Architecture of managing big data of mixed transportation of passengers in agglomerations

N A Zhuravleva¹³ and M Poliak²⁴

¹Emperor Alexander I St. Petersburg State Transport University, 190031 Saint Petersburg, 9 Moskovsky pr., Russia
²University of Žilina, 010 26 Žilina, Univerzitna 1, Slovak Republic
E-mail: ³zhuravleva_na@mail.ru, ⁴milos.poliak@fpeds.uniza.sk

Abstract. The aim of this study is to develop a big data architecture that can provide the formation and management of competitive mixed passenger traffic in agglomerations in real time, taking into account the optimization of their cost, speed and new services. The work is based on studies of development trends of transport systems in agglomerations (SmartCitiesWorld); analysis of the methodology and use of systems for managing relational databases of structured data, as well as non-relational databases, information processing experience AIIM. For the analysis of passenger flows and decision-making on optimal routes, spatial databases OGS that implement standards have been used. The study substantiates the conclusion that the Big Data technology, implemented according to the cascade principle of information support for the transportation process, ensures the growth of monetization of all its components (transport infrastructure, vehicles and their management system). On this basis, a big data architecture was built, which implies the sharing of structured and unstructured data in the management of passenger traffic in agglomeration. This architecture made it possible to take into account the influence of the most important changes in the agglomeration on the mobility of its population and to optimize the financial performance of transport organizations due to the competitiveness of the allocated mixed routes.

1. Introduction
The way out of the global crisis caused by the pandemic will be accompanied by unprecedented competition in the markets for goods and services. Unfortunately, this will not bypass the transport services market. The current trend of moving away from competition between different modes of transport towards interspecific competition based on multimodal passenger transportation will undoubtedly receive new development. The effectiveness of multimodal (intermodal or multimodal) transportation in agglomerations is confirmed by a number of studies [1, 2], since it is precisely such passenger transportation that optimizes cost and travel time.

The post-crisis economics will change the gravity of travel in agglomerations, as there will be serious changes in the labor market and this will affect passenger behavior. The need for a “door-to-door” trip at an optimal price for the agglomeration resident, minimum travel time and a certain quality of services will force transport organizations to modify their business models, which have been written about a lot in previous years [3].

It is impossible to implement the proposal of new “values” (speed, price, service) of passenger transportation in the metropolitan area without the inclusion of digital technologies, Big Data
technology primarily. Previously, we researched “Data Privacy and Security Vulnerabilities of Smart and Sustainable Urban Space Monitoring Systems” [4].

Today, for each passenger carrier, or even the industry as a whole, we can find a sufficient amount of structured data to judge their specialization, capacities, transportation technologies, competitive advantages and disadvantages. But, when the question arises of what kind of transportation will be competitive, what kind of transportation the consumer is waiting for and how much he will be ready to pay for it, there is no information practically. There is no complete data on how exactly the labor market and the associated mobility of the population will change, how household incomes and travel tariffs will be related. Important information such as travel motives and personal information about the passenger are not structured at all. A completely new set of data appears, generated by various devices with the participation of a person, for example, from social networks. The spread of new, more powerful mobile devices, combined with increased access to global networks generates a completely unknown set of data and their sources.

Despite the fact that the existing technologies for searching for individual data sources and their accumulation are clear and sufficient today, however, the practice of using aggregators capable of analyzing and processing diverse data from incompatible sources is practically absent. We are talking, for example, about data from photographs, social networks, Internet resources, case studies, official statistics, etc. This means that traditional methods of processing and managing data will not be able to provide reliable analytics when working with huge arrays of heterogeneous information, without which it is impossible to build an optimal route network of competitive passenger traffic in agglomerations.

The object of this study is the routes of mixed transportation of passengers in urban agglomerations. It is agglomeration that is characterized by high mobility of the population within the territory uniting various settlements, primarily to its center, which is usually a large city. This mobility is ensured by all types of competing city and suburban transport, depending on the geographical location of the agglomeration.

The aim of this study is to develop architectural solutions for big data management to create an efficient, real-time routed transportation network in the metropolitan area, taking into account cost, time and new services optimization. At the same time, the architecture of big data should allow avoiding interspecific competition of certain types of transport in the competition of multimodal transport on routes.

2. Methodology and Empirical Analysis

By means of SmartCitiesWorld [5] the development trends of transport systems in agglomerations were investigated, which allowed us to conclude that the predominant development of mixed (intermodal and multimodal) transportation. An analysis of the relational database management systems (RDBMS) made it possible to generate consistent and complete information about the basic characteristics of the agglomeration and the dynamics of their changes, the parameters of socio-economic development, budget and cost indicators, as well as transport development trends with the corresponding characteristics of passenger flows. Based on non-relational databases (NoSQL class), the architecture was supplemented with unstructured data (polls, complaints from the population, data from social networks) about passenger requirements for routes, services and transportation services.

