УДК 330.43
JEL classification: C44

Shperlov R. V.
ORCID ID: 0000-0003-2430-1011

Zhukovska O. A.
Cand. physics and mathematics, associate professor
ORCID ID: 0000-0003-1110-9696

National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"

ECONOMIC AND MATHEMATICAL MODELING OF THE LOAD INCREASE OF HIGH-LOAD SYSTEMS

ЕКОНОМІКО-МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ПРИРОСТУ НАВАНТАЖЕННЯ HIGH-LOAD СИСТЕМ

An analytical and forecasting simulation of the main indicator of the information infrastructure load measurement - the number of requests to the system per second was carried out in the article. The statistics of this indicator are investigated in order to find and identify the trend of movement of its time series for ensuring the reliability and accessibility of services of IT-company. The relevance of this research is getting some resonance in the information technology industry at this time, because expanding information systems to the ability to transmit petabytes of data per second requires a clear plan for their infrastructure. Providing a minimal forecast and data estimate can not only guarantee the stability of a software product or technology, but also directly affect the cost of production and the amount of profit received. The timely detection of system crashes can prevent the domino principle, which can cause harm to other synergistic processes of the company. The long-term outlook provides companies with some retreat paths to scale their infrastructure size and provides some time to make decisions about purchasing equipment or expanding their capabilities at "Cloud Services".

To this end, the Holt forecast was constructed on the basis of statistics data, the statistical detection of abnormal emissions and their removal by the moving average method were performed. Data were grouped to increase the forecast periods and the possibility of using the Holt method was tested. The residuals were analyzed, which showed the adequacy of the model. The results of the forecast are presented in the form of generated reports of the program mathematic product. In the final stage, the results covered 16 forecast periods, which are generally equivalent to two full days. Estimates of cost savings for the average physical configuration of the receiving party and the minimum client software have been announced. On the basis of these data, it was concluded that the importance of conducting research data using such methods of forecasting and analysis of data for different vectors of activity of an IT company.

Keywords: trend analysis, Holt method, scalability, QPS, server infrastructure, resiliency, internet traffic.

У статті було проведено аналітико-прогностичне моделювання основного показника вимірювання навантаження інформаційної інфраструктури - кількість запитів до системи в секунду. Досліджено статистичні дані цього показника з метою знаходження та

виявлення тенденції руху його часового ряду для забезпечення надійності та доступності сервісів IT-компанії. Актуальність даного дослідження отримує певний резонанс у індустрії інформаційних технологій на даний час, адже розширення інформаційних систем до можливості передачі петабайтів даних в секунду потребує чіткого плану їх інфраструктуризації. Забезпечення мінімального прогнозу та оцінки даних може гарантувати не тільки стабільність програмного продукту чи технології але й неопосередковано вплинути на собівартість продукції та об'єм отриманого прибутку, а своєчасне виявлення падінь системи може запобігти принципу «доміно», що уможливлює нанесення шкоди і іншим синергетичним процесам всередині компанії. Прогноз на довготривалий період забезпечує компанії певні шляхи відступу для масштабування її інфраструктурних розмірів та надає певний запас часу на прийняття рішень щодо закупівлі обладнання чи розширення потужностей у «Cloud Services".

З цією метою на основі статистичних даних був побудований прогноз методом Хольта та проведено статистичне виявлення аномальних викидів та їх видалення методом ковзної середньої. Було здійснено групування даних для збільшення прогнозованих періодів та перевірена можливість застосування методу Хольта. Проведено аналіз залишків, що показав адекватність моделі. Отримані результати прогнозу представлені у вигляді сформованих звітів програмного математичного продукту. В кінцевому етапі результати покривали 16 прогнозованих періодів, що в цілому еквівалентні двом повноцінним добам. Були оголошені приблизні оцінки заощаджених витрат для середньої фізичної конфігурації приймаючої сторони та мінімальної програмної із сторони клієнта. На основі цих даних був зроблений висновок, щодо важливості проведення даних досліджень із використанням таких методів прогнозу та аналізу даних для різних векторів діяльності IT-компанії.
Ключові слова: трендовий аналіз, метод Хольта, масштабованість, QPS, серверна інфраструктура, відмовостійкість, інтернет-трафік.

