THEORY OF THE DEVELOPMENT PROCESS OF A SOFTWARE MODULE THAT ANALYZES AND ENSURES THE SECURITY OF BIG DATA IN INTELLIGENT TRANSPORT SYSTEMS

Nozima Akhmedova[1], Komil Tashev[2]

[1,2]Tashkent university of information technologies named after Muhammad al-Khwarizmi

https://doi.org/10.5281/zenodo.7858400

Abstract. *Intelligent Transport Systems (ITS) have revolutionized the way we move and interact with our transportation networks. Leveraging cutting-edge technologies such as Internet of Things (IoT), sensors, and data analytics, ITS enable intelligent traffic management, predictive maintenance, and optimized routing, among other benefits. However, the increasing reliance on big data in ITS raises concerns about data security, as sensitive information such as traffic patterns, vehicle data, and user information need to be protected from cyber threats. To address this challenge, the development of a software module that analyzes and ensures the security of big data in ITS is crucial. In this article, the architecture of such a software module is analyzed based on a review of related works and best practices in software development and cybersecurity.*

Keywords: *Intelligent transport system, Big data, information security, ITS elements, components.*

INTRODUCTION

Intelligent transport systems (ITS) are designed to implement a number of global technologies for the transportation of goods and passengers. Examples of such intelligent transport technologies are as follows. The technology of "intelligent cargo", which, in the process of transportation, "automatically reports its properties." Technologies based on the principles of "cargo tracking" logistics - information and telematic technologies and systems that take into account the requirements of interoperability or their elements. Technologies for providing automatic control of moving units, etc.

Intelligent technologies for transportation processes contain a number of elements for automatically collecting data on transportation conditions, modeling processes, comparing with templates or with standards, recognizing emergency situations or conditions and the possibility of their occurrence, forecasting the states of transport systems and planning transportation, etc.

In order to implement the above processes and operations, it will be necessary to process a large amount of data. This requires Big Data technology. [1] And the main issue for the continuous operation of the system, and this is very important, it will be necessary to ensure information security.

ELEMENTS OF INTELLIGENT TRANSPORT SYSTEM

Before proceeding to the issues of information security of the ITS, it is necessary to understand what components it consists of, what objects should be protected and from whom.

For any ITS, the following types of elements are characteristic:

- on-board facilities installed on mobile ITS objects (means of remote monitoring, measurements, etc.);
- means installed on stationary infrastructure facilities (means of remote monitoring, measurements, etc.);

- remotely controlled actuating and indicating devices (devices, components and assemblies);
- servers for processing and storing information;
- situational, dispatching and operational centers;
- communication means - Internet, GSM/GPRS network, satellite communication;
- information and telecommunication means providing secure information interaction with external information systems.
- The ITS technological complex may include a variety of technical systems and means:
- systems and means of coordinate-time, meteorological, etc. types of security;
- systems, means, lines and networks of communication and data transmission;
- systems and means of remote monitoring;
- systems and means of collecting, accumulating and processing information;
- automated systems and controls;
- systems and means of displaying and communicating information;
- other technical and software and hardware means.

Most of the systems and tools are used to form a feedback channel both with the human operator and with the controlled technical components of the transport system.

The object of attack can be any of the listed elements. However, in general, all elements of ITS can be classified into one of three categories (Figure 1):

- Data processing center (DPC),
- Periphery equipment,
- Communication system for data exchange. [2]

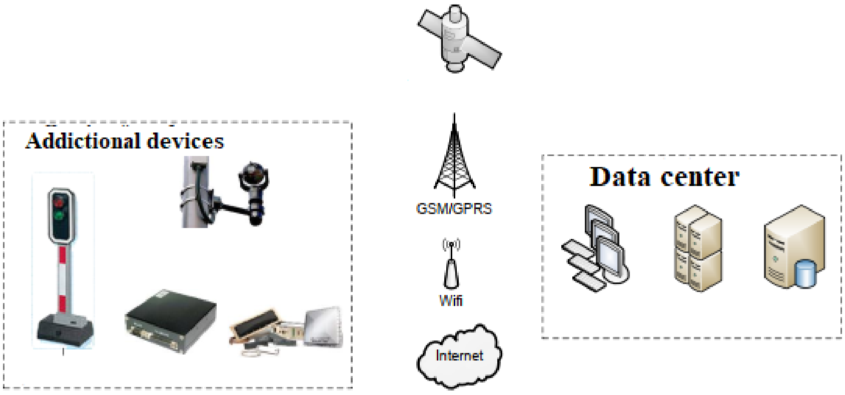

Figure 1. Categories of ITS elements

In a nutshell, Big Data is data that does not fit into a computer's RAM.

