

R in the cloud

Karim Chine

karim.chine@cloudera.co.uk

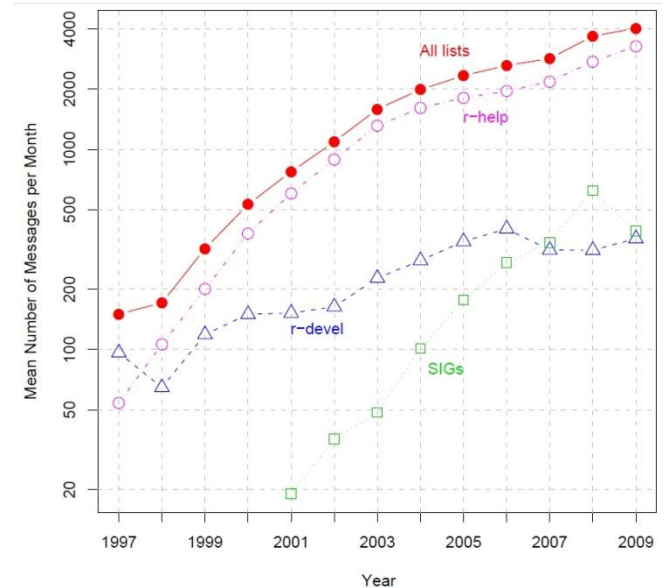
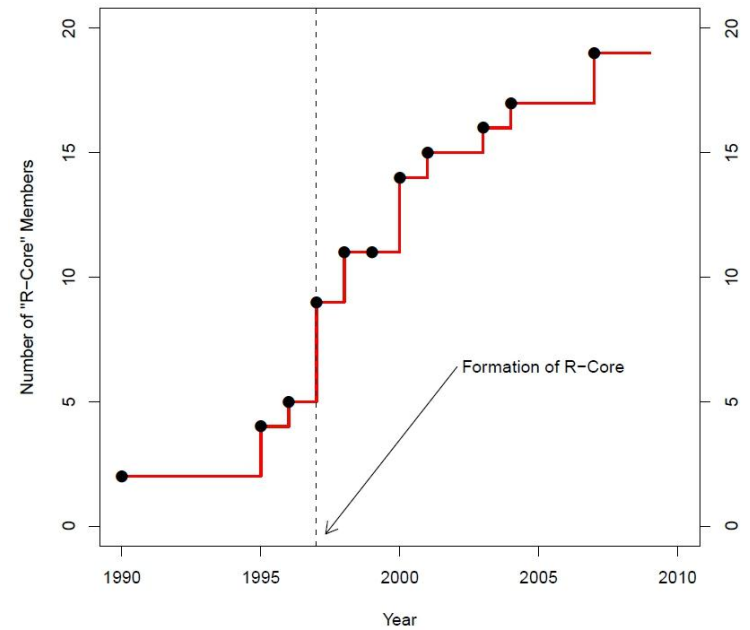
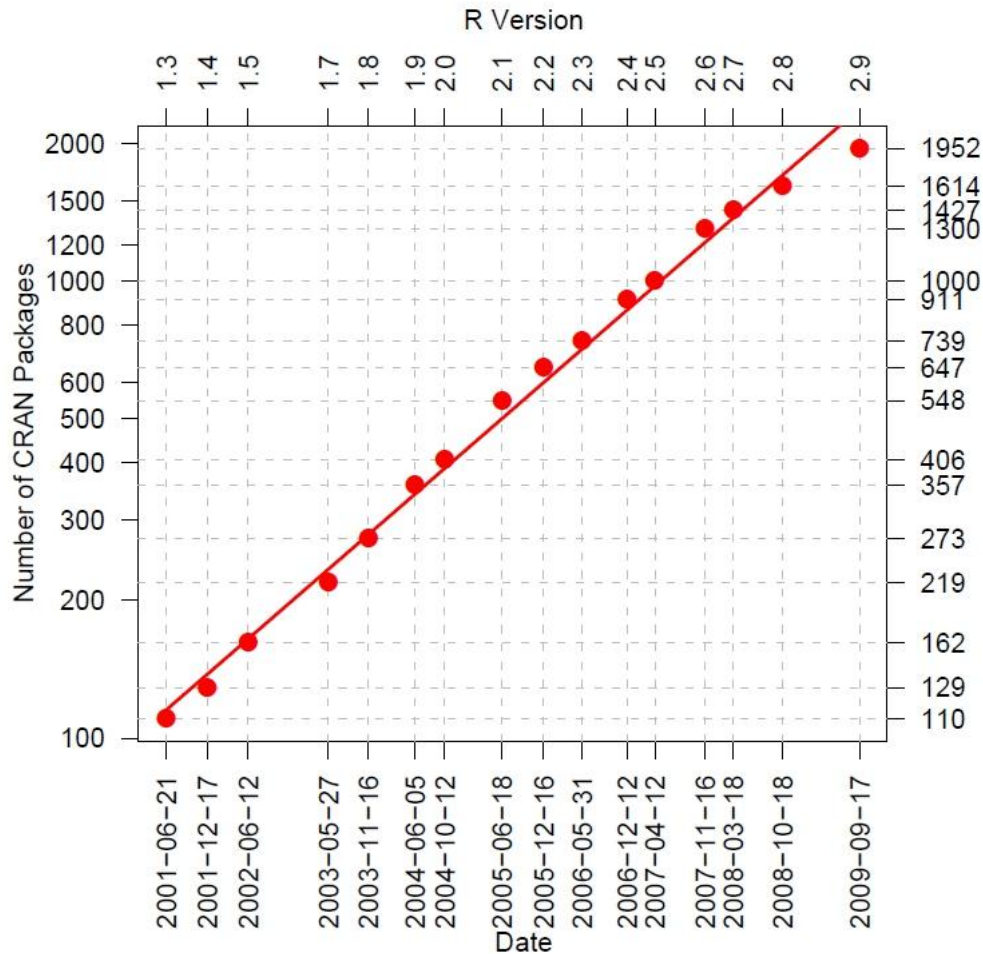
Cloud Era Ltd

