Accounting and e-Infrastructures

While some e-infrastructures included accounting in their design and operations from the start, others are now being asked or required to add accounting support to their existing systems. Typically accounting forms part of a relationship between the infrastructure and some other organisation – perhaps a funder, host or customer – rather than the infrastructure's relationship with its individual users. These organisations may be interested in usage statistics across particular categories: for example by subject, by time, by project or by origin. It might be assumed that infrastructures already have enough data to generate these statistics retrospectively, as a result of their authentication, authorisation and security activities. However a closer examination indicates that these may not, in fact, be sufficient and that specific data, processes and agreements may be needed to support reliable accounting.

This paper uses accounting infrastructures' experience to examine some of the situations where accounting may be required and the extent to which existing data may, or may not, support it. It then identifies some of the new issues that accounting requirements are likely to raise, and suggests questions for infrastructures that do not currently provide accounting to consider either when, or before, they are asked about it.

What is accounting?

In everyday life, the concepts of "accounting" and "billing" often go together. Typically we want to know what we paid, and what we got in exchange. Even if e-infrastructure use doesn't involve a transfer of money, the purpose of comparing expenditure (whether of money, time, effort or some other measure) against the resulting benefit remains the same. As the examples below indicate, this calculation will usually be done at the level of an organisation, a project or a discipline, rather than that of individual users; an accounting report will usually consist of summary statistics rather than detailed records. For some purposes, those reading the accounts may also want the ability to drill down and identify which of their groups or users incurred a cost or received a benefit; providing such customised access is likely to require significantly more effort from the infrastructure provider.

The most obvious requirement for accounting is where use of an infrastructure is chargeable: typically an organisation, project or funding body will pay the bill on behalf of its individual users. This obviously applies to pay-per-use commercial infrastructures such as cloud services, where users buy resources – such as CPU, storage or data transfers – against an account or credit limit. Very similar accounting statistics may also be required where a funding grant includes a quota of infrastructure resource that a project or other group can spend on its activities. But even where payment is based on a subscription, rather than specific use, accounting may be needed to resolve any refunds or credits that may arise, for example as a result of failure to meet a Service Level Agreement (SLA).

More generally, funders and supporting organisations may want to compare the results of their investment against targets: did an investment result in the expected quantity and quality of new research, were opportunities taken up by the intended subject domains or organisations, etc. Organisations not directly involved in infrastructure provision or use may also be interested in accounting statistics: universities may want to publicise how much use their researchers are making of a national facility, for example. Organisations that are hosting infrastructure components, or otherwise supporting their development or operation, may want reassurance that their contributions are producing the hoped-for benefits.

Finally, some collaborative infrastructures require that all participants contribute and use their “fair share” of the pooled resources. Here accounting statistics might be a way to demonstrate compliance with this policy or to rebalance future contributions based on actual use.

Existing records

E-infrastructures already generate significant volumes of logs as part of their routine operation. Most infrastructures will have records relating to authentication and authorisation, as well as information needed to detect and investigate security problems. Some will also keep logs of the performance of the infrastructure itself, to support performance tuning and inform development plans.

Authentication and authorisation logs focus on login accounts on the infrastructure, recording at least the time when each account began to be active and possibly also the time when activity ended. Some may also record what processes or commands were run during each authenticated session. However, depending on how the infrastructure is configured and used, an account may refer to a single user, or to a project or tool. These logs have two main uses: to allow detection and investigation of faults and misuse and, where necessary, to check usage quotas. For misuse investigations they are likely to be retained for several months; usage quotas may require them to be kept for the duration of each project.

Security logs concentrate on the operating system components that are most often the target of security incidents: network connections, application programs, etc. Network security logs include information such as when connections were established, between which addresses and ports, and the volume of data transmitted. Logs from applications such as web servers may record the requests made to the application and whether they were successful. On systems handling particularly sensitive data, security logs may also include records of which accounts accessed particular files. The main purpose of these logs is to allow detection and investigation of attacks on the infrastructure, though they may also be relevant to investigations of misuse. Typically they will be kept for several months.

Performance logs relate to computer hardware components: CPU, memory and disk. In the short term such logs may be used for system tuning, to extract as much performance as possible from existing hardware. Longer term, performance logs may be used to identify failing components, to plan which components need to be upgraded, or to inform the design of future systems. The retention period for these logs will depend on the purpose: for longer-term planning they may be summarised to reduce the volume of information that needs to be stored.

However these records of login, operating system and hardware activity are almost certainly not all that will be needed to answer accounting questions.

New data for accounting

Though it is tempting to think that accounting statistics could be reconstructed from existing authentication, security and performance logs, attempts to do so are likely to reveal gaps that make this inaccurate or impossible. Logs may not contain the values required to calculate the accounting statistics, or may lack the metadata required to group data into the required accounting categories, or there may be no process or tool available, or reconciling data or releasing the results may require a new agreement between parties or breach an existing one.

The most common lack is likely to be metadata. Most accounting requirements involve associating data with categories for which statistics are derived (for example what resources were used by each subject discipline). Unless those categories happen to be required for authentication, security or performance purposes then they are unlikely to be captured by the infrastructure, in which case it will be impossible to assign the data to the categories required for the accounting. A performance log may show that 80% of CPU was consumed by a particular program, but working out what subject area that corresponds to may be guesswork. Where there is a simple and constant mapping between the objects referenced in the logfiles – accounts, programs, hardware – and the accounting categories then it may be possible to reconstruct the missing metadata. However such mappings are often both multi-valued and dynamic: individuals may well use the same account to work on more than one project or discipline, and there is no guarantee that their institutional affiliation when accounting statistics are calculated is the same as when they used the infrastructure.

