Global Scientific JOURNALS

GSJ: Volume 7, Issue 11, November 2019, Online: ISSN 2320-9186
www.globalscientificjournal.com

CAPACITY FORECASTING IN AN ENTERPRISE STORAGE ENVIRONMENT

IP, HO YIN

Ho Yin Ip is currently pursuing masters degree program in Information Systems in Atlantic International University, Hawaii, United States. E-mail: patrick.hoyin.ip@gmail.com

KeyWords

Storage capacity forecast, Enterprise storage, Storage system, Direct Attached Storage, Network Attached Storage, Storage Area Network, piecewise linear regression, regression model

ABSTRACT

Digital transformation is profoundly changing the shape of business today. Companies continue to become "information-driven" in order to remain competitive, or even outrun, their market leaders. Digital transformation is not just about the evolution of devices, it is also about leveraging data to improve customer experiences, open new markets, make employees and processes more productive, and create new sources of competitive advantages. Inevitably, data created by both businesses and consumers would grow exponentially and is resulting in huge demand in enterprise storage for companies around the globe. Professionals in the field of Information Technology should be familiar with Enterprise Storage systems. Enterprise storage devices generally offer higher reliability, availability, and scalability. It is entirely conceivable that enterprise-level storage devices require specialized technical skills to maintain and support, and are more costly than those general storage devices for consumers and small businesses. Therefore, it is imperative for any organization to establish a statistical method to predict usage growth by analyzing usage trends, in order to avoid systems degradation or failure, and budgeting nightmare. By better planning of storage utilization and capacity forecasting, organizations may be able to avoid overbuying hardware or untimely upgrade interruptions and lead to economies of scale.

1 Introduction

In my early mid-career years, I was a Senior Systems Analyst in the regional hub IT department of a global financial institution. The department was managing over 1,000+ systems to provide IT services to the local offices as well as other branch offices across eighteen countries in the Asia-Pacific region. I was once tasked to develop a method to predict the date of full capacity of the storage systems.

Without formal learning in statistics, I developed a simple approach by exporting history usage data from a storage system and plug into an Excel spreadsheet to graph usage trends on a weekly basis. Other than showing a gradual usage growth after weeks, the predictions generated from this simple method were quite far from being accurate. There was plenty of room for improvement in order to receive recognition and appreciation from stakeholders in various business units. What even worse was, in a few occasions when I had to explain the method I used to non-technical management executives, my content was inevitably crafted in a technical manner and communication was commented as ineffective and rambling.

The purpose of this research paper is to devise and develop a storage capacity forecasting methodology based on my previous work assignment aforementioned. It is written in the hope that it will provide a more complete and comprehensive study, in a tone appropriate for non-technical audience by applying the knowledge I have acquired from the course of Business Statistics.

2 Forecasting Methods

There are a combination of methods used in this research of storage capacity forecast – Quantitative data & Qualitative data, i.e., the mixed-method research that involves collecting, analysing and integrating quantitative and qualitative research. The major advantage of this approach is that it will "provide strengths that offset the weaknesses of both quantitative and qualitative research [1]". "Forecasting capacity needs is part intuition, and part math. It's also the art of slicing and dicing up your historical data, and making educated guesses about the future [2]".

In short, quantitative analysis is used to predict storage consumption growth rate as forecasting basis, whereas, qualitative analysis is used to improve the accuracy of the forecast by adding insights from future real events.

2.1 Data Collection

The following table illustrates the typical quantitative metrics that are commonly used in analysing storage resource consumption and determining future resource needs:

Table 1: Typical Metrics Needed for Storage Capacity Forecasting.

Category	Metric	Description	Interested Parties
Primary Storage	Primary Storage Capacity	Number of gigabytes or terabytes required by a storage user. It's important to capture this in terms of storage tiers as well as location (if the organization has multiple data centres).	Users, Storage Administrators
Backup & Restore	Tapes & Tape Drives	Number of additional tapes and tape drives required to support backup and restore for the additional storage capacity.	Backup Administrators
Network	Connectivity ports required (IP & Fibre Channel)	Additional connectivity may be required based on the addition of new servers.	Network Administrators
	Replication Bandwidth (MB/sec)	Additional long-distance bandwidth maybe required for environments where data is being replicated to remote sites for disaster recovery purposes.	Network Administrators
Costs		This includes costs for all resources required to provide the additional storage capacity.	Finance, Senior Managers

As shown in Table 1, there are four metric categories that can be analysed to determine the future storage resource needs. Generally speaking, corporate enterprises often find little or no interests to predict consumptions of physical magnetic data tapes as the unit costs are generally favourable and, contributions to IT budget are insignificant. Hence, the metrics in the category of Backup & Restore will not be collected for analysis.

In prior to data collection, the types of storage systems deployed in the network environment must be identified. In general there are three types of storage technologies: Direct Attached Storage (DAS), Network Attached Storage (NAS), or a Storage Area Network (SAN). As DAS is not scalable and has limited sharing capabilities, capacity forecasting is not applicable to this storage type. Thus, there will be two sets of time series data to be collected – historical data of storage resource consumption of a NAS appliance and a SAN appliance in the network environment.

For the purpose of capacity forecasting, two independent variables are required at each point in time:

1. Total physical space used by all NAS's
2. Total physical space used by all SAN's

2.2 Data Cleaning

In order to ensure data integrity, physical space used by non-production data will be excluded in forecasting the storage capacity consumption because:

- Its characteristics may vary from production data as it is of value to IT developers only and,
- It usually is not bound by any data retention policy in various industries and jurisdictions - it can be erased as it served its purposes.

2.3 Predictive Model

Before choosing a regression technique for predictive modelling, one must first identify any trend or pattern in the time series in order to achieve a reliable forecast.

