Capacity Forecasting in a Backup Storage Environment

Mark Chamness
EMC
Mark.Chamness@emc.com

Abstract

Managing storage growth is painful [1]. When a system exhausts available storage, it is not only an operational inconvenience but also a budgeting nightmare. Many system administrators already have historical data for their systems and thus can predict full capacity events in advance.

EMC has developed a capacity forecasting tool for Data Domain systems which has been in production since January 2011. This tool analyses historical data from over 10,000 back-up systems daily, forecasts the future date for full capacity, and sends proactive notifications. This paper describes the architecture of the tool, the predictive model it employs, and the results of the implementation.

Tags: storage, predictive modeling, case study, capacity planning, forecasting, machine learning.

1 Introduction

Data storage utilization is continually increasing, causing the proliferation of storage systems in data centers. Monitoring and managing these systems requires increasing amounts of human resources and therefore automated tools have become a necessity.

IT organizations often operate reactively, taking action only when systems reach capacity, at which point performance degradation or failure has already occurred. Instead, what is needed is a proactive tool that predicts the date of full capacity and provides advance notification.

Predictive modeling has been applied to many fields: forecasting traffic jams [2, 3], anticipating electrical power consumption [4], and projecting the efficacy of pharmaceutical drugs [5]. Within the IT field, capacity management of server pools has been studied [6]. Ironically, there seems to be little previous work discussing applications of predictive modeling to data storage environments.

During the past year a predictive model has been employed internally at EMC to forecast system capacity and generate alert notifications months before systems reach full capacity. The ultimate purpose of this tool is to provide customers with both time and information to make better decisions managing their storage environment.

2 Data Collection

Data Domain systems are backup servers that employ inline deduplication technology on disk. All Data Domain back-up storage devices have a “phone-home” feature called Autosupport. Customers can configure their Data Domain systems to send an email every day with detailed diagnostic information. In addition, they can send email when specific events are encountered by the operating system. Once these emails are received at EMC, they are parsed and stored in a database.

Sending of diagnostic data via email to EMC is voluntary by the customer. Often, in secure environments, customers choose to disable the feature. In order to monitor their systems, customers have the ability to configure the autosupport emails to be delivered to internal recipients.

Most customers choose to send autosupports to EMC because the historical data enables more effective customer support. Given the more than 10,000 autosupports received daily, EMC has a statistically significant view across the Data Domain install base.

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

1. Total physical capacity of the system
2. Total physical space used by the system

For Data Domain systems, the total physical capacity changes over time because they generate an index which slightly decreases the amount of physical capacity available for data storage.

3 Data Cleaning

In order to ensure data integrity there are two issues to be addressed: data artifacts and elimination of non-production data.

Data Artifacts: In order to prevent bad data from entering the analysis, the tool assesses the quality of every autosupport and applies rules to guarantee consistency. These artifacts may arise due to an error in parsing the autosupport, data corruption during the transport of the autosupport, or both.

Non-Production Data: All internal Data Domain lab systems and QA systems send autosupports which are parsed and loaded into the database. These systems may be under development and therefore their performance characteristics may vary dramatically from production systems being used in the field. While this data is of value to internal teams, for the purposes of capacity forecasting, the data from these systems is excluded.

4 Predictive Model

One of the most common methods employed in predictive modeling is linear regression. Unfortunately, application of regression to storage capacity time series data is challenging because behavior changes. System administrators may add more shelves to increase capacity, change retention policies, or simply delete data. Therefore blind application of regression to the entire data set often leads to poor predictions.

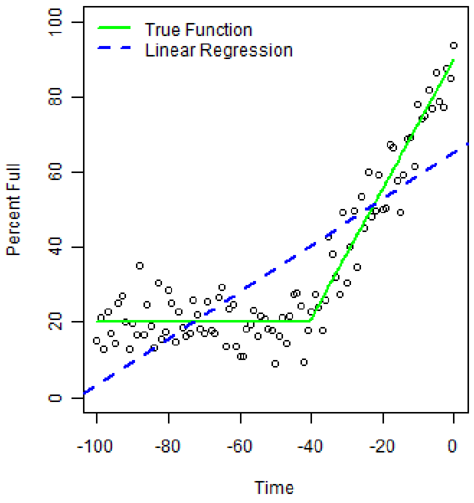

Figure 1: Example capacity data for the prior 100 days. (Time = 0 is the most recent data.) The standard deviation is 6 thoughout the data. The blue line shows the result of applying linear regression to the entire data set.

