Models and data quality in information systems applicable in the mining industry

Yordanka Anastasova[1,*], and *Nikolay* Yanev[1]

[1]University of Mining and Geology „St. Ivan Rilski“, Department Mathematics and Informatics, Studentski Grad, “prof. Boyan Kamenov” Str., Sofia 1700, Bulgaria

Abstract. The purpose of this article is to present modern approaches to data storage and processing, as well as technologies to achieve the quality of data needed for specific purposes in the mining industry. The data format looks at NoSQL and NewSQL technologies, with the focus shifting from the use of common solutions (traditional RDBMS) to specific ones aimed at integrating data into industrial information systems. The information systems used in the mining industry are characterized by their specificity and diversity, which is a prerequisite for the integration of NoSQL data models in it due to their flexibility. In modern industrial information systems, data is considered high-quality if it actually reflects the described object and serves to make effective management decisions. The article also discusses the criteria for data quality from the point of view of information technology and that of its users. Technologies are also presented, providing an optimal set of necessary functions that ensure the desired quality of data in the information systems applicable in the industry. The format and quality of data in client-server based information systems is of particular importance, especially in the dynamics of data input and processing in information systems used in the mining industry.

Introduction

Modern databases, which are the basis of information systems, operate with different data models. The aim is to represent them to describe the described real objects as accurately as possible, and the challenge is at the same time for the data form to allow their online processing in real time (Fig. 1).

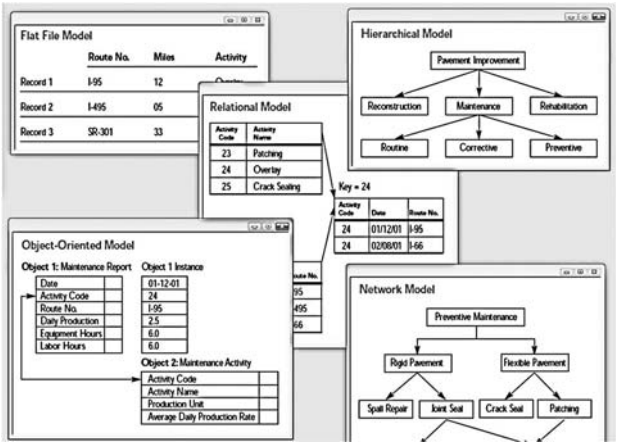

Fig. 1. Database models.

In general, the evolution of database management systems (DBMS) can be described in three stages:

- Navigation systems - those were used in the 1960s and represented hierarchical and network models of data description;
- Relational - those were created in the 1970s and are used to this day. They are based on set theory and on relational algebra. The objects are described in the form of two-dimensional tables allowing for connections (relations) between them. They use the SQL programming language;
- Post-relational - this category comprises a wide variety of data description methods. The object-oriented model was introduced in the 1980s, and the NoSQL and the NewSQL models have become popular in the recent decade.

Over the past 10 years, NoSQL and NewSQL models have become popular, which are targeting for a specific problem, such as short-term OLTP (Online Transaction Processing) operations.

At the same time, the information in them should be as high-up to date, accurate and sufficiently comprehensive as possible to enable maximum effective solutions.

In the mining industry, the processes are in continuous dynamics, mutually connected, and each of them can affect the operation of the whole system, depend also on the natural resources and require large investments in resources and funds. Moreover, the majority of the tasks in the modern mining industry are characterized by a pronounced uncertainty, non-linearity and multifactorial. [1] In this case, an unfortunate decision taken based on poor quality information can lead to huge losses for the particular enterprise.

In order to avoid such situations, it is especially important to obtain quality data, i.e. data meeting the

* Corresponding author: yordanka.anastasova@mgu.bg

requirements of the specific information system. The format and quality of the data is directly dependent on the purposes for which they will be used [2], and from the point of view of information systems the format and quality of the data is part of the whole process of data management.

Modern data models

Standard relational databases were not designed to handle the scale (Big Data), flexibility and real-time operation that are required by modern information systems. In addition, they do not take full advantage of the low cost of storage devices, nor of the high performance of the machines we have at our disposal nowadays.

NoSQL encompasses a wide variety of database technologies that have been developed in response to the increasing amount of data stored for users, objects and products, the frequency with which this data is accessed, as well as the need of high performance in their processing.