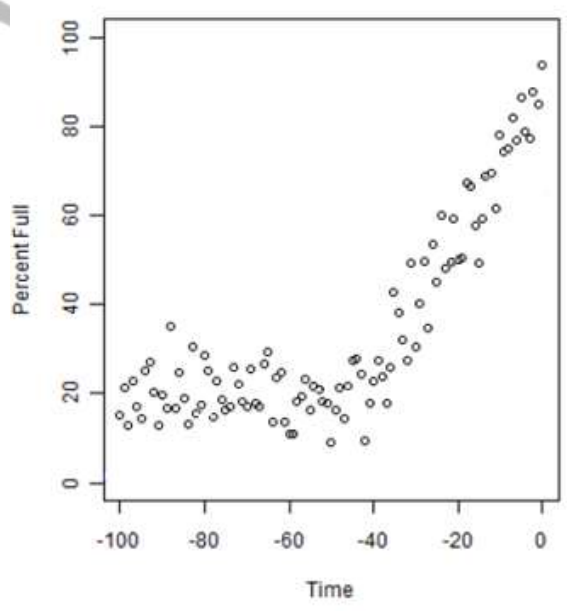

Figure 1. A Time series data plot diagram of a storage system.

As shown in Figure 1, the time series data may illustrate the physical space used in an organization of all storage systems for the prior 100 days. (Time = 0 is the most recent data.) The standard deviation is 6 throughout the data. Consider the following observations and assumptions:

- One dependent variable and one independent variable,

- overall storage consumption increased gradually, and
- similar storage trends should be observed in organizations in the same industry

There is more than one regression technique can be used to generate forecasts. Between a sophisticated model and a simple mean model, we will start with a simple linear regression technique with the following reasons:

- The dataset has only one independent variable,
- Although there are turning points in the diagram, it still exhibits linear growth in the time interval, and
- It is a simple, easier-to-understand model for technical users with little or no knowledge in business statistics.

After all, a storage capacity forecasting doesn't have to be a hyper-complicated process that involves high-level mathematics and projections.

2.4 Subset Selection

Applying linear regression modelling to storage capacity time series data is quite challenging because total storage capacity changes over time and it makes an impact to the consumption percentages. From time to time, system administrators may add additional physical disks or deploy new storage systems to increase capacity, data may be moved to tape storages as it is no longer accessed frequently or simply deleted due to a change in retention policy.

Therefore, it is crucial to select the optimal subset of data which indicates perfectly linear characteristics.

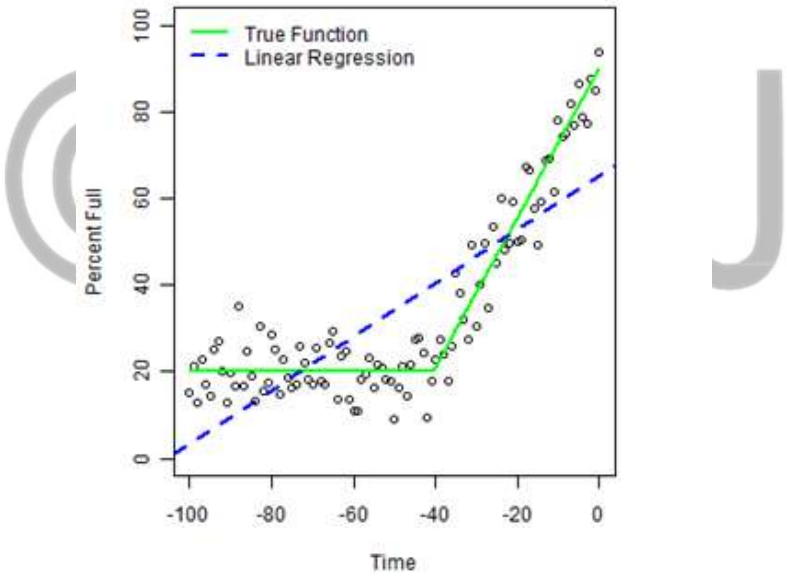

Figure 2: A Time series data plot diagram of a storage system

As shown in Figure 2, the predictions of the linear regression are very poor as the true function (nonlinear) indicates the storage capacity pool is going to reach 100% in a few days, but the linear regression line predicts in more than a few weeks (a false negative). For mitigation, a subset of recent data, for instance, the prior 30 days, would be chosen to eliminate the influence of older data and improves the accuracy of the model's predictions.

2.5 Piecewise Linear Regression (PLR)

By implementing PLR, the error rate of the linear regression model can be significantly reduced. To do this, we will apply the regression to a data subset that best represents the most recent behaviour. And to find the best subset of data, the boundary must be determined where the recent behaviour begins to deviate.

$$R^2 = \frac{SSM}{SST} = \frac{\sum_i [f(x_i) - \bar{y}]^2}{\sum_i [y - \bar{y}]^2} \quad (1)$$

SSM = Regression Sum of Squares
SST = Total Sum of Squares

Properties of R^2

- $0 \leq R^2 \leq 1$
- R^2 =1 indicates perfectly linear data

To calculate the boundary, we will start with a small subset of data and then apply regression to incrementally larger subsets to find the regression having the maximum value of R^2:

1. Regress $\{(X_{-10}, Y_{-10}), (X_{-9}, Y_{-9}), \ldots\ldots, (X_0, Y_0)\}$
2. Calculate R^2 for regression
3. Regress $\{(X_{-11}, Y_{-11}), (X_{-10}, Y_{-10}), \ldots\ldots, (X_0, Y_0)\}$
4. Calculate R^2 for regression
5. …
6. Regress $\{(X_{-n}, Y_{-n}), (X_{-n+1}, Y_{-n+1}), \ldots\ldots, (X_0, Y_0)\}$
7. Calculate R^2 for regression
8. Select the subset with maximum R^2