The predictions of the linear regression in Figure 1 are very poor. Intuitively, the data indicates the system is going to reach 100% capacity within a few days, but the regression line predicts far later (a false negative).

Select a Subset of Recent Data

The simplest method to mitigate the issue illustrated in Figure 1 would be to choose a subset of recent data such as the prior 30 days. This eliminates the influence of older data and improves the accuracy of the model's predictions. Unfortunately, using a fixed subset to model all systems results in poor linear models for many systems. Significantly more accurate models can be obtained by finding the optimal subset of data for each system and applying linear regression to only that subset of the data.

4.1 Piecewise Linear Regression

The error rate of the linear regression model can be significantly reduced by applying the regression to a data subset that best represents the most recent behavior. This requires implementing piecewise linear regression [7].

In order to find the best subset of data, the boundary must be determined where the recent behavior begins to deviate. The method described here analyses the quality of many linear regressions and then selects the one having the best fit.

The goodness-of-fit of a linear regression to experimental data can be measured by evaluating the coefficient of determination R^2. It is defined as the regression sum of squares ("SSM") divided by the total sum of squares ("SST") [8].

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

Properties of R^2

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

Calculating the Boundary: Start with a small subset of the data, such as the prior 10 days, and then apply regression to incrementally larger subsets to find the regression having the maximum value of R^2.

1. Regress $\{(x_{-10}, y_{-10}), (x_{-9}, y_{-9}), \ldots, (x_0, y_0)\}$
2. Calculate R^2 for regression
3. Regress $\{(x_{-11}, y_{-11}), (x_{-10}, y_{-10}), \ldots, (x_0, y_0)\}$
4. Calculate R^2 for regression
5. …
6. Regress $\{(x_{-n}, y_{-n}), (x_{-n+1}, y_{-n+1}), \ldots, (x_0, y_0)\}$
7. Calculate R^2 for regression
8. Select the subset with maximum R^2

The boundary is the oldest data point within the subset of data determined in step 8. The predictive model is generated by applying linear regression to that subset.

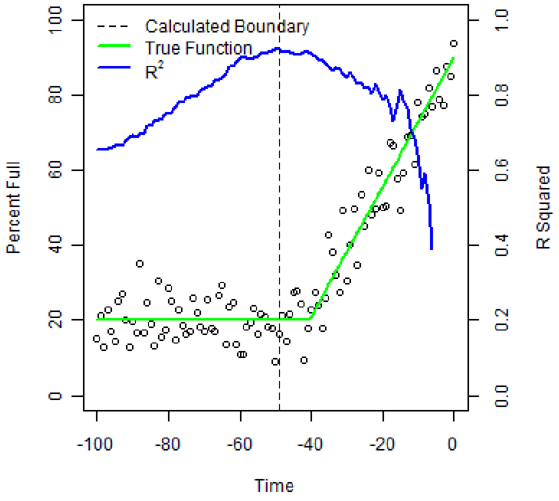

Figure 2: The same data used in Figure 1 with R^2 plotted for each subset of data. The date when R^2 reaches its maximum value is the "calculated boundary" and occurs near the discontinuity of the true function. Maximum R^2 = 0.92 at -50 days and the true boundary is -40 days.

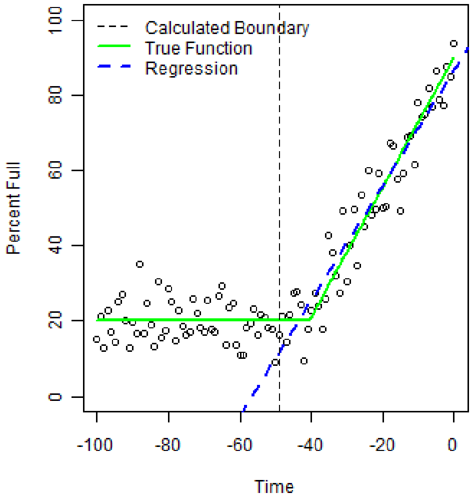

Figure 3: The same data from Figures 1 & 2. Piecewise linear regression results in a better fit to the data. This model was generated using the subset $\{(x_{-42}, y_{-42}), \ldots, (x_0, y_0)\}$.

