Real-Time High-Load Infrastructure Transaction Status Output Prediction Using Operational Intelligence and Big Data Technologies

Solomia Fedushko [1,*] , **Taras Ustyianovych** [1] **and Michal Gregus** [2]

[1] Social Communication and Information Activity Department, Lviv Polytechnic National University, 79013 Lviv, Ukraine; taras.ustyianovych.dk.2017@lpnu.ua
[2] Faculty of Management, Comenius University in Bratislava, 83103 Bratislava, Slovakia; Michal.Gregus@fm.uniba.sk
*   Correspondence: solomiia.s.fedushko@lpnu.ua; Tel.: +38-0322-582-595

check for updates

Abstract: An approach to use Operational Intelligence with mathematical modeling and Machine Learning to solve industrial technology projects problems are very crucial for today's IT (information technology) processes and operations, taking into account the exponential growth of information and the growing trend of Big Data-based projects. Monitoring and managing high-load data projects require new approaches to infrastructure, risk management, and data-driven decision support. Key difficulties that might arise when performing IT Operations are high error rates, unplanned downtimes, poor infrastructure KPIs and metrics. The methods used in the study include machine learning models, data preprocessing, missing data imputation, SRE (site reliability engineering) indicators computation, quantitative research, and a qualitative study of data project demands. A requirements analysis for the implementation of an Operational Intelligence solution with Machine learning capabilities has been conducted and represented in the study. A model based on machine learning algorithms for transaction status code and output predictions, in order to execute system load testing, risks identification and, to avoid downtimes, is developed. Metrics and indicators for determining infrastructure load are given in the paper to obtain Operational intelligence and Site reliability insights. It turned out that data mining among the set of Operational Big Data simplifies the task of getting an understanding of what is happening with requests within the data acquisition pipeline and helps identify errors before a user faces them. Transaction tracing in a distributed environment has been enhanced using machine learning and mathematical modelling. Additionally, a step-by-step algorithm for applying the application monitoring solution in a data-based project, especially when it is dealing with Big Data is described and proposed within the study.

Keywords: IT Operations; IT Management; decision making; machine learning; operational intelligence; application performance; Big Data Analytics (BDA)

1. Introduction

The development and growth of academic and enterprise projects, which include the usage of Big Data processing technologies and intelligent systems, contribute to the spread of distributed computing, complex architectures construction for data storage and processing, and clusters creation. Despite this, it is difficult for key decision-makers to find the right solution and qualitatively evaluate infrastructure needs, develop and deploy certain IT processes, and implement monitoring solutions to acquire operational insights. The purpose of this article is to facilitate data-driven IT Operations management decisions based on BDA, Machine learning and Operational Intelligence.

Identifying the key tools and indicators for data workflow, infrastructure, user satisfaction monitoring in order to increase user experience, team-working, and project success level is equally important. Additionally, the study and monitoring methods will help to correctly implement techniques and ways of collaboration within the organization on the basis of successful use-cases and correct development of the IT solution implementation strategy.

Research goals are to identify suitable algorithms, methods, and indicators based on a solution requirements analysis, architectures and clusters development for logs data collection and load measurement, support decision-making on IT operations based on data-driven operational insights. The research objectives are to describe the most commonly used existing metrics for analyzing information technology project activities; their ability to predict and/or identify certain patterns, tools and methods to improve metrics calculation processes and obtain the best possible result; and to develop a machine learning model to predict transactional request results to identify defects or possible system errors of a particular information product and perform further root-cause analysis.

The application of Big Data and Machine learning technologies to a variety of data projects has always been challenging, as it required an architecture development capable of withstanding high loads through huge real-time datasets and complex computing processes used for intelligent systems and other functions. However, the successful application of these technologies can lead to successful data mining, certain patterns identification and bring new information to organizations. The study of Gloria J. Miller [1] investigates Big Data and Business Intelligence (BI) projects classification paying attention to similarities and differences between project types, success factors identification using quantitative methods, which is quite crucial for making correct decisions and inferences. This research allows us to observe what factors might increase productivity, accelerate a project and how real-time monitoring practices, intelligent systems usage can increase all of the above.

Additionally, Papadaki D., Bakas D.N., Karamitsos D., and Kirkham D. [2] were able to identify relationships between project success criteria and risk and project management using Big Data technologies, scientific literature databases, and social media data. This study helps classify potential uncertainties that can minimize the overall performance of a technology project, and it an important point to take into account when doing research and development activities, including monitoring implementation.

A study conducted by Gunasekaran A., Papadopoulos T. [3] showed that Big Data and Predictive Analytics technologies usage correlated with successful decision-making of organization management, and described an algorithm for successful implementation of a data-driven decision. Key business decisions and support for operations using data-driven methods is an efficient way to increase productivity, embedding analytics and requirements analysis to help assess project needs and deal with risks [4–6]. The study of Pugna, I. B., Duțescu, A., and Stănilă, O. G. [7] researches the organizational challenges raised by Big Data technologies, its impact on the business environment and performance management. The most crucial set of skills to develop a data-driven culture in the organization are as follows: goal setting, assessing benefits and limitations, learning to trust data and commitment to data discovery [8].