Introduction. The rapid growth of digital technologies and the virtualization of doing business have led to an increased scalability of the infrastructure architecture of information systems, both horizontal and vertical. Sustainable high-load systems have radically changed the vector of doing business, which has grown into a kind of arms race for competitors. The main qualitative indicators of which are not accuracy and strength of destruction, but reliability and productivity. The efficiency of operating an IT-based business depends essentially on the capacity of the IT infrastructure and the productivity of its resources. Any long-term decline of such systems entails a number of significant problems: large financial losses, loss of reputation of the company and loss of its position in the market as a reliable partner, which is the main catalyst for the development and movement of the company's activity.

The term "high load system" hides a complex, interconnected structure of servers or services characterized by a large number of requests per unit of time to be processed and transmitted. Overloading such systems, that is, the inability to process a massive data stream, causes a collapse, a drop in productivity, and renders the entire business process in a state of disrepair, causing a number of economic problems mentioned above.

With such risks, any information system controls two basic criteria as it increases in scale and increases its computing power: fault tolerance and accessibility [1]. One of the main indicators of system load is the so-called QPS (query per second) - the number of system requests per second [2]. The value of this indicator is a heart rate monitor for a high-load query-type system. The quality of monitoring and control of which is inversely proportional to the fault tolerance of the system.

This indicator also has an economic meaning, because the maintenance of its value is embedded in the cost of the technology produced by the company. Dynamically changing it from the expected, you can either reduce the cost of maintaining the IT infrastructure (banal by reducing the number of servers), or respectively increase them to expand technology for the consumer, which will lead to increased profits. Optimizing a game with this metric has two sides to the coin:

– positive-qualitative impact on the cost of technology and consolidation of competitive position of the company in the market;

– unexpected financial losses and bankruptcy risk.

In order to stabilize the changing of these possible scenarios while optimizing this indicator on a good quality condition, forecasting future values of the indicator is an integral part, which necessitates a great practical need for research on a number of data and finding optimal, adequate and theoretically sound methods.

In accordance with the urgency of this problem, different views on the solution of this problem with possible internal deviations of the input parameters of information systems have been repeatedly highlighted. Recent works covering this topic were the works of Kashin M.M. "Development of congestion management methods in SIP networks based on signal traffic forecasting" using Hirst parameter estimation, analysis of variance, as well as ARIMA and FARIMA models [6]. The work of Bagmanov V.H., Komissarov A.M., Sultanov A.Kh. "Prediction of teletraffic based on fractal filters" is based on the Kolmogorov-Wiener filter [7]. Also relevant are statistical methods for detecting traffic anomalies to prevent unauthorized attacks on the information system, this issue is considered in the work of Sudarikov R.A. "Analysis of information features in the problems of detecting traffic anomalies by statistical methods" [8]. Domestic scientists Kuchuk G.A, Mozhaev A.A. in their work "Traffic forecasting for congestion management of an integrated telecommunications network" was based on the estimation of the Hirst parameter and approximation using the Pareto distribution [9]. Making certain conclusions, we can emphasize that the study of statistics and the construction of a forecast model of system loads are necessary and carried out in accordance with the needs of the objects of influence, in our case the needs of the IT company.

Setting objectives. The purpose of this article is to analyze the data and apply the method of forecasting the load on the infrastructure of the IT-company, which will reduce the financial losses of the company and consolidate competitive positions in the market.

Methodology. The basis of the study is the work of domestic and foreign scientists to predict the load of power grids [3], as they are a functional analogue of a high-load system with input parameters of influence. The main tools used are the medium moving model, the Holt model.

