Analysis of Operational Data to Improve Performance in Service Delivery Systems

Yixin Diao and Aliza Heching
IBM Thomas J. Watson Research Center
P.O. Box 704, Yorktown Heights, NY 10598, USA
Email: {diao|ahechi}@us.ibm.com

Abstract— **Enterprises and service providers are increasingly challenged with improving the quality of service delivery. Toward this end, service delivery organizations are collecting large volumes of operational data. However, it is often difficult to effectively extract insights from this data that can be used to guide decisions in the service delivery environment. In this paper we study an array of analysis methods that were performed against service delivery operational data and that can be used to provide managerial insight into a complex service delivery system. We demonstrate the applicability of our approaches in a large IT services delivery environment.**

I. INTRODUCTION

In recent years, the IT services industry has faced continual pressure to improve quality of service provided to its customers while simultaneously driving down the cost of service delivery. These apparent conflicting objectives have lead the industry to explore innovative methods for managing its business. Common metrics for measuring quality of service, including equipment availability, time to resolve incidents, and mean time between failures, are measured on a regular basis as part of the standard management and operating process. Such metrics are required to measure the provider's performance against contractual service level agreements. However, in an attempt to improve quality of service while driving down cost of delivery, service providers are adopting a more continual and measurable focus on additional metrics and measures of interest. These include exploring the nature of the workload driven by each customer, the profile of agents supporting each customer, agent skills, agent and team performance, level of agent cross-training, and the like.

We describe an IT services delivery environment where a provider provides IT support from either on-shore or off-shore locations to customers who may be globally located. A customer may have one or more IT needs ("request types") and these request types may require different skills ("complexity of request") in order to respond to the request. The service delivery provider has one or more service delivery locations. As customer requests arrive, they are assigned to a service delivery location; agents in the service delivery locations are responsible for handling and responding to the requests. Although the agents in these service delivery locations respond to the requests, they do not directly interact with the end customers.

In this paper we describe how data collected by a service delivery provider can be used to create metrics to gain managerial insights into these new metrics and measures of interest. We highlight actions that can be taken based upon the analysis. We also discuss challenges related to the quality of the data that is collected and methods that can be used to identify and correct the data anomalies. The approaches described in this paper have been implemented at a large services delivery provider with worldwide delivery locations and global customers.

A vast amount of literature has been dedicated to describing various aspects of service delivery. We focus here on literature dedicated to three areas of service delivery: (i) performance measurement in service delivery, (ii) estimating parameters based on service delivery data and (iii) optimal decision making in a services delivery environment. We mention a few papers in each of these research streams. Parasuraman et al. [1] describe a theoretical model for measuring quality of service in a services environment. Cardoso et al. [2] discuss methods for monitoring and predicting quality of service for workflows based upon the component services. Larson [3] discusses metrics that can be defined and monitored to help achieve service level agreements. Finally, Bose et al. [4] describe a framework for continually measuring performance of service in an IT service delivery environment. In the area of estimating parameters based on service delivery data and related challenges, Brown et al. [5] suggest that service time distributions are more appropriately modeled as being generated from a lognormal distribution. Chandra et al. [6] discuss methods for estimating arrival rates from historical workload data, using time series analysis. We refer the reader to Gans et al. [7], [8], Mandelbaum and Zeltyn [9] and the references therein for additional discussions. Finally, many papers discuss optimization in the services delivery environment. Of particular focus is staffing optimization in this environment. We cite just a small number of papers in this vast research area. Wallace and Whitt [10] and Robbins et al. [11] discuss cross training in delivery environments. Gurvich and Whitt [12], [13] discuss optimal staffing levels in a delivery environment where agents have different skill levels and requests require different skills and service levels are relevant. However, while the above analytical methods has been studied to address different needs of service delivery, none of them have dealt with thorough analysis of the operational data, which is of practical importance to improve the performance of service delivery organizations.