Preprocessing data by applying a smoothing function can increase R^2, but has limitations. Filtering out noise while maintaining the signal is easier said than done. Too much smoothing and it becomes too difficult to determine the boundary point.

4.2 Other Models

Many other models can be applied to time series data, such as weighted linear regression, logarithmic regression, and auto-regressive (AR) models. In the current implementation, a simple linear model has shown to effectively model many systems (see Section 5: "Results of Predictive Modeling"). It is an open question whether the remaining systems can be modeled by other methods.

4.3 Model Validation

The model needs to be able to say, "I don't know." Sometimes there is no pattern in the data. Before employing a model to predict future behavior, it should be evaluated to determine if it is reasonable model for the data set. In the current implementation, validation rules are applied to the results of the linear model to determine if capacity forecasts should be published.

Goodness-of-fit: When the R^2 value from piecewise linear regression is too small, it indicates the model is a poor fit to the data. In the current tool, linear regression models with $R^2 < 0.90$ are not used.

Positive Slope: Linear models having a zero or negative slope cannot be used to predict the date of 100% full.

Timeframe: Forecasts for systems to reach full capacity far into the future are extrapolating the current behavior too much to be practical. The current model limits forecasts to less than 10 years. The expectation is that within 10 years the storage technology will be significantly different than it is today.

Sufficient Statistics: Storage systems that have been recently deployed lack enough historical data to produce statistically sufficient regression models. A minimum of 15 days of data is a reasonable threshold for the size of the data set.

Choosing a smaller minimum value may result in fitting the model to noise. Linear regression can achieve a very good fit to a handful of data points, but the results are not statistically significant.

Space Utilization: Experience has shown that systems which are less than ~10% full tend not to produce reliable predictions. For this situation, the current tool does not generate capacity forecasts.

Last Data Point Trumps All: Recent changes in system capacity must be taken into account to evaluate the linear fit. When systems are nearing maximum storage capacity, the administrator often takes action which results in drastic changes in the amount of available capacity. If the administrator reduces the amount of data stored on a device, the capacity prediction of the model is no longer valid. Assessing this error is a simple form of cross-validation.

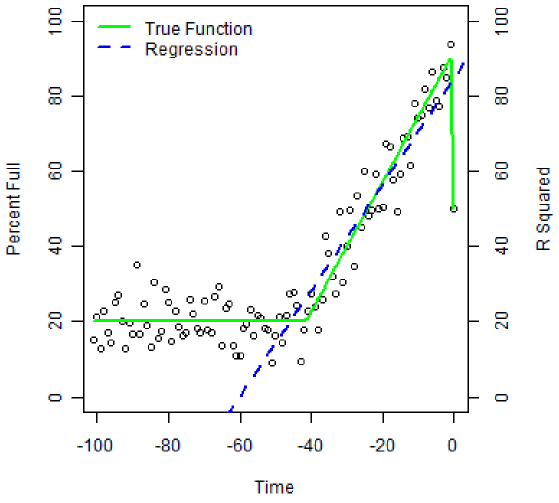

Figure 4: System capacity dropped from 94% to 50%. The linear model generated has a high goodness-of-fit ($R^2 = 0.88$) but the prediction for the most recent data point has 35% percent error. The model predicts the system is 85% full at Time=0, but it is only 50% full.

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 and therefore the model is no longer valid.

5 Results of Predictive Modeling

5.1 Analysis of Linear Regression Fit to Past Data

If historical data does not demonstrate linear growth, then obviously linear regression would be a poor model to employ. To investigate this issue, the piecewise linear regression algorithm described in section 4.1 was applied to the historical dataset from Data Domain storage appliances and the maximum R^2 was calculated for each system.

