integris APPLIED

Data Collection:

The Foundation of Fact-Based Decision Making

By Chris Payne

July 2012

Executive Summary

One key element in the preparation of any request for proposal (RFP) for outsourced services is the collection of data: both service volumes and financial data. This information will provide the foundation of the proposals prepared by the service providers and will be used in the business case that is a key part of the decision to either outsource or retain a service. The accuracy of the information gathered will have a significant impact on the success or failure of the initiative. Data collection errors can lead to a decision to retain a service that could be provided more efficiently through a third-party service provider, or vice versa. Additionally, erroneous data can lead to significant cost overruns after the signing of an outsourcing agreement that affect both the service provider and client. This paper highlights the keys to a successful data collection initiative and provides insight on ways to test and ensure the accuracy and completeness of the data gathered.

Table of Contents

1. Overview of Data Collection Process

The process of data collection requires detailed planning and experienced personnel, as well as project management oversight to ensure progress and resolve issues. The process generally requires a significant effort from a large group of resources representing operations, finance, and administration. One of the keys to success is proper documentation and training of the specifics related to the data that needs to be collected. The process tends to be cyclical, with an initial data collection surge followed by interrogation of the data and refinement of the information to reach a point where the various data elements fairly represent the environment that is in scope of the RFP.

The financial data is collected by either taking an extract from a general ledger system(s) or by utilizing budget information. The data often needs to be reorganized to align with how the market provides this type of service (i.e., organized by functional service tower, such as server, network, or applications) rather than by business units or cost centers. Additionally, in some cases, costs related to a portion of work that is intended to be in-scope are buried in a cost center that is out-of-scope (referred to as shadow support); or, in reverse, certain functions are determined to be out-of-scope and the cost needs to be excluded from the financial data.

The collection of service volume data can be more complicated than the collection of financial data. While in many organizations the financial data is in a centralized system, the volume data often needs to be collected from a variety of sources. In some organizations, the information technology (IT) data is maintained in a perfectly up-to-date configuration management database (CMDB). However, most organizations struggle to keep their asset inventory information current and require inputs from multiple subject matter experts by discipline to complete the effort.

For physical inventory counts, the most accurate way to collect asset data is to perform a wall-to-wall inventory; however, this can be both expensive and time consuming and is usually not practical. In circumstances where a wall-to-wall inventory is not conducted, it is necessary to leverage the available data and supplement it with discrete physical inventories. In addition to asset data, certain consumption volumes will need to be collected (e.g., mainframe CPU hours, pages printed, etc.). The volume data that is available in most organizations is a “best efforts” assembly of data from various sources and pulled together by people who often don’t fully understand the purpose of the collection. It is important to treat this information carefully and test the accuracy of the data to ensure an accurate portrayal of your environment.

2. Timing is Everything

A common challenge with a data collection effort that involves consolidation from multiple sources is the timing of the data. The three primary timing-related challenges are:

1. **Alignment of time periods between financial and service volume data**

   One of the dangers during RFP construction is that financial data and service volume data do not represent the same period of time. For example, if the financial data is based on a budget forecast for a future period and the service volumes are based on an asset inventory performed in the previous year, it is easy to get a misrepresentation of the true costs. The problem is exacerbated when an organization is rapidly growing (or shrinking). Additionally, data from a merger, acquisition, or divestiture may be included in (or excluded from) the financials but not the volumes, or vice versa.

2. **Annualizing partial periods**

   Another timing issue related to financials and consumption data arises when using less than a full year of data. In such cases, it is necessary to annualize a period less than a full year (e.g., most recent quarter, last six months). The issue with annualizing partial periods is that the data may have seasonal service or cost fluctuations that would be inappropriate to extrapolate to the entire year. It is important to identify these elements and normalize the annualization process.

3. **Matching spend with life of service elements**

   Depending on the source of financial data, it is important to align the period of time the financials represent (generally a year) with the useful life of the underlying components. This is a concern primarily with hardware, software, and maintenance. As an example, in some cases, either a software maintenance contract or hardware purchase has a useful life of several years but the cost is reflected entirely in a single period. As a result, in a multi-year financial model that is both provided to the service providers as part of the RFP and used internally for business case purposes, the spend may be misrepresented as annual amounts when actually they are incurred every three or four years.

It is important to note that these are all common timing situations, and one or more of them may occur on any given project. However, what differentiates quality data from poor data is a normalization process of the data. As an example, if there is a five-year Oracle maintenance contract purchased upfront on a one-time basis and the financial model is accrual-based, the normalization adjustment should back out 80% of the cost. Other timing-related normalization adjustments that should be considered are:

- If using next year's forecast and last year's asset inventory, what are the growth assumptions included in the forecast?
- What is the historical growth rate for service volumes that supports what percent of financial growth?
- Do all (or most, or none) of the software renewals occur in the same quarter used for the financial base case?