Results of the research. In order to cover both the significant outcome of the forecast and its practical method, it is necessary, first of all, to have really real data that was created by the activity of the high-load IT infrastructure. One of the main sales-house companies in the Internet advertising market of Ukraine was obtained for the research. On the Fig. 1 shows the dynamics of change in the value of the aggregate QPS average for the server infrastructure over the 6 months of 2019.

Further manipulation of this time series requires some manipulation to help identify and eliminate anomalies that are not specific to the time series that distort the overall picture. One such method of cleaning a number of abnormal values is the moving average method. Using the simple average method, smoothing is performed on the basis of averaging the existing data over a period of time, which is the so-called smoothing window.

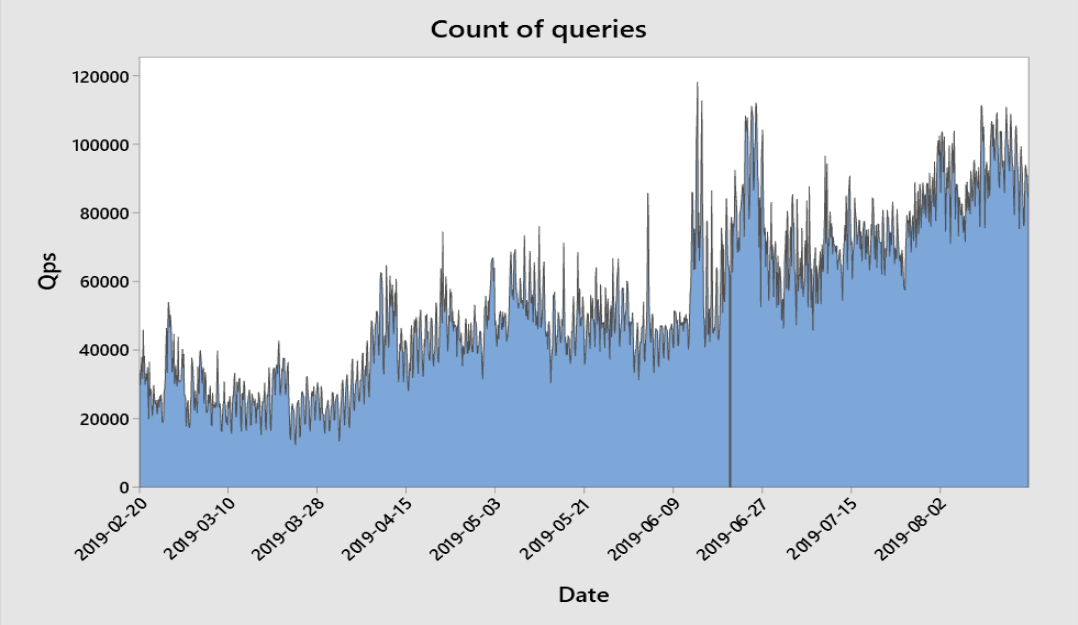

Figure 1 - Graph of the dynamics of QPS change

As you increase this window, the row will gradually move to the average of the entire row. In determining each new smoothed value, the moving average eliminates the last value from the averaging window and captures the following (1).

$$SMA = \frac{\sum_{i=1}^{n} q_i}{n}, \quad (1)$$

where q_i – query per seconds; n - the magnitude of the smoothing window.

With different variations of the row window, its optimal index was chosen, which eliminates unnecessary anomalous emissions and minimally distorts a number of features (Fig 2).

In the case of this series, the window with n = 3 is the most optimal smoothing window, since it minimally distorts the series while clearing out emissions. As the series is ready for further work with the forecast, we will determine the presence of a trend, which is a basic condition when applying the Holt forecasting method. The trend in this time series can be detected without the use of appropriate methods - just by looking at the formed trend line, which is gradually increasing (Fig. 3).

