

FORECAST MODELLING & PERFORMANCE LOAD VOLUME ANALYTIC TECHNIQUES OF AN APPLICATION

Vishal Anand¹

Advisory Technical Services Specialist, IBM

Abstract— Forecasting algorithms play crucial role in providing the static data for creating work load model which directly helps in simulation of production usage pattern of an application in testing environment well ahead of any potential failure. There is various model available to forecast the projected number of an application. Few models are quite complex that considers the various factors (e.g. Investment, GDP, Rain, Acquisitions) apart from the historical record of the application which could influence the projected volume. These factors are accessed through to fit to be independent variable and later used in the forecasting techniques (ARMA and ARMAX) model.

Incremental percentage model has been proposed as one of the forecasting techniques to calculate the future volume. Incremental percentage model leverages the independent variables so that there would be minimal chance of deviation from projected number to actual number.

Paper also explain the various approach to build up the work load modelling with an illustrative data table and tried to include all the possible measures to design the robust WLM (Work Load Model). It further goes with recommendations and challenges to make the error proof WLM and forecasting techniques which is basis of the performance testing domain.

Key Words: - ARMA, ARMAX, Incremental Percentage Method, WLM, Regression Analysis, Front End Key Indicators.

1. Introduction

Performance testing an application is the way to emulate real world workload and its behaviors in pre-production environment, with continuous evolving marketing strategies to attract the customers and new technical advancement, posed a new challenge in building real user patterns. There are many significant factors which are important in consideration of forecasting techniques which predicts the future load on the system. [1]

In recent media reports, company like (Amazon, Best-Buy and Tesco etc.) has been down under for hours, Occasionally Indian Banking application (SBI, PNB, HDFC etc.) goes down or IRCTC responds very slowly during peak time which has resulted huge revenue loss. All these down times occurred because of unexpected surge in users. These events emphasize the missing significant factors in forecasting techniques which acts as inputs to create the world work load model of the application to simulate in testing environment of the upcoming actual load in production environment.

Forecasting algorithms play crucial role in providing the static data for creating work load model which in turn directly helps in simulation of production application usage pattern in performance testing environment well ahead of their actual failures.

Selection of good forecasting techniques and adding the additional factors would conceive the much realistic statics and this will help in overall business strategy and inadequate performance evaluation can lead to bad decisions, which result either in inability to accomplish its objectives. However, a good evaluation study can be instrumental in the design and realization of an efficient and useful system.

Implementing suitable forecasting techniques and complementing its shortcoming in broad ways. We have come up with niche way to consider the new factors for future trending applications and explore the way to create the accurate performance workload models that matches to the realistic workload than the legacy-based workload modelling process in current performance testing practice.

2 MOVING AVERAGE METHOD.

Most common time series forecasting techniques is a Moving Average Method and estimates the future number by averaging the past number of Business-Critical Transactions.

Business-Critical Transactions and its past records are provided by business team to the testing team and it varies according to the production usage and domain hence generic transaction name (Transaction_0X) are given below and can be changed with

the requirement.

Let's assume that historical records of an application are available for the 3 years (2007-09) and wanted to forecast the number in 2010.

It's always recommended to have more number of data interval. Data interval of 2 year (2007 to 2008) predicts the number for the 2009 and average of 2008 to 2009 would estimate for 2010.

Estimation of Transaction_01 for the 2009 = $(1000+1100)/2=1050$

Estimation of Transaction_01 for the 2010 = $(1100 + 1200)/2 = 1150$

Data:

Business Critical Transactions	2007	2008	2009	Prediction in 2010
Transaction_01	1000	1100	1200	1150
Transaction_02	200	400	600	500
Transaction_03	4000	6000	7000	6050
Transaction_04	500	700	800	750
Transaction_05	300	500	600	550

2.1 LIMITATIONS OF MOVING AVERAGE METHODS:

Moving averages are considered a "smoothing" forecast technique because taking an average over time, would exclude the irregular occurrences within the data. [2] Based on the total hits made in the past financial year 2007 to 2009, number of hits to Transaction_01 would be 1150 in 2010 (explained above) which is irrelevant as it is less than the true number of 2009.

Weighted Moving Average is the way to lower the MAD (Mean Absolute Deviation) value by assigning more weights to more recent data and low weight to old data to forecast the realistic future value but should find out the ways to assign the correct weights and method is cumbersome to find out the suitable weight so we came up with novel approach "Incremental Percentage method" to keep the MAD to minimum. defect assignment can delay resolution time by hours or even days.