In essence, this definition means that the property "to be large" is not an independent property of the data, but depends on the characteristics of the system used to process them.

For example, it is difficult for an ordinary person to remember exactly what temperature was in our city every day over the past month. Thus, three dozen values could well be an example of Big Data. However, here is a person confidently reporting "the past month was cold". This message carries information about the processed data: according to the interlocutor, the average temperature for the past month was lower than usual this month for several decades.

Another example would be data on objects that theoretically carry important information, but are of such size that it is almost impossible not only to process or store this data, but even to collect it. Consider, for example, a dataset containing the coordinates and velocities of molecules in an air column over an airport. There is also metadata describing when the measurement was taken and what kind of molecule it is. Such a dataset carries information about the weather conditions over the airport, including temperature, pressure, humidity, cloud cover, special weather conditions - a passing tornado or a falling hail. On the other hand, for correct processing, the data for all molecules must be sufficiently complete and representative for statistical processing.

As a result of such a thought experiment, we understand that in order to work effectively with big data, we need a data model that allows us to form methods for working with data.

TYPES OF DATA PROCESSED AND THE ADVANTAGE OF BIG DATA

The data can be of various types. Information obtained as a result of accounting or measuring any objects or parameters is called master data. For example, taking into account the quantity, measuring the coordinates and velocities of specific molecules are master data.

Transactional data(in the English literature, the terms Transactional Data, Application Specific Data, Operational Data are used) are data that reflect the result of performing any operations. For example, data on the interaction of molecules with each other, namely, on crossing the boundaries of the area under consideration, on the trajectory of a particular molecule, on the evaporation of raindrops, are transactional data.

Historical data(Historical data) is data with timestamps.

Reference data(reference books, reference data, reference information, Reference Data, Lookup Data, Dictionaries) are basic immutable data known in advance from external sources, such as regulations, abbreviations, acronyms, dictionaries, standards.[3]

Data format. Structured data has a predefined format. Semi-structured or semi-structured data is data that is often collected from various sources. The data structure is documented, but depending on the data source, the specific format for presenting information may be different. Unstructured data requires mandatory processing and subsequent validation before use.

Data in which coordinates are measured in different units of measurement, numbers are sometimes written in words, sometimes in Latin numerals, and sometimes in the form of a scanned image of a laboratory assistant's handwriting, is unstructured data.

Typically, Big Data is described using the following characteristics.

1. Volume (Volume) - the amount of generated and stored data. The size of the data determines the value and potential of the data, and whether it can be considered Big Data.

2. Variety - data type. Big data can consist of text, images, audio, video. Big data, when compared with each other, can complement the missing data.

3. Speed (Velocity) - speed. This refers to the speed at which data is generated and processed. Very often, Big Data is used in real time.

4. Variability - the inconsistency of data sets can interfere with their processing and management.

5. Reliability (Veracity) - the quality of the data directly affects the accuracy of the data analysis. [4]

To store and process Big Data, distributed data storage systems have been created, including distributed file systems that allow using the external file space of the storage system to process data on the nodes included in the computing cluster.

It is often convenient to use distributed file systems leased as a separate cloud service, for example, Google Colossus5, Amazon S36, Yandex Disk7.

Processing of data located on distributed storage systems is carried out in parallel on computers that make up nodes (nodes) of a computing cluster. To organize calculations, developers of processing systems use distributed frameworks. Most frameworks are available under the Apache license and are designed to run on Linux-based clusters. There are also cloud frameworks that are rented as a separate cloud service.

Setting up a Big Data infrastructure is not an easy task, and deploying new applications in a Big Data cluster is the responsibility of Big Data engineers. They largely automate the installation and configuration of Big Data components.

Big Data security tools allow you to centrally control access to data. Big Data security has become a discipline in its own right, and data scientists usually only deal with it as data consumers. Big data security is handled by information security experts.

It is proposed to develop a software module for data processing to ensure continuity and security. First of all, it will be necessary to determine the architecture and algorithms of this software module.

PROGRAM ARCHITECTURE

To work with Big Data, complex systems are used in which several components or layers (Layers) can be distinguished. Usually, four levels of components of such systems are distinguished: reception, collection, analysis of data and presentation of results (Fig. 2). This division is largely conditional, since, on the one hand, each component, in turn, can be divided into subcomponents, and on the other hand, some functions of the components can be redistributed depending on the task being solved and the software used, for example, they separate data storage into a separate layer.