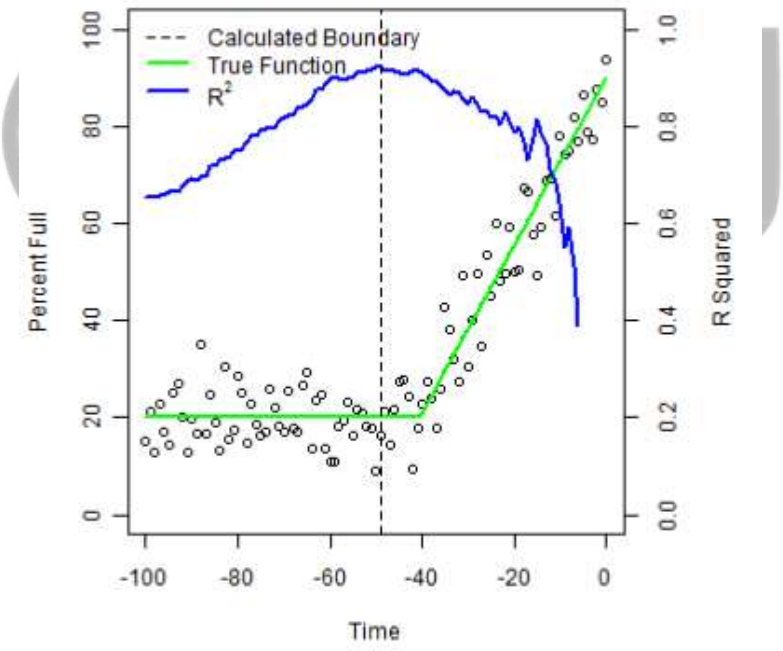

Figure 3: Calculating the boundary for the predictive model

As shown in Figure 3 above, for the same data used in Figure 1, the boundary is the oldest data point within the subset of data determined in step 8 and the predictive model is generated by applying linear regression to that subset of data. The date when R^2 reaches its maximum value is the "calculated boundary" and occurs near the discontinuity of the true function. Maximum R^2 = 0.92 at -50 days and the true boundary is -40 days. The resulting PLR model in a better fit to the data generated using the subset $\{(X_{-42}, Y_{-42}), \ldots, (X_0, Y_0)\}$, is shown in Figure 5 below:

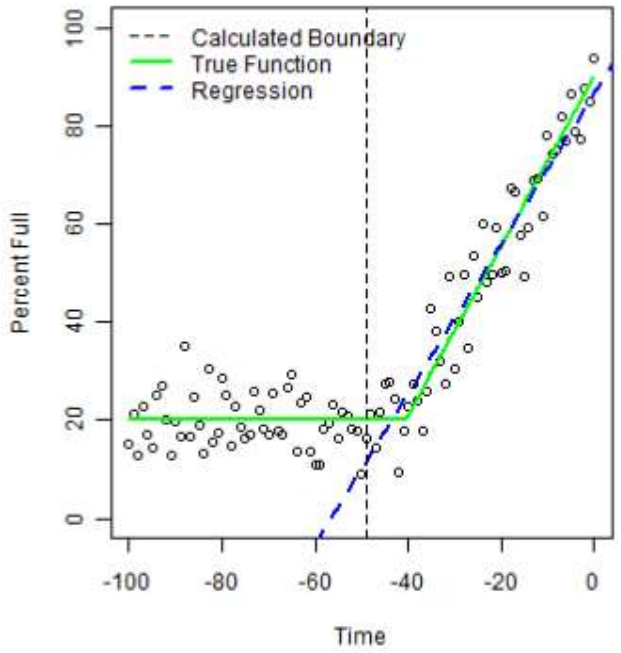

Figure 4: The Goodness-of-fit of the Piecewise Linear Regression model

2.6 Regression Model Validations

To determine if you should publish capacity forecasts calculated based on the linear model described under Section 3.5 above, the results shall comply with the following validation rules.

- **Goodness-of-fit:** When the R^2 value from PLR is too small, it indicates the model is a poor fit to the data. Hence, linear regression models with $R^2 < 0.90$ will not be used.
- **Positive Slope:** Linear models with Slope ≤ 0 cannot be used to predict the date of 100% full.
- **Timeframe:** The linear model shall only be used to forecast date of 100% full of storage systems to less than 12 months because forecasting capacity far into the future is extrapolating the current behaviour too much to be practical. Also, it should be expected that in the next 12 months, technologies (data storage and file system compression, etc.) will be significantly different than it is today.
- **Sufficient Statistics:** Storage systems recently deployed lack enough historical data to produce statistically sufficient regression models. It is recommended to have a minimum of 20 days of data as we will be starting at a subset of 10 days' data in calculating the boundary for computation of maximum R^2.
- **Storage Utilization:** We have learnt from the past experience that underutilized storage systems tend not to provide reliable predictions. Hence, our regression model will not be applicable to storage systems which are less than 10% full.
- **Prediction Error:** Recent changes in storage systems capacity must be taken into account to evaluate the linear fit. System administrators often take proactive actions when systems are nearing maximum storage capacity which results in drastic increase of available capacity. In Figure 5 below, system capacity dropped from 94% to 50% on Time=0 as obsolete data was deleted.

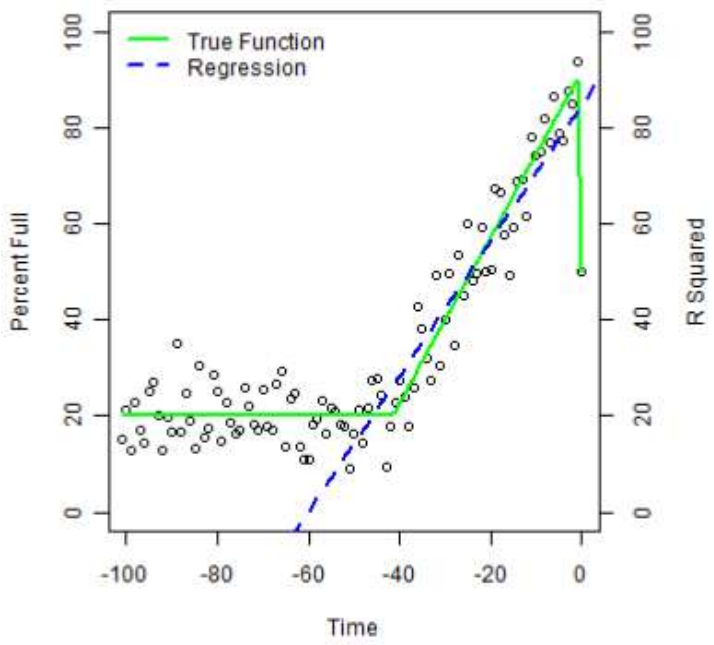

Figure 5: A drastic drop in storage capacity

A high goodness-of-fit (R^2 = 0.88) is generated using the piecewise linear regression model defined which predicts the system should reach 85% full at Time=0, but in fact, it is only 50% full. If the error between the predicted value and the actual value of the most recent data point exceeds 5%, it is a good indication that the recent data diverges significantly from the model. Therefore, the model is no longer valid (and hence the declaration previously derived as the model would become invalid when R^2 < 0.90).

