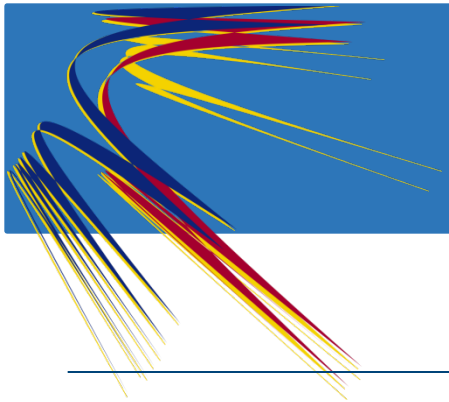


horizon
DIGITAL ECONOMY RESEARCH

Becoming Dataware

Richard Mortier
Horizon Digital Economy Research
University of Nottingham

With thanks to James Goulding, Derek McAuley,
Anil Madhavapeddy (CUCL)

The University of Nottingham

RCUK Digital Economy

- Digital Economy is defined by the Research Councils as:

  The novel design or use of information and communication technologies to help transform the lives of individuals, society or business.

- The Digital Economy progamme is:
  - Cross-Research Council (EPSRC, ESRC, AHRC)
  - Funded 2008-2011 for:
    - £80m research – including 3 x £12m hubs
    - £36m training – 8 x DTCs
  - Aimed at realising the transformational impact of ICT for all aspects of Business, Society and Government

What is Horizon?

- A Digital Economy Research Centre at the University of Nottingham comprising:
  - A Digital Economy Hub
    - £20m from RCUK and university
    - Spokes at Cambridge, Reading, Exeter, Brunel
  - A Doctoral Training Centre
    - £15m from RCUK and university
    - 20 PhD students per year for 5 years
  - Now 120+ partner companies, from 40 in initial bid
  - 3 TEDDI projects
  - … + future Digital Economy projects

Our Digital Footprints

"Every time we register for a new web service, or upload our photos and videos, we are enlarging our own digital footprints"

- Whether "informed" or not
  - Facebook, Google
- Growing digital footprint poses major societal challenges
- …but also forms a key basis for the digital economy's growth