Cambridge – UK

Department of Statistics and Mathematics Seminar
Wirtschaftsuniversitaet Wien

29 April 2010

lingua franca of data analysis



R + Elastic-R : « plug & play » computing environment

Computational Components

R packages : CRAN, Bioconductor, Wrapped C,C++,Fortran code
Scilab modules, Matlab Toolkits, etc.
Open source or commercial

Computational Resources

Hardware & OS agnostic computing engine : R, Scilab,..
Clusters, grids, private or public clouds
free: academic grids or pay-per-use: EC2, Azure

Computational User Interfaces

Workbench within the browser
Built-in views / Plugins / Spreadsheets
Collaborative views
Open source or commercial

Computational Data Storage

Local, NFS, FTP, Amazon S3, Amazon EBS
free or commercial

Computational Scripts

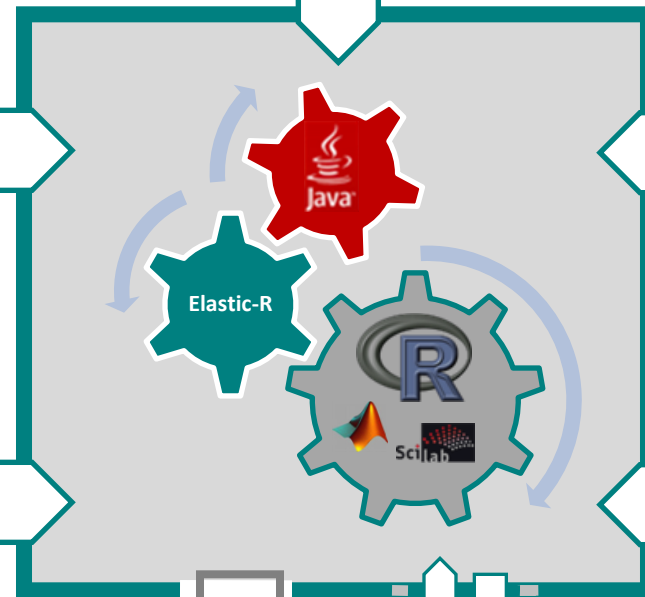
R / Python / Groovy
On client side: interactivity..
On server side: data transfer ..

