Big data analytics of the technological equipment based on Data Lake architecture

Ilya Kovalev^{1,*}, Ramil Nezhmetdinov¹, and Denis Kvashnin¹

¹MSTU STANKIN, 127055 Vadkovskiy per. 3a, Moscow, Russia

Abstract. Currently, more and more managers of medium and large industrial companies are thinking about conducting a digital transformation of their enterprise. Each company is forced to strive to find an approach to optimizing production in order to remain competitive in the market. For industrial enterprises, this approach may be a digital transformation using the ideas of Industry 4.0. The digital transformation of an enterprise is a complex and multifaceted process that affects almost all levels of production. At the head of this whole process are data. Data on the work of production must be collected, stored, aggregated, transferred to various levels. Existing methods for storing data are not always suitable for working with BigData, new solutions are needed. The paper shows a comparison of the traditional approach to data aggregation and a promising direction based on the architecture of Data Lake.

1 Introduction

Data and mathematical models in Industry 4.0 should be the basis for decision making and changes to industrial enterprises based on economic or technical optimization [1]. Modern control systems have a fairly developed monitoring system and potentially have a huge amount of data at the time of work, on the basis of which it can be concluded that components and mechanisms are worn out and notify about it in time [2, 3]. The collection of data on the functioning of the enterprise and the integration of management systems at different levels is of strategic nature for the researcher, because identification of current problems, as well as building models, may be required much later and within the framework of a wide variety of strategic and technological problems. The unification of control systems at various levels, CNC machines, robotic systems goes hand in hand with the Internet of things and cloud computing according to “needs”. The main tasks solved on the basis of data analysis are:

- Reduced downtime
- Optimization of the technological cycle
- Cost reduction

To be able to collect data, for example, from a line of machines, it is necessary that this target equipment be connected to a single information network. Currently, with a trend in the concept of Industry 4.0, machine tool builders use either ready-made solutions from

* Corresponding author: ilkovalev@mail.ru

developer companies to solve such a problem, or use CNC system developers' solutions ("option 4.0" for FAGOR, "FOCAS2" for Fanuc, "SDK Remo Tools" for Heidenhain and etc.) It is also possible to use solutions based on OPC servers and SCADA systems, depending on the tasks [4].

Fig. 1. Combining various technological equipment into a single network.

The presented diagram (Fig. 1) shows the technological equipment integrated into a single information network, the data from which are transmitted to the OPC UA server or SCADA, and then to the aggregation and provisioning information Web server, or to the upper level of enterprise management (ERP) . In the presented work, a solution was considered when the data arrives at its own information collection server [5]. This means that the server must include all the functionality of the intermediate nodes and supports the work on various communication protocols used by technological equipment [6]. The server in question is an advanced version of the class of systems MDC (Machine Data Collection).

Database scaling is based on the principle of separation: partitioning data within a table (partitioning), allocating data into groups and transferring them to separate servers (sharding), as well as creating a full copy of the database (replication).

2 Standard application building DB model

The traditional approach to building an application with repositories is to use relational database management systems. This approach is due to the huge number of implementation examples, excellent documentation and a huge community (Fig. 2). A typical database creation scenario consists of several steps:

1. Conceptual design or problem statement - collection, analysis of data requirements
2. Logical design, includes the construction of an object model, normalization and subsequent optimization. As a rule, schemes are always "Star" or "Snowflake" schemes.

Fig. 2. Three-tier architecture of a classic web application.

This approach can have significant drawbacks. We will design a small database that will store data about the machine (type), machine parameters (number of control channels and software axes) and information received from various sensors (temperature, vibration).

Fig. 3. Scheme of the database for storing technological information.

Even for such a small database, the complexity of the model makes it necessary to have a sufficiently high qualification in writing SQL queries for a data researcher and engineers using information stored in logs [7,8]. The diagram contains duplicate tables, such as "history_machine_fitting_raw" and "history_machine_fitting_value", they completely repeat one another with the exception of the "value" and "raw" fields - the first is intended to store integers, i.e. in the context of industry and control systems, these are digital control signals, the data type raw - is a "raw" data type, mainly intended for analog signals (Fig. 3).

