Research Statement and Ph.D. Thesis Proposal

An Architecture Proposal for High-Performance and General Data Processing System on Big Data Cluster

Ph.D. Candidate:	**Martin Stufi** martin.stufi@solutia.cz Solutia s.r.o. Prague, Czech Republic
Advisor:	**Leonid Stoimenov, Ph.D., full professor** leonid.stoimenov@elfak.ac.rs Computer Science Department Faculty of Electronic Engineering University of Nis, Serbia

November 2020

1. Motivation

My professional and research goals are to enable everyone to succeed by developing novel computer system architecture or technological platforms, especially in the Big Data domain, AI, ML, and modern DevOps procedures, which are more accessible, useful, and efficient. Throughout our life journey, I desire to make our life in civilization more pleasant, comfortable, and of better quality. In the same way, information systems can give us that desire and make our lives better.

2. The subject of scientific research

Information systems used for data processing are often used as systems built on the "silo" architecture principle. Such an architecture is characteristic of systems that have several layers in their architecture, very often three: the application (Application Layer), the middleware (Business Layer), and the database where the databases are (Database Layer). Tuning such systems in terms of performance has many limitations, primarily based on the limits in adjusting the performance and capacity of their hardware corresponding scalability. Such restrictions, which exist, relate mostly to the maximum number of CPUs, RAM, disks, network cards, and other components installed in such computer systems.

In addition to this approach, the recent trend is that information systems are designed according to their architecture based on a larger or large number of individual, traditional, or conventional servers connected into one unit, called a Cluster. Such systems have the feature of horizontal scalability, which allows them to increase capacity and connect, based on which there can be a significant increase in performance and cluster resources.

Today, an additional requirement for information systems is a severe increase in the amount of data. In the last ten years, the information society has experienced a significant turnaround in terms of an exponential rise in the amount and types of data. As a consequence, in the field of information technology, the current and well-known phrase that appears very often is Big Data. This term is very often used as a term for large-scale data, those data that can not be stored and processed supposed Traditional Systems. As standing, computer systems face the challenge of acquiring or reading large-scale data, storing it, and processing and displaying it. In addition to the problems related to their storage and transformation, it is often necessary to ensure the processing and analysis of such data in the shortest possible time and at the same time to provide further operations on them to reach certain conclusions. An example of such procedures can be at the level of reading data, their acquiring, cleaning, and transformation into a specialized part of the data storage system on a specialized file system, operative memory and database. However, the operations that can be applied can refer to various more complex data processing in terms of application, e.g. analytical and aggregation operations. It is further possible to use several machine learning algorithms over such data. It is very often necessary to display or visualize large-scale data for their visual perception.

Such a dramatic increase in data leads to new requirements for information systems in terms of location, availability, efficient processing, and visualization. Horizontal adjustment systems are increasingly favored, increasing in popularity in terms of performance and resources and the cost of their construction, implementation, and efficiency in processing various types and kinds of data.

The main characteristic of a large amount of data is the supposed property 7V which describes their difference concerning the traditional data. Property 7V refers to *(1) Volume*, which is about how much data we have, *(2) Velocity*, data generation, and accessible speed *(3) Variety. Data* can be different, structured, non-structured, and semi-structured data. Recently, we have come across additional features of large-scale data such as *(4) Variability*, data value achieved after their interpretation; *(5) Veracity*, which refers to the data reliability in terms of their accuracy, i.e. data quality**;** *(6) Visualization*, which refers to the method of consumption of information by users and *(7) Value*, is about getting value from data.

Large-scale data arise, among other things, as the importance of society's digitalization and the increasing use of information systems in modern society in various domains. One of the vital domains concerning health, where large-scale data can be of excessive importance, leads to a critical increase in the quality of medical services to patients. Using them can decrease costs based on significantly higher work optimization.

The analysis of large-scale data in traditional systems becomes a bottleneck, primarily due to their quantity, i.e., massiveness, complexity, diversity, precision, speed, and manner of generation.

The subject of scientific research in this doctoral dissertation relates to the design of eSystem Big Data architecture. It enables efficient data acquisition, optimal data placement, processing of large amounts of data, and various algorithms for concluding and displaying or visualization of data. Additionally, the research subject will be the possibility of generalizing this architecture. The basic idea is that such an architecture is used in various domains with the necessary specific adaptations. A notable review in scientific research is related to evaluating the proposed architecture in the health field and its practical use, specifically in the health of the Czech Republic.

3. The goal of scientific research

The aim of the scientific research in this doctoral dissertation is to research and propose a solution that answers the following question: whether and how it is possible to build a modern and generalized general architecture related to a cluster of information systems for processing large amounts of data, including acquisition, transformation, and storage of large-scale data, their availability as well as their processing in a very efficient way? Additional goals relate to the problems of how to efficiently and optimally determine the capacity of the cluster information system, how the performance of the cluster architecture increases with the increase of individual servers (nodes), and how to achieve efficient performance measurements of such solutions.

