Account Tidying

Project description

Following on from the our data tidy-up and implementation of prometheus lifecycle, we now have a large number of accounts which are disabled. We need to tidy up these and establish, and automate where possible, the processes for what happens in the future.

In the following text, "downstream accounts" refers to entries in the KDC, AFS-PTS, AFS-VLDB (and home directory data) and "System" LDAP (rfe 2307).

The following steps should be taken:

- Firstly, communicate our account deletion policy to all users.
- Establish, and fix if necessary, what discrepancies we have between prometheus data and downstream accounts.
- Decide what to do with user data which is not in home directories.
- Delete all downstream accounts which predate implementation of lifecycle.
- Ensure we have adequate tools and reporting for dealing with post-grace/post-suspension accounts.
- Decide what data, if any, we keep in prometheus following account deletion.

It's envisaged that this project will be cross-unit, involving infrastructure, user support and services.

Initial thoughts

Fleshing out the original project submission...

We should clarify that this project will apply to user accounts only. The management of host-based KDC principals, and any corresponding AFS identities, will not be explicitly covered.

Similarly, the management of multiple identities (e.g. accounts such as user/cron) is also outwith the scope of this project, but should be borne in mind where relevant.

This project is concerned with the very final stage of an account's lifecycle, specifically after the grace and suspension periods have expired.

Adding detail to the topics in the original project submission:

Firstly, communicate our account deletion policy to all users.

This page explains our account closure policy:

http://computing.help.inf.ed.ac.uk/account-closure

This page explicitly notes our policy on archival:

http://computing.help.inf.ed.ac.uk/account-archival

i.e.

```
First a definition of "Account archival" - the taking of a specific
copy of a users home directory, on some medium, with the intention
that it will last forever and could be restored at any point in the
future.

With this definition of account archival, then the simple fact is
that we do not do it for any computer accounts or general data. We do
take regular backups of accounts for disaster recovery purposes, but
we do not "archive" someone's home directory prior to deletion.

Anyone (staff/students/visitors etc) who is leaving Informatics, is
encouraged to take their own copy of their files before they
leave. Blank DVDs can be supplied if required. If their data is to be
shared/used by someone else after their departure, then it should be
moved to some communal area, eg group file space, or a subversion
repository.
```

This page is also explicit about the same policy for taught student data:

http://computing.help.inf.ed.ac.uk/archiving-taught-student-files

A couple of thoughts re all of the above:

- We should be more explicit about what happens (or, more accurately, what we want to happen) to files that are not in home directories. See below for more thoughts on this.
- Although these pages exist and are linked from http://computing.help.inf.ed.ac.uk/guidelines, which is itself linked from the main computing.help page, we don't know if people read them - they would have to go looking for the information. Perhaps we should send a yearly email with these links?

Establish, and fix if necessary, what discrepancies we have between prometheus data and downstream accounts.

...and...

Delete all downstream accounts which predate implementation of lifecycle.

A reminder that prometheus data for a person is held in an Entities->Identities->Accounts tree structure. A diagram explaining this is here:

Prometheus Entity Model

Put simply, identity objects typically correspond to principals in the KDC and account objects correspond to entries in LDAP, AFS-PTS and AFS-VLDB. Creation of these downstream entries are governed by possession of the appropriate entitlements. See here for more information:

PrometheusDesignDetail#Roles_Entitlements_in_Prometheus

Since the introduction of lifecycle (January 2015), accounts which have subsequently expired should be in a consistent state - disabled (in the KDC) but still with LDAP, AFS entries and with all data untouched. Accounts which expired before that are likely to be in a more inconsistent state with respect to downstream entries. We last did a mass deletion of accounts in 2012.

For accounts which predate lifecycle, there is likely to be quite a disconnection between data in prometheus and existing downstream accounts. In particular, we know that there are a large number (more than 3000) accounts which have AFS-PTS entries only (i.e. no AFS-VLDB, KDC or LDAP) and have no identity objects in prometheus.

We can iterate through all 4 downstream stores (either using prometheus's interface or through more specific tools) to gather data on discrepancies.

To clarify, deleting an account entails deleting the following:

- KDC entry
- LDAP entry
- AFS-PTS entry
- AFS-VLDB entry
- Home directory (and all files contained therein)
- Identity and Account objects from Prometheus (leaving just a top-level Entity object)

Decide what data, if any, we keep in prometheus following account deletion.

