Web Consortium (W3C). It has a big set of libraries and developers are familiar with it. XML can be validated on both client and server. It supports hierarchical data and can be transformed into other formats. Relatively large but with correct use of attributes, namespaces and closing tags, its size can be significantly reduced. Since it is a markup language, XML has the potential to present data. It supports many different character encodings. Derivatives of XML that are used in the mining industry are: GML, XMML, CMXML, SVG. Databases which use XML are: eXist-db, MarkLogic, etc.

Data quality

From the point of view of information technology, data quality is defined as a set of qualitative or quantitative criteria. The key features of data quality are accuracy, completeness, consistency, uniqueness, relevance, and timeliness (Anastasova, 2019).

The following indicators must be considered so that a successful strategy for improving data quality can be developed and executed. (Naydenova, 2009):

Data completeness. Completeness as a measurement for data quality is dependent on the business environment, the specific task, and the expected results. It refers to the presence of all the needed data attributes. Most problems with data completeness originate from the fact that companies often have not done the proper preliminary analysis, have not specified the mandatory data, which is needed, and they end up storing important attributes in their databases too late which leads to many empty or incomplete records.

Data accuracy (precision). Even if this component of data quality looks obvious, it is of great importance. When the data is collected, it is often integrated in different systems which transform it according to different data standards (Cai, 2015). This can lead to loss of accuracy when transforming data types or due to the sensitivity of the tool used to make the initial measurements. Therefore, data must be corrected promptly to be accurate. There are two aspects – the first one states that the data input must be correct, without any typos. The second one states that the data must be presented in a consistent and unambiguous form. The preliminary analysis is very important. Before any data input or manipulation, each field must be studied carefully to make sure that its values are meaningful and to ensure that there are no incorrect records in the database as they can lead to wrong decisions.

Data compatibility. In the same organisation data can be stored in different data warehouses with different structures and from different sources. Various types of software are used to manage this data, along with others to manage employees, customers, financial data, etc. Compatibility and synchronisation of this data is crucial.

Data validation. For specific data attributes, it is not enough to be complete and accurate. Without additional validation models, this data cannot be processed. For example, if the attribute is an e-mail it must have “@” in it. The validity check can easily be done in many cases with regular expressions. After invalid data has been identified, it becomes a problem for the data completeness which can be solved with the methods described above (Ziad, 2020).

Data relevance. The older the data is, the more irrelevant it is. Relevance is measured in time needed for the generated data from an event to be processed and made available in the database management systems according to the chosen goals. The timeliness of the data set depends on the data integration conveyor which is responsible for the data generation. This can be done in real-time when data is presented very soon after the event’s occurrence, or it is processed in sets, which means that data is “frozen” until the next refresh. Changes in the way this conveyor works give the opportunity to access data in real-time or to react to recent events (Farnworth, 2020). It is important for organisations to work with relevant data to avoid making critical decisions based on data that is weeks or months old.

Data uniqueness. Data uniqueness suggests that in data sets, there are no duplicate values and that for every event, there is only one record. Duplicates undoubtedly lead to incorrect data processing and making the wrong decisions. To avoid this problem, it is necessary to choose unique identifiers and proper primary keys.

Data integrity. There must be precise and clear rules of how to achieve data integrity according to its specification. Also, when modeling the data warehouses, it must be guaranteed that the data could be integrated from one system to another, so that it does not scatter attributes or lose important data.

Conclusion

Using Big Data in mining industry makes the workflow easier by:

- Guaranteeing steady data flow from the very beginning of the extraction to the processing factory.
- Minimising the time between operations – unplanned maintenance, delays, idle time, and losses.
- Increasing ore mining.
- Providing current results from the analysis.
- Allowing fast decision making in production processes from extraction to the factory deliveries.

For all of this to happen, industrial data must be prepared for integration in the appropriate information systems. This process requires evaluation of the sources and the quality of the data and analysing the models and formats of the data used.

References

Anastasova, Y. 2019. Technologies for ensuring data quality and security in industrial information systems. - *Journal of Mining and Geological Sciences*, 62, Number4, 5 – 8.

Cai, L., Z. Yangyong. 2015 Challenges of Data Quality and Data Quality Assessment in the Big Data Era. - *Data Science Journall*,14, DOI: 10.5334/dsj-2015-002.

Farnworth, R. 2020 “The Six Dimensions of Data Quality — and how to deal with them”. Available on: https://towardsdatascience.com/the-six-dimensions-of-data-quality-and-how-to-deal-with-them-bdcf9a3dba71.

Naydenova, I., Z. Covacheva, K. Kaloyanova. 2009 A Model of Regular Sparsity Map Representation. - *Analele Stiintifice ale Universitatii Ovidius Constanta-Seria Matematica*, 17 Issue, vol: 17, issue: 3, 197-208, ISSN (print):1224-1784, ISSN (online):1844-0835.

SAP. 2017 The importance of Big Data analytic. Available on:https://www.sap.com/latinamerica/insights/what-is-big-data.html.

Shefali, P. 2017 "General Electric". [Online]. Available: https://www.ge.com/digital/sites/default/files/download_assets/Unlocking-Business-Value-Through-Industrial-Data-Management-whitepaper.pdf.

Violino, B. 2018. Many executives lack a high level of trust in their organisation's data, analytics, and AI, Red Ventures Znet. Available on: https://www.zdnet.com/article/most-executives-dont-trust-their-organizations-data-analytics-and-ai/.

Yanev, N. 2019. Contemporary approaches in storage and dataprocessing. - *Journal of Mining and Geological Sciences,* 62, Number4, 96 – 99.

Ziad, Z. 2020. “Does Your Organisation Measure Up to the 6 Critical Dimensions of Data Quality?”. Available on: https://dataladder.com/measure-up-to-6-critical-dimensions-of-data-quality/.