2.7 Equation of the Linear Model Defined

When the linear model passes all validation rules described in Section 3.6, the future date the system reaching full capacity can be solved:

$$y = \alpha + \beta x \qquad (2)$$

Definitions:

- *y* is capacity
- α is the intercept term (α = y when regression intercepts the True Boundary)
- β is the slope
- *x* is the date

Assuming the slope is positive (β>0), the future date for the system reaching full capacity can be computed by setting the capacity y = 1 (100%) and solving for *x*.

$$\text{Forecast Full Date: } x = \frac{(1-\alpha)}{\beta} \qquad (3)$$

3 Case Study

In this section, we will use the historical datasets from the organization I previously worked in, to evaluate the methodology by back-testing for each week over the period of year 2017. Two different type of storage systems will be selected for model validations and for each one, we will calculate 1-month, 3-month and 6-month forecasts, and compare with the actual values in the first two quarters of year 2018.

Our case study research commenced with the formulation of the following hypotheses:

3.1 Hypotheses Formulation

Research Hypothesis 1: PLR (Piecewise Linear Regression) model may not fit for both NAS & SAN types of storage systems.

NAS and SAN are two different technologies. Their costs of implementation, administration, and upgrades vary significantly. Their use cases are quite different as well and the types of data (or the combinations) that are stored on these systems are normally different and thus, having different data retention policies bound to them. Therefore, it is expected that the two types of systems do not have identical usage growth trends and the PLR model we defined may not be applied to all.

Research Hypothesis 2: Efficient use of capital.

In general, disk prices are falling so quickly that it's often cheaper to buy only the storage needed in the immediate future. If a storage administrator were to make a bulk disks purchase, by the time that additional disk space was actually utilized, the cost of the disks would be lower than what they were bought for. Hence, it will be less expensive to delay the purchase of additional storage until it's absolutely needed.

3.2 Scope of Quantitative Analysis

a. Historical data will be collected from a NAS system and a SAN system

b. There will be 2 Storage Utilizations reports generated from:

   i. Historical data of a NAS system

   ii. Historical data of a SAN system

c. Utilization history

   i. The history data of January 1, 2017 to December 31, 2017, will be analysed to predict future values

   ii. The history data in of January 1, 2018 to June 30, 2018, will be used as actual values to validate the predicted values generated by the model

d. Interval: Every 7 Days at 12:00AM

e. Reporting values:

Date	Week *[x]*	Raw Capacity (TB)	Used Raw Capacity	Unused Raw Capacity	Utilization (%) *[y]*
1/1/17 12:00 AM	1	10.00	4.50	5.50	45.00%
1/8/17 12:00 AM	2	10.00	4.55		
1/15/17 12:00 AM					

Figure 6: Reporting values for quantitative analysis

f. To calculate the boundary, we will find the maximum R2 value starting with subset data of I = 4 weeks. Then, we will regress an increment of 1 week until I = 70 weeks:

$$R^2 = \frac{\{[\, i * (\boldsymbol{Sum\ of\ xy})] - (\boldsymbol{Sum\ of\ x}) * (\mathbf{Sum\ of\ y})\}^2}{[i*(\boldsymbol{Sum\ of\ x^2}) - (\boldsymbol{Sum\ of\ x})^2] * [i*\boldsymbol{Sum\ of\ y^2} - (\boldsymbol{Sum\ of\ y})^2]} \quad (4)$$

Week [x]	1	2	3	.....	70
Utilization (1=100%) [y]	0.45	0.46	0.46	.....	0.46
i	4	5	6	.....	70
R Squared	0.80763	0.88321	0.92325	.....	0.449547475

Figure 7: Calculating the maximum R2 by regression

3.3 Analyse Current Storage Trend and Forecast Utilization

The storage utilization history of the two systems were analysed and details of projected forecasts are discussed as follows.

3.3.1 Application of Regression Model for Storage System NAS01

The historical dataset of the NAS system is represented by histograms as shown in Figure 8 below.