3. Align Data and Definitions

Another challenge with ensuring accuracy of the data collection process and alignment with representations in the RFP is consistency between the process used to collect the data and the definitions of service volumes in the RFP (and ultimately contract). Proper alignment between the volume counts and the definitions requires training of both the client data collection team and the service provider pursuit teams.

Imagine a theoretical IT infrastructure RFP, with a simple resource unit (RU) structure as follows:

- Mainframe
  - CPU Hour: Count of computing hours on the mainframe, normalized to a single platform, excludes service provider overhead, tools, etc.
- Server
  - Server Instance: Count of instances running client applications, excludes service provider server infrastructure for patch management, backup, monitoring, DNS, DHCP, etc.
- Server Storage
  - Allocated Gigabyte (GB): Count of GB allocated to billable server instances

- LAN Services
  - Active LAN Ports: LAN ports turned on and authorized for use in the environment

In this simplified example, the mainframe CPU hours and the active LAN ports are consumption RUs, while the server instances and storage GBs are based on asset counts. All of these RUs will have portions that count as billable and portions that are excluded as non-billable (generally because they are overhead and considered service provider-controlled).

The definition of billable mainframe CPU hours generally includes end-user application time, database time, and time sharing option (TSO) time but will exclude CPU time related to job scheduling (JES), capacity management (SMF), and other service provider overhead functions. As part of the RFP data collection process, it is important to develop the list of billable and non-billable applications, and then properly exclude the non-billable applications' CPU time. It is important to align the inclusions and exclusions with the definition in the RFP documents. Additionally, the hours of the various mainframe engines in the environment will need to be normalized to a standard platform. It's important to be consistent and document the assumptions in the RFP, as misinterpretation of this conversion to a standard platform can have a meaningful impact on the cost of the solution.

For the server instance RU, it is important to make the distinction between physical servers and virtual instances (generally the billable RU is virtual instances). Often excluded from this RU are:

- Host instances for ESXs
- Load balancers
- VPN terminators
- Appliances
- DNS
- DHCP
- WINS
- Enterprise backup
- Monitoring
- Software distribution

These are just examples. It is necessary to spend time with the technical personnel who are performing data collection and communicate the importance of distinguishing the primary function of the instance and how it impacts the solution and price of the proposal. It also pays to think through conflicts, for example:

- How will an application server that also runs DNS be treated?
- If software distribution servers are non-billable (as they are service provider overhead) and server computing is in scope but desktop computing is out of scope, how will servers that distribute software to the desktop be treated?
- If categorizing instances by Intel and UNIX, how is Novell or Linux or AS/400 categorized?

It is important to communicate these distinctions to all of the people doing the data collection and document them in the RFP language.

Storage is generally measured and billed on the basis of storage gigabytes (GBs) allocated to billable server instance. These GBs are measured after the discs have been formatted, mirrored via RAID, carved into LUNDs, and then allocated to the end-user instances. There are usually exclusions for any storage used for the operating system and other system files. Because storage volumes depend on identification of billable server instances, completion of the storage data collection effort will follow completion of service instance

counts. The storage data collection team often struggles with determining the portion of storage that has been both allocated to a virtual server and is actually used by the billable instances on that server. Issues like this require the storage team to work closely with the server team to make an accurate determination. The storage team may also have difficulty precisely identifying exclusions, and it may be necessary to develop rules of thumb to make estimates. When estimates or assumptions are required, consider documenting these in the RFP.

LAN services are generally billed using active LAN ports as the RU. In the RFP documentation, active LAN ports are typically defined as ports in use by some device. In many cases, the client LAN team will consider "active LAN ports" those that are available on the switch. It is important to work with the technical team to communicate this definition of billable LAN ports to ensure the proper volumes are collected.

If the pricing structure includes a pool of hours or full-time equivalents (FTEs), it is important to align the way this data is collected with the RFP definitions. In most cases, time for travel, training, vacation, and administrative tasks is excluded as non-billable. Additionally, an FTE is defined as a fixed number of hours per period. For example, one FTE may equal 1,920 billable hours per year. Overtime hours may or may not be billable against the pool, depending upon the exclusions in the RFP definition. It's important to consider all of these factors and assess how much productive work will occur under a pool mechanism as compared with how much work is performed by internal staff. A normalization factor in the business case may be required to perform a like-for-like comparison (e.g., a pool of 50 FTEs may not be equivalent to an in-house staff of 50).