Machine Learning and Business Intelligence [9] has been successfully used to make the best decisions for organizations. New trends and specific solutions for the complementary application of Business Intelligence practices with machine learning are being developed. A combination of these two techniques helps reveal valuable patterns and bring value to the organization. The trend for using machine learning to augment and improve existing techniques' performance is increasing.

The practical applications of Machine Learning for Operations Management and Digital Marketing are well-studied in the research of Wang Q. [10], who presents approaches to four real-world academic/industrial problems and their data-driven solution using Machine learning. Specifically, methods for Credit Debt Collection and contact center staffing are described, namely, the use of machine learning to predict the likelihood that a person will be able to repay a debt if a bank employee communicates with him or her on a particular day of the week, which has allowed for a 14% increase

in debt repayments and improvements management operations in an organization. Using data mining to analyze a large number of log files and application/transaction data is very useful for gaining key knowledge about the infrastructure, understanding its data flow and service architecture, and identifying and predicting the effects of each data pipeline in real-time.

In particular, Stephany S., Strauss C., James A., Calheiros P. [11] developed a new data mining approach for near-real-time weather monitoring for such countries as Brazil, which lacks weather radar coverage. Data analysis and insights extraction from information arrays can cut down costs and increase productivity if the measures are accurate.

Wang Y. [12] and others proposed a data mining method based on unsupervised learning for unlabeled sheath current data monitoring that leads to effective earlier unknown patterns revealing. Therefore, applying data analysis and statistical methods for monitoring are discovered and used to solve problems related to infrastructure and hardware.

Abghari S., Boeva V., Brage J., and Johansson C. [13] and others describe higher-order mining to monitor heating substations operations and behavior. The proposed solution allows performing data mining over patterns rather than raw data using cluster analysis, sequential pattern mining, minimum spanning tree (MST) and other methods. This method shows use-cases of applying analysis to gain real-time insights as well as collect historical heating substations data to improve industrial processes. Because of the various technologies for data mining usage [14,15], different visualization types and analytics help domain specialists in understanding a substation's operational data. Machine learning models used for prediction pupils' successfulness level at educational institutions helped reveal new insights about the difference between various student/pupil groups and how it affects studying processes and results. Therefore, creating a mathematical and statistical model for this gives new knowledge to a particular field of science.

Operational Intelligence is well-suited to interact with Big Data, as it is also a kind of a large information amount analytics but is more focused on working with real-time data than traditional Business Intelligence. The tools used for Operational Intelligence can be successfully applied to BDA, one of which is Splunk [16]. Splunk offers ingesting data from multiple sources, performs analysis and monitoring directly in the application itself. Custom add-ons bring more insights and help define the strategy for operational intelligence application. In particular, many use-cases have been already developed and even successfully implemented with Splunk Enterprise and its Add-on for Machine Learning (Machine learning Toolkit) that allows a statistician or a data scientist to execute all necessary steps for statistical/machine learning model development.

Therefore, we claim that a successful combination of Operational Intelligence, Big Data, and Machine learning is capable of delivering many benefits with a reliable and correct application.

IT operations are one of the most crucial technology project success components, which is why the correct implementation of these processes will allow an organization to achieve the best result and create a reliable technology basis and infrastructure for further action. Thus, this research leads to IT operations optimization by defining the project goal implementation strategy, building the right software architecture solutions, using BDA, Operational Intelligence, and Machine Learning to monitor and compute key indicators. Developing successful use-cases will help guide organizations to the right path and successful project implementation, research, task completion, and more. Operational Intelligence including the use of Machine Learning is only just gaining its popularity and emerging as an integral component of IT operations management; consequently, many institutions are unable to use these technologies properly due to lack of experience and expertise in this field.

Many research and industrial projects that use analytics and Big Data are incapable of implementing the solutions qualitatively due to lack of experience and inability to find usage of the acquired insights. For example, in the medical field, about 84% of healthcare providers using Big Data technologies do not receive any benefit from this technology because of a lack of experience in implementing data-driven decisions and scarcity of analytic across a wide range of functions experience. In addition, 85% of

institutions [17,18], regardless of the type of activity involved in Big Data projects, fail to apply Big Data solutions and use all the benefits of it.

Thus, organizations that have not been able to take full advantage of Big Data through an improper approach are immediately moving to other modern technologies. This creates a temporary mainstream benefit from which has only a small number of institutions and departments.