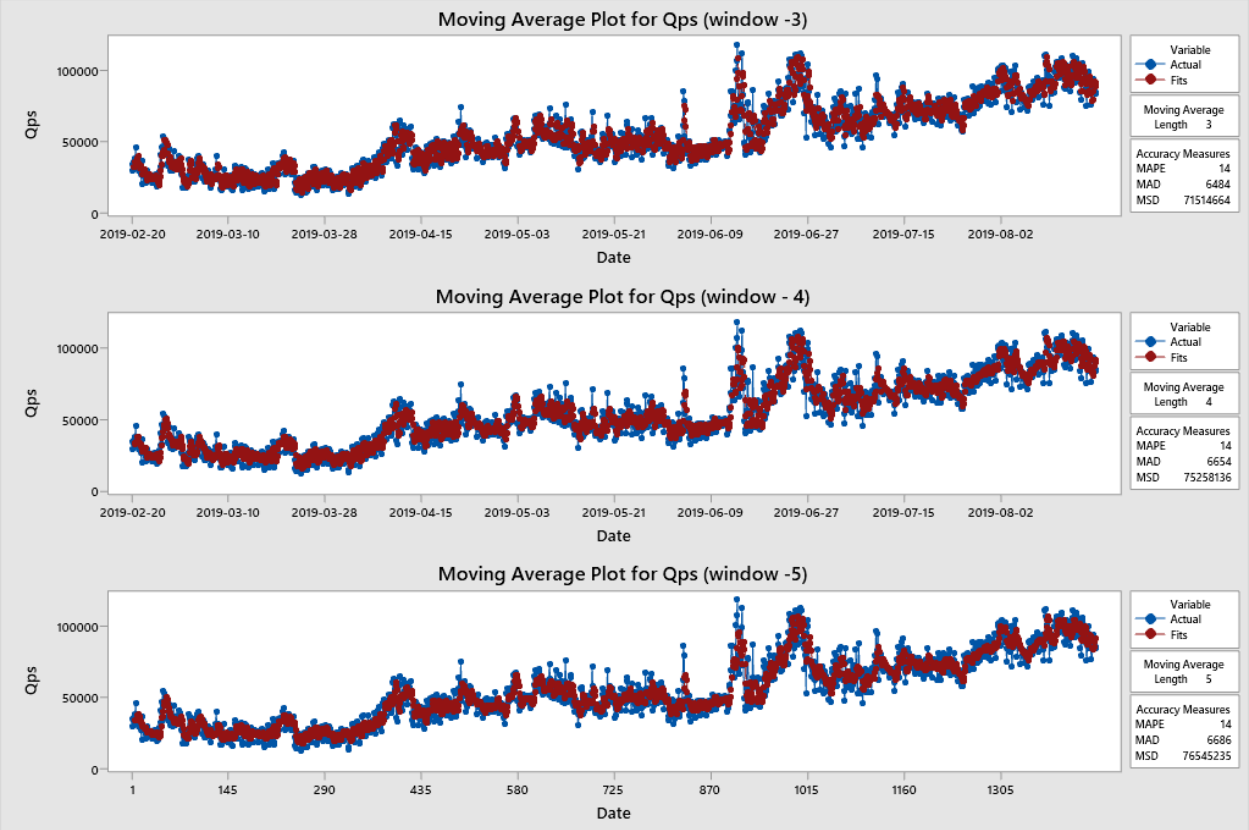

Figure 2 - Graph of Smooth row with different averaging window

Confirming that the trend is present on average, you can go to the implementation of the Holt method. The Holt method is used to predict time series when there is a tendency for the time series to rise or fall. And also, for rows where data is not for the full cycle and seasonality is not yet highlighted (for example, for a part-time year for the monthly forecast).