The ability to correctly assign a defect, based on automated analysis of the underlying patterns of defects will enable the raiser to quickly take preventive actions. This helps ensure that defects are being addressed by the most appropriate team, thereby eliminating any cycle time.

2.2 INCREMENTAL PERCENTAGE METHOD:

Incremental Percentage Method is evolved to keep the MAD (Mean absolute deviation) value to minimum which shows error in predictability. It's purely based on the percentage increased/decreased year to year basis.

Calculate the percentage increased/decreased year to year basis so Inference1 column is derived with the following formula.

Inference 1= $(\text{Number of hits for the variable1 in year (i)} - \text{Number of hits for the variable1 in year i-1}) * 100 / \text{Number of hits for the variable1 in year (i)}$

Same formula is applied across all the transaction for the provided year to get the inference1 column and Inference2 and Average is simple mean to get the average of two respective columns.

Average Number Increased/Decreased = $(\text{Percentage Increased/decreased variable i in Inference1} + \text{Percentage Increased/Decreased variable i in Inference1})/2$

Final number to be forecasted in year f = (Number in penultimate Year * Average percentage)/100+ Number in penultimate Year.

Implementing the above formula, Transaction_01 which is business critical transaction, forecasted for 2010 around 1305.

Data Table: Data table is attached below to illustrate the Incremental Percentage method.

Business Critical Transaction	2007	2008	2009	Inferenece1	Inferenece2	Average	2010
Transaction 01	1000	1100	1200	9.09	8.33	8.71	1305
Transaction 02	200	400	600	50	33.33	41.67	850
Transaction 03	4000	6000	7000	33.33	14.29	23.81	8667
Transaction 04	500	700	800	28.57	12.5	20.54	964
Transaction 05	300	500	600	40	16.67	28.33	770

Limitations of Increment Percentage Methods:

This method is based on the change in the events of previous successive years and don't explicitly include the number of independent variables which might change the forecast for future years and work load model designed on the above business requirement would not emulate the actual number and fail the purpose of forecasting.

There are others factors which also impacts the forecasting apart from the historical data called as Independent variables which are required to carefully access and combine through ARIMAX/ARMA forecasting techniques such as Economic situation of the country where company is operating, Company profitability are the factors which are assumed as the Independent Variables [3] and only be included in calculation if the p- value of the independent variable is less than .15 and explained below to calculate the p-value for any independent variables. either a code or no-code type of defect using a pre-trained Random Forest model.

2.2.1: ECONOMIC FORECASTING OF COUNTRY:

This type of forecast is important to understand the general economic trend through a careful study of Five Year Plans, Gross national products, National income, Government expenditure, Unemployment, Consumer spending habits etc. and selecting the independent variable is quite important and should be factually correlated though below explained method in ARMA model.

Let's assume the below data records of country in ARMA model calculation (table 1).

Independent Variables	2007	2008	2009
Gross national products	7	7.4	10
National income	10k	11k	12k
Government expenditure	10%	11%	13%
Unemployment	2%	3%	4%

2.2.2: COMPANY PROFITABILITY:

Next consideration goes to future strategies of the organization and based on chosen marketing plans (kind of investment being planned for the next financial year and expected increase in number of user's due acquisitions or expansion of the business) within the firm are taken in formulating forecasts.

Any of the below variables is showing significant deviation from previous year's record and should be considered potential independent variables in ARIMAX which is as followed.

Following factors (Investment, Acquisition, Portfolio Sold, Expenses on advertisement) and its number are assumed to explain

the ARMA model calculation which are data records.

Independent Variables	2007	2008	2009
Investment	10%	15%	30%
Acquisition	NA	NA	Yes
Portfolio Sold	NA	NA	Yes
Expenses on advertisement	2%	3%	16%

Real number can be sourced from the market relative to their domain and follow the explained procedure [4]

Regression Analysis: When an independent variable is considered while estimation, called linear regression and if multiple independent variables are considered, called multiple regression as depicted in the data table.

ARMA (autoregressive moving average), ARIMA (autoregressive integrated moving average), ARMAX (autoregressive moving average with exogenous variables) and ARIMAX (autoregressive integrated moving average with exogenous variables) are the most often used classical time series methods for the complex estimation like of GDP, Metrological or agricultural productions. [3,4]

ARMA (autoregressive moving average) -

When multiple independent variables (e.g. GDP, Advertisement, Investment) are considered while forecasting the number of the dependent variable (future estimates).