We have reviewed the best practices of AIIM information processing (Alphanumeric and Iconographic Information Association (USA) [6] and determined that this tool is most consistent with the strategies and methods used to collect, store, deliver content and receive operational documents, as well as manage them. This tool can provide the necessary reporting, management of work processes, web content and analysis of the feasibility of the formation of multimodal transport on the basis of all types of transport in the metropolitan area (road, rail, including rail, bus, trolleybus) on competitive routes.
The big data analysis methodology is based on reports and dashboards of data on mobility in the agglomeration, aggregating information from various sources and provides for the visualization of dynamic data interconnected with each other (changing diagrams, maps, route diagrams).

This makes it possible to show the processes in motion. For analysis, spatial databases were used that implement the OGC (Open Geospatial Consortium) standards [7].

We studied the experience of analyzing big data based on the Google Prediction API interface, and suggested that it could be used to study passenger behavior based on posts in Facebook, Twitter. Thus, we will form mixed routes provided with new services, and we will also be able to offer virtual testing of new routes. The results of the study are fragmentarily tested on the basis of the St. Petersburg metropolitan area and confirm our research model.

3. Results

1. The use of Big Data technology in the management of passenger traffic allows implementing the cascading principle of information support for processes and projects, while ensuring the growth of their monetization.

   This result was obtained on the basis of a study of the evolution of data processing, indicating that the technology of data analysis and management changed every time new types of data appeared. This is especially true for Big Data technology, which provides management of large amounts of information from various sources and carriers with a speed that allows them to be analyzed in real time and timely response regime. This is important for managing dynamic processes, such as the transport of goods and passengers. At the same time, the dynamic process of transportation is determined, first of all, by the transport infrastructure - the most conservative and expensive element of transportation, the dynamics of which are lagging behind all other processes for many years. This means that the methods of managing data on transportation, no matter what analytical tasks they correspond to, must combine new methods and methods that were applied earlier, i.e. match the cascade processing principle.

   This statement is confirmed by the evolution of the introduction of IT technologies and database management systems. So, the advent of relational databases has led to the creation of a set of technical tools that allow us to study connections between data elements. Further, the accumulation of unstructured data has led to the emergence of new natural language-based analysis tools. Internet search engines have led to the emergence of asymmetric analysis methods. Big data adds fundamentally new ways to manage information based on digital technologies for data collection and storage, new networks and computing models, virtual and cloud computing.

   In general terms, the data processing environment depends on its cost (storage, calculation, protection), or rather, the specified cascade optimizes the price and result, which are complexly affected: large amounts of data; high speed of processing and variability of decisions. Thus, monetization of the use of the big data medium is achieved only if they adequately meet the needs of solving the business problems of the organization, industry or a dedicated design solution.

2. The efficiency of using Big Data is related to their architecture. The created architecture of big data is determined by the functional requirements for the project and is a cyclic process of collecting, systematizing, generalizing, analyzing and making managerial decisions. This cyclicity should not only satisfy the functional requirements of the problem being solved, but also provide sufficient performance in the preparation of management decisions.

   To solve the problem of managing optimal multimodal transport routes in the metropolitan area, we described and presented in Table 1 the following components of the required Big Data architecture.

   Interfaces and data flows are of paramount importance because they provide the collection of a large amount of information from multiple sources. They are constantly updated taking into account constant changes in the system, in our case, changes in the development of the transport network and a number of routes. At the same time, big data cannot function without integration and security services.

   The most important element of the architecture is the source of operational data. In our case, for the most part this is highly structured data related to the agglomeration transport system and managed
using relational databases. At the same time, architecture allows you to include a variety of unstructured sources, such as a social portrait of an agglomeration resident, data from social networks. Managing the entire set of data involves new approaches, including textual, tabular, columnar and geospatial database architectures, which together form NoSQL (open source distributed non-relational databases).

| Table 1. Enlarged Big Data Architecture Components. |
|-----------------------------------------------------|
| **Power**                                           | **"Organization" of data storages and tools** | **Analytics (traditional and advanced)** |
| interfaces and streams for exchanging data over the Internet | operational databases: structured | analytical data storages and data marts |
|                                                    | unstructured                           |                                        |
|                                                    | poorly structured                      |                                        |
| interfaces and data flows from internal applications and to internal applications security infrastructure | Big Data Applications | analytical data processing methods |

In this study, we do not consider data organization services and tools. The most important issue for us is the performance of the architecture, since it should allow us to make decisions on the management of multimodal transportations on busy routes in real time. First of all, it is connected with the speed of calculations. In particular, in our case, it should ensure the management of new routes of competitive multimodal transport. This requires real-time data on the number of people at stops, interchange nodes, the speed of reading ticket information, etc. The right performance - making decisions within minutes can be provided by a distributed computing model. The second performance factor is associated with the choice of the type of database, in our case, it is a graph database in combination with column databases.