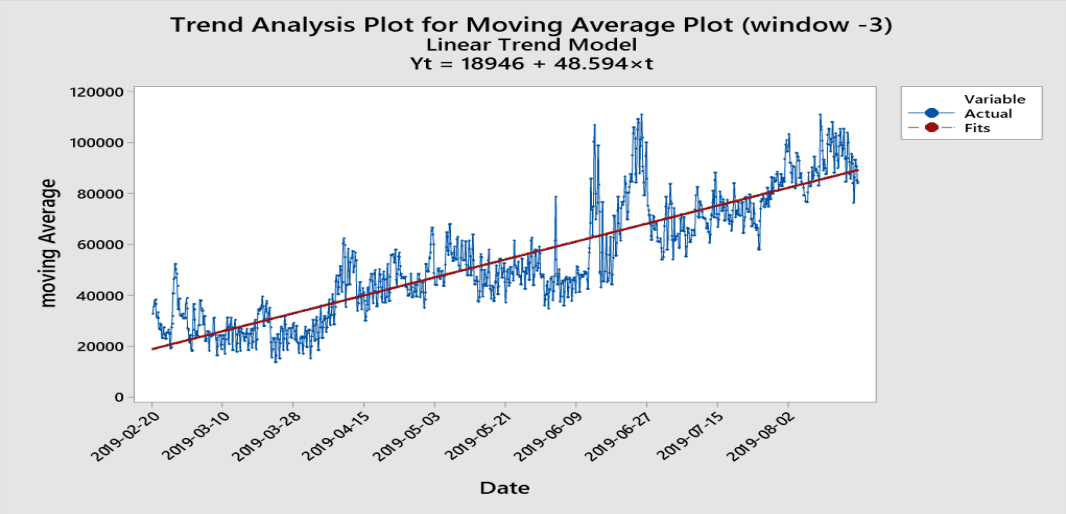

Figure 3 - Graph of trend

In the proposed algorithm, the level and trend values are smoothed out by exponential smoothing. If the time series tends to rise or fall, then a trend should be distinguished together with an estimate of the current level of the series (as in simple exponential smoothing). To control the level and slope of the model Holt introduced 2 coefficients of smoothing - the coefficient of smoothing series and trend [4]. Moreover, they have different smoothing options:

$$\hat{y}_{0k} = \alpha y_k + (1 - \alpha)(\hat{y}_{0k-1} - y_{k-1}), \quad (2)$$

$$t_k = \beta(\hat{y}_{0k} - \hat{y}_{0k-1}) + (1 - \beta)t_{k-1}, \quad (3)$$

$$\hat{y}_{k+p} = \hat{y}_{0k} + pt_k, \quad (4)$$

where

$\hat{y}_{0k}$ - smoothed amount for the given period,

α – row smoothing factor,

y_k- current row value,

$\hat{y}_{0k-1}$- smoothed amount for the previous period,

t_{k-1} - the value of the trend for the previous period,

β – trend smoothing factor,

$\hat{y}_{k+p}$ – forecast for p-periods,

t_k – recent trend.

The first equation describes a smoothed series of general levels. The second equation is used to estimate the trend. The third equation determines the prognosis for p-periods ahead. Their selection is formed on the basis of algorithms "Double Exp Smoothing" mathematical package for data analysis Minitab.

One of the major drawbacks of the Holt method is the forecast for one or two future periods [5]. With such arrays of data, it is difficult to predict value values that would virtually help plan the number of servers that will be used to process requests. In this case, it is possible to increase the forecast period by dividing the time series. The full time series, in our case, is a reflection of the state of the system during the

day with a three-hour range. That is, every day for every three hours the QPS values are averaged into one. Time series grouping will allow you to forecast future metrics over the same period for the next few days. The algorithm for smoothing and detecting the presence of a trend will be repeated for each group of time series data (Fig. 4)

Using the built-in algorithms of the mathematical package Minitab to predict the following indicators (Fig. 5), (Fig. 6).