Generated Computational Web Services

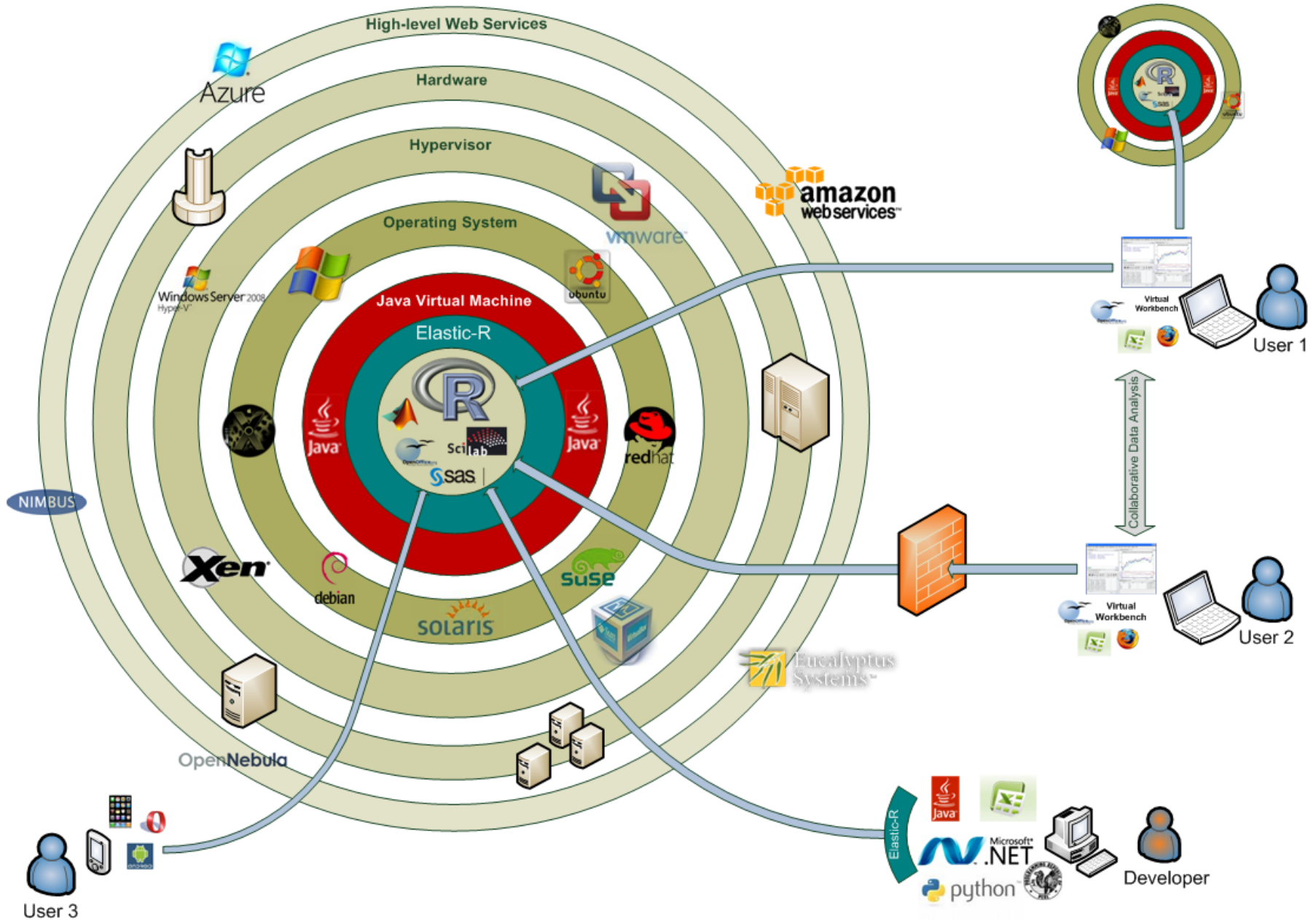
Stateful or stateless, automatic mapping of R data objects and functions