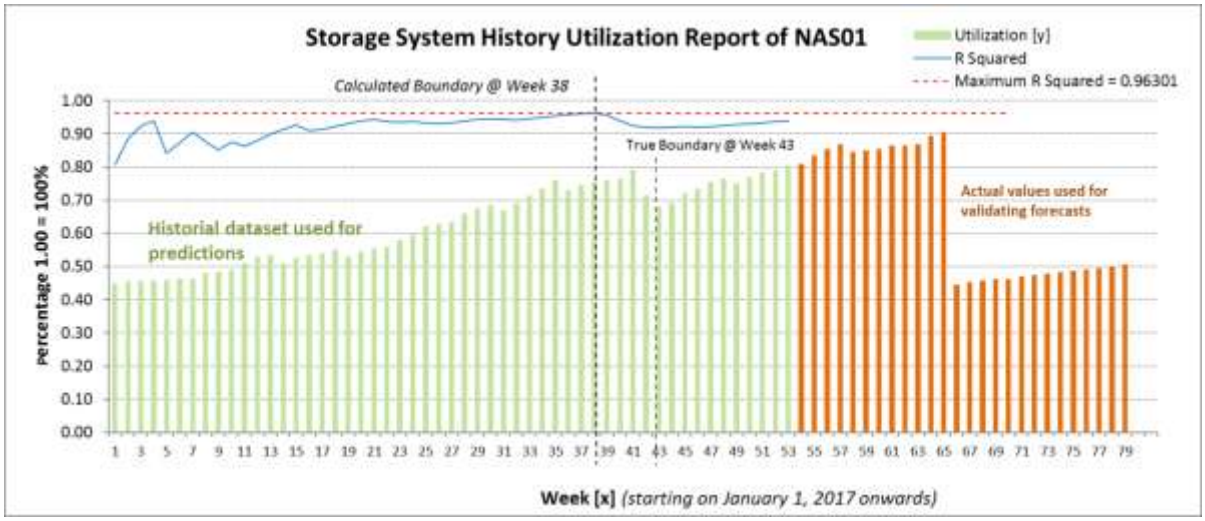

Figure 8: Calculating the maximum R^2 by regression

The R^2 values were first computed to obtain the Calculated Boundary and True Boundary using the maximum R^2 value. The Optimal Subset of Data was determined and shown in Figure 9 below:

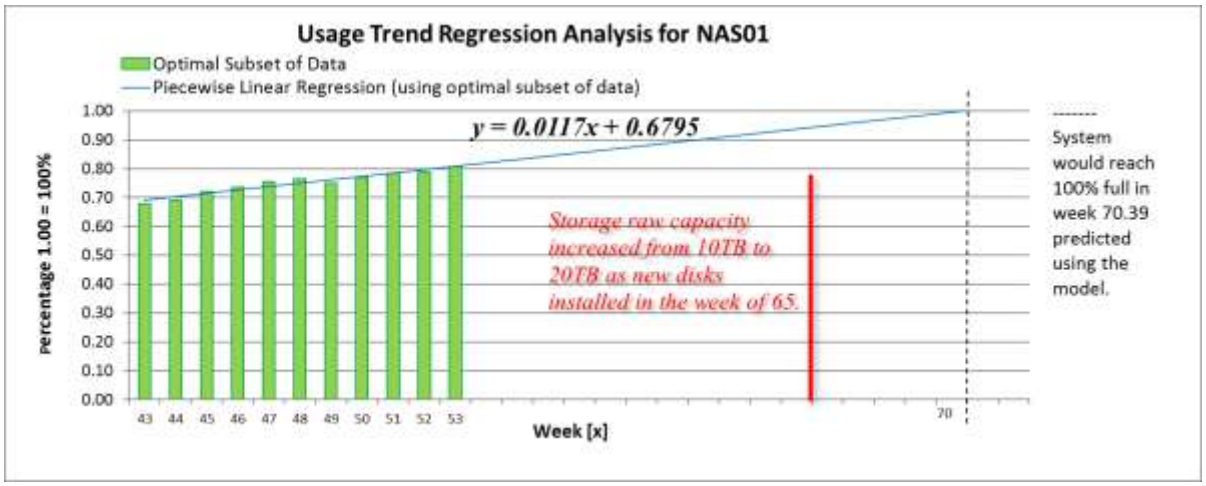

Figure 9: Piecewise Linear Regression model for the optimal subset of data

A. Deriving Equations:

Equation of the Linear Model:

$$y = 0.0117x + 0.6795 \quad (5)$$

α (intercept term) = 0.6795

β (slope) = 0.0117

y – Capacity

x – Date (week)

The equation for capacity forecast:

$$x = \frac{(y-0.6795)}{0.0117} \quad (7)$$

B. Forecasting Results for NAS01:

The forecasting results are displayed in Table 2 below:

Table 2: 1-month, 3-month and 6-month forecasts for NAS01.

	1-Month Forecast	3-Month Forecast	6-Month Forecast
Forecast Date:	January 31, 2018	March 31, 2018	June 30, 2018
Week#	57	In between week 65 & 66	In between week 78 & 79
Predicted value(s) [y]:	y = 0.0117(57- 43) + 0.6795 = 0.8433 (84%)	For Week 65 (Mar 25, 2018): y = 0.0117(65 - 43) + 0.6795 = 0.9369 (94%) For Week 66 (Apr 1, 2018): y = 0.0117(66 - 43) + 0.6795 = 0.9486 (95%)	For week 78 (Jun 24, 2018) y = 0.0117(78 - 43) + 0.6795 = 1.089 (109%) For week 79 (Jul 1, 2018) y = 0.0117(79 - 43) + 0.6795 = 1.1007 (110%)
Actual value(s) [y]:	y = 0.87 (87%)	y = 0.89 (89%) for week 65* y = 0.89 (89%) for week 66*	y = 10.02 (100%) for week 78* y = 10.11 (101%) for week 79*
Error %:	3%	5% - 6%	9% - 10%
Comment:	Error is less than 5%, therefore, the model is **valid**.	Error is more than 5%, therefore, the model is **invalid**.	Error is more than 5%, therefore, the model is **invalid**.

**As capacity was doubled in the first week of April 2018 and this prediction is made based on Jan 1, 2018, we will use the old capacity to compute the actual utilization (%) for analysis.*

C. Full Capacity Forecast for NAS01:

For the forecast date when system reaches full capacity:

$$x = \frac{(1 - 0.6795)}{0.0117} = 27.4 \quad (8)$$

So, Week = 43 + 27.4 = 70.4

Therefore, system will reach full capacity in between the week 70 (Apr 22, 2018) and week 71 (Apr 29, 2018) as predicted by the model.

However, looking at the history dataset, the actual storage capacity during the forecasted week was captured as 46.25% and, we realized the significant gap is due to new disks were installed in the first week of April 2018 which doubled the system capacity. In Figure 10 below, portions of the utilization report illustrate the drastic increase in total capacity and the actual values for week 70 and week 71.