Most of the studies referenced are primarily focused on historical data, high-level/overview monitoring practices without the application of mathematical modeling and machine learning, which can help to solve a huge range of problems related to industrial data projects. A need for site reliability practices is needed to augment APM (application performance monitoring) and analytics. The novelty of the study is to obtain a highly accurate model for transactions tracing in a real-time distributed environment based on machine learning and mathematical modeling techniques together with Operational intelligence and SRE to enhance technology, especially data-based projects' success levels, get extra insights into what is actually happening in a particular tier/layer of a data/IT operation process. A proper way of implementing an Operational Intelligence solution within a technology project is proposed and described. This allows project members to maintain a system availability at a high level, monitor particular operations, and identify particular problems early. Real-time monitoring practices are crucial and do not only focus on historical data, which is valuable, but also on real-time insights to perform well-arranged two-level (historical and current datasets) analysis. The rest of the paper is structured as follows: Section 2 contains materials and methods for the correct implementation of Operational Intelligence best-practices, data preprocessing and processing methods used, and key metrics used for APM and management; Section 3 contains the results with a proposed machine learning model for real-time transactions tracing, dataset and environment specifications and challenges; Section 4 highlights the main characteristics that the research is based on, requirement analysis for real-time monitoring solution including Splunk, a novelty approach for dealing with huge amounts of real-time data; Section 5 contains a conclusion of the main study focus and findings, and states a specific aspect that needs to be studied more within further researches.

2. Materials and Methods

2.1. Main Research Methods Used

The main focus in the study is done on real-time data and ingesting logs to obtain operational intelligence insights with the application of machine learning and mathematical modeling to the continuous data stream, performing data-preprocessing on the fly, and making use of site reliability metrics to predict about 10 classes of transaction output. The methods include traditional data processing (statistical analysis, feature engineering, data ranges, ratios calculation) as well as the one with more focus on APM and service level objectives.

Particular methods include traditional data preprocessing: data ranges of certain features; normalization as a way to improve model accuracy, PCA (principal component analysis) used to select the best features and reduce dataset dimension, imputation, and removal of missing data; the research benefited from statistical analysis and feature engineering; and application performance monitoring has been conducted to get a more accurate view of the system functionality. SRE techniques were used to define and enhance service level indicators, calculate them thoroughly and make modifications with the metrics, for example, showing the remaining budget for errors, downtimes not only as standard quantity but also as a percentage or in minutes. Applying real-time monitoring for system observability enhanced data pipeline functionality and transaction workflow. Use-cases study of applying machine learning models to improve and accelerate real-time monitoring has been studied to develop a strategy for certain technology implementation; quantitative methods have been processed to get an overview of data projects in industry; qualitative analysis of steps to meet project requirements through monitoring, and transactions tracing has been done.

To conduct the study, Splunk as data ingestion and storage, and the its Add-on for Machine learning has been used. It is beneficial because data pre-processing is made quite carefully with Splunk search processing language; real-time monitoring dashboards inside Splunk allow the observation of the needed metrics and conduct all the steps in a single product. Additionally, Splunk has a good data compression algorithm which allows it to quickly process huge volumes of data. Nonetheless, there are other options to be used for the above-described methods and steps—for example, Elasticsearch or a specific programming language with the ability to process real-time data stream, and using machine learning/data modeling—but it depends on the amount of data to be indexed and requirements.

The problems that have been encountered during the research are getting a lot of messy data with missing values, ingesting huge amounts of data in real-time, selecting not only the best features but also the proper for model training data, and giving detailed attention to dataset collection because it should include various transaction outputs with a certain ratio.

2.2. Operational Intelligence Analytics and Engineering in a Technology Data-Based Project

Data analysis projects have specific specifications and requirements, such as the ability to collect and store data, information security, environment for data storage availability, data quality and data manipulation engineering [19,20], a strategy for data usage (which includes goals and aims that are related to the business problem), tools and software for data processing. [21]. In order to meet all the above requirements, the article proposes to apply Operational Intelligence Monitoring using machine learning and BDA. As shown in Figure 1, the initial stage of any Data project is to identify a problem that needs to be solved with data and information technology and requires the use of root–cause analysis to identify, discuss, and determine the severity of the business problems. After that, in certain projects due to lack of time and experience, they often miss the second and third steps and immediately move on to data collection and analytics. However, the article argues that existing or similar solutions analysis needs to be done, an effort to successfully implement a similar data-project may have already been done at a certain institution and find out what benefits it has brought to those organizations, whether it is able to deliver it to those who have encountered a particular business problem. After conducting a detailed use-case analysis, it ought to be found out what information is needed, how to collect it, and where to store it. The next key step is Data Engineering. In case it is enough to be limited to a small dataset, a single-cluster environment will be deployed that will allow data storage and collection and making its analytics. Very often, in this case, open-source solutions are used.

If it is needed to collect huge amounts of data but also with the use of real-time data stream mining, a complex architecture solution for data ingestion and distributed computing is developed. In such cases, special attention is paid to the usage of cloud technologies capable of processing and storing large amounts of data or local data centers [22]. In parallel, the processes of building an Operational Intelligence solution must occur. This is required in order to provide data workflow monitoring, avoiding missing data and gaps [23], service reliability monitoring, and predict service downtime, information security, data privacy, quality and integrity checking; compute application performance specific metrics; implement system reliability engineering practices; and to prevent data and infrastructure issues. The next step after completing data engineering is analytics, data science, especially exploratory data analysis, statistical models' development and selecting the most accurate one that will be used.