Computational Application Programming Interfaces

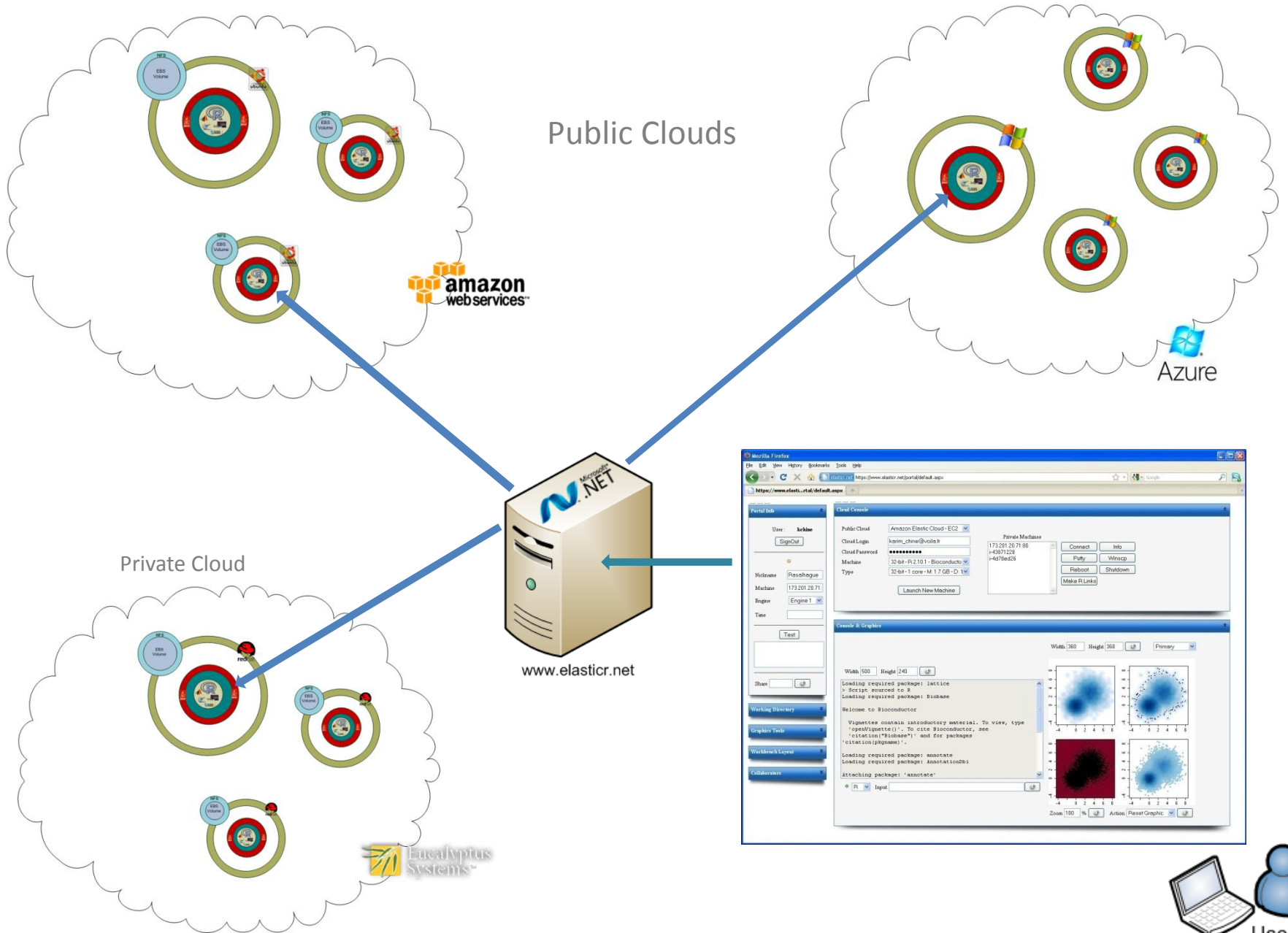
Java / SOAP / REST, Stateless and stateful