The first NoSQL software appeared in the early 21st century: MongoDB (2009), Redis (2009), Cassandra (2008), etc. Today there is a wide variety of data models used in NoSQL systems. The most popular are shown in Fig. 2:

- Key-value: here, information is stored in records of the "key-value" type and complex data structures, including XML, can be stored as "value". The search is performed via a key. Dynamo, Riak, Azure, Redis, Cache are such NoSQL databases;
- Document: the work data and related information are stored in documents, most often in the XML or JSON formats. This model resembles the key-value model, with the "value" being the document itself. Such models are MongoDB, CouchDB, Raven, BaseX, etc.;
- Wide column stores: again, a "key" is used, but this may point to a family of columns. Each record can have a different number of columns and can be placed in other columns called super columns. BigTable, Hbase, Cassandra, Accumulo are popular examples of column family database software;
- Graph: this works with graph structures. Data is modeled as a network of links between particular elements. Neo4J, Allegro, Virtuoso, Bigdata are such models;
- Multidimensional: Globals, SciDB, Minim DB.

Among the main advantages of NoSQL databases are:

- flexibility - they do not work with static schemes;
- scalability - they also allow for horizontal scaling;
- facilitated database transfer across multiple servers.

The biggest drawback to NoSQL systems is that they are not transitive.

Typically, NoSQL databases are used in distributed systems information systems, where the emphasis is on productivity in processing large volumes of data, which makes them applicable to information systems in the mining industry.

In such systems, the CAP theorem (Brewer's theorem) is observed [3]: "In a distributed system, at most two of the categories can be satisfied:

- Consistency (C): all database clients see the same information, even with competitive updates;
- Availability (A): all database clients can access any version of the information;
- Partition tolerance (P): The database can be partitioned over multiple servers.

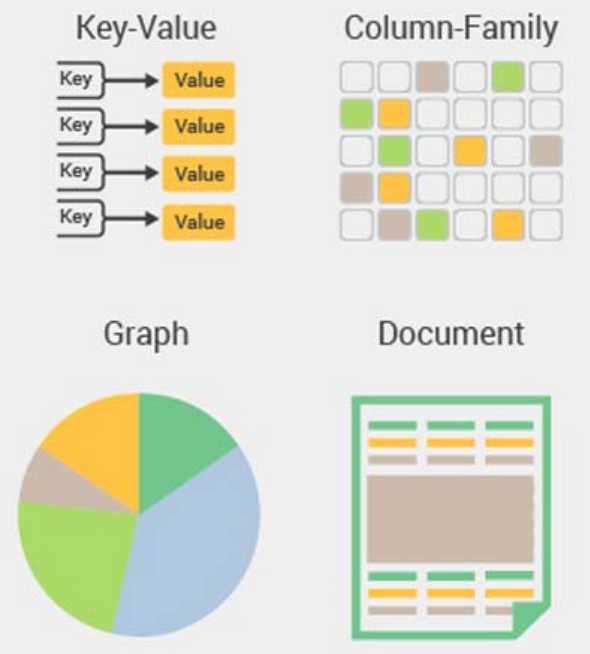

Fig. 2. Popular NoSQL models of data.

The simultaneous provision of all three guarantees is impossible (Figure 3).

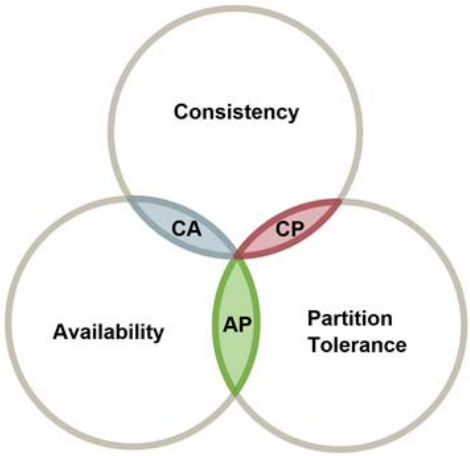

Fig. 3. CAP theorem.

The theorem proves that only two of the three pillars can be used to create such a system, i.e. we may have a system with high consistency and expandability, a system with high data availability and expandability, or a system with high consistency and high availability, but without expandability.