A fragment of the operating database model in the specified Big Data architecture is presented in Table 2.

| Table 2. The model of the operational database in the architecture of big data (fragment). |
|----------------------------------------------------------------------------------------|
| **Highly structured data (fragment)** | **Unstructured data** | **Analysis results** |
| Dynamic indicators of changes in socio-economic parameters of the development of agglomeration | Informational “portrait” of real household incomes | Virtualization of passenger flows on routes |
| Agglomeration Investment Development Indicators | Expert assessments of the impact of investment decisions in the agglomeration on population mobility | Assessing the monetized effect of new mixed routes and directions |
| Indicators of compliance of transport development trends of the territory with the parameters of technological progress of the “value” of the passenger transportation service | Data from surveys and social networks about the demand for existing routes and “desired” routes. Reviews of transport waiting times, services, etc. | Parameters of the monetized effect of new (digital) services |
| Indicators for assessing trends in the impact of transport tariffs on population mobility. | Forecasts of the cost of mixed trips on various routes and their competitiveness, surveys of the population about the desired options for choosing travel routes, etc. | Parameters of the monetized effect of the “value” of a mixed trip along the route |

The implementation of such a fragment of this big data architecture has one main goal - the joint use of structured and unstructured data in the management of passenger traffic in the metropolitan area, which allows you to take into account the impact of the most important changes in the metropolitan area on the mobility of its population and the optimization of the financial result of passenger traffic.
4. Discussion
The results of this study are relevant to specific projects to increase the efficiency of passenger transportation in growing Russian metropolitan areas and optimize the activities of transport companies. The presented architecture of the big data of the project for the management of optimal routes for mixed passenger transportation provides the monetization of decisions that provide the "value" of transportation, i.e. time reduction, cost optimization and providing the passenger with new services, such as, for example, a single ticket.

Discussions on the problems of using Big Data in transport, as well as the theory and methodology of applying this technology, are presented in many studies. In particular, a recent study by Du, G., et al. connects the theory of innovative development of transport organizations and innovative technologies of big data, which is fully consistent with the logic of our study [8].

In the same way, we relied on the model of Saki, M., Abolhasan, M., Lipman, J., using the Internet of Things (IoT) data classifications, as well as critical and non-critical data sets for servicing urban rail transport systems. [9]

Our conclusion about the influence of transport infrastructure on the cascade of information processing is confirmed by Viri, R., et al. At the same time, the authors are building a data architecture taking into account the growing role of multimodal transportation. By differentiating the types of structured and unstructured data from different sources, by using interfaces for APIs application programming, they note the low effect of BD on the state of the transport infrastructure, but note their high efficiency in increasing traffic volumes and the quality of services for passengers [10].

The methods of information processing that we use can be expanded taking into account the results of the Ding, L. study. The author emphasizes that when designing multimodal transport systems at the level of the transport system, priority is given to the degree of conformity and coordination of port facilities and transport information platforms [11].

Considering the results of creating a system for forecasting passenger flows on the basis of rail transport in the agglomeration obtained by Zhu, K., et al., we find confirmation of the reliability of the data architecture design proposed by us. In particular, combining the concept of deep learning, AMD hardware virtualization and the support vector machine (combines the deep learning (DL) theory and support vector machine (SVM) into the DL-SVM model) into a single passenger flow tracking model. (URT) [12].

Passenger flow modeling parameters, such as hourly passenger flow directed to departure stops; daily fluctuations in the population of areas due to inter-district passenger traffic; the impact of competition on the functioning of the municipal transport system; the choice options of stopping destinations in accordance with estimates of the attractiveness of stops, justified by the work of Krushel, E., Stepanchenko, I., Panfilov, A., Berisheva, E. can significantly improve our data architecture [13].

The experience of using unstructured data from various sources, taking into account the influence of public transport congestion and travel delays on the accessibility of agglomeration jobs, expands our understanding of the results of big data analytics [14].

The importance of the search for optimal solutions for the formation of competitive multimodal transport, implementing new services for passengers based on big data and optimizing the activities of transport organizations, is evidenced by a large number of publications. In particular, the study of the impact of the new high-speed passenger transportation service, the construction of an innovative marketing model based on big data and the increase in the organization of passenger transportation in the megalopolises discussed in Tan, H. allowed us to expand the data architecture of our study [15]. The results of the study Ma, Y., Yin, W made it possible to justify the introduction of indicators of the influence of the age of passengers on the seasonality of their trips into our databases [16].