The remainder of this paper is organized as follows. Sec-

978-3-901882-48-7

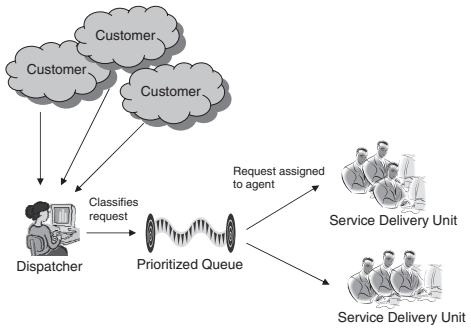

Fig. 1. Illustrative Process and Operation Flow of A Service Functional Unit.

tion II discusses the background for service delivery systems. Section III describes some anomalies often observed in the data and approaches adopted to preparing the data prior for data analysis. Section IV presents analysis of the service delivery data as well as conclusions that can be drawn based upon the data analysis. Our conclusions are contained in Section V.

II. Service Delivery Systems

Service delivery involves customers contracting with a services provider on a menu of IT services such as security patch management, network management, and data backup and restore management. The customer contract specifies the scope of services (e.g., number of servers, number of users), the locations from which services will be provided (customer site, provider located), and the measures of quality of service (i.e., service level targets). The service delivery provider responds by assigning each contracted service to a delivery location and assigns a team of agents to respond to the customers' requests. These teams of agents are typically, (though not necessarily) shared across multiple customers. The agents typically are differentiated with respect to their depth and breadth of skills, where the breadth of skill refers to the range of IT areas the agent can support and the depth of skills refers to the level of knowledge mastered by the agent in each of these IT areas. Agents are grouped into *service delivery units* where all agents in a service delivery unit have common breadth and depth of skills.

We now provide a more detailed description of the workload management process after customers contract for service and once customer requests begin to arrive to the provider. Figure 1 illustrates the process and operational flow. Requests are routed to a *service functional unit* at a service delivery location. A service functional unit is a group of service delivery units with similar IT knowledge, i.e., common breadth of skills. A dispatcher is assigned to the service functional unit whose role is to review all incoming requests assigned and determine the priority and the complexity of the requests. Priority of the requests will determine the order in which the requests will be serviced; complexity assigned to the request will determine which service delivery unit is capable of handling the request.

In addition to the scope of service requests that the provider will service for the customer, customer contracts specify service levels associated with each of these service requests. Service levels are a measure of quality of service delivery. Although many types of service level agreements exist, the most common service level agreements specify the following main terms regarding response to a service request: (i) scope of agreement, (ii) target time, (iii) percentage attainment, and (iv) time frame over which service will be measured. For example, a service level agreement may state that 95% (percentage attainment) of all severity 1 tickets (scope) that are opened over each one month period (time frame) must be resolved within 3 hours (target time). One will typically find a large number of service level agreements associated with each customer contract.

Customer service requests can be broadly classified into two types: primary requests and project requests. Primary requests are characterized by relatively short service time (typically, minutes or hours) and short target time (typically, hours or days), and in most cases require a single agent to complete the request. In addition, these requests typically involve a single activity in order to complete the request. Examples of primary requests include incident tickets, change requests, and maintenance work. In contrast, project requests are characterized by requests that are composed of a sequence of tasks and may require the coordination of a number of different service delivery units where different units are responsible for different tasks in the overall project request. There may be dependency relationships between the different tasks. Tasks within a project often take weeks or months to complete. In many cases, the agents who service the project workload are different from those who service the primary workload. This is due to the different skills required. (In some cases, the agents handling project requests are separated into different service delivery units due to the differences in cadence and arrival processes as compared with the primary request workload as well as for ease in management that is introduced by separating these two types of workload.)

Due to the varied nature of project requests, in this paper the metrics and analysis that focuses on analysis of requests is limited to primary request types. A more extensive and focuses analysis of project requests would be required in order to provide meaningful insights into that workload. However, the metrics that are focuses on pool performance or agent cross skilling consider all activities performed by the agents, as will be described later in this paper.

III. Data Pre-analysis

We now describe the data that was used for the purposes of the analysis described in this paper. We consider a number of categories of data that are collected by a services provider that can be used to provide business and managerial insight into performance of the service delivery units and to improve quality and reduce cost of service delivery. *Workload data* includes records of all requests that arrive to the service delivery unit and details including the date and time of request

creation, the request type, the customer name, and request description. *Customer attribute data* provides information regarding the attributes of the customers supported by the service delivery unit including the menu of requests supported for the customer, customer business hours, and customer-specific service level agreements. *Service delivery unit attribute data* provides information about the service delivery unit including the customers and menu of request categories supported by the service delivery unit. It also provides information regarding the working hours in the service delivery unit and the shift schedules against which agents may be assigned. *Agent data* provides information regarding the skills of each of the different agents in the service delivery unit including the range of different requests to which the agent can respond as well as the level of skill the agent possesses. The level of an agent's skill will dictate the level of complexity of a request to which an agent may be assigned. The agent data also contains information regarding the shift to which each agent is assigned. Finally, *activity data* contains detailed information regarding each of the activities performed by each agent on a daily basis as well as the start and stop time of each of these activities. This activity data can be used to measure time allocated to different types of requests and different complexities of requests.