We need to decide what we do with data we have for people whose accounts have fully expired (i.e. at the end of their grace period). The top-level Entity object left behind would still have the following:

- username
- name (givenname, surname, fullname)
- contact information (email, telephone, room) as last seen
- home directory path

We should probably delete these Entity objects entirely. We need to carefully consider GDPR for any data retention policy.

If we do, as assumed, delete all data for expired accounts, we lose the ability to know if we ever had an account for that person. This is obvious, but is worth explicitly stating.

Decide what to do with user data which is not in home directories.

For accounts we want to delete completely, we need to consider the following:

- Files owned by that user outwith home directories - e.g. group space, home pages. If there are other places, we should note them explicitly.
- ACLs including these users.

For both of the above cases, there is no way of locating these files/directories other than trawling through file systems. This is likely to be very time consuming, particularly for large volumes, e.g. group space. Locating these may be something we want to do as an infrequent, but regular, job.

We would assume that we would remove old users from ACLs, but what to do with their files needs to be decided.

Ensure we have adequate tools and reporting for dealing with post-grace/post-suspension accounts.

Expanding on this, we need to be able to obtain, at any time, a list of user accounts eligible for deletion. This should be relatively trivial once the initial tidy up has taken place.

The script we have previously used to delete a user is `/afs/inf.ed.ac.uk/user/t/toby/coutils/prometheus/delete-user` (formerly known as 'hacky-delete-user'). This hasn't been used in some years, so should be carefuly checked, brought into the prometheus source and properly distributed.

We should establish procedures for account deletion and document them, e.g.:

- How often do we do it?
- Who is responsible for doing it?
- Should accounts be deleted as soon as they are eligible?

Summary

Considering the issues raised above, it seems there are three distinct parts to this project.

1. User documentation and communication of policies concerning the end stage of an account.
2. A one-off data-tidying and deletion process for Prometheus, downstream stores (KDC, AFS-PTS, AFS-VLDB, LDAP) and filesystems.
3. Establishing procedures, tools and documentation for computing staff to manage account expiry in the future, automated where possible and advisable.

Update 2018-01-31

Email sent...

```
From: Toby Blake
Date: Wed, 31 Jan 2018 11:41:30 +0000
To: Jennifer Oxley <joxley@inf.ed.ac.uk>, Craig Strachan <cms@inf.ed.ac.uk>

As an initial bit of data-gathering for the one-off tidy-up, I've put a
couple of files in /afs/inf.ed.ac.uk/group/cos/account-tidying/ ...

entities-status.2018-01-30 reports on all entities in prometheus and
reports on various aspects.

Initially I want to concentrate on entities which have no identity object
(this corresponds to a kerberos principal).  These are reported as
"no-identity" in the file above.

There are 9836 of these.  The file entities.no-identity.2018-01-30.dss
reports on the downstream store summary of all of these.

e.g. an active account would look like this:

[bolt]toby/admin:tests/downstream-store-summary toby
toby: kdc afs-pts afs-vldb sys-ldap
[bolt]toby/admin:

If you look through the file entities.no-identity.2018-01-30.dss, you
can see there are a large number which have an afs-pts entry only (3214).
These would seem like good candidates for deletion from the AFS pts
database.  However, the thorny issue is what do we do about any files
they may have left behind, or ACL lists they may be on?

Toby
```

Meeting 2018-02-08

cms, joxley, toby

One-off data tidying

We agreed that all prometheus entities with no identities and no downstream store (kdc, afs-pts, afs-vldb, ldap) entries should be deleted completely from prometheus. There are 6616 of these.

We also agreed that all prometheus entities with no identities and only afs-pts downstream store entries should be deleted from both prometheus and afs-pts (but see "File system implications for account deletion" below). There are 3213 of these.

Toby to check with gdmr about what GDPR says about the above, i.e. whether we have to delete entity data (although we would probably still want to anyway, even if not compelled).

Communication of policies

We agreed that a regular email should be sent to sys-announce summarising all policies which affect accounts. This mail should be sent, at a minimum, yearly; possibly every 6 months, early in each semester.

What to delete on account expiry

We have agreed that we will delete the following on expiry of an account (to clarify: expiry of an account refers to when both the grace period and suspension period have passed):

- KDC entry
- LDAP entry
- AFS-PTS entry
- AFS-VLDB entry
- Home directory (and all files contained therein)
- All data held in prometheus (including Entity object)

We should also consider deletion of:

- Home pages

In the longer term we need to consider all aspects of data retention (and punt it up to any GDPR project), e.g.:

- Retention of uun information in mail forwarding

Ultimately decisions on what happens on account expiry are policy matters, so are likely to be undertaken by CEG.