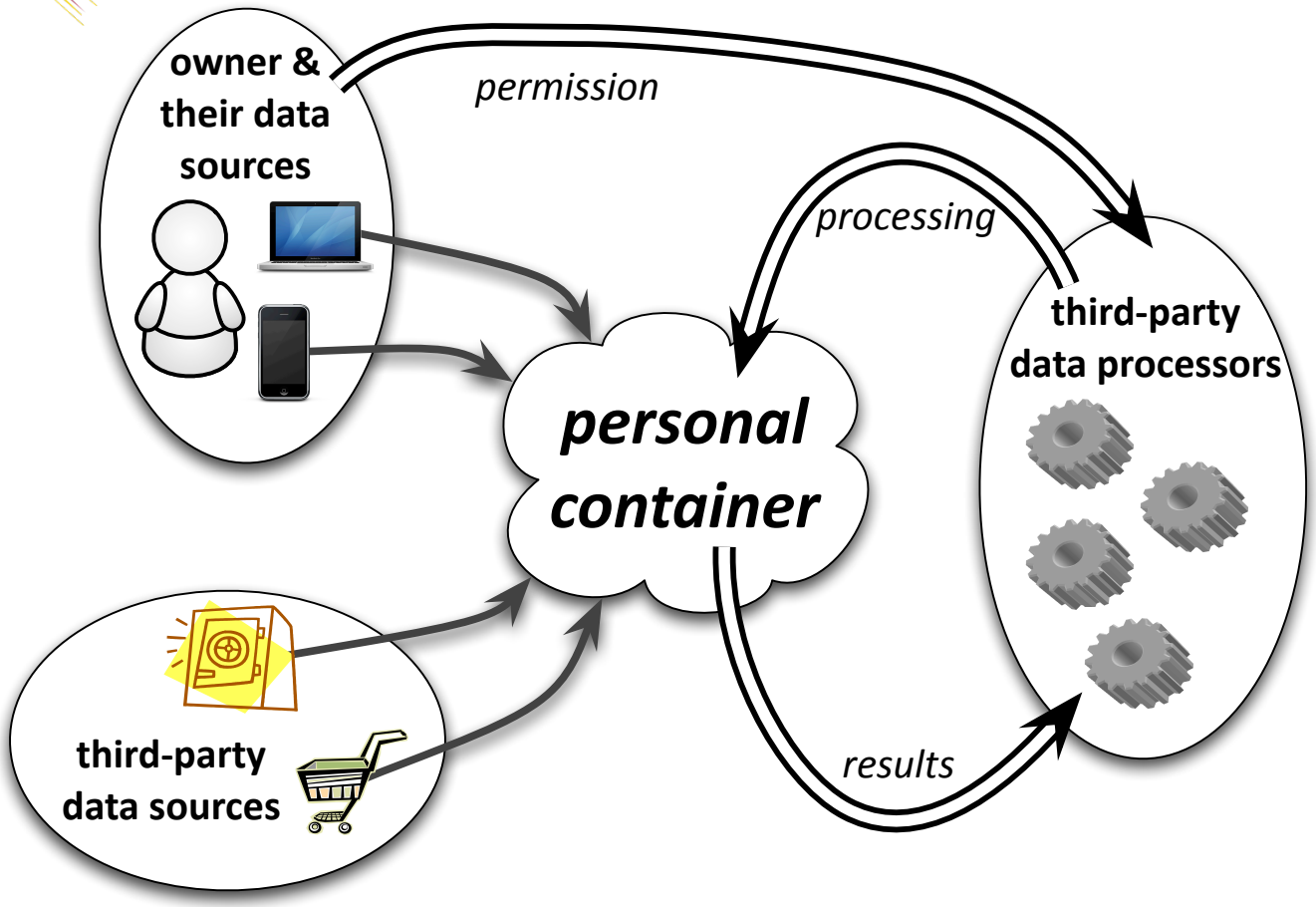
In More Detail

- We are making more and more information about our lives available, digitally
  - Whether or not we realise it
  - Often in very large, very rich data silos (e.g., Facebook)
  - "Contextual footprint"
- Simultaneously, there are more and more opportunities for mutually beneficial exploitation of digital personal data
  - E.g., Shopping basket optimizer;
    Boots prescription conflict detector

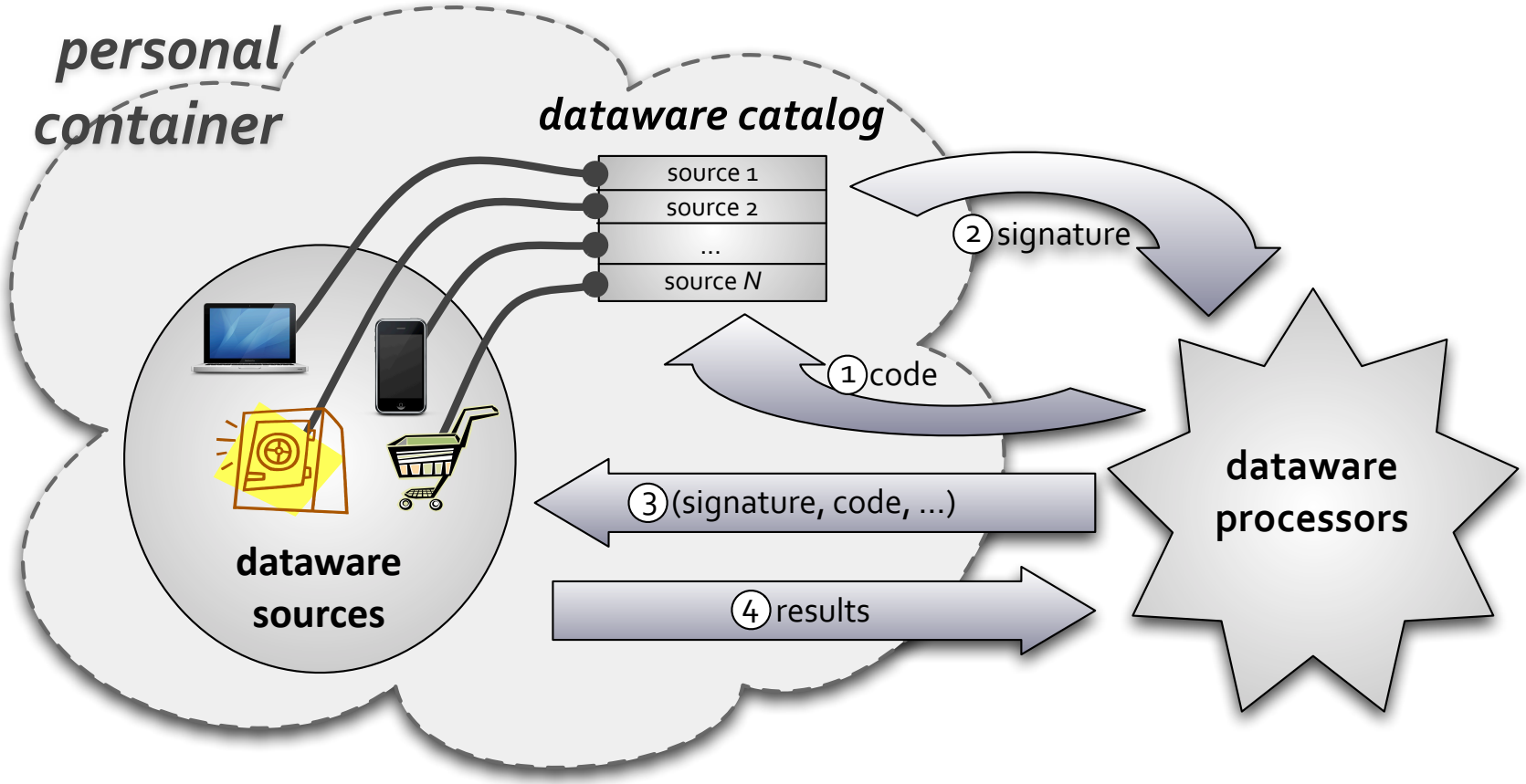
Key Challenge:

How do we enable individuals to control collection and exploitation of both *their data* and *data about them*?

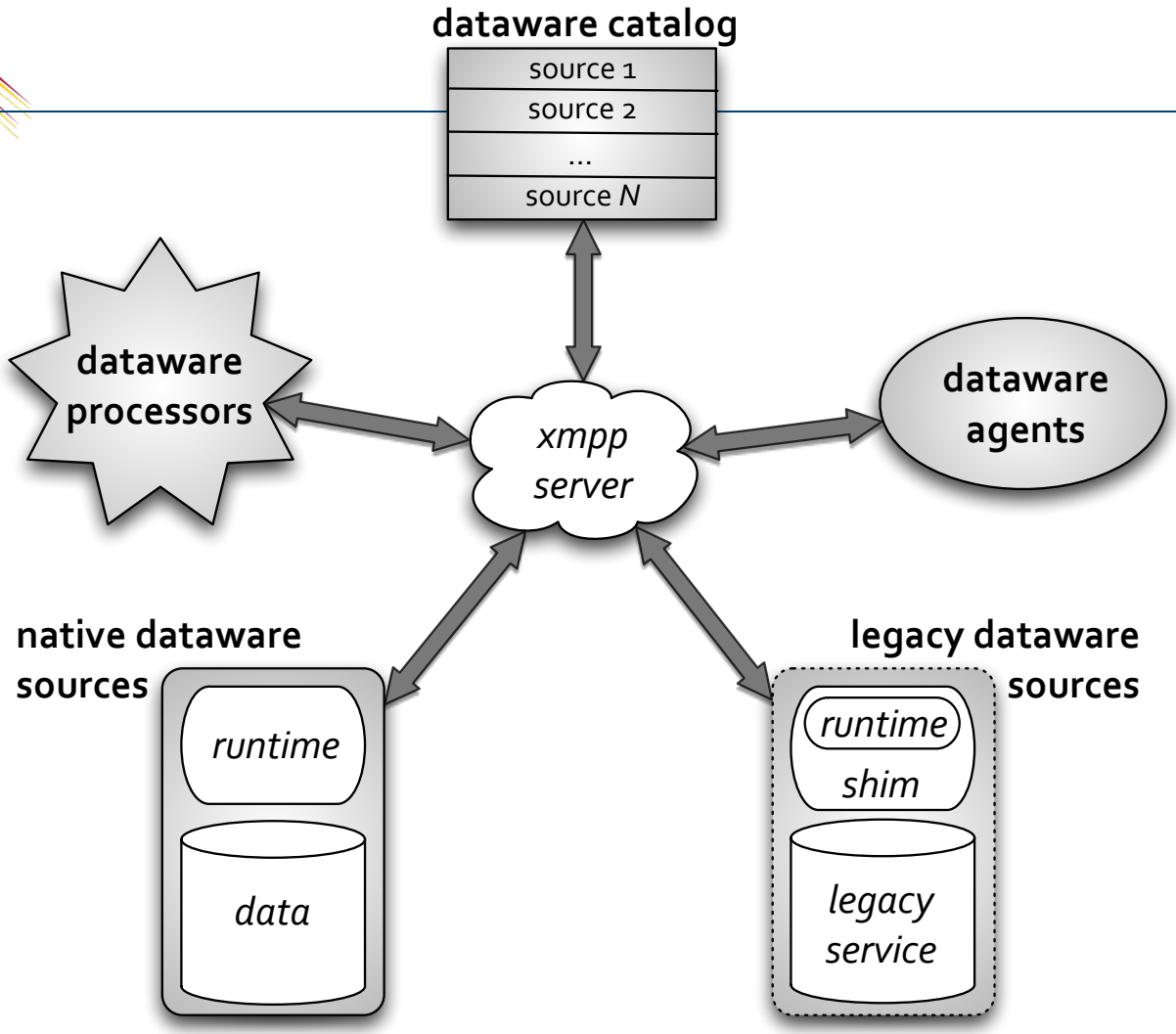
Dataware


owner &
their data
sources
permission
processing
third-party
data processors
personal
container
third-party
data sources
results