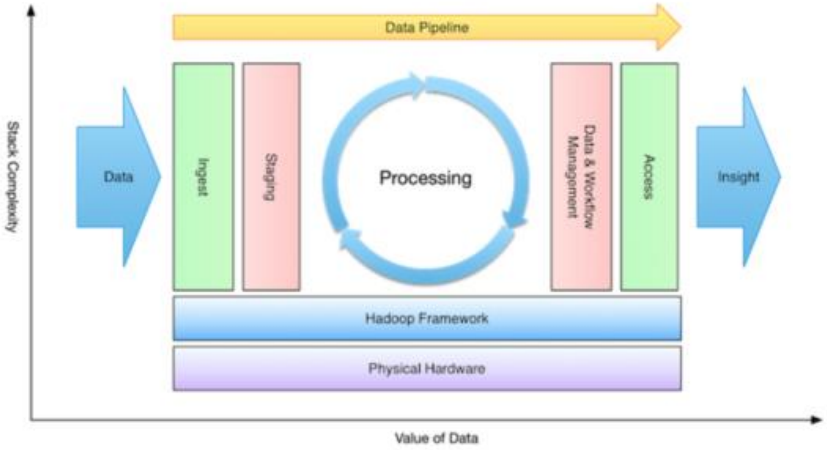

Figure 2. Big data stack.

To work with Big Data, system developers create data models that are meaningfully related to the real world. The development of adequate data models is a complex analytical task performed by system architects and analysts. The data model allows you to create a mathematical model of the interactions of real world objects and includes a description of the data structure, methods of data manipulation and aspects of maintaining data integrity.

Distributed systems of various types are used for data storage. These can be file systems, databases, journals, mechanisms for accessing shared virtual memory. Most storage systems are focused exclusively on working with Big Data, they have an extremely limited number of functions (for example, it may not be possible not only to modify, but also delete incoming data), which is explained by the internal complexity of creating highly efficient distributed systems.

In order to work with data faster, data storage and processing systems are parallelized in a cluster (cluster, a group of computers connected by a network to perform a single task). However, according to Brewer's hypothesis, it is impossible to ensure the simultaneous consistency (consistency) of data, data availability and system stability to the separation of individual nodes. Data sources have various parameters, such as the frequency of data from the source, the size of the data portion, the data rate, the type of incoming data and their reliability.

Data sources need to be established for effective data collection. These can be data warehouses, providers of aggregated data, APIs of any sensors, system logs, human-generated content in social networks, in corporate information systems, geophysical information, scientific information, legacy data from other systems. Data sources define the original data format.

The data collection stage is characterized by direct interaction with data storage systems. A collection point is established at which the collected data is provided with local metadata and placed in storage or transferred for further processing. Data that for some reason did not pass the collection point is ignored.

For structured data, the transformation from the original format is carried out according to predetermined algorithms. This is the most efficient procedure if the data structure is known. However, if the data is presented in binary form, the structure and relationships between the data are lost, then the development of algorithms and data processing software based on them can be extremely difficult.

For semi structured data requires the interpretation of incoming data and the use of software that can work with the used data description language. A significant advantage of the semi-structured data is that they often contain not only the data itself, but metadata in the form of information about the relationships between the data and how to obtain them.

Software development for semi-structured processing data is quite a challenge. However, there are a significant number of ready-made converters that can, for example, extract data from an XML format into a generated table view.

The greatest amount of work requires processing unstructured data. To convert them to a given format, it may be necessary to create special software, complex manual processing, recognition and selective manual control.

During the collection phase, data type checking is performed and basic data validation can be performed. For example, the coordinates of the gas molecules contained in any region cannot lie outside this region, and the velocities cannot significantly exceed the speed of sound. In order to avoid typing errors, it is necessary to check whether the units of measure are set correctly. For example, one dataset might measure altitude in kilometers and another in feet. In this case, it is necessary to convert the height into those units of measurement that are accepted in the model used.

When collected, data is organized and provided with metadata stored in related metadata. When there are many data sources, data collection management may be required to balance the amount of information coming from different sources.

Collected data is either stored in storage systems or (especially for streaming data) transmitted for real-time analysis.

Data analysis, unlike data collection, uses the information contained in the data itself. The analysis can be carried out both in real time and in batch mode. Data analysis is the main task in terms of labor intensity when working with Big Data.

There are many data processing techniques: predictive analysis, queries and reporting, reconstruction using a mathematical model, translation, analytical processing, and others. Methods use specific algorithms depending on the goals. For example, analytic processing may be image analysis, social networking, geographic location, feature recognition, text analysis, statistical processing, voice analysis, transcription.