The proposed architecture will ensure the planned system's efficiency and application. Due to various failures, circumstances based on the system's needs will swiftly expand and change based on performance requirements, data storage capacity, and availability and resilience. The proposal of such a new architecture with horizontal adaptability would have to be different from the architecture of the traditional systems. The essential feature and lack of traditional systems are related to the limits of increasing their capacities. The hardware limitations described to the architecture's limitations are related to, e.g. the maximum number of processors, RAM, disk capacity, and network cards.

The additional requirement of the proposed architecture relates to its application in different domains based on their specificity. Within the doctoral dissertation, an evaluation of the proposed architecture performs an example of designing and implementing such a system in the healthcare field in the Czech Republic.

Furthermore, the research within the proposed doctoral dissertation will include an analysis of achieving the required performance when designing the Big Data architecture as an analytical platform, which should be completed based on the TPC-H benchmark. TPC-H performance test to support decision-making when processing large quantities, i.e. large-scale data. As it is a large volume of data, this dissertation will present how it is possible to build a system based on predefined performance parameters, which the new system meets about required conditions. One of the goals of this research, within the proposed architecture, is to provide the possibility of adjustment in terms of performance and achieves the required results consistent with the TPC-H benchmark. Furthermore, the possibility of platform expansion is due to the exponential growth of data in the coming years, as indicated by all trends in large-scale data processing.

4. Expecting the results of scientific research

The expected results of scientific research within the doctoral dissertation focusing on the proposal of a generalized model of platform architecture enable its construction while providing an optimal and generalized way of downloading data as well as their integration, efficient and optimal placement, efficient data processing, including advanced principles and algorithms, their analysis as well as data visualization. Besides, the platform should adhere to specific standards for such high-volume data processing systems in terms of performance, TPC - H benchmark test based on predefined requirements, of which there are more than 100.

The expected results of scientific research include the fulfillment of the following goals:

- The proposed methodology for creating a cluster architecture of a system for processing large amounts of data
- A model of universal architecture that encompasses from data acquisition to their storage and presentation
- Proposal of the concrete architecture of cluster system solutions for achieving high performance for data processing
- The implementation manner of the new system architecture as well as its organization in terms of its adaptability
- Evaluation of the proposed solution in terms of efficiency and performance
- The proposed procedure based on which such a cluster system is created enables an efficient achievement of set goals with the lowest possible cost of its construction

Furthermore, the expected results will refer to the evaluation and analysis of information related to the proposed cluster architecture solution based on predefined measurements, TPC-H performance test, and the following analyses:

- The result of measuring the performance of the Cluster when using the given volume 1TB, 3TB
- Influence of Cluster performance with 3, 4, and 5 nodes concerning defined requirements

- Comparison of results based on queries Q1, Q2… Q22 based on the TPC-H performance test
- Results enable the conclusion that the Cluster performance increases when the number of nodes in the cluster increases

5. Applied scientific methods

To prepare the proposed doctoral dissertation, various research methods are used to fulfill the set goals, such as characterization, hypothesis, hypothesis-based assumption, experimental method, evaluation, and confirmation of conclusions.

Initially, the characterization method used for predefined performance parameters must be achieved by correct sizing. The next method is a hypothesis of solution architecture that could meet the expected requirements based on which the method of assumption could lead to the determination of performance parameters on the proposed Cluster, i.e. its architecture. Finally, a comparative method will be used, including comparing the results obtained during measurements based on different amounts of data 1TB, 3TB. The final method is a method of synthesis by which the individual works, individually obtained, will be combined into an appropriate methodology to create the proper cluster size for processing an enormous amount of data.

6. Bibliography of scientific and professional papers

- Štufi M, Bačić B, Stoimenov L. Big Data Analytics and Processing Platform in Czech Republic Healthcare. Applied Sciences. 2020 Jan;10(5):1705. https://doi.org/10.3390/app10051705
- Stufi, M., Janković, D., Stoimenov, L. Healthcare Information Systems Supported by RFID and Big Data Technology. In: Konjović, Z., Zdravković, M., Trajanović, M. (Eds.) ICIST 2016 Proceedings Vol.1, pp.211-215, 2016
- Stufi, M., Veljković, N., Bogdanović-Dinić, S., Stoimenov, L. Implementing an Effective Public Administration Information System: State of PAIS in the Czech Republic and its potential application in the Republic of Serbia. In: Zdravković, M., Trajanović, M., Konjović, Z. (Eds.) ICIST 2014 Proceedings Vol.2, pp.371-375, Belgrade, Serbia, 2014
- Stufi, M., Rančić, D., Ćirić, M., Stanimirović, A., Stoimenov, L. Practical experience with the change management process in software development. In: Konjović, Z. (Eds.) ICIST 2013 Proceedings, pp.183-188, 2013