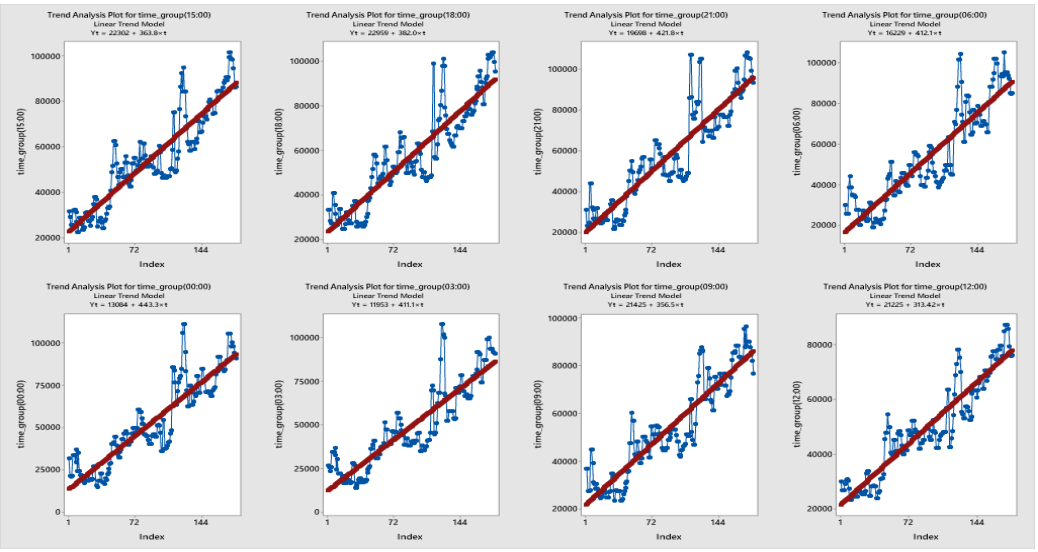

Figure 4 - Graph of trend in grouping data

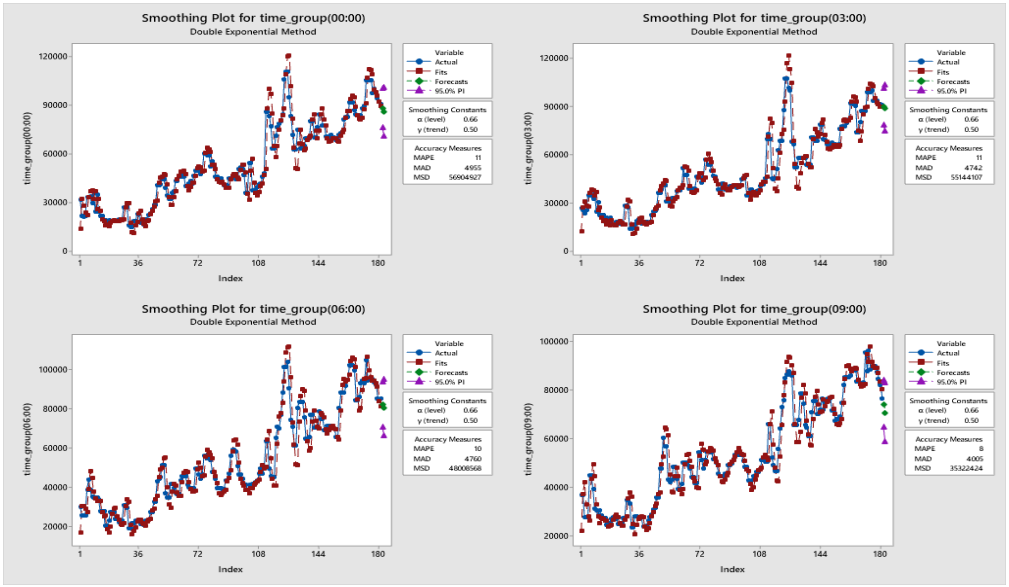

Figure 5 – Forecast for group (00:00), (03:00), (06:00), (09:00)