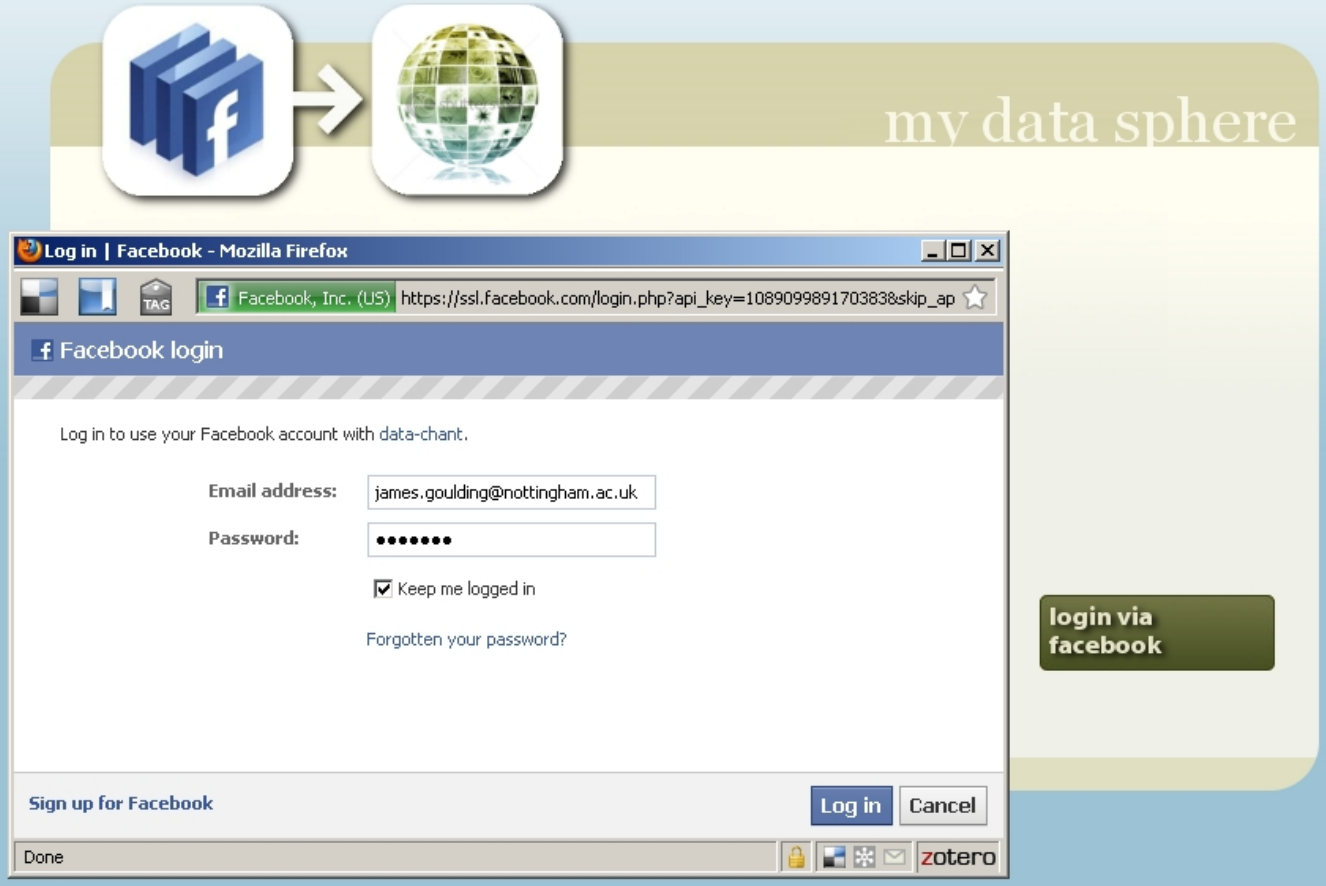
Components


personal container
dataware catalog
source 1
source 2
...
source N
2 signature
1 code
3 (signature, code, ...)
4 results
dataware sources
dataware processors