Data from sensors can come at different frequencies, and even if you interrogate once or twice a second (which is not enough to get the correct historical data), then this table will quickly take up large volumes. The complexity of building a SQL query (for example, which machine is running which program) and the serious amounts of data stored in such tables, when using classical table join algorithms, will significantly increase the load on the database and lead to data swapping to the hard disk.

3 Architecture for implementing application logic on a database server

When processing even small amounts of data, a web application (application server) spends considerable time on serializing and transmitting data over the network, as well as deserializing data and converting it to the desired database format. When using stored procedures, batch processing, as well as object-oriented programming tools on the database side, it is possible to: reduce network costs; hide the data structure, which will allow you to encapsulate the scheme and work with the database in the form of providing a high-level API; use the optimal data processing algorithms for a particular DBMS; aggregate data at the database level.

Currently, the most popular technologies that implement this architectural pattern are the PL / SQL (Oracle) programming language, as well as the Transact-SQL (Microsoft) language used in SQL Server and Sybase. We will design a small database schema for storing technological information in which the following tables exist:

DWH_MACHINE_TOOL - stores equipment information

DWH_MACHINE_TOOL_PARAMETERS - a table storing information about the measured parameters

DWH_HISTORY_WORK - stores data on the operation of technological equipment

One of the traditional ways to solve the problem of processing big data was partitioning (partitioning) - dividing database objects (tables, indexes, sequences, etc.) into logical groups (parts) according to some partitioning criterion. Partitioning primarily allows you to process information in several independent streams, which significantly affects the performance of database management systems [9]. A typical scheme for implementing vertical partitioning is partitioning a table, for example, DWH_HISTORY_WORK into several independent tables implemented by the inheritance mechanism provided by the databases (Fig. 4). In this case, the identifier of a specific equipment (machine_tool_id) was selected by the partitioning criterion.

Fig. 4. Partition of dwh_histrory_work table.

Each separate partition should be located in a separate data file (datafile), which can be located on different media, which increases fault tolerance in the event of failure of one of the drives on which the data file is located.

However, this approach works up to a certain amount of data and often serves as an additional way to increase productivity, and not as an architectural pattern of highly loaded solutions. It is used in patterns of logic storage near data and horizontal partitioning (sharding).

Another of the main ways to increase DBMS performance and fault tolerance is replication - a mechanism for synchronizing the source database with the receiver database / databases.

There are various architectural patterns of database replication: master - slave, master - master, variable master, etc. In turn, the types of replication are divided into synchronous and asynchronous, each of which has significant disadvantages. As part of the task of collecting technological information, data for various tasks should have a different degree of relevance [10]. For example, the field level refers to web services that write data to the master and there are consumers who need real-time information, such consumers should read and write only from the master, because When replicating data to a slave database, there is a temporary delay in the presentation of data that exists in the wizard. At the same time, research tasks can access data in a slave database, because they don't need to have real-time data [11].

The main limitation is replication, the specific technology of a particular database and the implementation depends on the technology selected, but at the same time, replication can be considered as an architectural style and transferring logic to the application level, in this case the databases will not know about the existence of a replication mechanism, which allows you to mix different technology, use a cache server for storing operational data, NoSQL databases, etc.

Each of the solutions discussed above for organizing the storage of a large variety of technological information solutions have one significant drawback, which is the high cost of connecting additional data sources specific to industrial enterprises to the technological information storage (analog signals, unstructured data, unrelated data, etc.). The implementation of such an object model at the database level forces clients (including web applications) to implement similar schemes to work at their level, which increases the cost and time to implement a new business logic or correct an existing one.

4 Proposed solution

Based on the existing deficiencies in the solutions for organizing data warehouses and the need for research tasks, including industrial ones, it seems obvious to turn to solutions from the world of big data, including the implementation of the architecture of a data lake. David Loshin defines the term Data Lake: "The idea of a data lake is to store the raw data in its original format until needed".

In the case when enterprises begin to conduct end-to-end analytics of processes through many systems, then the obtained data cannot be analyzed using traditional methods that have become established in business management. Data, and most importantly, processing methods, united under the general name Big Data, are described by the following basic principles:

Fig. 5. Principle of Big Data.