In general, a significant amount of research on the impact of big data on the formation of a new competitive transport service indicates the high relevance of this topic.

5. Conclusion
The results of our study are a fragment of the complex process of creating competitive agglomeration transport systems based on the formation of a new value for the transportation service, namely, increasing speed, optimizing the price and availability of services important to the passenger. The search for an adequate Big Data architecture, which allows one to describe the patterns of database formation from various sources as broadly as possible, then form analytics to solve specific business problems and ensure that effective decisions are made in real time is only part of the solution to the problem. The study of high-speed flows of various types of large volumes should enhance the effectiveness of using this technology in order to increase the efficiency of agglomeration transport systems.

References

[1] Yang K, Yang L, Gao Z 2016 Planning and optimization of intermodal hub-and-spoke network under mixed uncertainty Transportation Research Part E: Logistics and Transportation Review. 95:248-266, doi: 10.1016/j.tre.2016.10.001.

[2] Kordnejad B 2016 Stakeholder analysis in intermodal urban freight transport Transportation Research Procedia 12:750-764, doi: 10.1016/j.trpro.2016.02.029

[3] Vatin N I, Usanova K Y 2019 BIM end-to-end training: From school to graduate school Advances and Trends in Engineering Sciences and Technologies III- Proceedings of the 3rd International Conference on Engineering Sciences and Technologies ESaT 2018 651-656

[4] Lun Y H V, Shang K C, Lai K H, Cheng T C E 2016 Examining the influence of organizational capability in innovative business operations and the mediation of profitability on customer satisfaction: An application in intermodal transport operators in Taiwan International Journal of Production Economics 171(2) pp179-188, doi: 10.1016/j.ijpe.2015.02.019

[5] Zhuravleva N A, Cadge K, Poliak M, Podhorska I 2019 Data Privacy and Security Vulnerabilities of Smart and Sustainable Urban Space Monitoring Systems Contemporary Readings in Law and Social Justice 11(2):56–62, doi:10.22381/CRLSJ11220198

[6] SmartCitiesWorld. Sharing Ideas to Solve Urban Challenges

[7] AIIM. www.aiim.org

[8] OGC. Making location count. www.opengeospatial.org

[9] Du G, Zhang X, Ni S 2019 Discussion on the application of big data in rail transit organization Smart Innovation, Systems and Technologies. 2nd International Conference on Smart Vehicular Technology, Transportation, Communication and Applications, VTCA 129:312-318, doi: 10.1007/978-3-030-04582-1_36.

[10] Saki M, Abolhasan M, Lipman J 2019 A novel approach for Big Data classification and transportation in rail networks IEEE Transactions on Intelligent Transportation Systems 21(3):1239-1249, doi: 10.1109/TITS.2019.2905611

[11] Viri R, Aunimo L, Aramo-Immonen H 2019 Connected and multimodal passenger transport through Big Data analytics: Case Tampere city region, Finland IFIP Advances in Information and Communication Technology. Springer, Cham568:527-538, doi: 10.1007/978-3-030-28464-0_46

[12] Ding, L. Multimodal transport information sharing platform with mixed time window constraints based on big data // Journal of Cloud Computing. 2020. 9(1) p 11, doi: 10.1186/s13677-020-0153-8.

[13] Zhu K, Xu P, Li W, Li Z, Zhou R 2019 Prediction of passenger flow in urban rail transit based on Big Data analysis and deep learning IEEE Access. 2019 7:142272-142279, doi: 10.1109/ACCESS.2019.2944744

[14] Krushel, E., Stepanchenko, I., Panfilov, A., Berisheva, E. Big data in the stochastic model of the passengers flow at the megalopolis transport system stops // Communications in Computer and Information Science. 2019. 1083:118-132, doi: 10.1007/978-3-030-29743-5_9.
[15] Arbex R, Cunha C B 2020 Estimating the influence of crowding and travel time variability on accessibility to jobs in a large public transport network using smart card big data Journal of Transport Geography 85(102671), doi: 10.1016/j.jtrangeo.2020.102671

[16] Tan H 2019 The upgrade of passenger service management for high-speed railway station in the era of big data ICTE 2019. Proceedings of the 6th International Conference on Transportation Engineering pp 854-860, doi: 10.1061/9780784482742.097

[17] Ma Y, Yin W 2020 Analysis of bus line passenger flow based on big data-a case study of Guangzhou bus line IOP Conference Series: Materials Science and Engineering 790(1): 012098, doi: 10.1088/1757-899X/790/1/012098