In the process of our analysis of the data we found that data preparation and data cleansing was required for most of the data sets. We now provide a description of some of this data preparation. Note that we analyze the data in hourly intervals. This provides a reasonable granularity for data display and issue identification, while the work assignment and processing process is still occurring within an hour, finer granularity is available for inspecting the spikes.

Although the workload data is typically collected using automated systems (i.e., most fields in the request creation system are automatically populated once the customer creates a request; the primary field that is manually populated is a description of the problem), one key problem that has been identified is misalignment of timezones between the different data sources. As an example, consider a service delivery unit in India servicing a customer located on the east coast of the United States. The request system, which is shared across service delivery units in multiple geographies, converts the request arrival timestamp to GMT. The shift data provided by the teams is provided in local service delivery unit IST. All agents in this service delivery unit work a common shift: 17.5h–2h five days per week. Figure 2 plots the volume of requests per hour (Monday through Friday) as compared with the number of agents scheduled to work daily. One observes the misalignment between the agents working hours as compared with the request arrival patterns. However, discussions with the lead of this service delivery unit indicated that this misaligned was due to misalignment in the timezones used to record data rather than misalignment in the agent working hours. Figure 3 displays the same data with the request data corrected so that both the agent shift hours and the request data are both in a common time zone. One now observes that

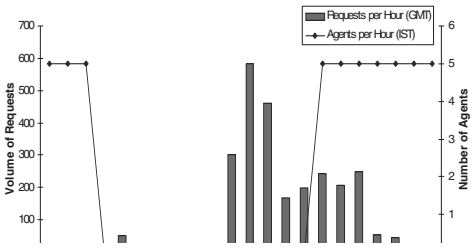

Fig. ligned
Tim

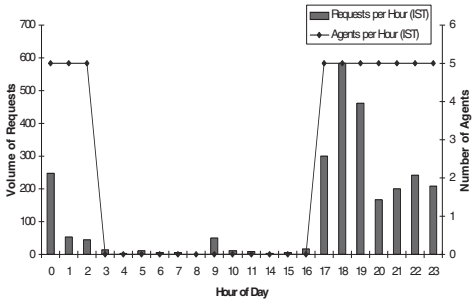

Fig. 3. Comparison of Request Data and Agent Shift Data after Time Zone Corrections.

the agent shifts (working hours) are aligned with the hours when workload arrives.

We also analyze the workload data to observe statistically significant shifts in volumes of requests over time. We explore these shifts to determine their cause. In some cases, volumes of workload may be shifting due to changes in scope in the services that are supported for the account. For example, the customer increases the number of servers in scope or changes the number of services that the provider is supporting under the terms of the contract. In other cases, the scope of services remains constant but the customer behavior changes over time. For example, the volume of requests for an online retail customer may increase during popular shopping periods such as (US) Labor Day shopping period, end of year Christmas season, and other popular shopping times. Volume of requests for financial customers may increase at end of quarter or end of year when financial reports are due. Such patterns observed in request volume are important as these patterns may indicate a need for the provider to modify staffing levels or modify the hours in which agents support the account. In other cases, unusual patterns or changing patterns in volume of requests are not reflective of customer behavior. Rather, these patterns may reflect the manner in which customer requests are recorded or may indicate (provider) system generated requests that do not require action.

We provide, as an example, Figure 4. This figure plots total hourly arriving request volume for one customer. For this service delivery unit, each request that arrives to the service delivery unit is recorded by the agent handling the request. In

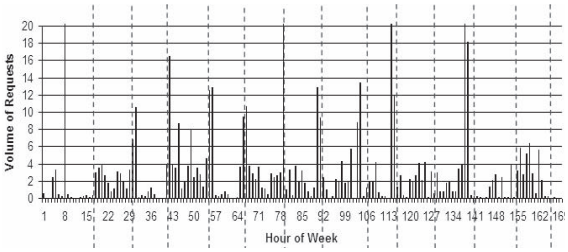

Fig. 4. Volume of Requests per Hour Reflecting Spikes in Request Volume.