A model that is trained on a pre-collected and refined dataset is deployed on a production environment and Operational Intelligence is used again not only to monitor all the above but also to validate and evaluate the model's accuracy during its usage. Additionally, some mathematical modeling and machine learning tasks for various industry purposes are likely to be shown on real-time dashboards, which can be a good way to monitor model performance as well as show its output to the users. If accuracy does not meet the minimum threshold, project members responsible for the model training and data science get notifications and alerts for accuracy which is not high enough

to meet project requirements. If all is well and the data project has allowed resolving the initial task, monitoring continues until the completion of data engineering and data science parts.
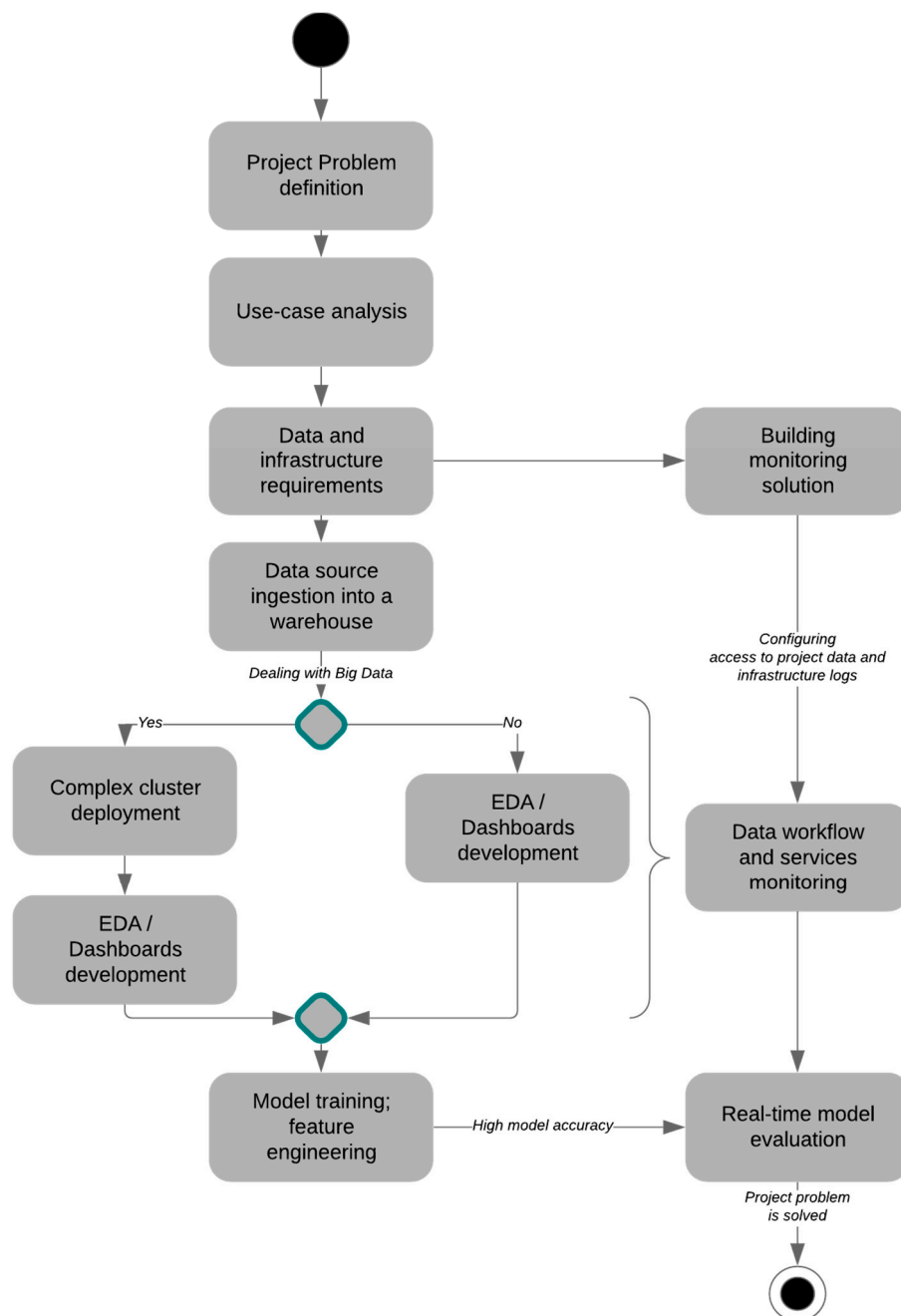
Figure 1. Data project with Operational Intelligence Practices usage process.

Thus, the use of Operational Intelligence occurs in a way that monitors the system and research and development processes and depends on the amount of data collected and processed, as well as the load on the infrastructure to prevent its failure or incorrect functionality.