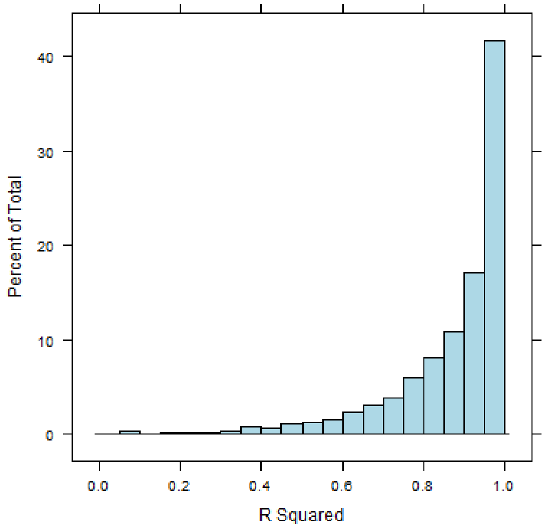

Figure 5: Histogram of R^2 across all systems using a minimum 15 days of data. This illustrates that most of the regression models generated for storage systems have R^2 close to 1.0.

Summary of results:

1. The median R^2 for all systems was 0.93
2. Models for 60% of systems had $R^2 \geq 0.90$
3. Models for 78% of systems had $R^2 \geq 0.80$

These results indicate that the majority of systems exhibit very linear behavior since the linear model had a very good fit to the datasets.

5.2 Forecasting Full Capacity

After the model is generated from historical data, the next step is to apply the validation rules described in section 4.3. For models that pass validation, the final step is to solve for the future date the system will become 100% full. The linear model:

$$y = \alpha + \beta x$$

Definitions:

- y is capacity
- α is the intercept term
- 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 calculated by setting the capacity y = 1 (100 %) and solving for x:

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

5.3 Analysis of the Quality of Forecasts

False Positives

False positives frequently originate from unforeseen future human activities which cannot be predicted by the model. It is difficult to construe such false positives as flaws in the model per se given that the only input provided to the model is historic behavior of the system.

When a system is on a linear trajectory to full capacity but never reaches 100% full, it is may be due to external or internal events. An external event may originate from a significant change in the amount or rate of data placed into primary storage. An event internal to the system may be caused by the system administrator taking action to implement configuration changes. These can include:

1. Hardware changes
   a. The system was entirely replaced
   b. A shelf was added , increasing capacity
   c. Internal disk drives were replaced
2. Software changes
   a. Retention policy was changed
   b. Data was deleted and/or moved

A specific example may help elucidate the issues concerning false positive capacity forecasts. Even with visual inspection of the data by a human, it is extremely difficult to assess a false positive a priori.

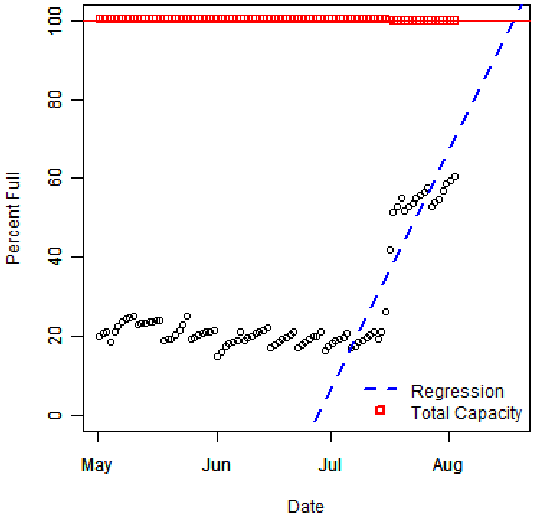

Figure 6: System exhibiting several changes in the rate of storage utilization. At this point in time the regression may be a *false positive*.

Visual inspection of the storage capacity of the system in Figure 6 indicates that the rate of storage is on a trajectory to reach full capacity in September. However, the most recent data in August might be an early indication that the trajectory is changing. This recent data may imply the system is stabilizing near 60% of capacity, but at this point in time there is insufficient data to establish a new trajectory.

From a statistical perspective, it is unknown whether the recent data points are signal or noise. This illustrates how allowing the use of small data sets has the risk of fitting the model to noise.