Figure 4, the x-axis represents the hours in the week, where hour 0 represents midnight on Sunday night, hour 6 represents 6am on Monday morning, etc. The y-axis represents the number of requests arriving to the service delivery unit each hour. The plot indicates in increase of volume of approximately every 12 hours. For example, increases in volume of requests are observed in hours 30, 42, 54. However, further analysis of the composition of the agents in the service delivery unit and shift schedules revealed that this service delivery unit operated under 12 hour shifts. The shift changes have been marked on Figure 4 which vertical dotted lines. We note that the increases in request volumes that appear on the plot correspond to end of shifts. Discussions with the agents in the service delivery unit revealed that agents did not record the requests as they arrived at the service delivery unit, as designed. Rather, agents waited until the end of their shifts to record the majority of the work that they had performed during their shift. Thus, the peaks in workload are reflective of requests that had arrived during the past 12 hours. Smoothing techniques were required in order to properly estimate the request arrival patterns for this service delivery unit.

As another example, we discuss the data cleansing that is required on the activity data that is collected. The agents use the activity data collection system to indicate when they start each activity and then stop the record when they stop performing the activity. In the case that an agent temporarily stops working on a task, the agent ”pauses” the activity in the activity data collection system. However, agents sometimes forget to stop or pause activities in the activity data collection system resulting in observed long handling times in the activity data collection system. In other cases, agents erroneously create records in the activity data collection system and immediately close the records, resulting in records with very short handling times in the activity data collection system. Methods for statistically detecting both of these types of records must be devised and these records must be eliminated so that the distribution of the activity handling times that are estimated based upon the data in the activity data collection system accurately reflects true handling times for the different activities performed by the service delivery unit. We use statistical methods to identify the erroneous records. First, we classify requests into different categories, where requests in each request category (e.g., request type, complexity) are expected to follow the same handling time distribution. Next, we identify and eliminate

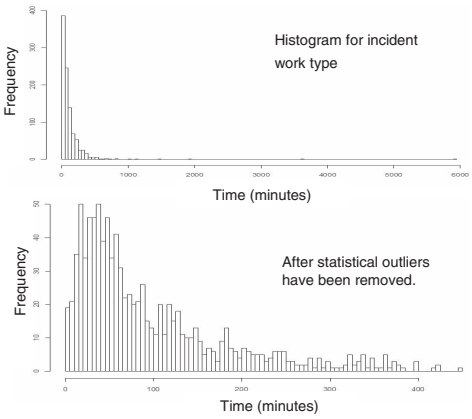

Fig. 5. Histograms of Request Handling Times Before and After Removal of Statistical Outliers.

statistical outliers in each request category. Specifically, we use the box plot approach to consider the quartile information of the collected data. Note that by removing statistical outliers we may also remove the legitimate ones and skew the actual observations. However, this separation helps us to focus on the major statistical behavior of service delivery, and we can always study the removed outliers separately.

Figure 5 provides an example. The upper plot displays a histogram of the data that was collected in the activity data collection system. The x-axis represents the handling time (in minutes) and the y-axis represents the number of requests that required that number of minutes to complete. While the majority of activities in this request class take on average 120 minutes to complete, we observe some requests that take between 1000-6000 minutes to complete (i.e., in excess of one working week). The lower plot in this graph displays a histogram of the handling time after outliers have been eliminated.

As a final example, in some cases it is of interest to study the distribution of activity time expended per day across the team. In this case, one looks at the data recorded across all activities and across the entire time period. However, one sometimes observes errors in the dates for which activities are recorded. In most cases, this is due to the fact that agents are permitted to change the time stamps on recorded activities. (Although the timestamp associated with each activity is automatically populated by the activity data collection system, agents are provided with the option to modify this time stamp. Such option is provided due to cases where, for example, agents forget to record activity at the time that it is performed and therefore must modify the activity start time at the time that they record the activity. As another example, agents may not have access to the activity data collection system at the time that they perform an activity –e.g., if a high priority activity is “paged out” to an agent. – The agent must modify the activity start time at the time that he records the activity in the activity data collection system. In the case where there are errors in the dates, creating distributions of effort expended based upon