To measure the infrastructure load level, specific metrics help to do this possible. In order to implement an operational intelligence solution inside a technology/data-based project, some indicators, especially those that are shown in the Table 1, need to be calculated or event ingested with the data itself. Those metrics help get insightful statistics about each component of a system, detect vulnerabilities and transform raw data into knowledge and business intelligence. Monitoring availability and performance increases user-satisfaction level. Thus, the metrics can potentially help observe the most vulnerable

and weak application components, perform system analysis, and accomplish a service. The metrics are also related to the field of SRE and defining service level objectives, which is another deep topic for research. These metrics are very log/application-specific because they show transaction/request output and describe human–computer interaction with the system. That is what makes them valuable for an analyst to be processed and project members to monitor, get knowledge and valuable information. Additionally, machine learning can enhance this experience through intelligent forecasting of these metrics, applying them as features during data modeling. In other words, a machine learning model for application performance specific tasks might be improve one of those metrics or get more accurate than it has been using them. For a detailed description of the indicators used in application performance monitoring and management, see Table 1.

Table 1. Application Performance Monitoring Metrics.

S. No	Metrics Explanation	
	APM Metric/Indicator	**Description**
1	Apdex Score	The metric, also known as an Application performance index, is an open-source standard, which allows splitting user requests into three groups (satisfied, tolerated, frustrated) by providing a static/dynamic thresholds which should be met, and measures their satisfaction/dissatisfaction level. To measure the satisfaction level, for instance, number of satisfied requests is summed up with half of the number of tolerated ones and divided by the total number of requests.
2	Response time	Amount of time a request or a transaction takes to get processed by a system. It is the elapsed time between the time when client has sent a request and a server receives it. Usually displayed in milliseconds.
3	Throughput	The rate at which a request is processed. Usually measured as requests per minute (rpm). To calculate it, the number of requests received need to be divided by their response time sum.
4	Error/Availability rate	Error rate stands for percentage of errors that occur during a period. Availability rate is opposite to error rate, as it displays number of successful transactions per a time span. It is a ratio of certain request group.
5	Error budget	The metrics provides a determination of how unreliably a system of a service is allowed to be within a particular time range. Is measured by defining service level objectives or thresholds which need to be met and according to them a budget for these is created. It can be measured in various quantities: the remaining number/costs/percentage/minutes.
6	Magnitude	Correlation between current maximum and historical average values (response time, number of errors, unique users). This metric compares application specific metrics and shows their difference before and after a system upgrade; helps system engineers get more observability and collect metadata about application specific information.
7	Bounce rate	Usually shows number of visitors to a website who enter the site and leave it after showing the first page. The metric might be applied to show ratio of users with more than one unique IP Addresses or number of single/multiple actions per transaction.

These metrics allow Operational Intelligence engineers to do high-quality and reliable system and infrastructure monitoring on which data will be stored and processed, user interaction, and data collection and integrity verification should be checked. Very often, not only management but ordinary engineers or technology project developers are devoid of operational insights, not always able to observe infrastructure load and due to this cannot observe and prevent possible downtimes. That is why the implementation of operational intelligence methods and metrics calculation for application

performance monitoring will allow project members to collect application logs, queries, perform their analytics in specialized software, perform security enhancement operations, service troubleshooting and root-cause analysis.

Each of these metrics and indicators provides user interaction with the infrastructure, data flow monitoring and certain vulnerabilities and malware detection. For example, response time for operational intelligence engineering is a very crucial factor to have because it allows complex indicators computation. Keeping an eye on them for a management level helps make new decisions and leads to their proper implementation. Application performance management is one of the key components of Operational Intelligence that corresponds to determining infrastructure load, user satisfaction level, calculating various ratios, collecting log files and query information in real-time. It is used for site/service reliability engineering, alerting and monitoring purposes.

A no less important indicator of an application stable operation is downtime. The metric displays the amount of time when a service component (node, cluster) or even the entire service was unavailable to users or did not perform its functions as expected. It is identified by receiving the 5XX HTTP response status code, which means that there is a server-side error in an application. Downtime might not allow a service to index information, conduct analytics, cause service unavailability for users. It should be thought of its prevention in the early stages of application architecture development, since downtime itself causes a lot of damage to the project, suspending the transactions, data flow and more. In addition to using standard HTTP status codes, we recommend that a system owner uses custom system status codes to help easily identify issues and their causes.

Therefore, using Operational Intelligence will allow it to be identified and minimized. The effect of downtimes on the information system in healthcare is studied in the research by Wang Y. and others [24]. Network downtime minimization techniques are being developed for future use in data centers and data warehouses [25].

3. Results

3.1. Problem Definition and Tools that Were Used

This article describes the development process and the model for predicting the transaction/request outcome based on various factors and conditions and specially features, such as request type, method, transaction duration, error rate, main application, backend or the system, hostname, request message, environment that is split into testing and production, user-agent, and timestamps including day of the week, hour and minute of the request as separate features. The data source used for this research is important as well. A detailed attention should be paid to its description. The data source is a high-load distributed system with three main applications and about 20 correlated backends and throughput of about 700 requests per minute. Each transaction goes through all applications and is processed by one backend which depends on the transaction purpose that is selected by a user/test executor. Therefore, the transactions are processing in a pipeline and an error might occur on one of the stages. The objective is to identify on a certain stage the error or prevent it as soon as possible using machine learning and operational intelligence and avoid similar circumstances during the next requests processing. Using the Operational Intelligence and Big Data processing Splunk tool with its custom Add-On Machine learning toolkit (MLTK) the data was indexed, preprocessed and processed, and a model was trained.