Ironically, in spite of the intuitive uncertainty, the fit to data is very good: R^2 = 0.90 and the prediction error is only 4.5% on the most recent data point. This example is potentially a good candidate for the model to fail validation and report, "I don't know." There is a trade-off between eliminating reasonable models versus generating false positives. By requiring more data for models, we gain higher confidence in their predictions, but reduce the advanced notification for true positives.

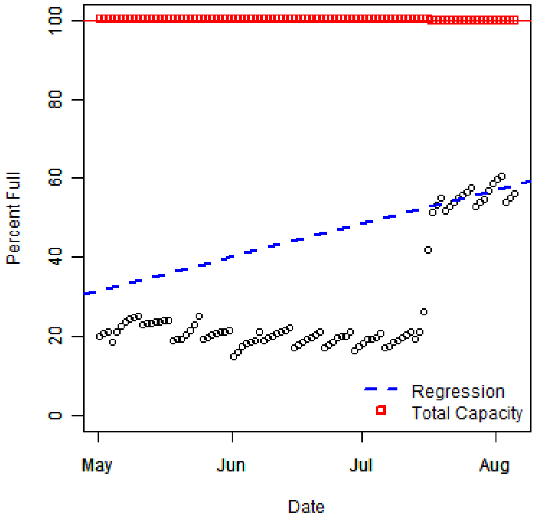

Figure 7: Same system shown in Figure 6 with additional data points.

After a few more days, the piecewise regression model fits the recent behavior of the system in Figure 7. Only after obtaining more data can we determine that the model in Figure 6 was a false positive. It is often the case that false positives can only be observed with the benefit of hindsight (addition data).

No Forecast for Full Capacity

When a model fails validation (described in section 4.3) no forecast should be made. On a typical day the current model does not publish forecasts for approximately 40% to 50% of all systems. This is not a surprising result. Most systems are expected to be efficiently managed by their administrators. The model is only considered valid for systems which are on a trajectory to full capacity in the future.

It is an operational decision to determine the quantity of forecasts to be published. The percent of systems for which forecasts are published can be easily adjusted by tailoring the validation rules for each environment.

5.4 Analysis of Forecasts across Install Base

Application of the model described to the entire install base results in a number of observations.

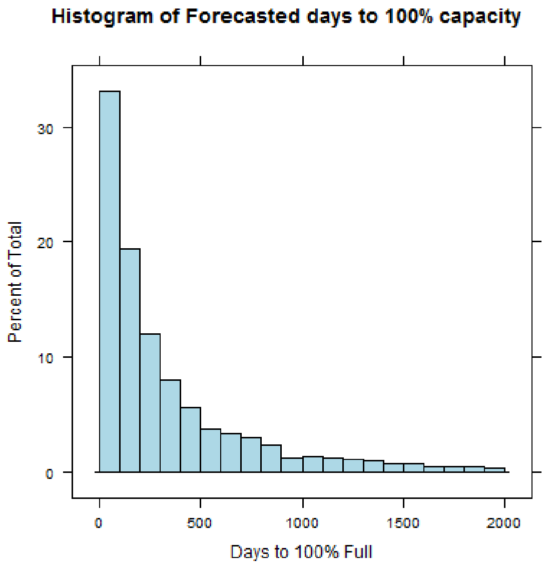

Figure 8: Histogram of forecasts for systems to reach full capacity. The median time to 100% full is 197 days. Therefore, for systems with valid models, the forecast is half of them will reach full capacity within approximately six months.

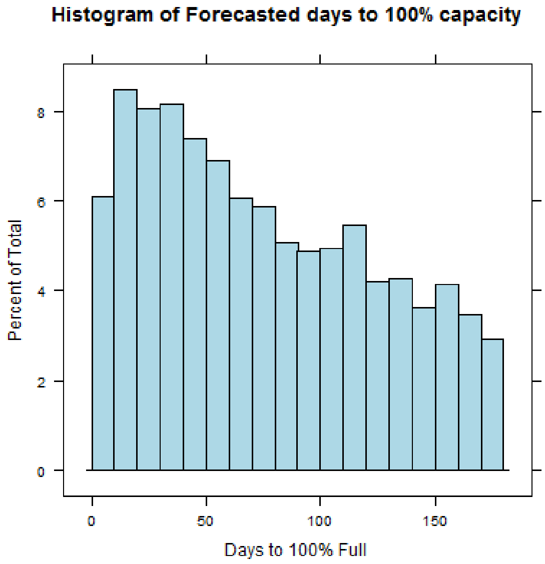

Figure 9: Greater detail (6 months) of the data used in Figure 8.