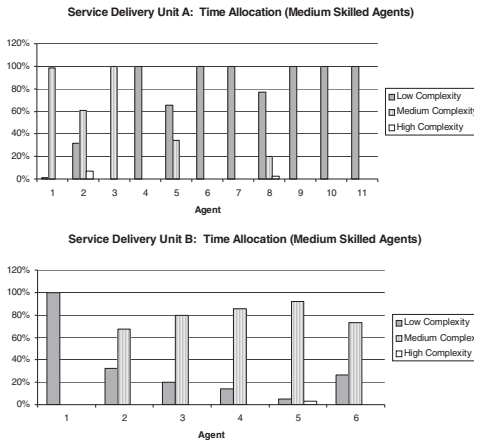

Fig. 6. Percentage of Time Medium Skill Level Agents in Servi Units A and B Allocate to Requests Requiring Different Skill Le

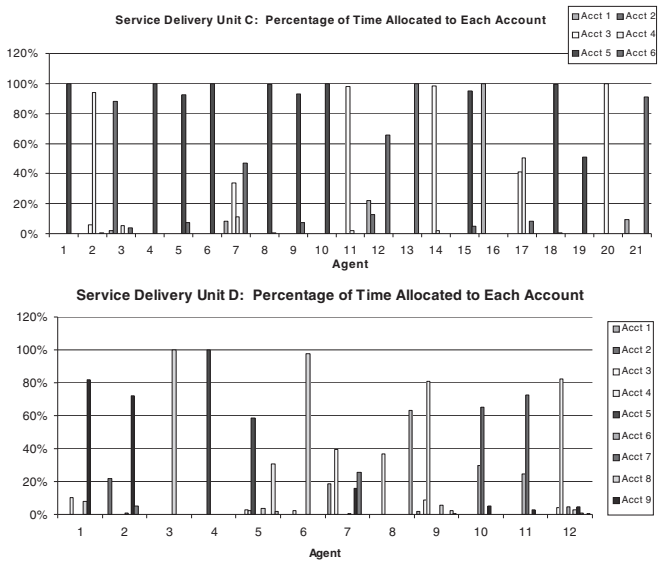

Fig. 7. Percentage of Time Agents in Service Delivery Units C and D Allocate to Different Customer Requests.

the data observed in the activity data collection sy lead to inaccurate estimates. We use statistical m identify outliers. The method in this case must be to the weekend versus weekday patterns so that in where weekend volume is lower than weekday volun is often the case) the method does not falsely detec of the data collection period once a weekend is rea daily activity volume drops).

IV. Data Analysis

In this section we describe data analysis metho service delivery systems. The purpose of the analysi described in this section are to provide insights for ment teams in the service delivery environment. Thes can be used to improve service delivery and to r cost of delivery. For the purposes of this analysis, data from service delivery units at a large services delivery organization. Although we conducted this analysis on a large number of service delivery units, we provide in this section a representative sampling of the analysis.

Staffing is a key cost driver for service delivery organizations. Staff are organized by level of skill, where more highly skilled agents receive higher compensation. Providers thus wish to ensure that more highly skilled agents devote the majority of their time to more complicated tasks, rather than spending time on simpler tasks that could be handled by less skilled (and lower cost) agents. We use the activity data to understand how agents allocate their time to tasks of varying complexities and compare these results with the agent data to understand how agents are utilizing their skills. We provide here a comparison of results for agents of medium skill level in two different service delivery units. Service delivery unit A has 36 agents, of which 11 have medium skill level. Service delivery unit B has 15 agents, of which 6 have medium skill level. Figure 6 contains plots of the percentage of agent time allocated to requests of different skill level.

The x-axis shows the different agents in the service delivery unit. For each agent, we plot the percentage of the agent's time spent on requests of different complexity level. We observe that in service delivery unit A (the upper plot) most of the medium skilled agents spend the majority of their time handling low complexity requests. This indicates that although these agents are higher skilled (and compensated accordingly), they are not spending their time on the more complex requests. The provider may wish to transfer more complex workload to this team. A different pattern is observed in service delivery unit B, where, with the exception of agent 1, agents of medium skill spend the majority of their time handling tasks of medium complexity.