File system implications for account deletion

Deletion of a user from LDAP and the corresponding AFS-PTS entry has implications for data stored on our file systems. We would like to make sure that data is not left unowned in the conventional unix sense (i.e. being owned by a user who no longer exists).

We also want to ensure that the user is removed from all AFS ACLs. We can use the 'fs cleanacl' command for this purpose, but it would involve a complete trawl of our file systems. This is perhaps something which could be done infrequently but regularly - to be taken for consideration by the services unit.

In a more general sense, it would be useful if we could map ownership of parts of our filesystem directly to named users, e.g. for group space. It might be possible to glean information from ACLs at the top level of volumes and store this data somewhere. We can obtain a list of all AFS volumes, however it doesn't seem trivial to find the mountpoint given volume name. fs commands of this kind of syntax may prove to be useful:

```
find /afs/.:mount/inf.ed.ac.uk:user.cms -type d ...
```

Craig to investigate further.

What to do next

We have created a directory `/group/cos/account-tidying/` to gather data for this project. We should be aware that data held there is likely to be GDPR sensitive.

We should begin the process of one-off data tidying. First by deleting entities and PTS entries as identified above. After this, we should identify further areas of inconsistency between downstream data and what is held in prometheus. Once we are in a position where all downstream data matches that held in prometheus, we should move onto creating tools and procedures.

Update 2018-02-14

Following discussion at the ops meeting...

- GDPR and retention of account data - toby - 2018-02-13
  - For expired and deleted accounts in prometheus, we currently keep a small amount of data (username, full name, last contact details). We would like, and GDPR probably compels us, to completely remove this data. Is there any reason for retention?

... we should go ahead with deleting a big batch of entities and afs-pts entries. The data protection issue is being punted to George's GDPR project.

The issue of what to do with any files in our filesystems which are owned by any of these users, or any AFS acls which contain these users, is deferred for careful consideration (by Services).

The accounts in question are:

`/group/cos/account-tidying/entities.nothing`

... which contains uuns for which entities should be deleted

`/group/cos/account-tidying/entities.just-afs-pts`

... which contains uuns for which entities and afs-pts entries should be deleted

Although we should generate this data anew just prior to deletion (see plan below).

To find out the basic status of all entities in prometheus:

`/group/cos/account-tidying/scripts/entitytreestore-all-basic-status`

... this reports the following for each person entity: *if the entity has no identity object
- if the entity has an account granting entitlement (prometheus/localIdentity)
- if the entity has the active-person role
- if the entity has an accountend date
- if the entity has the following flags:
  - disableAccount
  - initialPassword

    - requirePasswordChange

To find out the downstream store status of a user (i.e. whether they have kdc, afs-pts, afs-vldb or ldap entries):

```
/group/cos/account-tidying/scripts/downstream-store-summary
```

To delete entities:

```
/group/cos/account-tidying/scripts/delete-entity-by-uun <uun>
```

To delete from afs-pts:

```
pts delete -nameorid <uun>
```

We should keep a list of all uuns we delete.

We should also dump all entity data before we delete it, just in case we want to retrospectively keep something. I'm not sure what we'd want to keep, but we could stick a dump of all of prometheus's data somewhere safe (and secure) and deal with it later. Or run prometheus-get-info on all uuns to be deleted.

Here's a vague plan:

- run entitytreestore-all-basic-status
- run downstream-store-summary on all entities with no identity
- take dump of data
- for entities with no identity and no downstream stores: delete entities
- for entities with no identity and only afs-pts downstream store: delete entities and afs-pts

Note that any script which has to run on 9000 or so entries can take a long time. It will also take a lot longer if the same script is being run 9000 times with a uun argument.

Update 2018-03-28

We have now deleted the following:

- all entities in prometheus which have no identity object and no downstream store objects
- all entities in prometheus which have no identity object and only an afs-pts downstream store object

Details of these are in `/group/cos/account-tidying/backup`, including a snapshot of prometheus data prior to deletion and `pts examine` output for all deleted afs-pts entries. The list of uuns deleted is recorded in files, as referenced in `/group/cos/account-tidying/backup/README`

Totals:

```
[bolt]toby: wc -l entities.delete.uuns entities.delete.and.afs-pts.uuns
 6613 entities.delete.uuns
 3213 entities.delete.and.afs-pts.uuns
 9826 total
[bolt]toby:
```

In total, we have deleted:

- 9826 prometheus entities
- 3213 afs-pts entries

There were 5 accounts which have no identity object and no downstream store objects which were not deleted. These have `future-*` roles.