The so-called 5-Vs concept describes the main features of BigData. Volume - the data should be in sufficient quantity, and too small and too large volume can affect the final result of the evaluation of these data. Speed-speed of data processing and speed of their accumulation should constantly increase. Reliability - data must be accurate, reflect real

processes. Value - analysis of the data obtained should ultimately bring results. Variety - data can come in various formats, stored on various media. As applied to technological production, the diversity of data is one of the bottlenecks if one wants to implement the approaches of the Industry 4.0 concept [13]. Aggregated data can vary from equipment to equipment (only discrete or only analog signals), stored in a structured or simply in text form (logs), data can go to various types of storage (MongoDB or MySQL), etc. Given the variety and amount of continuously received data from equipment, problems may arise in the analysis of these process data. A data lake is one of the ways to work with Big Data, and most importantly, an architectural style that allows you to store such data (Fig. 6).

In general, we can say that each element in the data lake is assigned a unique identifier, as well as additional information about the data (metadata). The term data lake is often associated with the Hadoop stack and its object storage. In fact, Hadoop is the standard that supplies companies such as Cloudera, MapR, Microsoft (Azure HDInsight), Amazon, and others.

Fig. 6. The proposed option for organizing data storage using the Data Lake architecture

Within an industrial enterprise, like any other organization, there are many sources, such as databases, text data, excel tables, csv files, etc. Aggregating data within a data lake, in contrast to traditional analytics, provides such important advantages:

1. Data is processed in its original form, in contrast to the “fit” of data to a storage model.
2. Processing takes place over the entire data array, including in parallel using map-reduce technology.
3. Data can be analyzed in real time, including end-to-end analytics, using data from different systems about the same process of “knowledge” about a process that reflects hidden connections between processes, typical for the industry can be solving problems of associative rules (the state of a technological object depends on from related events), decision tree (understanding of the causes of marriage, downtime, as well as incidents at enterprises), genetic algorithms ces), neural networks and much more.

As a practical implementation, we can give a test example of obtaining data from heterogeneous equipment in the laboratory of the Department of Computer Control Systems of MSTU "Stankin" [13, 14]. The laboratory consists of several rooms: a testing ground

with planing-milling and milling and engraving machines under the control of the AxiOMA Control CNC system, a laboratory class with AxiOMA Control CNC emulators, Fagor Nc, BoschRexroth MTX, and a laboratory for testing PLC and PAC systems (Siemens, BoschRexroth, NCT). Data collection is carried out using OPC UA technology and a developed data collection server. With a wide variety of management systems, it was difficult to structure the data and bring it to a single format for storing in the MySQL database - such a process took time. The first tests using the Hadoop stack showed positive results - the data was stored in various formats, but processing in a single system was possible [15, 16]. Of the minuses of this approach, one can name the complexity of the initial setup and the competent prescribing of various scenarios. In the test example, the aggregation and analysis of data was aimed at obtaining information about the operating time of the equipment [17]. It was found that milling machines perform shaping tasks about 25% of the total time, 65% - technological downtime, 10% - unplanned downtime. This is due primarily to the fact that the machines are experimental and take preparatory operations for a long time, but for real production this is a very serious indicator of equipment downtime. In the PLC test laboratory, the situation is similar, due to the fact that it is intended for the educational process. In the laboratories of CNC system emulators, on the contrary, the operating time (emulating the execution of control programs with various shaping tasks) takes about 80% of the time - the time between failures is being checked.

4 Conclusion

Based on the approaches of the Industry 4.0 concept, it is possible to minimize production costs by fully monitoring the product life cycle, predicting the occurrence of various emergency situations, determining the need for unscheduled maintenance of equipment, and much more.

Received data from technological equipment (control systems, sensors, actuators, etc.) must be stored and processed. Existing database organization models for business processes are not suitable for a number of reasons. Within the framework of unclearly formulated data requirements and the need for technological information research, it is not possible to determine in advance the exact storage model in the RDBMS, and also to predict the need for scaling both vertical and horizontal, which in turn affects the cost estimate of the project being implemented, which usually leads to a significant overvaluation. Technological data are often not structured, stored in different formats. The proposed approach to organizing Data Lake for technological data related to big data solutions will minimize data processing time, store data in an unstructured form, and conduct real-time analysis.