Service delivery units are typically structured such that each unit manages the various types of requests (e.g., incidents, changes) for a number of customers. The provider typically wishes to ensure that the agents are sufficiently cross trained across the various types of requests for the different customers. This reduces risk for the provider and ensures continuity of service in the case that agents are absent or in the case of attrition. We use the workload data to observe how agents allocate their time to perform activities for the different request types and accounts supported by the service delivery unit. We provide here a comparison of results for two different service delivery units. Service delivery unit C has 21 agents and supports six accounts. Service delivery unit D has 12 agents and supports 9 accounts. Figure 7 contains plots of the percentage of agent time allocated to the different accounts.

We observe that for service delivery unit C (the upper plot) the majority of the agents spend most of their time supporting a single account whereas for service delivery unit D (the lower plot) the majority of the agents (with the exception of agents 3, 4, and 6) are cross-skilled across multiple accounts. However, this difference in cross-skilling across accounts may also be due to the difference in ratio of number of agents versus number of accounts supported that is observed between service

TABLE I

ACTIVITY TIMES FOR DIFFERENT REQUEST TYPES AND COMPLEXITIES.

Problem	Low Complexity	Medium Complexity	High Complexity
Average	52.34	80.50	115.24
Median	47.61	58.04	81.00
Std. Dev.	40.06	99.99	99.49
Change	Low Complexity	Medium Complexity	High Complexity
Average	65.98	106.03	109.36
Median	45.84	88.50	101.56
Std. Dev.	65.40	74.56	68.12
Maintenance	Low Complexity	Medium Complexity	High Complexity
Average	63.75	74.01	171.28
Median	47.50	62.07	126.00
Std. Dev.	60.79	78.22	172.63
Service Request	Low Complexity	Medium Complexity	High Complexity
Average	92.37	91.33	119.67
Median	59.00	83.28	99.94
Std. Dev.	127.42	52.36	84.75

delivery units C and D. The large number of agents in service delivery unit C as compared with the small number of accounts supported affords the management with the opportunity to allow agents to gain deep customer knowledge whereas service delivery unit D supports a large number of customers relative to the number of agents. Agents are therefore required to be cross-skilled and have limited time to invest in acquiring deep customer knowledge as compared with service delivery unit C. A provider must consider these tradeoffs, considering as well the relative complexities of the workload and the demands and the expectations of the customer.

Next, we consider the average activity times for different request types. Such analysis is of interest as it can impact the pricing scheme adopted by a provider. For this purpose, we collect data for 42 accounts and analyze the activity times by request type and complexity of request. Table I provides the mean, standard deviation, and median of the activity times by request type and complexity of request. (Data is modified to retain confidentiality of the data but the nature of the results is retained.)

We observe the following: (i) the expected activity time increases with the complexity of the request. This indicates that the provider must consider the distribution of the complexity of the requests when pricing the contracts. (ii) There is significant variation in the request activity times across the different accounts, even within a single request class (request type, complexity combination). This indicates that a provider may wish to consider account specific attributes that lead to this variation when pricing contracts. For existing accounts, the provider may explore the source of this variation to identify best practices and eliminate poor behaviors. (iii) Activity times vary by request class. When pricing contracts a provider will wish to consider the distribution of workload across the different request classes as this will impact agent productivity.

We now describe how workload data may be combined with

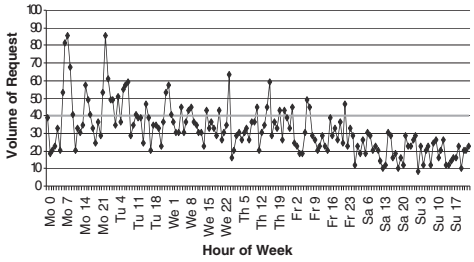

Fig. 8. Volume of Request.

activity data to estimate the total load for a service delivery unit. We refer the reader to the significant volume of literature including, for example, Gans et al. [7], Taylor [14], and Bassamboo and Zeevi [15] for discussions on estimating arrival rates from data. Service delivery systems can be distinguished based upon whether arrival rates are time varying or stationary. A varying arrival rate may be accomodated by fragmenting time into small intervals over which a stationary arrival rate is applied. The historical data can be used to derive the arrival rates of the different request types over time. Figure 8 plots the volume of requests per hour (for each hour of the week) for a single request class arriving to a service delivery unit. The x-axis marks the hours of the week, where, for example, "Mo 0" corresponds to the midnight hour between Sunday and Monday. The horizontal line on the chart indicates the average hourly volume. One observes that the average volume of arriving requests decreases over the weekend.