Test Data:

Year	Business Critical Transactions	Number	GDP	Advertisement	Investment
2000	Transaction_01	2000	4	1000	500
2001	Transaction_02	2200	5	2000	600
2002	Transaction_03	2300	5.5	3000	700
2003	Transaction_04	2400	6	4000	750
2004	Transaction_05	2500	6.6	4200	800
2005	Transaction_06	3000	7	4300	900

$$Y = \text{constant} + B1*X1 + B2*X2 + \dots + Bn*Xn$$

3.1 Steps involved to Check the Predictability of Independent Variables through Excel:

1. First enable the Analysis Tool Pack Add-in from the File -> Option -> Add-In and click on the GO button.
2. Select the Analysis tool pack from the pop up and click on the Ok button and Data Analysis button will be visible on the Data section.
3. Now select the regression option from the Data -> Analysis Tool -> Regression and select the deponent variable in the Y range and independent variables in the X ranges.
4. Click Ok button to extract the p-value, t value and coefficients as given below which will be used on the formula.
5. p-value is defined as the probability of obtaining a result equal to or "more extreme" than what was observed and corresponding p-value must be less than .15 to be used in the formula.
6. p-value for the GDP is more than .15 (Table 1) and would exclude the GDP from the calculation and repeat the above-mentioned steps (1-5) once again.

Table 1: Data table 01 is attached below to illustrate the ARMA modal

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-96.380	440.823	-0.219	0.847	1993.087	1800.328	1993.090	1800.328
GDP	2.194	266.797	0.008	0.994	1145.739	1150.127	1145.740	1150.127
Advertisement	-0.269	0.107	-2.509	0.129	-0.731	0.193	-0.731	0.193
Investment	4.679	1.761	2.656	0.117	-2.900	12.258	-2.900	12.258

p-value for the advertisement and Investment fell significantly in (Table 2) which could strongly predict the future value.

Table 2:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-94.216	288.831	0.326	0.766	-1013.405	824.973	-1013.410	824.973
Advertisement	-0.269	0.075	3.566	0.038	-0.509	-0.029	-0.509	-0.029
Investment	4.692	0.711	6.595	0.007	2.428	6.956	2.428	6.956

Now fit the above values in the following formula.

$$Y = \text{constant} + B1 \cdot X1 + B2 \cdot X2 + \dots + Bn \cdot Xn$$

B value of Advertisement for the 2006 = 4500

B value of Investment for the 2006 = 1000

$$Y (\text{Predicted number of Transaction}_{01}) = -94.216 + (-0.269 \cdot 4500) + (1000 \cdot 4.692) = 3387$$

Increment percentage method explained above calculates the Transaction₀₁ as 3229 for the same data set in year 2006 which is less than 3387. Any of the above explained forecasting model could be used to derive the projected number and is suggested to use the model which predicts higher number.

In this way, rest of the transactions could be calculated for an application.

Recommendation:

Companies makes most of the profit during certain period (peak time) so any down time during peak time would lead to loss of million dollars. Hence select the higher predicted number derived from multiple forecasting techniques.

Include the variance of various factors while forecasting for unavailable historical data and sources.

4.1 Workload Modeling on Forecasted Numbers:

Load distribution of the production volumes across all the identified scenarios is called the Workload. Workload model of an application depicts how the application will be used in the production environment. To get performance test results which portrays the true behavior of production environment, Workload model should replicate the projected number of business transactions in test environment. [6]

4.2 Steps Involved to build WLM :

1 Application Understanding:

Complete understanding of the AUT (Application Under Test) with all its features is the basic step in any testing activity. A tester can't thoroughly test an application unless have its complete understanding of application.

2 Navigation Paths of identified Scenarios:

Once all the scenarios have been identified, next step is to figure out each scenario's all possible navigational paths which a user can opt to successfully complete it.

3 Identify Unique Test Data

Test scenarios with improper test data would fail the test results thus correct test data is integral part of any performance test.

4 Relative Load Distributions across Identified Scenarios

After understanding the application, identifying its key test scenarios, their navigation paths and required test data for each identified scenario, next step is to figure out the distribution of all the identified scenarios in workload model.