Given the peak values of these histograms, a majority of the systems are predicated by the model to reach full capacity in the near future. There are at least two conjectures that may explain the patterns of Figures 8 & 9:

Hypothesis 1: Efficient use of capital: Since the cost of storage (dollars per GB) drops quickly over time, the majority of storage devices are intended to only have enough space for the near future. It's cheaper to delay the purchase of additional storage until it's absolutely needed.

Hypothesis 2: Capacity Exceeded Expectations: System administrators forecasted their capacity needs for the long-term, but they underestimated the rate of growth.

6 Capacity Forecasting Examples

The application of capacity forecasting may be illustrated by examining a few examples of production Data Domain storage systems.

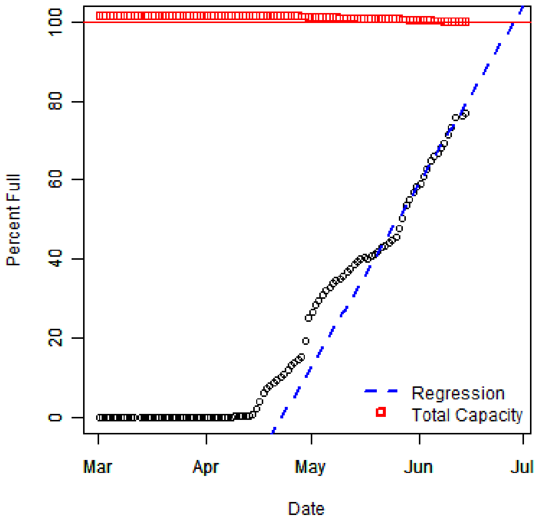

Figure 10: System exhibiting linear segments.

This type of behavior was the motivation for developing the piecewise linear regression algorithm. The data prior to May is useless for prediction since it significantly different from the current behavior of the system. Application of piecewise linear regression correctly found a model that fits the data from the beginning of June to the last data point.

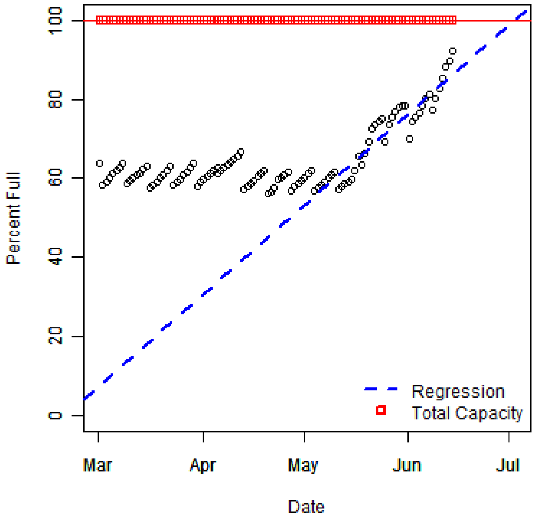

Figure 11: A behavioral change in the rate of storage utilization occurred at the end of May, but the piecewise linear regression model correctly fit the most recent behavior despite noisy data.

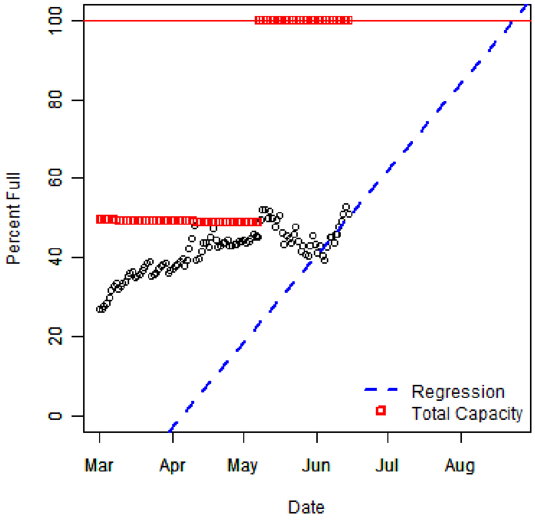

Figure 12: Shelf was added to an existing Data Domain system.