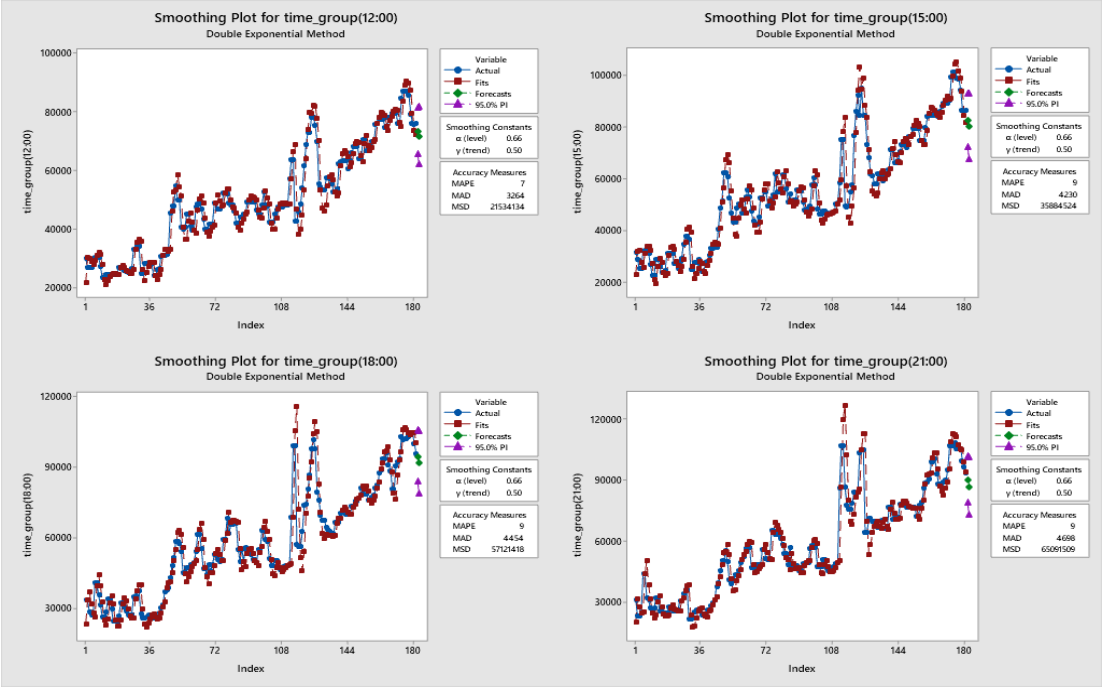

Figure 6 – Forecast for group груп (12:00), (15:00), (18:00), (21:00)

According to the groups of data, we can compile the above forecasts to reflect the picture of the next two days (Fig. 7)