5 Business Critical Transactions Identified from the Forecasting Method:

Primary requirement to execute the performance testing on any provided application, should have the production statistics for the identified business critical transactions so that work load model can be created to simulate the exact pattern of production load volume in the testing environment. [6]

Business critical transaction and its expected numbers are provided by the business team to the testing team who are responsible for building the WLM. One can have different business critical transaction and its number depending upon application hence generic transaction name (Transaction_0X) are given below and can be changed with the requirement.

Below sheet is assumed to explain how to design the WLM.

Data Sheet 1.0:

Users Count: 1000

Business Critical Transaction - X Domain	Business Critical Transaction- Retail Domain	Business Critical Transaction - Banking Domain	Business Critical Transaction - Insurance Domain	Required Number
Transaction_01	Home Page	Home Page	Home Page	2000
Transaction_02	Login	Login	Login	400
Transaction_03	Click On Left Navigation Tab	Click On Account Details	Click on Policies	4000
Transaction_04	Search Product on Left Navigation Tab after login	Transfer fund Same bank	Select Policy	1000
Transaction_05	Search Product on Left Navigation Tab prior login	Transfer fund Different bank	Fill details	2000
Transaction_06	Search Product after login	Transfer fund Same Bank NEFT Mode	Select Payment Online option	2000
Transaction_07	Search Product prior login	Transfer fund Same Bank Instant Mode	Click on Payment	6000
Transaction_08	Add to Basket after login	Transfer fund Different Bank NEFT Mode	Search Policy by Number	1000
Transaction_09	Add to Basket prior login	Transfer fund Different Bank Instant Mode	Cancel the policy	1500
Transaction_10	Continue to Checkout	Click On Pay Bill Utility	Click on to extend Policy	2000
Transaction_11	Place Order	Click On Credit Card Bill button	logoff	2000
Transaction_12	Click On promotion	Click On Credit Card Pay Bill	NA	8000
Transaction_13	Cancellation Order	Click On Add Beneficiary	NA	400
Transaction_14	Registration	Raise Service Request	NA	300
Transaction_15	Logoff	Logoff	NA	500

5. Implementation of WLM Concepts:

5.1 WLM for the 1 step script

WLM is quite simple for the one transaction script (e.g. web services) and following formula is used to derive the pacing given the number of users and transaction per second.

$$TPH = (3600 / \text{Total Journey Time}) * \text{Number of users}$$

Where Journey Time = Pacing + Think Time + Average Response Time of Transaction

$$\text{Total Journey Time} = (3600 * \text{Number of users}) / TPH$$
$$\text{Pacing} = \text{Total Journey Time} - \text{Response Time}$$

Given think time = 0 as there is only a transaction

$$\text{Journey Time} = \text{Pacing} + \text{Average Response Time of Transaction}$$

5.2 WLM for the Multi Step Script

There are many existing ways to create WLM for the multi-step requests/ transaction script. It would have the various parameter (think time, pacing, session length, number of sessions and basket size) to be considered while building up the WLM. [7,8]

Requirement is already forecasted above and to translate those target numbers into the scenarios, It is required to visualize the application flow which should makes sense and that should be functionally possible.

Following points are significant in WLM-

- Performance Test Eng. should have deep knowledge about the application so that one can create the Business scenarios.
- Business step should not be hit more in loop to avoid the caching.
- Number of sessions and session length can be matched as per requirement through varying the number and journey time of the script respectively.
- Number of sessions and session length can also be increased avoiding the log off after login.
- Number and journey time of the script are tweaked to match the number of connections on the web server.
- Number of items can be accordingly added to the basket to match the basket size of the requirement.

Section 1.2 Requirement is divided into 5 scripts to achieve the target number and can be further increase or decrease the number of scripts if the requirement is stringent to sessions and session length.

Business Critical transactions –

Data sheet 1.1 (Business Critical Transaction) have been assumed to explain how to design WLM and accordingly can be designed for any application provided business transaction by the business or one must identify it and later their own set of transactions and its target number to be inserted in Data sheet 1.1 . and follow the explained procedure (Approach 01) to derive the WLM.

Here, an example of WLM explained assumed to have no restriction on number of script and achieved through dividing the requirement in 5 scripts as given below in WLM1.1.