The total capacity exhibits a discontinuity in May. In this figure, the system reached full capacity and then a shelf was added. The model fits the recent data and predicts the system will reach 100% capacity in approximately three months.

7 Conclusions and Future Work

The role of automated predictive modeling for managing IT systems will become more pervasive as the complexity and size of data centers continue to grow. [9]

This paper describes a model that uses historical data to predict when Data Domain systems will reach full capacity. Advance notice of storage systems reaching full capacity allows system administrators to take necessary measures to avoid performance degradation and/or failure. It was demonstrated that many storage systems can be modeled using a piecewise linear regression model. Furthermore it was shown that for the systems that could be modeled, they were able to generate a forecast of the date of full capacity in advance.

Many questions still remain for future analyses which are natural extensions of the material discussed in this paper:

1. Are there other applications of predictive modeling within the existing data set? Could compression ratio, bandwidth throughput, load-balancing [10] or IO capacity also be predicted?
2. Why was the piecewise linear regression model not able to model some systems? Could the model be improved or could they be modeled by some other method?
3. Using the statistically significant view across the install base, could there be correlations between system variables or time series correlations for a single variable?

Capacity forecasting is a fundamental utility for system management, but it is only a starting point of the data analysis that can be explored for storage management.

Acknowledgments

I thank Stephen Manley, Fred Douglis, and Philip Shilane for both guidance and meticulous comments on early drafts. I especially appreciate the feedback and direction provided by my shepherd, Andrew Hume, as well as the anonymous referees.

8 References

[1] TheInfoPro. “Deduplication: A paradigm Shift in Backup”, *TheInfoPro (TIP) Research Paper*, January 2011. Available at: https://community.emc.com/docs/DOC-9720

[2] Andras Hegyi. “Model Predictive Control for Integrating Traffic Control Measures”, February 2004. Available at: http://www.dcsc.tudelft.nl/~deschutt/research/phd_theses/phd_hegyi_2004.pdf

[3] Eric Horvitz, Johnson Apacible, Raman Sarin, and Lin Liao. “Prediction, Expectation, and Surprise: Methods, Designs, and Study of a Deployed Traffic Forecasting Service”, *Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI-2005, Edinburgh, Scotland, July 2005. Available at: http://research.microsoft.com/~horvitz/horvitz_traffic_uai2005.pdf

[4] Eduardo Camponogara, Dong Jia, Bruce H. Krogh, and Sarosh Talukda. “Distributed Model Predictive Control”. *Control Systems, IEEE*, February 2002. Available at: http://www.ece.cmu.edu/~krogh/papers/CJKT02.pdf

[5] WM Watkins, EK Mberu, PA Winstanley. “The efficacy of antifolate antimalarial combinations in Africa: a predictive model based on pharmacodynamic and pharmacokinetic analyses”, *Parasitology Today, Volume 13, Issue 12*, December 1997, Pages 459-464, 1997

[6] Daniel Gmach, Jerry Rolia, Ludmila Cherkasova, Alfons Kemper, "Capacity Management and Demand Prediction for Next Generation Data Centers," icws, pp.43-50, IEEE International Conference on Web Services (ICWS 2007), 2007

[7] Robert Nisbet, John Elder, and Gary Miner. *Statistical Analysis and Data Mining Applications*. Academic Press, 2009

[8] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Pearson Education, 2006

[9] IBM. “The growing role of predictive analytics in data center management”, *IBM Developer Works*, December 2010. Available at: https://www.ibm.com/developerworks/mydeveloperworks/blogs/business-analytics/entry/the_growing_role_of_predictive_analytics_in_data_center_management

[10] Fred Douglis, Deepti Bhardwaj, Hangwei Qian, Philip Shilane. “Content-aware Load Balancing for Distributed Backup”, LISA 2011: Proceedings of the 25th Large Installation System Administration Conference (Dec 2011)