Most NoSQL databases operate on the BASE (Basically Available, Soft-state, Eventual consistency) principle: choosing availability and partitioning at the expense of consistency and looking for the fastest and

most reliable synchronization among individual servers.

NoSQL databases still have limited application in specific areas, but the fact that they are used by IT giants like Google, Facebook, Amazon, and LinkedIn is a proof about their potential.

Numerous comparative analyzes of the performance of RDBMS and NoSQL have shown that, in general, NoSQL systems perform better when recording, deleting, and updating Big Data sets than are common to information systems used to manage mining processes.

NewSQL databases have been talked about for the last few years. The term NewSQL was first proposed by Aslett [4]. These are actually databases that combine the advantages of SQL and NoSQL databases (Fig. 3), as NewSQL are transitive and horizontally and vertically extensible.

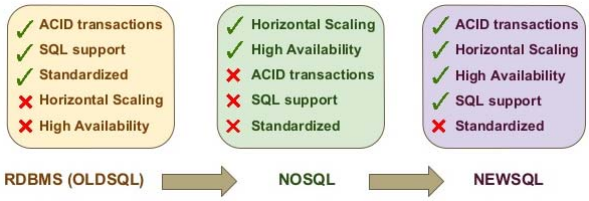

Fig. 4. Comparison between SQL, NoSQL, and NewSQL.

The products described as NewSQL databases are very diverse, but three main types can be classified:

- SQL engines: highly optimized storage engines for SQL (examples MySQL Cluster, Infobright, TokuDB);
- New architectures: databases that were designed to operate in a distributed cluster (examples Google Spanner, Clustrix, VoltDB, MemSQL);
- Transparent sharding: they provide a sharding middleware layer to automatically split databases across multiple nodes (ScaleBase).

The coal of NewSQL databases is to provide a high-performance and affordable solution for processing large volumes of data without compromising data consistency and high-speed transaction capabilities, making them very efficient and applicable to some processes in the mining industry, which are almost completely automated.

They are best used in the control of enrichment processes, where the data are very high frequency - the sensors (express analyzers) continuously provide information at intervals of up to 2 minutes, which requires the supply of appropriate reagents to obtain the desired content of ore concentrate.

Although in recent years many analytical comparisons have been made between SQL and NoSQL databases [5, 6], today the choice of which data model to use is determined mainly by the specific conditions and tasks.

Data quality assurance technologies

Data quality is a characteristic that shows the extent to which they are analyzed and meet the needs of the business to make informed and effective decisions. From an information technology perspective, data quality is part of the whole data management process.

The criteria determining whether we operate with quality data can be considered according to the requirements of information systems and from the point of view of their users.

The requirement for the use of high quality data in information systems is that they meet at least five main criteria [7] - completeness, accuracy, validity, consistency and timeliness (figure 5).

Unlike standard data collection (on paper), information technologies make it possible to ensure the completeness of data by using functions that allow the input and digital storage of information only, where all attributes for the object, activity etc. have been introduced.

To ensure full quality data, additional features are introduced that check not only the correctness of the data provided but also the exact implementation of the data entry format defined by the particular information system.

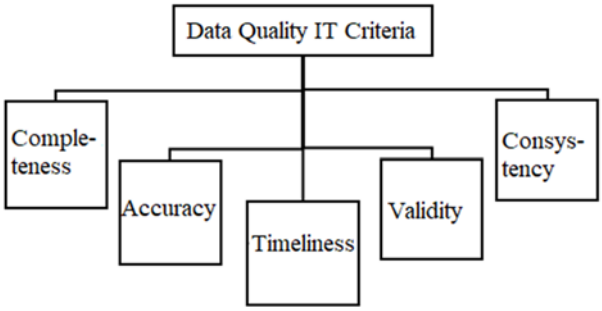

Fig. 5. Main criteria for data quality.

Accuracy of data criterion suggests that incoming data in the information system are correct and fully reflect the depicted object, process, etc. To avoid the risk of inaccurate data submission, the interference of the human factor in this activity should be minimised already at the design stage of a specific information system. Unfortunately, this is almost impossible, and therefore, the implementation of this activity must be done by competent and well trained specialists.