Date	Week *[x]*	Raw Capacity (TB)	Used Raw Capacity	Unused Raw Capacity	Utilization (%) *[y]*
3/25/18 12:00 AM	65	10.00	9.05	0.95	90.50%
4/1/18 12:00 AM	66	20.00	8.90	11.10	44.50%
4/8/18 12:00 AM	67	20.00	9.05	10.95	45.25%

Date	Week *[x]*	Raw Capacity (TB)	Used Raw Capacity	Unused Raw Capacity	Utilization (%) *[y]*
5/6/18 12:00 AM	71	20.00	9.39	10.61	46.95%
5/13/18 12:00 AM	72	20.00	9.48	10.52	47.40%
5/20/18 12:00 AM	73	20.00	9.57	10.43	47.85%

Date	Week *[x]*	Raw Capacity (TB)	Used Raw Capacity	Unused Raw Capacity	Utilization (%) *[y]*
4/29/18 12:00 AM	70	20.00	9.25	10.75	46.25%
5/6/18 12:00 AM	71	20.00	9.39	10.61	46.95%
5/13/18 12:00 AM	72	20.00	9.48	10.52	47.40%

Figure 10: Portions of storage utilization report for system NAS01

Assuming we were forecasting before the new disks were installed, we should be using utilization (%) based on old raw capacity (10TB). Hence, the actual utilization (%) for week 70 is:

$$y = \frac{9.25TB}{10TB} = 0.925 \; or \; 92.5\% \qquad (9)$$

As the error of this full capacity forecast exceeds 5% (100% - 92.5% = 7.25%), therefore, the model is not valid for full capacity forecast as it fails to comply with the validation rule of Prediction Error stated under Section 2.6, "Regression Model Validations".

3.3.2 Application of Regression Model for Storage System SAN01

The historical dataset of the SAN system is represented by histograms as shown in Figure 11 below.

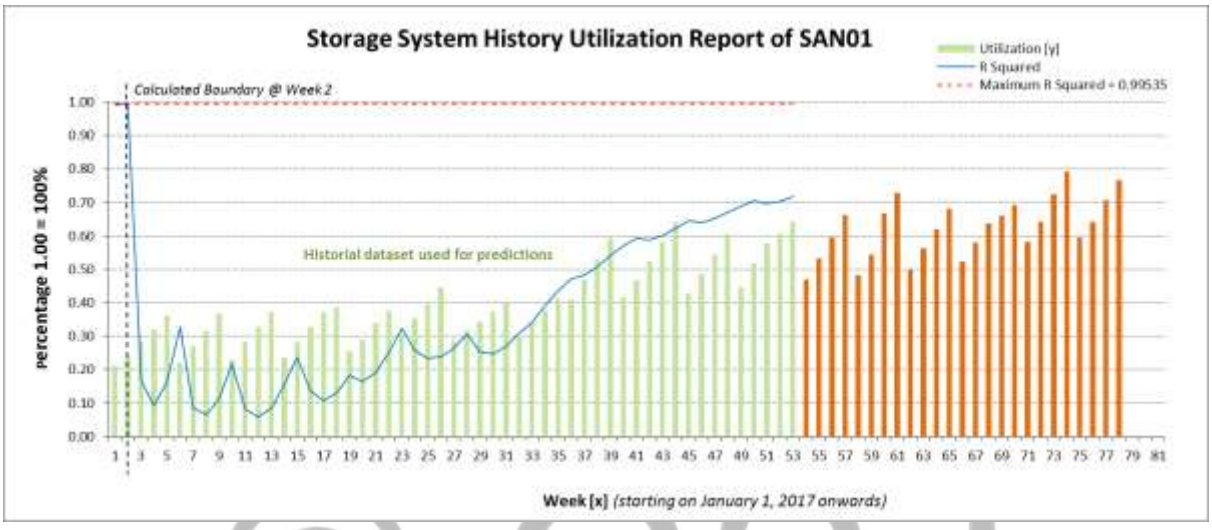

Figure 11: The Calculated Boundary & Maximum R2 for Storage System SAN01

In Figure 11, the calculated boundary is at week 2 and as it exhibits a positive slope, we conclude that this is also the true boundary. Hence, the optimal subset of data determined would be the histograms between week 2 (January 8, 2017) and week 53 (December 31, 2017), as shown in Figure 12 below.

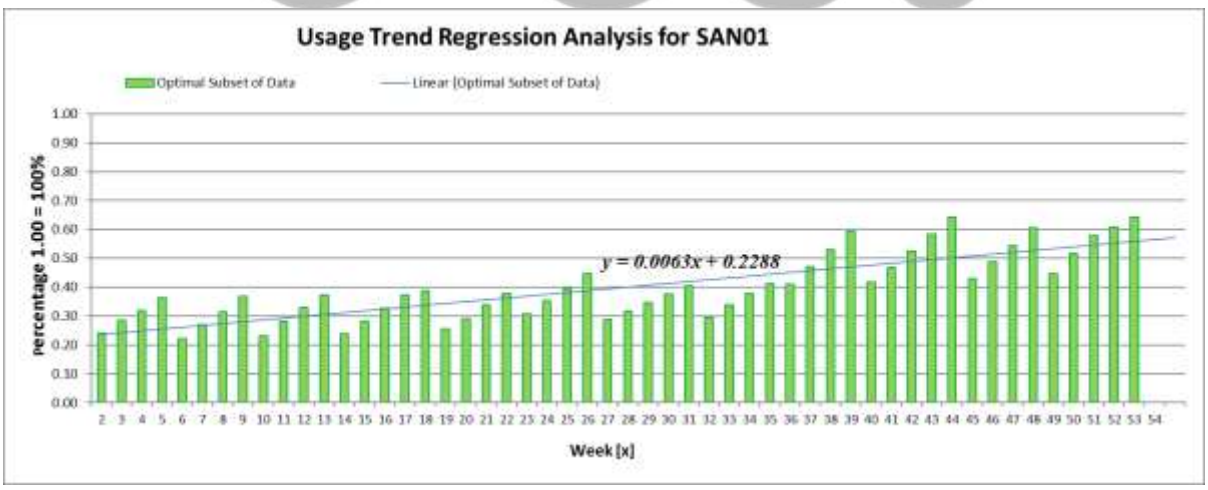

Figure 12: Piecewise Linear Regression model for SAN01 using optimal subset of data