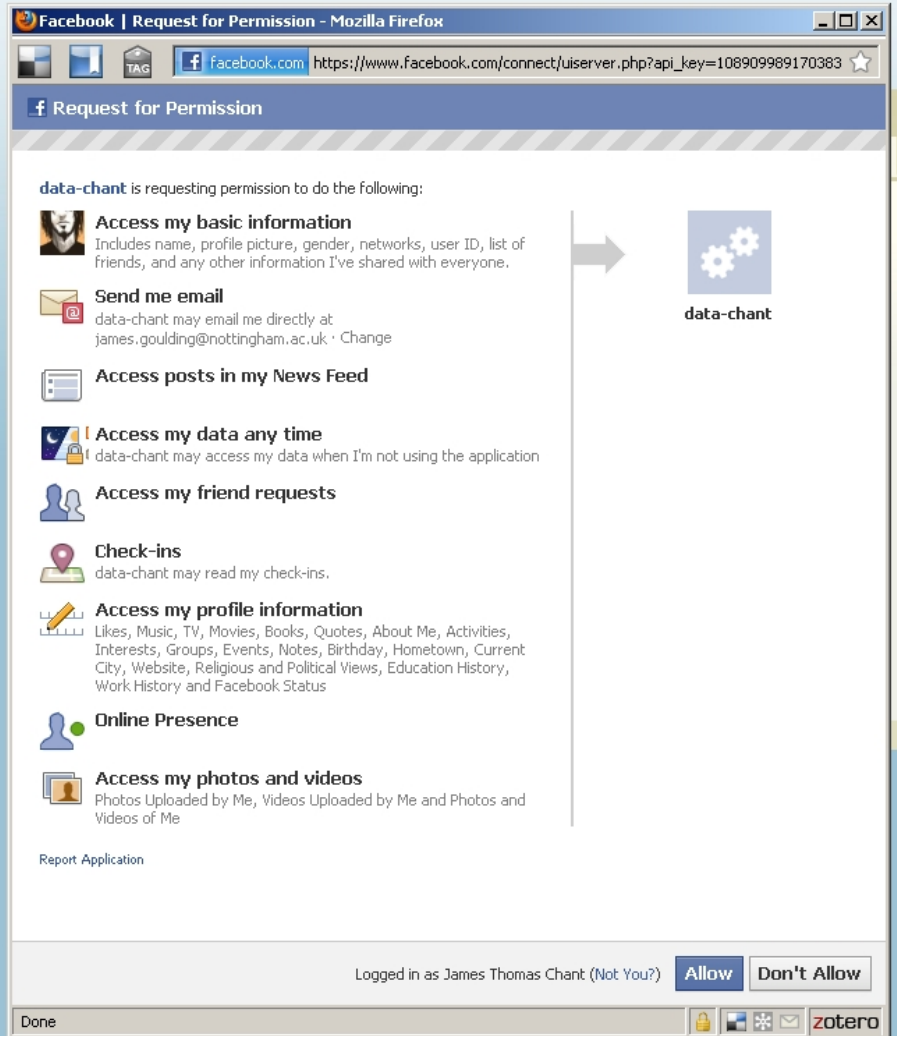
Implementation

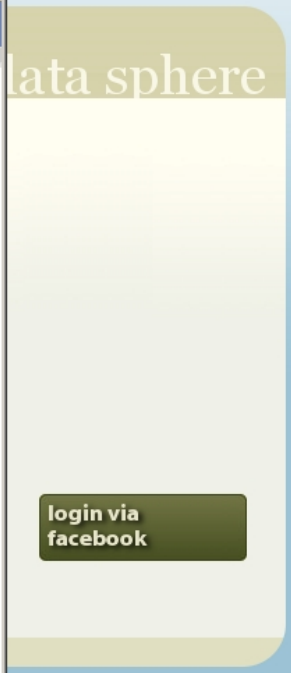


Shim Login

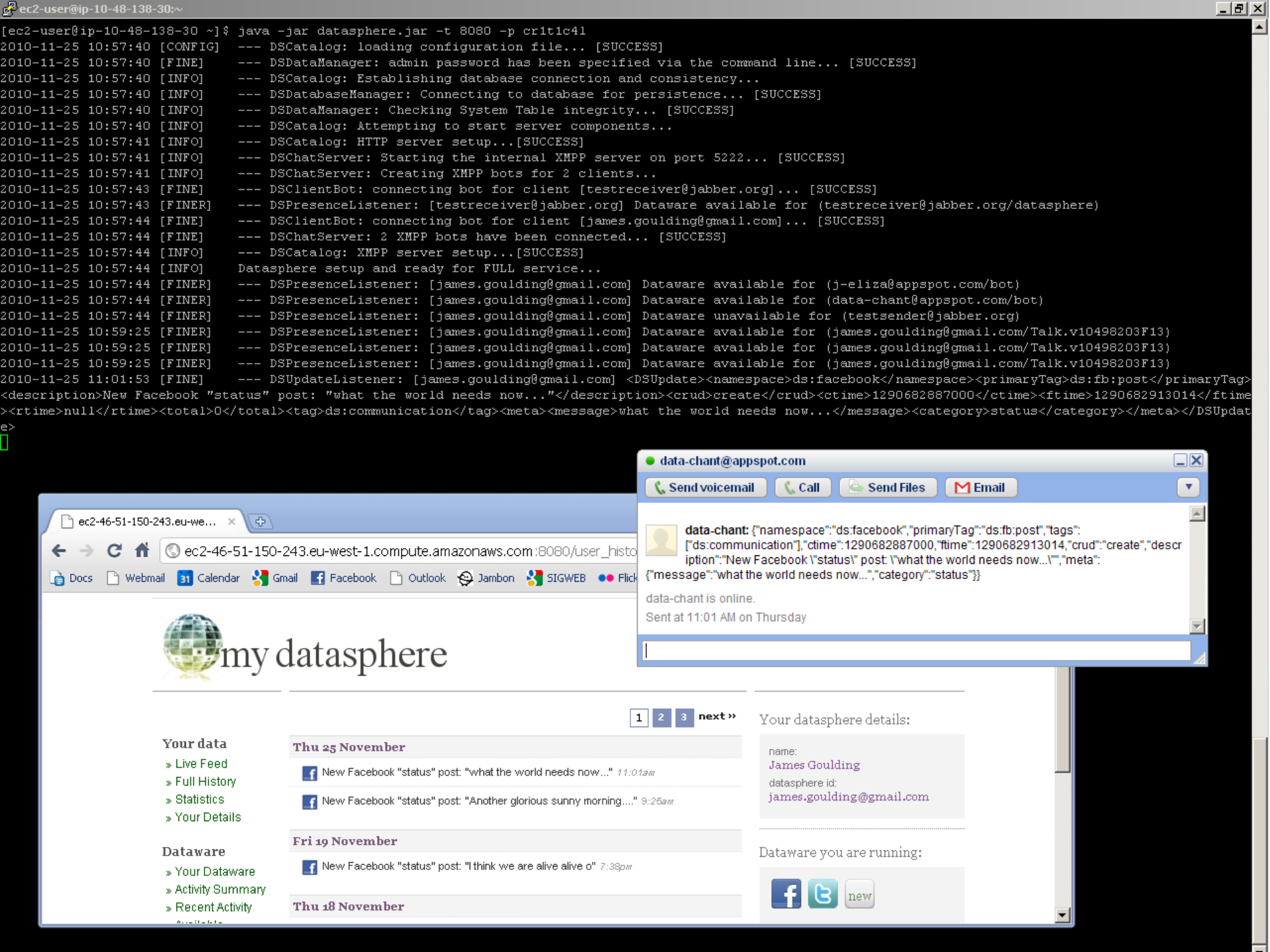

my data sphere
Log in | Facebook - Mozilla Firefox
Facebook, Inc. (US)
https://ssl.facebook.com/login.php?api_key=108909989170383&skip_ap
Facebook login
Log in to use your Facebook account with data-chant.
Email address:
james.goulding@nottingham.ac.uk
Password:
Keep me logged in
Forgotten your password?
Sign up for Facebook
Log in
Cancel
Done
zotero
login via facebook