This model allowed user-side or backend services error detection before they could arise for a user by predicting a request or transaction output. It also allows thousands of service performance behavioral test cases to identify possible system issues, errors, vulnerabilities and prevent them from being encountered by the user.

3.2. Model Development and Validation

The model was developed using the application and transaction log data collected by Operational Intelligence monitoring through various data inputs. To train the model, a dataset of 106,417 information

entities was used, cleaned and normalized using data pre-processing including missing data imputation, normalization and applying custom data ranges. Selected parameters have the maximum ratio and relationship to request output, namely HTTP Method, Back-end, application name, user-agent, request message, response time, hostname, exact timestamp, day of a week, and hour of the day. In total, 91% of selected for model transactions have a successful output, whereas the percentage of error transactions is the dataset is equal to 9%. The distribution of error transactions is as follows: 89.8% of failed transactions are due to server error; 10.2% is because of a user or client-side incorrect/failed requests. The split for training/test is equal to 80/20, accordingly. The metrics listed in Table 1 were also applied, which was a feature of the engineering task and helped to increase model accuracy. The algorithms, including Decision Tree Classifier, Random forest, linear regression, K-NN, were used to train the model. The Decision Trees model accuracy was 99.99% and allowed us precisely to know transaction output. The accuracy for this classification method is actual and predicted features, corresponding to the real and predicted labels, accordingly. The max depth of the tree is set to be expanded until all leaves of the node get pure. The Gini impurity criterion was used to measure the criterion also known as the quality of a split. This is also shown in Figure 2. In all cases, the result accuracy was quite high due to the use of the necessary factors for prediction, categorical and continuous values, as well as the sufficient amount of data that the model was "fed" with. Thus, the study of the system performance and the validation of receiving an error with a given set of parameters at the application development and testing stages were done.

Predict Categorical Fields for Predicting Transactions Response HTTP Status
Predict the value of a categorical field using the values of other fields in that event.

Experiment Settings　　**Experiment History**

i	Precision ⇕	Recall ⇕	Accuracy ⇕	F1 ⇕	Algorithm ⇕	Notes ⇕	User ⇕
>	1	1	1	1	DecisionTreeClassifier		admin
>	1	1	1	1	DecisionTreeClassifier		admin
>	1	0.99	0.99	0.99	DecisionTreeClassifier		admin

Figure 2. Accuracy after three conducted model evaluations in Splunk Machine learning Toolkit.

Figure 3 shows the confusion matrix for model accuracy after training and validation steps. Due to the presence of transactions with different status codes, huge and well-sampled dataset allows us to qualitatively design and train the machine learning model. Using the system for Operational Intelligence and Monitoring Splunk Enterprise enables automation of the processes of data collection and model training for transaction monitoring and predicting its final output in high-load applications.

Classification Results (Confusion Matrix) ↗

Predicted actual ⇕	Predicted 200 ⇕	Predicted 401 ⇕	Predicted 403 ⇕	Predicted 408 ⇕	Predicted 500 ⇕
200	348 (99.7%)	1 (0.3%)	0 (0%)	0 (0%)	0 (0%)
401	0 (0%)	2 (100%)	0 (0%)	0 (0%)	0 (0%)
403	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)
408	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)
500	0 (0%)	0 (0%)	0 (0%)	0 (0%)	31 (100%)

Figure 3. Confusion matrix of HTTP Status code prediction.

Machine learning models are getting complex and their monitoring is required for the reason of delivering highly accurate predictions for various industry purposes [26]. Transactions tracing

is an in-demand feature that allows a system engineering team to check every tier of the data flow, monitor information processing pipelines, conduct thorough performance testing and find out potential problems if any occur. This experience might be enhanced using Operational Intelligence technologies described in this paper as well as apply statistical/machine learning models to get new knowledge and insights about the application performance itself.

4. Discussion

In total, 77% of the analyzed literature in this research apply historical data for certain academic or industrial projects and research activities, make little use of complex monitoring using operational intelligence tools and, in rare cases, are not able to analyze and interpret insights, apply the best practices for data project processes and implementation, which are very crucial to meet certain requirements. The remaining 23% of the analyzed literature does not make use of machine learning and SRE in a real-time data stream, applying monitoring solutions for advanced transactions tracing and defining service level objectives. Whereas our research is focused on combining historical data analysis as well as continuous real-time monitoring with specific interest and application performance monitoring, Big Data-related processes improvements through continuous monitoring, and combining machine learning, mathematical modeling practices.

The research is based on working with real-time data, applying machine learning models for a tabular dataset using the Operational intelligence tool Splunk and SRE methods for application performance monitoring. The developed machine learning model has made a difference to improve transaction tracing. It helps identify errors, enhance operations, data pipelines to make a project requirement precise, identify use-cases and apply monitoring for project improvement. Continuous real-time monitoring combined with machine learning for a certain industrial operational use-case allows a system to increase availability which is one the factors that lead to higher user satisfaction levels.