A. Deriving Equations:

Equation of the Linear Model:

$$y = 0.0063x + 0.2288 \quad (10)$$

α (intercept term) = 0.2288

β (slope) = 0.0063

y – Capacity

x – Date (week)

The equation for capacity forecast:

$$y = 0.2288 + 0.0063x \quad (11)$$

Rearranging the equation () to solve for:

$$x = \frac{(y-0.2288)}{0.0063} \quad (12)$$

B. Forecasting Results for SAN01:

Table 3: 1-month, 3-month and 6-month forecasts for SAN01

	1-Month Forecast	**3-Month Forecast**	**6-Month Forecast**
Forecast Date:	January 31, 2018	March 31, 2018	June 30, 2018
Week#	57 In between week 57 & 58	In between week 65 & 66	In between week 78 & 79
Predicted value(s) [y]:	For Week 57 (Jan 28, 2018): *y* = 0.0063(57 - 2) + 0.2288 = 0.5753 (58%) For Week 58 (Feb 4, 2018): *y* = 0.0063(58 - 2) + 0.2288 = 0.5816 (58%)	For Week 65 (Mar 25, 2018): *y* = 0.0063(65 - 2) + 0.2288 = 0.6257 (63%) For Week 66 (Apr 1, 2018): *y* = 0.0063(66 - 2) + 0.2288 = 0.632 (63%)	For week 78 (Jun 24, 2018) *y* = 0.0063(78 - 2) + 0.2288 = 0.7076 (71%) For week 79 (Jul 1, 2018) *y* = 0.0063(79 - 2) + 0.2288 = 0.7139 (71%)
Actual value(s) [y]:	*y* = 0.87 (87%)	*y* = 0.68 (68%) for week 65* *y* = 0.52 (52%) for week 66*	*y* = 0.7675 (77%) for week 78* *y* = 0.6475 (65%) for week 79*
Error %:	29%	5% - 11%	6%
Comment:	Error is more than 5%, therefore, the model is **invalid**.	Error is more than 5%, therefore, the model is **invalid**.	Error is more than 5%, therefore, the model is **invalid**.

**As capacity was doubled in the first week of April 2018 and this prediction is made based on Jan 1, 2018, we will use the old capacity to compute the actual utilization (%) for analysis.*

C. Full Capacity Forecast for SAN01:

For the forecast date when system reaches full capacity:

$$x = \frac{(1 - 0.2288)}{0.0063} = 122.4 \quad (13)$$

So, Week = 2 + 122.4 = 124.4

Therefore, system will reach full capacity in week 122 (month end of April 2019 which indeed is a real future date at time of writing), as predicted by the model.

4 Discussions and Conclusion

In this section, we will bring together the trails of thought from previous sections, to discuss the forecasting results outlined in Section 3.3 and to draw conclusions. The objective is to respond to the research hypotheses (posed in Section 3.1), and to examine the

limitations of the methodology. We will also further study how qualitative analysis maybe extended to improve the accuracy of the forecasts.

4.1 Responses to the Hypotheses

To reiterate, for the first hypothesis, we made the assumption that the PLR model defined, is not a "one model fits all" methodology. Deriving an alternative regression model is not part of this research, but rather, to identify any limitations PLR model would have.
For the second hypothesis, we assumed capital efficiency can be maximized delay upgrade purchases as disk prices are falling quickly, i.e., buy new disks only when needed in the immediate future. We will then discuss if procurement should be initiated using 1-month, 3-month, 6-month or full capacity forecast.

4.1.1 Response to Research Hypothesis 1

There were two sets of forecasts generated using the PLR model. For the NAS system, we proved that the model is only valid for 1-month forecast. If we compare the predicted value against the actual value, this forecasting result seems to be convincing.
For the SAN system, the model was proved to be invalid for all forecasting results of 1-month, 3-month and 6-month respectively, whereas the full capacity forecast predicted a future date which we have no actual value to validate.
Therefore, we will accept **Research Hypothesis 1**, that we can generate accurate capacity forecasts for NAS systems using the PLR model but not for SAN type systems.

4.1.2 Response to Research Hypothesis 2

In July 2017, Backblaze, Inc. (a data storage provider hosting more than 750 Petabytes data for both personal and enterprises) released a report indicating the average cost per drive size and per GB has been declining over the years (as shown in Figure 13).

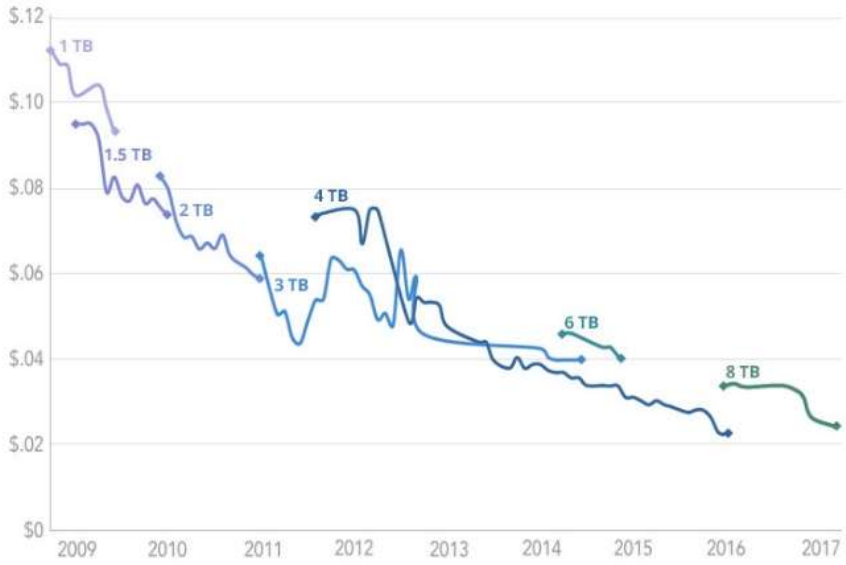

Figure 13: Backblaze Average Cost per Drive Size. Reprinted from "Hard Drive Cost Per Gigabyte", Backblaze Inc., July 2017.