The approach used by projects to manage access to e-infrastructures can create further difficulties for retrospective accounting. Mapping all project users to a single local username may make it easier to clear up disk space at the end of the project, but is likely to make it impossible to reconstruct which institution's users were responsible for use of CPU time. Many Principal Investigators use a group management system to allocate resources between different users and sub-projects, in which case the accounting process is likely to need access to that system's logs (if they exist) to assign the project's infrastructure activity to users or institutions. Some projects – particularly those whose main purpose is to provide data or services for others – have an open definition of "membership" that makes it hard to associate users and their activities with any specific project or accounting category.

Given the difficulties of reconstructing accounting statistics after the event, it is preferable to assign the appropriate categories to each activity at the time it takes place. This might be done by the individual user or, in some cases, by the Principal Investigator when allocating project resources. In a few cases, the infrastructure may be able to record the necessary metadata as part of the authorisation decision it makes to grant the individual access. The desired result is a log of raw accounting metadata listing the relevant resources used against the categories they will need to be reported against. Getting the right information in this file requires, however, that the resources and categories be known at the time of recording: in other words the questions to be answered by accounting need to be known in advance.

New tools for accounting

Some accounting requirements may involve the provision or development of new reporting tools, not otherwise needed for the e-infrastructure's operation.

Where a project has been allocated a certain quantity of resource to "spend", it is reasonable for project managers to want regular or real-time status reports of how much of their allowance has been used. This is particularly important if resources need to be used within a particular timeframe, where there may well be a need to check actual progress against that planned.

Where billing or accounting are based on cumulative use of resources such as disk space (e.g. "ten cents per Gigabyte-month") both new data and new tools may be required. Normal system commands typically only report the current usage of a resource: for a cumulative account these commands need to be run regularly to capture snapshots that can then be used to calculate the total charge. Again, users and managers are likely to want regular or on-demand "statements" of their current charging level.

Similar issues arise if charging rates vary over time, for example where an infrastructure offers "peak" and "off-peak" rates, or where charges for different resources can be combined or offset within a particular time window. If a job may run on different classes of CPU or storage, or if elapsed time is an accounting factor, then these ephemeral facts, too, must be captured and maintained as metadata to permit the calculation of the final account.

Finally, some accounting may require combining data from different sources, even from systems under different management. This is likely to require preparation to ensure that the different datasets are, and remain, compatible; agreements on retention, access to and use of data; and tools to do the required combination, calculation and reporting.

Things to consider

Although it may be possible to answer some accounting questions, to some level of accuracy, using data that are already collected for other purposes, the examples and discussion above suggest a number of questions that should be considered by any infrastructure that wants to have confidence in its ability to provide reliable accounting statistics.

What reports do you want to support?

Some accounting questions will be impossible to answer unless relevant information is captured at the time the infrastructure is used. This is most obviously the case where accounting uses categories that are not required, and therefore not collected, by the infrastructure's normal operating procedures. For other accounting questions, information may exist but in incompatible formats or subject to agreements that hinder its use.

Infrastructures should therefore review the likely sources of requests for accounting – notably funders, host organisations, user organisations, and projects – and try to identify the kinds of request they may make. An informed decision can then be made which of these to support, what granularity of reporting will be provided, and whether standard or customised reports will be offered.

What (extra) data/processes/agreements do you need?

With a list of the desired reports, the infrastructure can review what data and processes would be needed to generate them. If data do not currently exist then new processes will be required to capture that information; if they exist, but in the hands of some other party, then an agreement on access will be needed. Infrastructures should also check that calculating and disclosing accounting statistics will not breach any existing agreements, for example with funders or users.

This may involve policy choices. For example technology and infrastructure features that are designed to protect users' privacy may conflict with some accounting requirements. In some cases it may be possible to reconcile these using cryptography or trusted third parties, but infrastructures may have to choose between two incompatible requirements.

Where existing logfiles are to be used as an input to an accounting process, decisions are likely to be needed on what formats should be used. Accounting will be simpler if all logfiles use a common format, but changing these at source may break other systems. Alternatively, distributed infrastructures in particular may need to accept different logfile formats from different sources and parse them into a common database from which the accounting reports can be generated.

How accurate do reports need to be?

Perfect accounting may involve considerable effort, whether from infrastructure users required to provide additional information, principal investigators required to implement new processes, or operators required to develop new systems and respond to requests for customisation. If an accounting requirement appears likely to involve this level of disruption, it may be worth considering whether some approximation (for example using partial, estimated or sampled data) might provide a better balance between cost and benefit. The required accuracy is likely to depend on what the ultimate purpose of accounting is, and who (if anyone) suffers if it is less than 100% accurate.

A related question is whether accounting statistics are likely to be challenged, and how robust evidence needs to be to defend such a challenge. While it might be possible to produce accounting that would stand up in court, it may be cheaper to compromise on an occasional disputed "bill" and deal with any abuses of that privilege at organisational level. In some cases "accurate" may be hard to define anyway: if a user or system error results in a job using a project's entire disk or CPU budget then the "correct" accounting is likely to be a matter of discussion rather than fact. The increasing use of virtualisation may also make it harder to compare accounts against independent sources of truth such as network flow measurements.

Conclusion

Although all e-infrastructures need to provide Authentication and Authorisation, Accounting may have been viewed as optional or unnecessary by some. This discussion indicates that while Accounting is linked to the other two 'A's, and to other aspects of infrastructure operations, it nonetheless deserves its separate status. Treating it as a secondary issue, and assuming it can be addressed when required, may result in disappointment.

E-infrastructures that do not already provide accounting should therefore consider what requirements may arise and how they might address their data, process and policy implications. Knowing what accounting might involve should result in better informed discussions as and when it becomes a requirement.