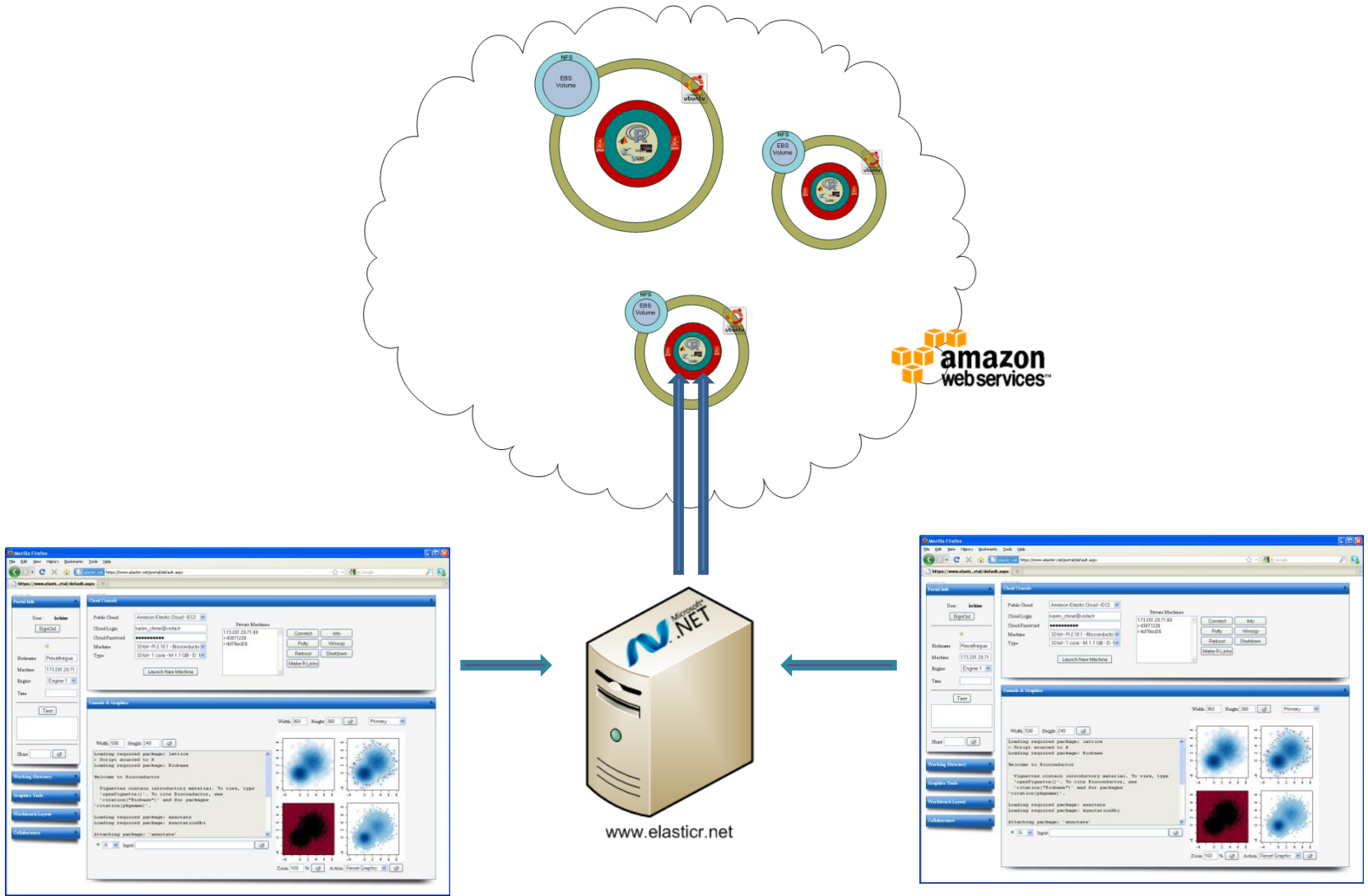
Elastic-R on IaaS-style clouds