Data analysis algorithms, as well as data processing algorithms, are based on a data model. At the same time, several models can be used in the analysis, defining a common data format, but modeling the meaningful processes in different ways, the data about which we process. When using artificial intelligence methods in the analysis, in particular neural networks, models are dynamically trained on various data sets.[6]

When analyzing data, the entities described by the data are identified based on the information available in the data and the models used. The essence of the analysis is an analytical engine that uses analytical algorithms, model management and entity identification to obtain new meaningful information that is the result of the analysis.

Data analysis results are provided at the consumption level. There are several mechanisms that allow you to use the results of big data analysis.

- Meta information monitoring.

Subsystem for displaying in real time the essential parameters of the system, the workload of computers, the distribution of tasks in the cluster, the distribution of information in storages, the availability of free space in storages, the flow of data from sources, user activity, equipment failures, etc.

- Data monitoring.

Subsystem for real-time display of the processes of receiving, collecting and analyzing data, data navigation.

- Generating reports, querying data, presenting data as visualizations on dashboards (Dashboard), in PDF format, infographics, pivot tables and quick references

- Data conversion and export to other systems, interface with BI systems.

CONCLUSIONS

The following has been done in this work:

1. Studied intelligent transport systems and technologies to ensure the security of data in the system;
2. The elements of an intelligent transport system are considered;
3. The types of data processed in the proposed software module are considered and the advantage of big data is considered;

The architecture of the program module for data processing is proposed.

In conclusion, the architecture of a software module that analyzes and ensures the security of big data in intelligent transport systems plays a pivotal role in safeguarding the integrity, confidentiality, and availability of sensitive information. The analyses conducted in this article shed light on the key components and design considerations that are essential for building a robust and secure software module for big data analysis in intelligent transport systems.

The article emphasizes the importance of a layered and multi-tiered architecture, where different components work together to handle data processing, storage, and security tasks. This includes components such as data ingestion, data processing, data storage, and security modules, each with their own specific functions and responsibilities. The use of advanced security technologies, such as encryption, authentication, and access control, are crucial in ensuring that big data in intelligent transport systems is protected from unauthorized access, tampering, and other security threats.

In summary, the architecture of a software module that analyzes and ensures the security of big data in intelligent transport systems is a critical aspect of building secure and reliable systems. The analyses presented in this article provide valuable insights into the key components and design considerations that should be taken into account when developing such a module. By leveraging a well-designed architecture that incorporates advanced security measures, intelligent transport systems can effectively mitigate security risks associated with big data and ensure the confidentiality, integrity, and availability of critical information.

REFERENCES

1. Xinhu Zheng, Wei Chen, Pu Wang, Dayong Shen, Songhang Chen, Xiao Wang, Qingpeng Zhang, and Liuqing Yang “Big Data for Social Transportation”. IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 17, NO. 3, MARCH 2016.
2. N. Akhmedova, J. Khamzayev “Problems of intelligent transport system and solutions with big data” International conference “Recent advances in intelligent information and communication technologies “ISPC-2022””, Tashkent-2022, -P. 34-42.
3. Danette McGilvray “Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information”. Morgan Kaufmann Publishers. Copyright © 2008 Elsevier Inc.
4. Akhmedova, N. (2020). A STUDY OF SECURITY PROBLEMS IN BIG DATA AND THEIR SOLUTIONS. Chemical Technology, Control and Management, 2020(4), 81-85.
5. Peng Yue, Liangcun Jiang “BigGIS: How Big Data Can ShapeNext-Generation GIS” September 2014. DOI: 10.1109/Agro-Geoinformatics.2014.6910649.
6. SANDRO FIORE, DONATELLO ELIA, CARLOS EDUARDO PIRES, DEMETRIO GOMES MESTRE, CINZIA CAPPIELLO, MONICA VITALI , NAZARENO ANDRADE, TARCISO BRAZ, DANIELE LEZZI, REGINA MORAES, TANIA BASSO, N`DIA P. KOZIEVITCH, KEIKO VERÔNICA ONO FONSECA, NUNO ANTUNES, MARCO VIEIRA, COSIMO PALAZZO, IGNACIO BLANQUER, WAGNER MEIRA, JR., AND GIOVANNI ALOISIO “An Integrated Big and Fast Data Analytics Platform for Smart Urban Transportation Management”. SPECIAL SECTION ON URBAN COMPUTING & WELL-BEING IN SMART CITIES: SERVICES, APPLICATIONS, POLICYMAKING CONSIDERATIONS. VOLUME 7, 2019.