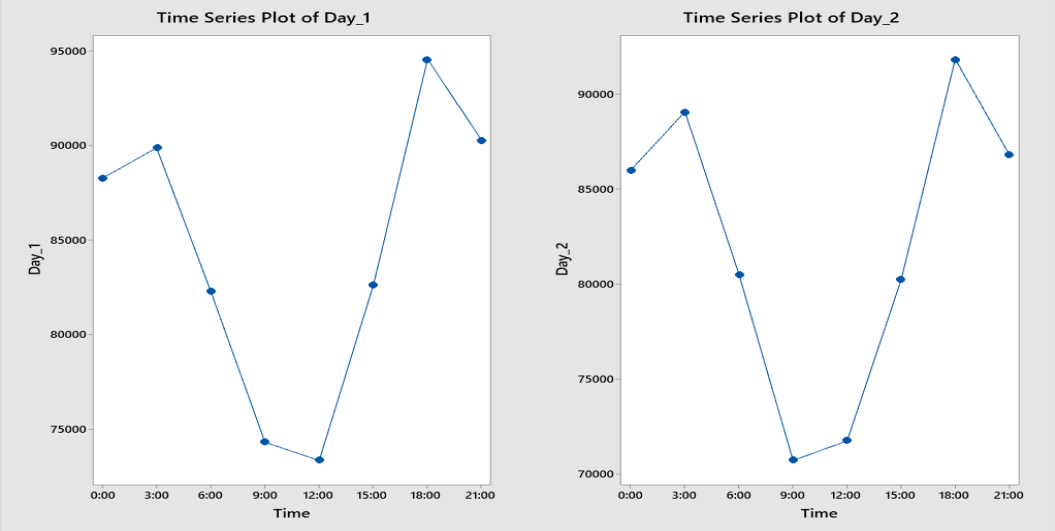

Figure 7 – Forecast for two days

Conclusions. The findings provide a vision of the future, not only in terms of technical support, but also in terms of cost savings. The overflow of the request processing process in ideal conditions, that is the rejection of possible down-time and idle periods, which are stochastic events with a low probability of occurrence, can

quickly transition the scalability of the infrastructure, which will save energy and internet traffic. Given the significant value of the difference between the smallest and highest load on the system during the day (~ 20000 requests in both cases), it is possible to disable the back-end of the infrastructure as it changes dynamically. According to the technical specifications of each information system, the percentage of inactive to the "live" part varies, but it is impossible to reject the fact of cost savings and high KPI of the constructed forecast with such results. With minimal, unobtrusive HTTP / HTTPS POST requests, this number of QPS requires two medium-sized configuration servers, which lease or support per month, translates into $ 1,100. If we drop half of that value (their shutdown time is 3:00 a.m. to 3:00 p.m.), We get $ 550 a month in savings.

A well-known method of forecasting and analyzing data can bring about a minimal cost reduction, but still a reduction that can produce even better results when expanding your business. Also, as the practice of using such forecasting methods shows, it is increasingly used in the fast-growing IT industry, but in a different software interpretation.

References:

1. Buyya R., Calheiros R. N., Son J., Dastjerdi A. V., Yoon Y. Software - defined cloud computing: Architectural elements and open challenges. *Advances in Computing, Communications and Informatics (ICACCI) 2014 International Conference*. September, 2014. P.5–12.
2. Loktev S. Problemy vnedrennya tekhnolohiy "klient-server". *CompUnity*. – 1996. №6. C.19–21
3. Zueva V.N. «Rehresyvni metody prohnozuvannya hrafichnykh navantazhuvachiv Elektrooboruduvannya». *Naukovyy zhurnal KubHAU*. 2017. № 126(02) C.1–12.
4. Ivanov S.A., Kvyatkovskaya Y.Y. Vykorystannya modeley Kholda dlya prohnozuvannya zmin temperaturnoho rezhymu v zakrytomu hrunti. *Vestnyk Saratovskoho Hosudarstvennoho Tekhnycheskoho Universyteta*. 2016. T.1., №1 (82). C.18–22.
5. Armstrong, J.S., Fildes R. On the selection of error measures for comparisons among forecasting methods. *Journal of Forecasting.* 1995. Vo.14, C.67–71.
6. Kashin, M. M. Metod borbyi s peregruzkami v seti SIP na osnove statisticheskogo analiza signalnogo trafika. *Infokommunikatsionnyie tehnologii*. 2011. T. 11., №3. S. 65–69.
7. Bagmanov V.H, Komissarova A.M , Sultanov A. Kh. Prognozirovaniye teletrafika na osnove fraktal'nykh fil'trov. *Naukovyy zhurnal "Vestnik UGATU"*. 2009. №6. S. 32–43.
8. Sudarikov R.A. Analiz informativnykh priznakov v zadachakh obnaruzheniya anomaliy trafika statisticheskimi metodami. *Naukovyy zhurnal "T-COMM"*. 2014. №5. S.41–49.
9. Kuchuk G.A., Mozhaev A.A. Prognozirovaniye trafika dlya upravleniya peregruzkami integrovannoy telekommunkatsionnoy seti. *Naukovyy zhurnal "Radioelektronni i komp'yuterni systemy"*. 2007. №3. S.36–42.