To ensure the data accuracy, especially in cases of a high volume or a continuous stream of data, additional features are being set in the information systems which check for inaccuracies at every step and eliminate admission of such.

The criterion validity of the data determines how data values are correctly measured according to the pre-set conditions. If we have received invalid data, this means that there is a problem in the process of collecting the data.

When you get values for specific data that are beyond the limits of the usual, it does not always mean that they are invalid. In such a case the values should be re-checked. In the flexible information systems this problem is easily solved by altering the defined limits for measured values and incorporating new values.

In information systems, especially in those with longer term of use, there are data about the same object, process, action, etc., that are introduced at certain periods and have different values. In other words, there are different versions of the data for an object or process.

The consistency criterion ensures that the data in the various versions are saved in the same format and in the most important this data format is not changed during processing.

In order for an adequate and efficient decision to be made, it is important that the data we need to analyse should be timely - i.e. there is no time interruption of the incoming data stream for various reasons.

The timeliness criterion is especially important in industrial systems, which manage continuous production processes because the lack of data for a specific segment of time can lead to incorrect management decisions.

From the point of view of data users, the criteria for data quality can be considered conditionally in four major groups - availability, usability, comprehensibility and security (figure 6).

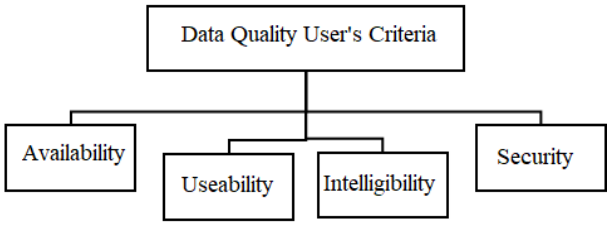

Fig. 6. Users' criteria for data quality.

Availability of data means that in every moment, when appropriate, users need, to have access to them and they are always available.

In information systems, basic characteristics about the availability of data are accessibility, authentication, authorisation, and timeliness of equivalence.

In the client-server technology used by modern information systems by design levels of access to a specific collection of data are defined and an access level is assigned to every particular user, which determines what kind of data to be submitted. Various collections (databases) available for specific levels may exist. An example in this respect are geographical information systems [8], where there is different accuracy (data quality) depending on the type and level of access.

Depending on the specific level of access, it is verified if that user has permission (authentication) to use the information resource (i.e. to a lower or higher quality data). Authorisation is performed by the information system itself, as it gives the user rights to perform the permitted set of actions.

Since a large part of the information systems, including industrial ones, are used by many users, and different users can enter information, the equivalence of data is of particular importance. It measures the extent to which equality (equal values) of the same data is guaranteed.

The timeliness guarantees users that data are timely (as timely as possible), which is essential in making effective decisions.

The usability criterion means that data incoming in the information system from different sources can be processed and analysed.

The data characteristics that determine their usability are documentation, validity, applicability, precision, flexibility and interactivity.

The most important feature of usability of incoming data is their ability to be converted into a digital format by the information system, i.e. they can be formalised by meeting their set conservation model [9].

The validity of the data is determined by comparing the relevance to the requirements set for the specific information system.

Applicability is a characteristic that determines how much data can be processed and analysed in support of specific targets. In order to have adequate solutions taken on the basis of the data it is necessary to have precise data – i.e. they need to have values in the range specified in the information system. Thus, the level of detail of the data, which is required by different groups of users and management levels, is defined. The too high level of refinement and detail of data often leads to difficulties in the operation of information systems and it is therefore necessary to find a level of balance that satisfies both these two characteristics.

Data security assures the users that they are provided with the requested information in an accessible form and the data origin is guaranteed.

The main features ensuring data security are standardisation, reliability, comprehensiveness, integrity, objectivity, comparability and stability.

Standardisation ensures that the data submitted and processed correspond to the rules set in each information system, which in some cases are valid for different information systems that share and exchange information. This data feature is set in the design process of the relevant information system and is monitored throughout its entire life cycle.

Nowadays, the reliability of data is a key feature not only for information systems but also for society as a whole. They give confidence about the source of the data and its reputation, which determines the degree of confidence in the data. Comprehensiveness is a complementary feature that determines to what extent the data is satisfactory and covers the user's request. Data integrity is one of the most important features of data, especially in an insecure environment such as the Internet, because it ensures that changes to data are made only by authorised users. The objectivity feature of the data ensures that the data are not modified under the influence of human emotions, i.e. only the specific facts about the data are reflected.