Looking back into January 2018, a storage upgrade project was initiated for the NAS system and the total capacity was doubled in April 2018. In conjunction with Backblaze's observation shown in Figure 13, we know that making the system upgrade decision was reasonably at the right time in order to avoid business interruptions while maximizing cost saving at the same time.
For SAN type storage systems, the costs of implementation and upgrade are generally more expensive than other system types. It is common for a storage administrator to perform more frequent housekeeping tasks on SAN systems, i.e., moving data that is used less frequently to lower tier storage systems and archiving old data to tape storage media, etc. Although the PLR model was proved to be invalid for the SAN storage system, Figure 11 still exhibits a gradual usage growth for the system but at a much slower rate than the NAS system. The best time to initiate a system upgrade project can still be analysed qualitatively to near immediate future for maximized cost saving.
Therefore, we will accept **Research Hypothesis 2**, that it will be less expensive to delay upgrades for additional storage capacity until it is absolutely needed.

4.2 Evaluation of the Case Study: Qualitative Analysis

In reality, storage capacity forecasting consists of two sub-processes:

- Resource Capacity Management – analysing the current resource utilizations and demands to determine a usage trend
- Business Capacity Management – obtaining business projections and forecasting the impact of the new demand on the existing resources

The most evident limitation of my research would be the restricted validity of the PLR model generating forecasts in the immediate future for NAS systems only. For the farther forecasts generated for the NAS system, error percentages were ranging from 5% to 10%. The dataset for analysis and the produced forecasts can still make meaningful predictions together with qualitative measures. To be specific, forecasts generated with error percentages out of the acceptable range, can still be interpreted by a storage expert (intuition and experience gained over time on the job) and stakeholders of other business units. Qualitative method anticipates future needs based on planned activities and information shall be gathered directly from users through regularly scheduled resource-planning meetings, when users should inform of new initiatives from their respective business units that may require more or less storage.
In my research, qualitative method is even more crucial when forecasting capacity for SAN type storage systems. Although the PLR generated has no meaning to the dataset, other than exhibiting a gradual growth, we can also see a pattern may possibly exist:

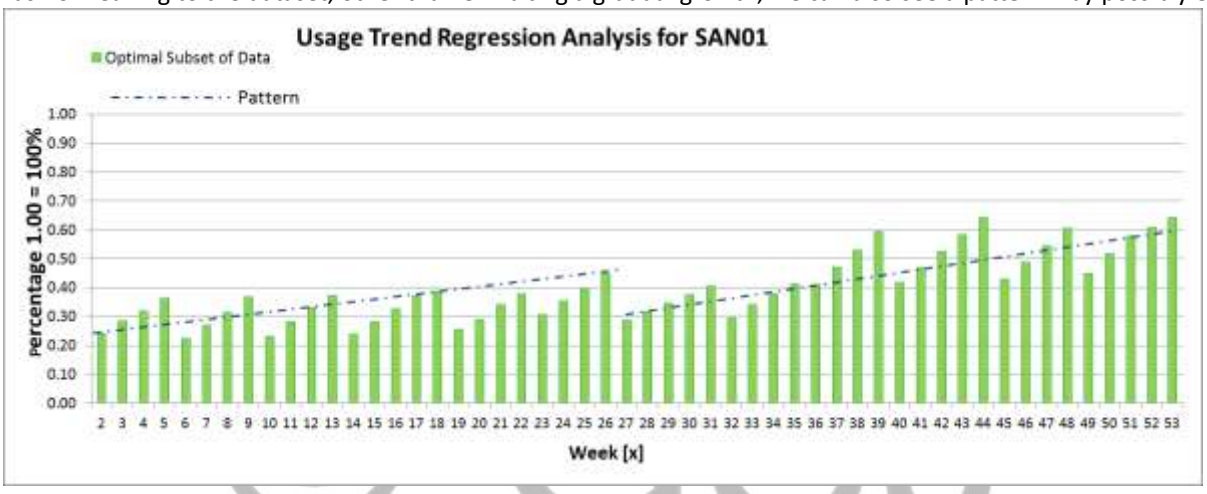

Figure 14: Time series data which exhibits a pattern

This best practice combining both methods should be well valid and extended to apply on SAN type storage systems as well.

4.3 Conclusion

Capacity forecasting is all about becoming proactive. In the context of this case study, we have illustrated that it is a good practice to combine the two methods of quantitative and qualitative when devising capacity forecasts for any storage systems.
IDC recently released a report on the ever-growing datasphere and predicts that the collective sum of the world's data will grow from 33 zettabytes this year to a 175ZB by 2025 as shown in Figure 15 [3].

Figure 15: Annual size of the global datasphere. Reprinted from "DatasphereSource: Data Age 2025", sponsored by Seagate with data from IDC Global DataSphere, Nov 2018.

For many companies, data storage management has become a very challenging task. More and more enterprise-wide transactional systems are being implemented, massive data warehouses are being deployed as we are entering the "Age of Big Data", explosive growth in email traffic, and all of these are contributing to the rising capacity demands. Having a storage forecasting process in place will definitely help an organization to achieve a predictable and controllable storage management environment. Data storage managers and administrators can better plan for demand in a more strategic mode to minimize business risks and be "ahead of the curve" in managing the resources.

Acknowledgment

None.

References

[1] "Mixed methods research," FoodRisC Resource Centre, 2016. [Online]. Available: http://resourcecentre.foodrisc.org/mixed-methods-research_185.html.

[2] J. Allspaw, "Predicting Trends," in The Art of Capacity Planning, O'Reilly Media, Inc, 2008. https://www.oreilly.com/library/view/the-art-of/9780596518578/ch04.html.

[3] A. Klein, "Hard Drive Cost Per Gigabyte," 2017. [Online] Available: https://www.backblaze.com/blog/hard-drive-cost-per-gigabyte/.