Close cooperation between the financial and technical teams as well as training for everyone involved are the keys to avoid the pitfalls of definition misalignment. If the RU definition makes it very difficult to collect the correct volumes, consider modifying the definition to better align with the data at hand. A small compromise in a definition may be better than erroneous data supporting the optimal definition. An important consideration when contemplating a change in definition is that any deviation from industry norms in RU definitions will make benchmarking more difficult.

4. Scope Impacts Data

As part of the process of creating an RFP, it will be necessary to make decisions regarding scope. The cleanest scope decisions are fully inclusive or exclusive by function. There are also real economies of scale that favor putting as much volume in scope as possible. However, in many environments, there are significant political or operational constraints that make it difficult to source complete functions. As a result, complicated scope boundaries make data collection difficult. In any event, the objective is to create a clear demarcation of what is in and out of scope. A determination by physical site, for example, makes this distinction much easier than determining that all of the applications owned by individual business units are out of scope.

In some cases, a group that performs two separate functions will be partially in scope. One common example is LAN support and desktop computing support. To source these functions separately, it will be necessary to determine the percentage of time spent supporting each function and separate the associated cost of each function. The functions may be so intertwined that it is difficult to source them as separate functional areas. For example, if a firm has many small sites and each site has one person supporting LAN and desktop computing, it may be inefficient to outsource LAN support as those individuals will still be necessary for desktop support.

5. Trust, but Verify

Once the initial data has been collected, it is important to perform several rounds of verification to ensure accuracy and completeness. There are two very common consistency errors in large data gathering efforts: internal redundancy within datasets and volume inconsistency between datasets.

When testing a list of data to determine if it is internally consistent, the first question should be, "What is the unique identifier of an element of this list?" Examples of unique identifiers are:

- For servers: serial number, asset tag, or server name
- For equipment leases: serial number of the associated equipment, or a lease number
- For voice services: phone number
- For WAN locations: circuit IDs or addresses

Use Excel or an appropriate database tool to test if these values are unique or duplicates across the data set.

A dataset with multiple "unique identifiers" could result from any number of reasons. If many people in a large environment were responsible for the collection of data, it is possible that overlap occurred and the same data was reported more than once. If the data is based on a database query, it is possible to create a query with insufficiently specific joins so that it creates duplicate records. In most cases, this problem manifests itself as a very large number of duplicates or triplicates.

It is also possible that there is a problem with the underlying data source. If each new row of a server inventory is populated by copying an old row and updating certain fields, many unique servers may end up with the same serial number. Any good application will have integrity checking to prevent this from happening, but, an internally developed tool or Excel sheet might simply rely on process to prevent this problem, which may not be sufficient.

Finally, the problem may not be with either the collection or the population of the data but the process by which the data is maintained. For example, a list of servers might have repeats of server serial numbers because each time a memory expansion is purchased for a new server, the server serial number field is populated with the serial number of the server in which the memory is to be installed.

In addition to checking individual datasets for problems, it is critical to check related datasets against one another to find gaps. For example:

- Do servers, LAN switches, network circuits, and voice stations exist at all sites?
- Do LAN port counts align with user and device volumes?
- Do phone station counts align with user counts?

As part of the initial data gathering process, it can be extremely helpful to standardize on a physical site identification number or name, which can then be applied to each physical asset record during the collection process. This way, volumes can be compared at a site level to identify obvious gaps. Additionally, even if persons are not in scope for the RFP, it may be worthwhile to gather employee and contractor counts to compare them with user-based service volumes.

In all cases, discoveries of this type of discrepancy lead to further interrogation. Perhaps what looks like a roll-up problem is just a quirk of a particular organization. For example, individuals in certain departments may have two phone lines for business purposes. Accuracy requires communication with the people who assembled the data and communication with the people who know the environment well.

6. Summary

Data collection is a cyclical process that evolves over time and requires increasing levels of attention to detail to refine the end result. It is key to ask questions throughout the process. At all steps of the cycle it is important to step back and ask, "Does this make sense?" It is very easy to get caught up in the detail of consolidating the data and assume that someone else has verified its accuracy.

The most important priority to performing a high-quality RFP data collection is to ensure that the data is aligned.

- The financials must be aligned to the volume data.
- The time periods must be aligned.
- The scope of these two data collections must be aligned with the RFP documents.
- The data collection methodologies must be aligned with the RU definitions.

In order to achieve this level of alignment, it is important to have a good training and communications plan. Large data collection initiatives involve many people, and there is a tendency to consider the effort as a distraction and "not part of my day job". But considering the size of most outsourcing transactions and the financial impact associated with even small errors in data collection, it makes good business sense to spend the time required to ensure a quality result.

Document Revision 2013-10-21: Update to company name and contact info.

Les Druitt
Founding Principal
+1 281 705 4895
les.druitt@integrisapplied.com