Naturally, one of the most important features of data is the ability to be permanently stored and accessible over a long period of time to ensure its stability.

Information technology uses many different techniques that guarantee the use of only high quality data. As particularly critical in this regard we can identify technologies that provide standardization, profiling, matching, control and cleaning of real-time data (Fig. 7), which are particularly important in ensuring the operation of information systems in the mining industry.

Data standardization is a technology that operates on the basis of established rules and criteria, ensuring the desired quality. The received data goes through various

transformation processes in order to comply with the rules set in the specific information system. [10] Additional functions must be included, allowing automatic correction in the presence of minimal inaccuracies or rejection of data in case of significant discrepancies.

Data standardization is especially important in ERP systems, where information comes from different sources. This technology is also essential when we have data exchange between different information systems with diverse databases, such as those used in the management of various processes in the mining industry.

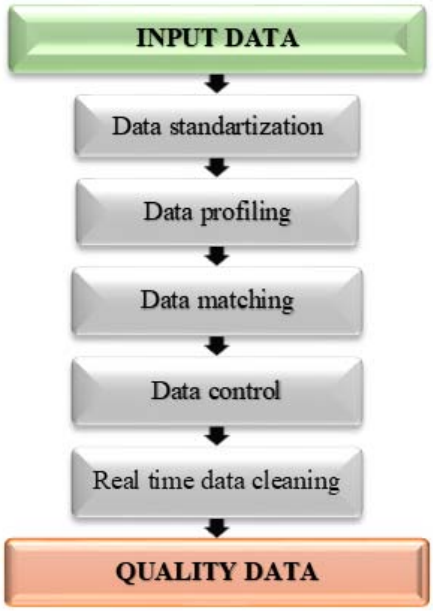

Fig. 7. Data quality assurance technologies.

Data profiling is a technology used to analyses the content, quality, and structure of output data, and is used in various criteria for data quality, such as determining their accuracy and completeness.

Data sources are considered, with an initial assessment of the data to identify potential and actual deficiencies. The goal is to find out the wrong areas in the data organization that can be found in user input, interface errors, data corruption when transferring, and so on. The use of this technique significantly improves data quality.

Data matching aims to find records that relate to the same object, process, individual, and so on. It can be done in many different ways, but the process is often based on algorithms or programmed circuits, where processors perform sequential analyzes of each individual data set, comparing it to each individual part of another data set or comparing complex variables for finding strings containing specific similarities. In the paper of I. Getova [11] present innovative test and evaluation model which gives a probability assessment obtained learning of the lectured material by the learners and provides information on how much the learner perceives the new material and how well the lecturer has presented it in a way can be readily understand it of students. The analysis in this article is performed on a set of data collected by two universities in Bulgaria using the IBM statistical analysis program - SPSS.

Data control is a set of technologies that monitor changes in data quality over time and report deviations from predefined quality indicators. The control of the data is realized through various software tools (drop-down menu, mandatory field, etc.), which monitor and guarantee the completeness, accuracy, validity, timeliness and other quality characteristics of the submitted data.

The timeliness of data in information technology is most easily ensured through cloud structures where all data about a particular object, process, individual, automatically transferred to the cloud once a process is complete and immediately available to all users authorized to work with them.

The data cleaning process monitors for incorrect, incomplete or inaccurate data and ensures that all obsolete or non-compliant data quality criteria are removed.

In modern information systems, software tools for quality control and data cleaning are built into the respective input modules, which allows them to work in real time. In this way, incomplete, inaccurate and outdated data are not allowed to enter, which maximally supports the making of the right management decisions.

This is the process that ensures that the data is correct, consistent and applicable. Data clearing is important because it improves data quality by removing any obsolete or incorrect data and leaving the highest quality information.

For the data to be used by different management levels (different user groups) and to be available on different devices (PC, Tablet, Smartphone), it is necessary to possess flexibility, which is particularly important in ERP systems in the mining industry. This means that they are subject to processes for different organizational changes or reengineering with minimal modification of the existing objects and relations in them. The use of information systems through the Internet or in a network mode requires the data to be interactive – that is, to have two-way communication between the data and users.