Update 2018-03-29

We need to find any other accounts which are inconsistent with prometheus. This means we need to check all objects in the downstream stores which are not in prometheus.

So, the next stage of the project can be summarised as dealing with the following (approximate number of accounts affected in brackets):

- any accounts which do not have a full list of kdc, afs-pts, afs-vldb and sys-ldap downstream objects (48)
- any entries in kdc which are not in prometheus (194)
- any entries in afs-pts which are not in prometheus (724)
- any entries in afs-vldb which are not in prometheus (193)
- any entries in sys-ldap which are not in prometheus (179)

Update 2018-04-16

When we come to automate the selection of accounts to be deleted (i.e. including data) we will need a way of indicating that an account should not be selected. In the interim, we'll use a 'noDeletion' flag.

Update 2018-09-14

A reminder that the following distinct areas were identified for this project:

- User documentation and communication of policies concerning the end stage of an account.
- A one-off data-tidying and deletion process for Prometheus, downstream stores (KDC, AFS-PTS, AFS-VLDB, LDAP) and filesystems.
- Establishing procedures, tools and documentation for computing staff to manage account expiry in the future, automated where possible and advisable.

User documentation and communication of policies concerning the end stage of an account.

The documentation we have is described above. We think this is adequate, but users need to be regularly reminded of our policies. See email to CEG below. The email also raises my more general concerns about deleting accounts, which might be worth discussing at the meeting. This will be the first time we delete home directories without archiving them first. We haven't, as yet, deleted any home directories as part of this project.

Email to CEG:

From: Toby Blake <toby@inf.ed.ac.uk>
To: compexec@inf.ed.ac.uk
Cc: Jennifer Oxley <joxley@inf.ed.ac.uk>
Date: Thu, 5 Apr 2018 16:00:09 +0100
Subject: Account tidying (deleting homedirs) and regular informational mailings

There are a couple of things which have come out of of the account tidying project (https://computing.projects.inf.ed.ac.uk/#349), which I'd like to run past CEG.

Firstly, we are soon going to start deleting old accounts, specifically entries in KDC, LDAP, AFS-PTS and AFS-VLDB. The last of these means we will be deleting users' home directories and all data held within them, as per our policies:

http://computing.help.inf.ed.ac.uk/account-closure
http://computing.help.inf.ed.ac.uk/account-archival

The last time we did a bulk deletion was in 2012 and we archived all data prior to deletion. Our policy is now *not* to do this, as per the second link above. I believe this policy has been in place since January 2015.

We also introduced prometheus lifecycle code in January 2015 and the account expiry mail we send out specifically refers to account deletion (see https://wiki.inf.ed.ac.uk/DICE/PrometheusLifecycleExpiryEmail).

I would, however, like to get a specific sign-off from CEG that we are going to start deleting data for expired user accounts. Should we, for example, send out a last warning "your Informatics home directory will be deleted" email to the last known email address for people; should we treat accounts from *before* this policy was put into place any differently, etc.? i.e. Would someone who left in 2010 have any reasonable assumption that we had maintained an archive of their home directory, as that was our policy at the time?

Following on from the above, we decided, as part of this project, that we should send regular (6 monthly) emails to sys-announce summarising our account policies (https://wiki.inf.ed.ac.uk/viewauth/DICE/Project349AccountTidying#Communication_of_policies). It seems that there may be other regular informational emails that we, as a computing body, should send. It seems appropriate that this is a matter to be discussed at CEG.

```
Toby
```

Reply from Craig (on behalf of CEG):

```
From: Craig Strachan <cms@inf.ed.ac.uk>
To: Toby Blake <toby@inf.ed.ac.uk>
Cc: compexec@inf.ed.ac.uk, Jennifer Oxley <joxley@inf.ed.ac.uk>
Date: Wed, 16 May 2018 15:41:45 +0100
Subject: Re: [compexec] Account tidying (deleting homedirs) and regular informational mailings

Yes, we discussed this on Monday.  Please go ahead and delete all the
relevant home directories.  We were of the opinion that we shouldnt
differentiate based on when the account holder left and that you should
go ahead and delete everything.

We've taken on board your point about periodic reminders but are still
mulling over what else the user base should be periodically reminded of!

Craig.
```

A one-off data-tidying and deletion process for Prometheus, downstream stores (KDC, AFS-PTS, AFS-VLDB, LDAP) and filesystems.