Above requirement can also be achieved through increasing/decreasing the number of scripts but following WLM is designed to have only 5 number of scripts

WLM 1.1- Please find the attached table below

Business Critical Transaction	Required Number	Script1	Script2	Script3	Script4	Script5
Transaction_01 (e.g. HomePage)	2000	300	200	1000	200	300
Transaction_02 (e.g. Login)	400	0	200	0	200	0
Transaction_03 (e.g. Click On Left Nav)	4000	0	1000	3000	0	0
Transaction_04 (e.g. Search Product on Left Nav after login)	1000	0	1000	0	0	0
Transaction_05 (e.g. Search Product on Left Nav prior login)	2000	0	0	2000	0	0
Transaction_06 (e.g. Search Product after login)	2000	0	0	0	2000	0
Transaction_07 (e.g. Search Product prior login)	6000	0	2000	4000		0
Transaction_08 (e.g. AddtoBasket after login)	1000	0	0		1000	0
Transaction_09 (e.g. AddtoBasket prior login)	1500	0	1500	0		0
Transaction_10 (e.g. Continue to Checkout)	2000	0	1500	0	500	0
Transaction_11 (e.g. Place Order)	2000	0	1500	0	500	0
Transaction_12 (e.g. Click On promotion)	8000	1000	2000	3000	1000	1000
Transaction_13 (e.g. Cancellation Order)	400	0	0	0	400	0
Transaction_14 (e.g. Registration)	300	300	0	0		0
Transaction_15 (e.g. Logoff)	500	300	100	0	100	0

Steps:

- Add the average response time (provided by business team) for all the transaction.
- Add think time for all the steps (Assume think time if it is not provided by business).
- Add Pacing, Total average response time and total think time which would be total journey time for 1 iteration with single user.
- Divide the 3600 (1 hour) by the total journey time which would reflect the number of iterations with single user.
- Multiply the designated number of users to the number of iterations.
- Average Response time in Total Journey Time cannot be set by us however think time and pacing could be increased or decreased to match the TPH which sets the speed of script execution.
- Think Time and Pacing would be calculated as per below formula for all the script.
- Number of iteration for single User= (3600/Total Journey Time)
- Where Journey Time =Pacing + Think Time+ Average Response Time of Transaction
- TPH = Number of iterations for single User in an Hour * Number of vusers
- One of the five script (Script 2) is devised to form below flow (attached below as WLM 1.1_Script02) and there could be multiple approach to achieve the target number (e.g. one business process, lets says search functionality can be kept in loop of 5 time to achieve the 1000 or can be called separately with in script).

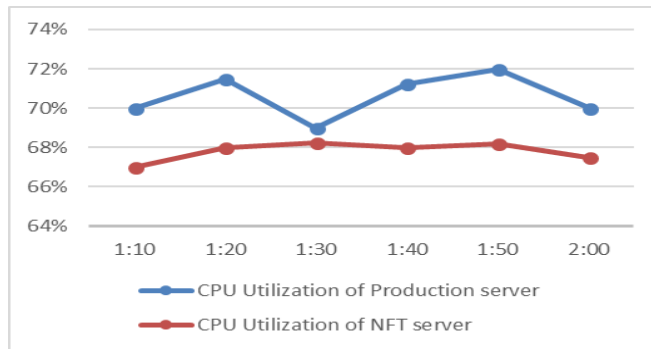
There are few transactions below (WLM 1.2) which is mentioned as .5, 3.5 that's would be handled through assigning the 50% to "1* Search Product and 7 *promotion ()" in run time setting of the script.

Experiments:

Load Test was conducted with proposed load model and It was found that variance was not more that 3.5% which proves the validity of the model and could be further used by others.

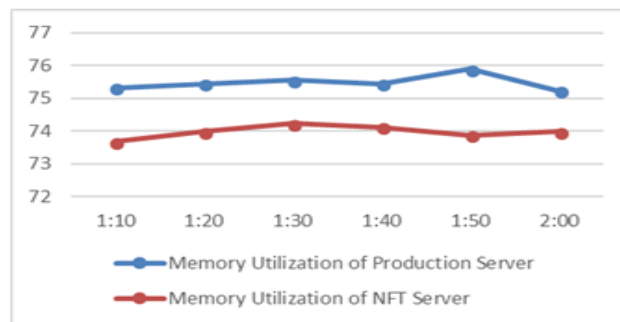
CPU Utilizations:

Web Servers: Below graph depicts the CPU utilization of web server in production to the testing environment.