Shim Permission

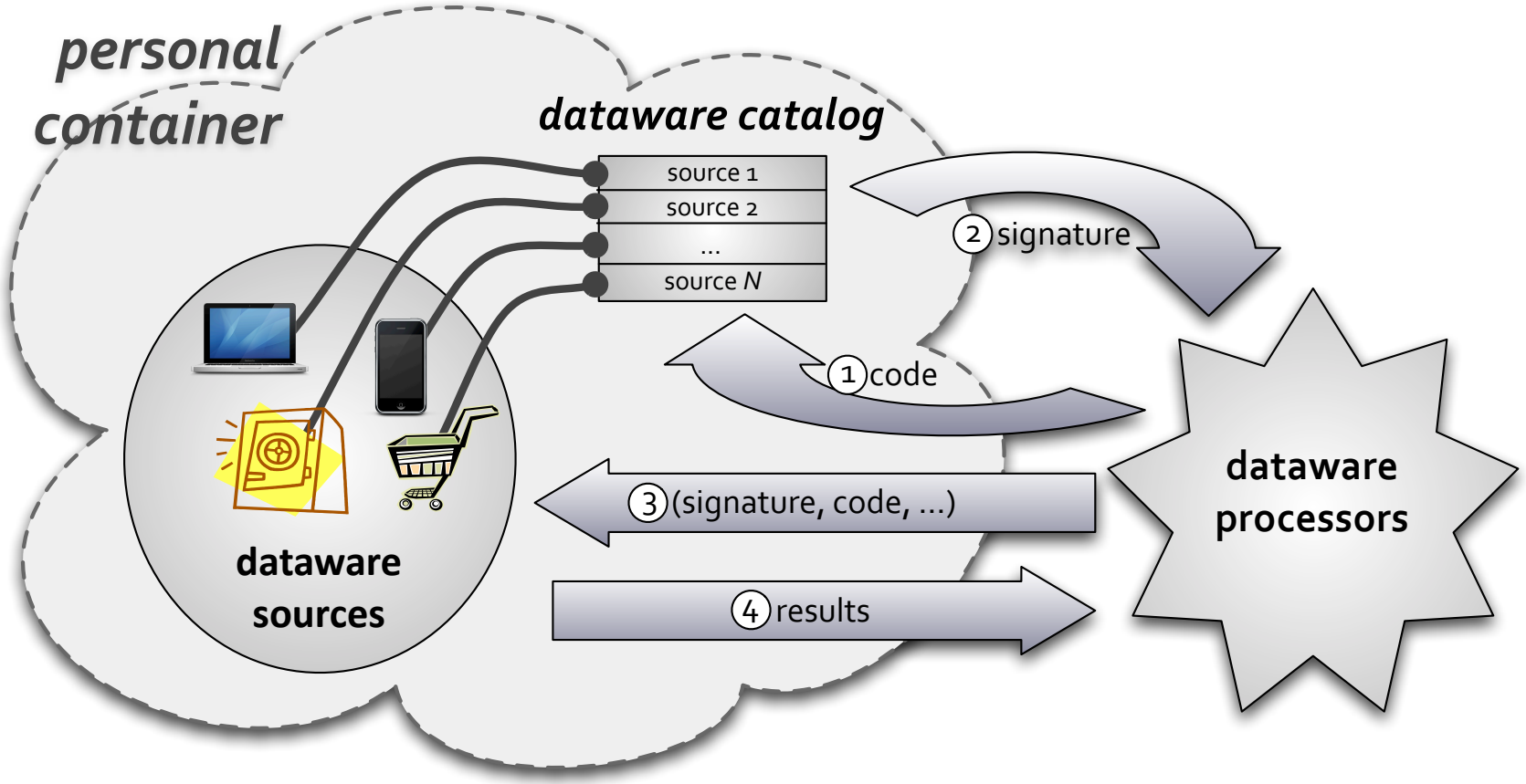




Web Interface


ec2-user@ip-10-48-138-30:~
[ec2-user@ip-10-48-138-30 ~]$ java -jar datasphere.jar -t 8080 -p cr1t1c41
2010-11-25 10:57:40 [CONFIG] --- DSCatalog: loading configuration file... [SUCCESS]
2010-11-25 10:57:40 [FINE] --- DSDataManager: admin password has been specified via the command line... [SUCCESS]
2010-11-25 10:57:40 [INFO] --- DSCatalog: Establishing database connection and consistency...
2010-11-25 10:57:40 [INFO] --- DSDatabaseManager: Connecting to database for persistence... [SUCCESS]
2010-11-25 10:57:40 [INFO] --- DSDataManager: Checking System Table integrity... [SUCCESS]
2010-11-25 10:57:40 [INFO] --- DSCatalog: Attempting to start server components...
2010-11-25 10:57:41 [INFO] --- DSCatalog: HTTP server setup...[SUCCESS]
2010-11-25 10:57:41 [INFO] --- DSChatServer: Starting the internal XMPP server on port 5222... [SUCCESS]
2010-11-25 10:57:41 [INFO] --- DSChatServer: Creating XMPP bots for 2 clients...
2010-11-25 10:57:43 [FINE] --- DSClientBot: connecting bot for client [testreceiver@jabber.org]... [SUCCESS]
2010-11-25 10:57:43 [FINER] --- DSPresenceListener: [testreceiver@jabber.org] Dataware available for (testreceiver@jabber.org/datasphere)
2010-11-25 10:57:44 [FINE] --- DSClientBot: connecting bot for client [james.goulding@gmail.com]... [SUCCESS]
2010-11-25 10:57:44 [FINE] --- DSChatServer: 2 XMPP bots have been connected... [SUCCESS]
2010-11-25 10:57:44 [INFO] --- DSCatalog: XMPP server setup...[SUCCESS]
2010-11-25 10:57:44 [INFO] Datasphere setup and ready for FULL service...
2010-11-25 10:57:44 [FINER] --- DSPresenceListener: [james.goulding@gmail.com] Dataware available for (j-eliza@appspot.com/bot)
2010-11-25 10:57:44 [FINER] --- DSPresenceListener: [james.goulding@gmail.com] Dataware available for (data-chant@appspot.com/bot)
2010-11-25 10:57:44 [FINER] --- DSPresenceListener: [james.goulding@gmail.com] Dataware unavailable for (testsender@jabber.org)