V. CONCLUSIONS AND FUTURE WORK

The services delivery business is highly dynamic and highly competitive, with thin profit margins. Strict service quality targets coupled with highly variable service request arrival patterns and ever increasing cost containment targets make it challenging for a service delivery provider to deliver consistent quality and remain profitable. Due to various Lean initiatives, there is little room for a provider to pilot alternative solutions whose resulting improvements in system performance are uncertain. In this paper we describe methods for analyzing data collected by a service delivery provider to improve efficiency in a services delivery organization. Such methods have been used in production environment to identify operation issues and have gained positive feedback that the analysis has provided valuable insights to the service delivery operation.

The initial results of this approach are encouraging, which indicates that there is significant opportunity that a provider could gain by creating an automated data analysis and visualization system to continually collect, analyze, and display this data. Such results, produced on a more frequent basis, could be displayed via a dashboard and reviewed by the service delivery management team. In addition, while some of the data cleaning and data analysis techniques studied in this paper for the primary requests are also applicable to the project requests, other analysis on task dependency, skill requirement, task transition time will be required.

REFERENCES

[1] A. Parasuraman, V. Zeithaml, and L. Berry, "A conceptual model of service quality and its implications for future research," *Journal of Marketing*, vol. 49, pp. 41–50, 1985.

[2] J. Cardoso, A. Sheth, J. Miller, J. Arnold, and K. Kochut, "Quality of service for workflows and web service processes," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 1, pp. 281–308, 2004.

[3] K. Larson, "The role of service level agreements in IT service delivery," *Information Management and Computer Security*, vol. 6, pp. 128–132, 1998.

[4] A. Bose, A. Heching, and S. Sahu, "Elements of system design optimization," *Proceedings of IEEE International Conference on Services Computing*, pp. 197–204, 2008.

[5] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao, "Statistical analysis of a telephone call center: A queueing-science perspective," *Journal of the American Statistical Association*, vol. 100, pp. 36–50, 2005.

[6] A. Chandra, W. Gong, and P. Shenoy, "Dynamic resource allocation for shared data centers using online measurements," in *IWQoS 2003 Proceedings of the 11th International Conference on Quality of Service*, 2003.

[7] N. Gans, G. Koole, and A. Mandelbaum, "Telephone call centers: Tutorial, review, and research prospects," *Management Science*, vol. 5, pp. 79–141, 2003.

[8] Z. Aksin, M. Armony, and V. Mehrotra, "The modern call center: A multi-disciplinary perspective on operations management research," *Production and Operations Management*, vol. 16, pp. 665–688, 2007.

[9] A. Mandelbaum and S. Zeltyn, "Empirical analysis of a call center," 2001.

[10] R. Wallace and W. Whitt, "A staffing algorithm for call centers with skill-based routing," *Manufacturing and Services Operations Managment*, vol. 7, pp. 276–294, 2005.

[11] T. R. Robbins, T. P. Harrison, and D. J. Medeiros, "Partial cross training in call centers with uncertain arrivals and global service level agreements," in *Proceedings of the 2007 Winter Simulation Conference*, M.-H. H. J. S. J. D. T. S. G. Henderson, B. Biller and e. R. R. Barton, Eds. Washington, D.C.: The Society for Computer Simulation International, 2007, pp. 2252–2258.

[12] I. Gurvich and W. Whitt, "Queue-and-idleness-ratio controls in many-server service systems," *Mathematics of Operations Research*, vol. 34, pp. 363–396, 2009.

[13] ——, "Service-level differentiation in many-server service systems: A solution based on fixed-queue-ratio routing," *Operations Research*, vol. 29, pp. 567–588, 2007.

[14] J. W. Taylor, "Density forecasting of intraday call center arrivals using models based on exponential smoothing," *Management Science*, vol. 58, pp. 534–549, 2012.

[15] A. Bassamboo and A. Zeevi, "On a data-driven method for staffing large call centers," *Operations Research*, vol. 57, pp. 714–726, 2009.