This research was supported by Moscow State University of Technology "STANKIN" and Saint Petersburg National Research University of Information Technologies, Mechanics and Optics (agreement N31-1/03-C18 from 01 August 2018).

References

1. P. Nikishechkin, I. Kovalev, A. Nikich, *An approach to building a cross-platform system for the collection and processing of diagnostic information about working technological equipment for industrial enterprises*, MATEC Web of Conferences, **129**, pp.03012 (2017)
2. R. Nezmetdinov, P. Nikishechkin, A. Nikich, *Approach to the Construction of Logical Control Systems for Technological Equipment for the Implementation of Industry 4.0 Concept*, In: 2018 International Russian Automation Conference (RusAutoCon), (2018)

3. O. Perez, S. Saucedo, J. Cruz, *Manufacturing 4.0: The Use of Emergent Technologies in Manufacturing*, Palibrio, (2018)
4. G. Martinov, N. Kozak, *Numerical control of large precision machining centers by the AxiOMA Control system*, Russian Engineering Research, **35(7)**, pp.534-538 (2015)
5. I. Kovalev, P. Nikishechkin, A. Grigoriev, *Approach to Programmable Controller Building by its Main Modules Synthesizing Based on Requirements Specification for Industrial Automation*, International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), pp.1-4 (2017)
6. G. Martinov, I. Kovalev, A. Al Khoury, *Construction of a Specialized CNC System for Thread Grinding Machines*, In: 2018 International Russian Automation Conference (RusAutoCon). Sochi: IEEE, (2018)
7. G. Martinov, N. Kozak, R. Nezhmetdinov, A. Grigoriev, A. Obukhov, L. Martinova, *Method of decomposition and synthesis of the custom CNC systems*, Automation and Remote Control, **78(3)**, pp.525-536 (2017)
8. S. Grigoriev, G. Martinov, *An Approach to Creation of Terminal Clients in CNC System*, In: 3rd Russian-Pacific Conference on Computer Technology and Applications. Vladivostok, pp.1-4 (2018)
9. Jian-Yu Chen, Kuo-Cheng Tai, Guo-Chin Chen, *Application of Programmable Logic Controller to Build-up an Intelligent Industry 4.0 Platform*, in Proc. CIRP, vol. **63**, pp. 150-155 (2017)
10. G. Martinov, S. Sokolov, L. Martinova, A. Grigoryev, P. Nikishechkin, *Approach to the Diagnosis and Configuration of Servo Drives in Heterogeneous Machine Control Systems*. In: 8th International Conference, ICSI. Fukuoka, Japan, pp.586-594 (2017)
11. G. Martinov, N. Kozak *Specialized numerical control system for five-axis planing and milling center*, Russian Engineering Research, **36(3)**, pp.218-222 (2016)
12. G. Martinov, A. Lyubimov, A. Bondarenko, A. Sorokoumov, I. Kovalev, *An approach to building a multiprotocol CNC system*, Automation and Remote Control, **76(1)**, pp.172-178 (2015)
13. L. Martinova, G. Martinov, *Automation of Machine-Building Production According to Industry 4.0.*, In: 3rd Russian-Pacific Conference on Computer Technology and Applications. Vladivostok, pp.1 – 4 (2018)
14. V. Chekryzhov, I. Kovalev, A. Grigoriev, *An approach to technological equipment performance information visualization system construction using augmented reality technology*, In: MATEC Web Conf. Volume **224**, 2018. International Conference on Modern Trends in Manufacturing Technologies and Equipment (ICMTMTE 2018), pp.1-7 (2018)
15. G. Martinov, P. Nikishechkin, A. Grigoriev, N. Chervonnova, *Organizing Interaction of Basic Components in the CNC System AxiOMA Control for Integrating New Technologies and Solutions*, Automation and Remote Control, Vol. **80**, No. **3**, pp. 584–591 (2019)
16. R. Pushkov, L. Martinova and S. Evstafieva, *Extending Functionality of Control System by Adding Engraving Capabilities*, In: 2018 International Russian Automation Conference (RusAutoCon), (2018)
17. L. Martinova, N. Kozak, R. Nezhmetdinov, R. Pushkov, A. Obukhov, *The Russian multi-functional CNC system AxiOMA control: Practical aspects of application*, Automation and Remote Control, **76(1)**, pp.179-186 (2015)