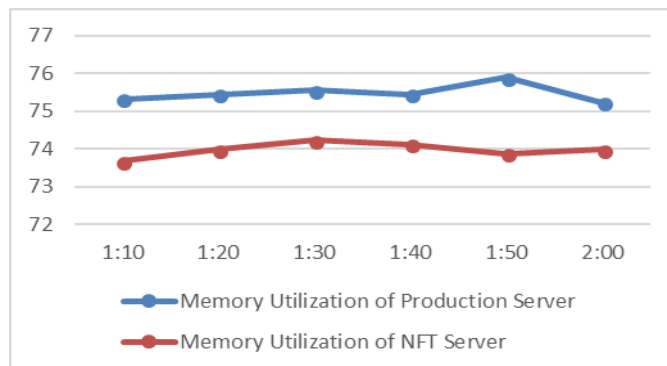
Application Servers:

Below graph depicts the CPU utilization of app server in production to the testing environment.



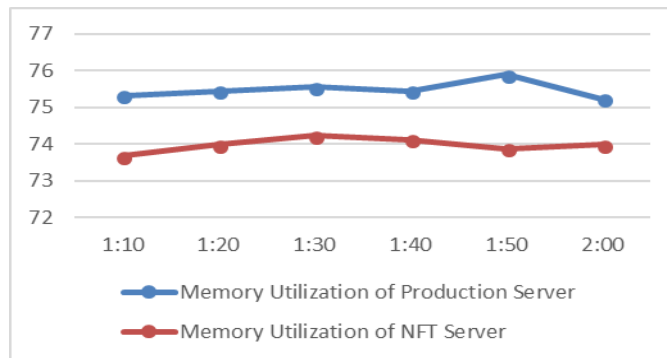
Memory Utilization:

Web Servers: Below graph depicts the memory utilization of web server in production to the testing environment.

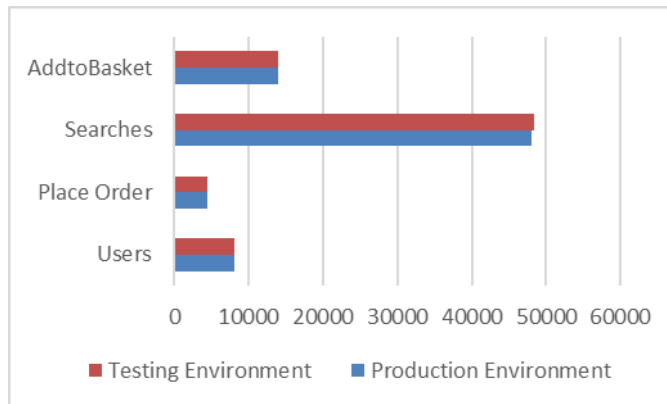


Application Servers:

Below graph depicts the memory utilization of app server in production to the testing environment.



Front End Key Indicators: It represents that all the front-end keys parameters were same during tests in the production and testing environments.



2.1 Conclusions

Application stakeholders are more concerned about their applications performance and are not reluctant any more for conducting performance testing activity on them. Work load modelling are main constituents of performance testing which simulate the future year the predicted number in pre-production environment.

Performance enhancement, Code fix, Configurational changes could only be successfully tested in pre-production environment provided WLM is appropriately designed and Implementing proposed approach for an application, one can avoid any performance bottlenecks getting deployed in production.

Model (Approch01) is validated and variance was not >3.5% in resource utilization of production to the testing environment which is significant booster to accept it. Approach 02 could also be implemented if there is no constraint to maintain the same level of session length and session number.

Getting all the required information (business critical transaction and its number) for developing an accurate workload model is a challenging task and performance test teams often collaborate with the Business team, application stakeholders, network, and marketing teams for obtaining this information.

Reference:

1. Dror G. Feitelson, The Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel

2. <https://hbr.org/1971/07/how-to-choose-the-right-forecasting-technique>
3. Hongmei Chen¹, Multiscale Forecasting Method using ARMAX Models, www.isye.gatech.edu/~brani/isyestat/04-30.pdf
4. FORECASTING FUNDAMENTALS, mech.at.ua/Forecasting.pdf
5. Performance Test Workload Modeling, www.agileload.com/agileload/blog/2013/06/.../performance-test-workload-modeling
6. Performance Workload Design, <https://www.ibm.com/developerworks/community/.../WorkloadDesign.pdf>
7. walderson.com/IBM/Practices/...performance.../workload_model_16458386.html
8. <https://en.wikipedia.org/wiki/Benchmarking>