As mentioned above, Splunk has been used to conduct the research. It offers multiple license types for ingesting a needed amount of data. Free development/research licenses for a non-production environment that allows ingesting up to 50 GB data per day for six months, and has been used for research and development activities. For small enterprises and organizations, there is also a free license with limitations to ingest less than 500 MB of data per day. The exact costs need to be clarified in the Splunk itself and they may be different for various regions.

Multiple calculators are available online to measure how much storage is required and what number of nodes is needed for ingesting a certain amount of data in a real-time monitoring tool. The Splunk Sizing calculator has been used to conduct requirement analysis for this research [27]. For Splunk, in our research, a daily data volume is met with the development license. Up to 50 GB of data has been ingested regularly, the total retention size is set to 45 days (the data is stored up to 45 days). For using Splunk Machine learning toolkit, it is recommended to have a separate node. Therefore, the total number of nodes suggested is two but might be decreased to only one. As an outcome, a total storage equaling 300 GB is recommended considering the specified above requirements. One of the most important steps to implement operational intelligence practices and meet project requirements is to have a clear strategy of what should be done in the scope of a certain data project, collect insights from operational intelligence to meet a technology project requirements, and even enhance it with Machine learning.

The current limitation is that the model has been trained on a collected and previously cleaned data set, missing data has been removed. In further researches, we will focus on the possibility to select proper data entities for model training in real-time. The main issue is that not all data is clean enough and dealing with noisy data entities is important as well. Additionally, time-series analysis, forecasting for application performance purposes needs to be considered for application performance purposes. Using cases of applying machine learning in infrastructure monitoring is advantageous to identify successful BDA project patterns.

5. Conclusions

Operational intelligence will allow you to deploy and refine IT Operations, ingest logs data for monitoring purposes, keep project members, managers informed about current application status. Developed model to determine the probable transaction status code allows technology project members to identify new insights about vulnerabilities and weaknesses in the infrastructure, application or software before the user encounters them. It also allows users to more efficiently communicate with application support services, perform causal analysis through Big Data mining, Machine learning, and perform multiple tests to analyze performance issues and prevent production errors through their identification on testing stages. A focus is made on adopting real-time monitoring to increase data or technology project operations, analyzing not only historical but also real-time insights, developing a proper strategy and implementing it using data-driven methods, using site reliability and machine learning to make infrastructure availability up to 100% availability.

The created machine learning model enhanced and made it possible to perform transaction tracing in a distributed environment; applying SRE metrics together with mathematical modeling is beneficial to improve various processes. A lot of technology projects lack operational insights, advanced infrastructure, data pipelines monitoring and applying mathematical modeling to prevent errors; the model can be used to execute various test-suites and scenarios, and monitor its output as well. As it has turned out, key data project success factors are not only infrastructure availability, a team of analysts who perform certain data engineering and analysis processes, but also the implementation of real-time application monitoring, certain indicators calculation to determine service's current state, develop operational intelligence best practices. Prediction models ought to be evaluated and accuracy monitoring with thresholds might be done in real-time as well. Therefore, Operational Intelligence, Machine Learning and Big Data Analytics is an evolving trend in the era of cloud and quantum computing and has the potential to lead to high-quality data project implementation.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Miller, G.J. Quantitative Comparison of Big Data Analytics and Business Intelligence Project Success Factors. *Inf. Technol. Manag. Emerg. Res. Appl. Lect. Notes Bus. Inf. Process.* **2018**, *346*, 53–72.
2. Papadaki, M.; Bakasand, N.; Ochieng, E.; Karamitsos, I.; Kirkham, R. Big Data from Social Media and Scientific Literature Databases Reveals Relationships Among Risk Management, Project Management and Project Success. Project Management and Project Success Symposium, September 2019. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3459936 (accessed on 26 September 2019).
3. Gunasekaran, A.; Papadopoulos, T.; Dubey, R.; Wamba, S.F.; Childe, J.; Hazen, B.; Akter, S. Big data and predictive analytics for supply chain and organizational performance. *J. Bus. Res.* **2017**, *70*, 308–317. [CrossRef]
4. Kryvinska, N. Building consistent formal specification for the service enterprise agility foundation. *J. Serv. Sci. Res.* **2012**, *4*, 235–269. [CrossRef]
5. Gregus, M.; Kryvinska, N. *Service Orientation of Enterprises-Aspects, Dimensions, Technologies*; Comenius University in Bratislava: Bratislava, Slovakia, 2015; p. 294. ISBN 9788022339780.
6. Kryvinska, N.; Gregus, M. *SOA and Its Business Value in Requirements, Features, Practices and Methodologies*; Comenius University: Bratislava, Slovakia, 2014; ISBN 9788022337649.
7. Pugna, I.B.; Duțescu, A.; Stănilă, O.G. Corporate attitudes towards Big Data and its impact on performance management: A qualitative study. *Sustainability* **2019**, *11*, 684. [CrossRef]

8.　Kaczor, S.; Kryvinska, N. It is all about Services—Fundamentals, Drivers, and Business Models. *J. Serv. Sci. Res.* **2013**, *5*, 125–154. [CrossRef]

9.　Reshi, Y.S.; Khan, R.A. Creating business intelligence through machine learning: An Effective business decision making tool. *Inf. Knowl. Manag.* **2014**, *4*, 5–75.