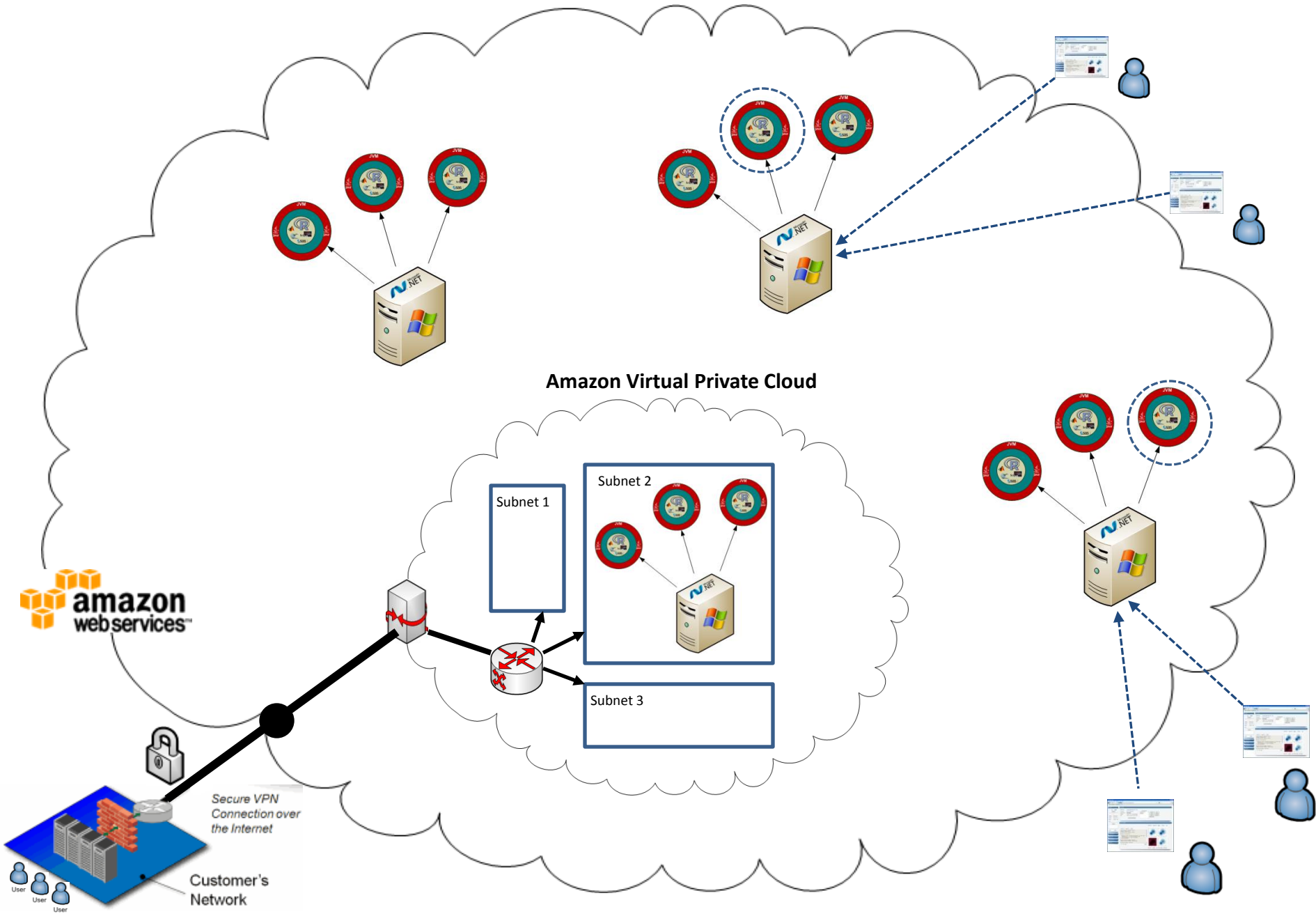
Elastic-R portal: single facade to public/private clouds