Deletions still that need to be done:

- delete all downstream account objects which are not in prometheus (mostly <= 2011)
  - 730 in total
- delete all accounts which expired prior to lifecycle (introduced 2015-01-24)
  - ~3000 in total
- delete all accounts which expired post lifecycle
  - ~4600 in total
  - ~3000 eligible for deletion for > 1year
- delete 'new' accounts which aren't needed (mainly students who briefly appeared in the database feed and then disappeared)
  - >50 in total

We also need to consider what we do with data outwith home directories, as discussed previously.

Establishing procedures, tools and documentation for computing staff to manage account expiry in the future, automated where possible and advisable.

Code (awaiting review) has been added to prometheus-lifecycle to identify accounts which are eligible for deletion. Beyond that, we will still need to consider the questions raised originally:

- How often do we do it?
- Who is responsible for doing it?
- Should accounts be deleted as soon as they are eligible?

One for user support?

Update 2018-10-08

Account deletions

We have now deleted the following:

- all downstream account objects which were not in prometheus (mostly <= 2011)
  - (729)
- all accounts which expired prior to lifecycle (introduced 2015-01-24)
  - (3052)
- all unactivated accounts which are no longer entitled to an account (all unused and with empty home directories)
  - (1035)

This purge should now mean that all accounts which are eligible for deletion expired after prometheus lifecycle was introduced (as documented here). This means that, for all these accounts:

- we have an account end, a grace end and a suspension end date
- we can identify these accounts using prometheus tools (`prometheus-lifecycle`)
- emails were sent to these account holders with details of account expiry (the email looks like this)

Given the above, it seems wise to involve user support in the next stage of deletions, particularly with a view to developing any documentation or tools required.

Regular emails concerning our account policies

As discussed at CEG, the Projects meeting and here. We should, in conjunction with user support, develop the text for this email and procedures for sending it. At the last projects meeting, it was suggested that mid-late October would be a good time to send this email.

Implications of account deletion

We need to consider again any other areas of our systems which are affected by the deletion of user accounts. Specifically:

- data outwith home directories
- file system implications
- any other areas of retention (briefly considered above)

Prometheus support for non-person identities/accounts

Prometheus should be able to support identities and accounts which are not connected to real people. This needs to be tested and then any such objects should be added to prometheus. This is outwith the scope of this project, but may need to be considered.

Update 2018-10-09

I've written up the procedure for deleting users from prometheus:

PrometheusDeletingUsers

Update 2018-11-27

A further 2439 accounts have been deleted. This means that all accounts which have been eligible to be deleted for more than a year have now gone. This still leaves more than 1300.

What I think is remaining in the project is:

- Handing over to user support for future management/deletions - make sure we have documentation/procedures in place; work out how we're going to do it (just US, or US/Inf), how frequently, etc.
- Implications of account deletion (see above)
- Regular emails concerning our account policies - see recent email to Alison/Jennifer - assuming we're agreed on the text, we need to decide whose remit it falls under.

Update 2018-12-04

Jennifer/Toby met to begin handover of deletions to user-support. 10 users were deleted. The plan is that Jennifer/Toby meet regularly to work through the backlog and to make sure procedure is properly documented.

Update 2019-01-16

The remaining aspects of this project:

- Handing over to user support - Jennifer/Toby will continue to meet, as availability allows, to work through backlog of user deletions and ensure that procedure and documentation is adequate.
- Implications of account deletion - this should be handed over to services, possibly as a subsequent project.
- Regular emails concerning our account policies - the text of this email has been agreed. This will be sent out twice-yearly (mid-October/mid-February, to be near the beginning of each semester). The sending of this email will be automated from the prometheus server, to ensure it doesn't get forgotten.

Update 2019-02-07

- Jennifer has now taken over deletions, to be done as and when scheduling permits.
- Implications of account deletion new project created - Account tidying - AFS/homepages/groupspace
- First automated email sent 7th Feb 2019.

-- TobyBlake - 16 Jan 2019

This topic: DICE > Project349AccountTidying
Topic revision: r17 - 07 Feb 2019 - 16:06:48 - TobyBlake