10.　Wang, Q. Machine Learning Applications in Operations Management and Digital Marketing. Ph.D. Thesis, Amsterdam Business School Research Institute, Amsterdam, The Netherlands, 2019.

11.　Stephany, S.; Strauss, C.; Calheiros, A.J.P.; de Lima, G.R.T.; Garcia, J.V.C.; Pessoa, A.S.A. *Data Mining Approaches to the Real-Time Monitoring and Early Warning of Convective Weather Using Lightning Data. towards Mathematics, Computers and Environment: A Disasters Perspective*; Springer: Cham, Switzerland, 2019. [CrossRef]

12.　Wang, Y.; Ye, H.; Zhang, T.; Zhang, H. A data mining method based on unsupervised learning and spatiotemporal analysis for sheath current monitoring. *Neurocomputing* **2019**, *352*, 54–63. [CrossRef]

13.　Abghari, S.; Boeva, V.; Brage, J.; Johansson, C.; Grahn, H.; Lavesson, N. Higher Order Mining for Monitoring District Heating Substations. In Proceedings of the International Conference on Data Science and Advanced Analytics (DSAA), Washington, DC, USA, 5–8 October 2019; pp. 382–391.

14.　Fedushko, S.; Ustyianovych, T. Predicting pupil's successfulness factors using machine learning algorithms and mathematical modelling methods. In *International Conference on Computer Science, Engineering and Education Applications*; Springer: Cham, Switzerland, 2019; pp. 625–636.

15.　Shakhovska, N.; Nych, L.; Kaminskyj, R. The Identification of the Operator's Systems Images Using the Method of the Phase Portrait. In *Advances in Intelligent Systems and Computing*; Springer International Publishing: Cham, Switzerland, 2017; pp. 241–253.

16.　Zadrozny, P.; Raghu, K. *Big Data Analytics Using Splunk: Deriving Operational Intelligence from Social Media, Machine Data, Existing Data Warehouses, and Other Real-Time Streaming Sources*; Apress: New York, NY, USA, 2013; p. 376.

17.　Tkachenko, R.; Izonin, I. Model and Principles for the Implementation of Neural-Like Structures based on Geometric Data Transformations. In *Advances in Computer Science for Engineering and Education*; ICCSEEA2018. Advances in Intelligent Systems and Computing; Springer: Cham, Switzerland, 2019; pp. 578–587.

18.　Assay, M. 85% of big data projects fail, but your developers can help yours succeed. *Big Data Tech. Rebuplic* **2017**, *11*, 1–5. Available online: https://www.techrepublic.com/article/85-of-big-data-projects-fail-but-your-developers-can-help-yours-succeed/ (accessed on 26 March 2020).

19.　Wang, R.; Strong, D. Beyond accuracy: What data quality means to data consumers? *J. Manag. Inf. Syst.* **1996**, *12*, 5–33. [CrossRef]

20.　Saha, B.; Srivastava, D. Data quality: The other face of big data. In Proceedings of the 30th International Conference on Data Engineering, Chicago, IL, USA, 31 March–4 April 2014; pp. 1294–1297.

21.　Ismail, A.; Truong, H.L.; Kastner, W. Manufacturing process data analysis pipelines: A requirements analysis and survey. *J. Big Data* **2019**, *6*, 1–26. [CrossRef]

22.　Skourletopoulos, G.; Mavromoustakis, C.; Mastorakis, G.; Dobre, C. Big Data and Cloud Computing: A Survey of the State-of-the-Art and Research Challenges. *Adv. Mob. Cloud Comput. Big Data 5G Era. Stud. Big Data* **2016**, *22*, 23–41.

23.　Fedushko, S.; Ustyianovych, T. Medical card data imputation and patient psychological and behavioral profile construction. *Procedia Comput. Sci.* **2019**, *160*, 354–361. [CrossRef]

24.　Wang, Y.; Coiera, E.; Gallego, B.; Concha, O.P.; Ong, M.S.; Tsafnat, G.; Roffe, D.; Jones, G.; Magrabi, F. Measuring the effects of computer downtime on hospital pathology processes. *J. Biomed. Inform.* **2016**, *59*, 308–315. [CrossRef] [PubMed]

25.　Wackerly, S.; Clark, C.F. Minimization of network downtime. U.S. Patent 10,601,701, 24 March 2020.

26.　Izonin, I.; Tkachenko, R.; Kryvinska, N.; Tkachenko, P.; Greguš, M. Multiple Linear Regression based on Coefficients Identification using Non-Iterative SGTM Neural-Like Structure. In *Advances in Computational Intelligence*; Rojas, I., Joya, G., Catala, A., Eds.; Springer Science+Business Media: Berlin/Heidelberg, Germany, 2019; pp. 467–479.

27.　Splunk Sizing. Available online: https://splunk-sizing.appspot.com/ (accessed on 10 April 2020).