2010-11-25 10:59:25 [FINER] --- DSPresenceListener: [james.goulding@gmail.com] Dataware available for (james.goulding@gmail.com/Talk.v10498203F13)
2010-11-25 10:59:25 [FINER] --- DSPresenceListener: [james.goulding@gmail.com] Dataware available for (james.goulding@gmail.com/Talk.v10498203F13)
2010-11-25 10:59:25 [FINER] --- DSPresenceListener: [james.goulding@gmail.com] Dataware available for (james.goulding@gmail.com/Talk.v10498203F13)
2010-11-25 11:01:53 [FINE] --- DSUpdateListener: [james.goulding@gmail.com] <DSUpdate><namespace>ds:facebook</namespace><primaryTag>ds:fb:post</primaryTag>
<description>New Facebook "status" post: "what the world needs now..."</description><crud>create</crud><ctime>1290682887000</ctime><ftime>1290682913014</ftime
><rtime>null</rtime><total>0</total><tag>ds:communication</tag><meta><message>what the world needs now...</message><category>status</category></meta></DSUpdat
e>
data-chant@appspot.com
Send voicemail
Call
Send Files
Email
data-chant: {"namespace":"ds:facebook","primaryTag":"ds:fb:post","tags":
["ds:communication"],"ctime":1290682887000,"ftime":1290682913014,"crud":"create","descr
iption":"New Facebook \"status\" post: \"what the world needs now...\"","meta":
{"message":"what the world needs now...","category":"status"}}
data-chant is online.
Sent at 11:01 AM on Thursday
ec2-46-51-150-243.eu-we...
ec2-46-51-150-243.eu-west-1.compute.amazonaws.com:8080/user_histo
Docs
Webmail
Calendar
Gmail
Facebook
Outlook
Jambon
SIGWEB
my datasphere
1
2
3
next »
Your data
» Live Feed
» Full History
» Statistics
» Your Details
Dataware
» Your Dataware
» Activity Summary
» Recent Activity
Thu 25 November
New Facebook "status" post: "what the world needs now..." 11:01am
New Facebook "status" post: "Another glorious sunny morning...." 9:25am
Fri 19 November
New Facebook "status" post: "I think we are alive alive o" 7:38pm
Thu 18 November
Your datasphere details:
name:
James Goulding
datasphere id:
james.goulding@gmail.com
Dataware you are running:
new