Conclusion

Although relational databases are still widely used in the mining industry, with the increasing volume of processed data distributed in the Web environment and the introduction of the Internet of Things, they are finding it increasingly difficult to handle large real-time data sets. NewSQL databases still offer partial solutions, but NoSQL has already established itself in certain areas as a better solution than classic RDBMS.

The information systems used in the mining industry are characterized by their specificity and diversity both for the type of mineral deposit (each deposit is unique) and compliance with the requirements of the specific company [9, 12, 13], which is a prerequisite for the integration of NoSQL data models in it due to their flexibility.

More and more mining companies plan, manage and control their activities, using specialized information systems adapted to their conditions and requirements. Since many large mining companies, incl. and in Bulgaria they are already building their own cloud structures, using

information from different types and models of databases, the technologies guaranteeing the processing of high quality data are of special importance.

However, due to the diversity of the software tools used, the implementation of all criteria for high quality data proves to be a difficult problem to implement, as an optimal balance between all criteria is sought.

For this reason, each mining company, depending on its requirements and available software tools, determines which quality criteria are most important for its work at a particular time, and this process is dynamic with the introduction of new information technologies.

References

1. Z. Eftimov, D. Anastasov, Scientific Aspects in Formation of Quality of Ore in Extraction Stage. Paper presented at the 22nd World Mining Congress, Istanbul, Turkey, 11-16 September 2011
2. H. Tudjarov (ed), *Upravlenie na danni* (Data Management) (Publishing house Asenevtsi, 2013), https://tuj.asenevtsi.com/Data/IndexD.html Accessed 2 November 2020
3. E.A. Brewer, Towards robust distributed systems, PODC '00, 7-2000, Portland OR https://doi.org/10.1145/343477.343502 (2000)
4. M. Aslett, What we talk about when we talk about NewSQL. (Publishing 451 Group, 2011) https://blogs.451researche.com/information_management/2011/04/06/what-we-talk-about-when-we-talk-about-newsql/ Accessed 2 November 2020
5. K. Fraczek, M. Plechawska-Wojcik, Comparative Analysis of Relational and Non-relational Databases in the Context of Performants in Web Applications, BDAS 2017 vol. 716, p. 154-163 (Springer, Cham, 2017) DOI: 10.1007/978-3-319-58274-0_13
6. W. Ali, M. Shafique, M. Majeed, A. Raza, Comparison between SQL and NoSQL Databases and Their Relationships with Big Data Analytics. Asian Journal of Research in Computer Science 4(2) p. 1-10 (2019) DOI: 10.9734/ajrcos/2019/v4i230108
7. Nektar, https://www.nektardata.com Accessed 2 November 2020
8. I. Kazandjiev, N. Yanev, K. Ivanov (ed.), Annual of UMG "St. Ivan Rilski", Vol. 55, Part 3, p. 123-127 (2012) http://mgu.bg/session/12/03/iknqki.pdf Accessed 2 November 2020
9. K. Kutzarov, D. Anastasov, Z. Eftimov (ed.), Journal Mining and Geology, is.2-3 p. 56-58 (2012)
10. N. Yanev (ed), *Metodologii I technologii za razrabotvane na informatsionni sistemi* (Methodologies and technologies for development of information systems), (Publishing house "St. Ivan Rilski, 2013)
11. I. Getova, Investigation and analysis of algorithms for evaluating the acquisition of knowledge on teaching students at the higher educational institutions, EDULEARN19, p. 9336-9342 (2019) doi: 10.21125/edulearn.2019.2313
12. Yo. Anastasova, D. Anastasov, Use of modern information technologies in the education of students from the University of Mining and Geology "St. Ivan Rilski". Paper presented at Vth Scientific and Technical Conference with International Participation Technologies and Practices in Underground Mining and Mine Construction, Devin, Bulgaria, 4-7 October 2016
13. J. Todorov, I. Starbanova, M. Trifonova, Information System for Planning, management and reporting of Open Cast Mines Production (Output) Paper presented at the First International Conference on Information Systems & Datagrid, Sofia, Bulgaria, 17-18 February 2005