Processing


personal container
dataware catalog
source 1
source 2
...
source N
2 signature
1 code
3 (signature, code, ...)
4 results
dataware sources
dataware processors

Applications

- An application is a list of source schemas and signed components
  - Each element defines processing occurring on some data
- Each component starts out as
  - *Code*, describing the computation
  - *Schedule*, against which the code is executed
  - *Result schema*, describing the results
  - *Result destination*, JID to which results are sent

Application Signing

- Processor presents application to catalog
- Catalog applies user policy to determine whether to sign
  - Static policy concerning data to be processed, source, &c
  - Or interacting directly with user via web, chat, &c
- To sign, replace each component with one or more:
  - *Result schema, schedule, code*, as before
  - *Source JID*, specifying data source
  - *Destination JID*, to which results are sent
  - *Signature*, an HMAC for the component
- …And then return to the submitting processor

Application Execution

- Processor distributes signed components to given sources
- Sources then:
  - Install components, notifying catalog
  - Execute components in runtime, as per schedule
  - Send results to specified destination
- During execution, the runtime monitors code
  - Enables periodic reporting back to user
  - Exceptions raised if code operates outside boundaries

JOINing Data

- A clear issue is how to deal with data JOINs
  - Recall: we're avoiding giving access to raw data
  - We must also avoid giving access to intermediate results
- Setup *trusted third-party sources*
  - Have no data of their own
  - Act as destination for other code
  - May host subsequent application components
- Makes sense that the catalog serves as at least one such

Summary

- So conceptually, a *dataware application* is:
  - A network of running components,
  - Each processing your personal data,
  - In a manner acceptable to you
- Your *dataware* implements your *personal container*:
  - Defining APIs to your data,
  - Enabling third-parties to process your data,
  - While ensuring *you* retain control,
  - And they *don't* get copies of your data
    - (unless you want them to!)

Evolution or Revolution?

- One way to put this in context is via Van Jacobson's *content centric networking*:
  - Telephones care about building paths (not calls)
  - Internet cared about connections (not data)
  - Content centric networking cares about data (not results)
- …so, finally,
- *Dataware*, cares about results
  - Computation is as mobile as data, if not more so
  - (Anil) cf. datacentre computing, map-reduce, &c

Status

- First version of catalog built
- First legacy source built
  - Facebook, using Graph API
  - Twitter, Google Gdata on their way
- Now prototyping
  - Dataware Processor and first applications, and
  - Native source
- Challenge: how to build protocols for trusted aggregation
  - Would like to guarantee *at-least-N* contributors