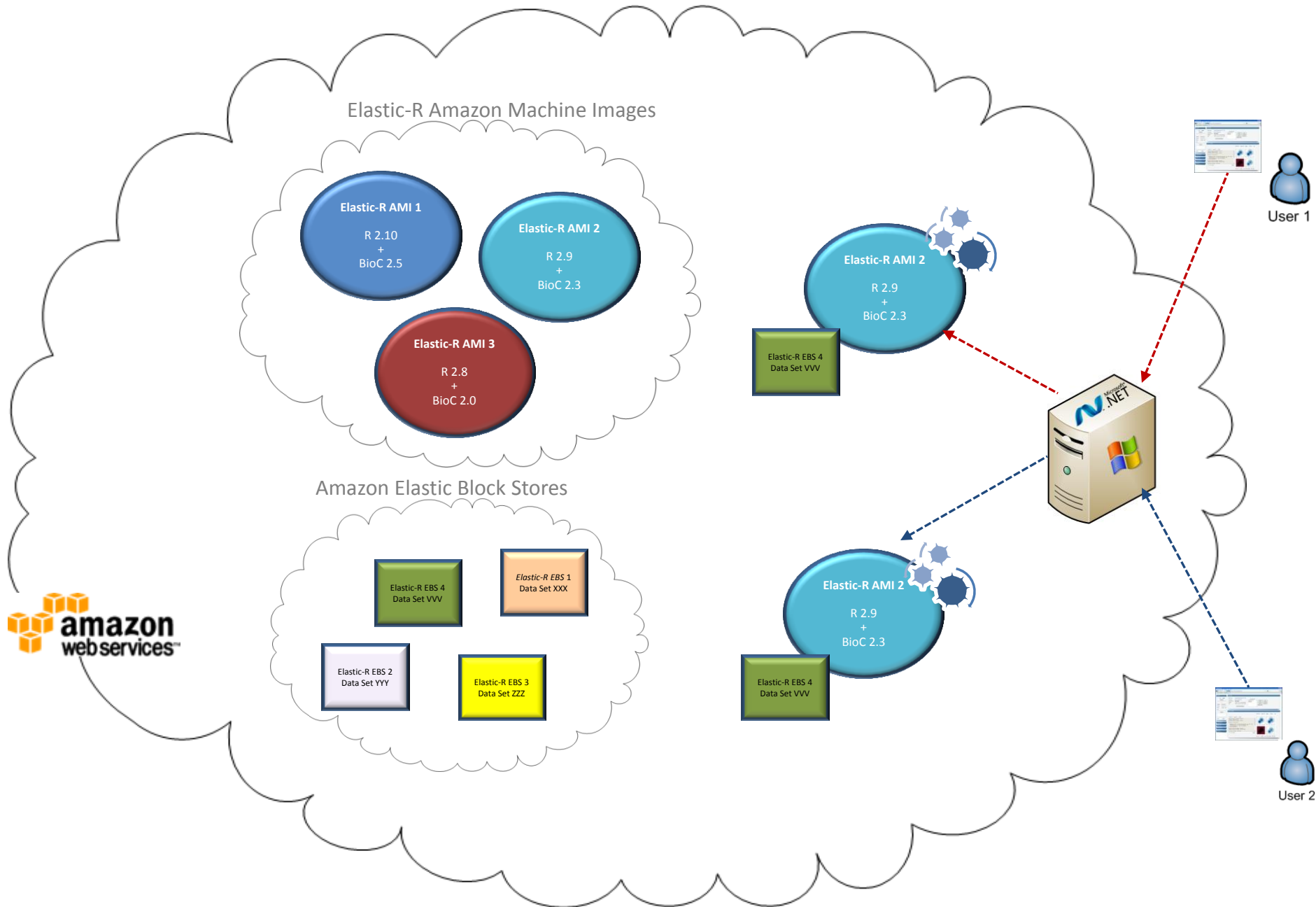
Elastic-R portal : collaborative Virtual Research Environment



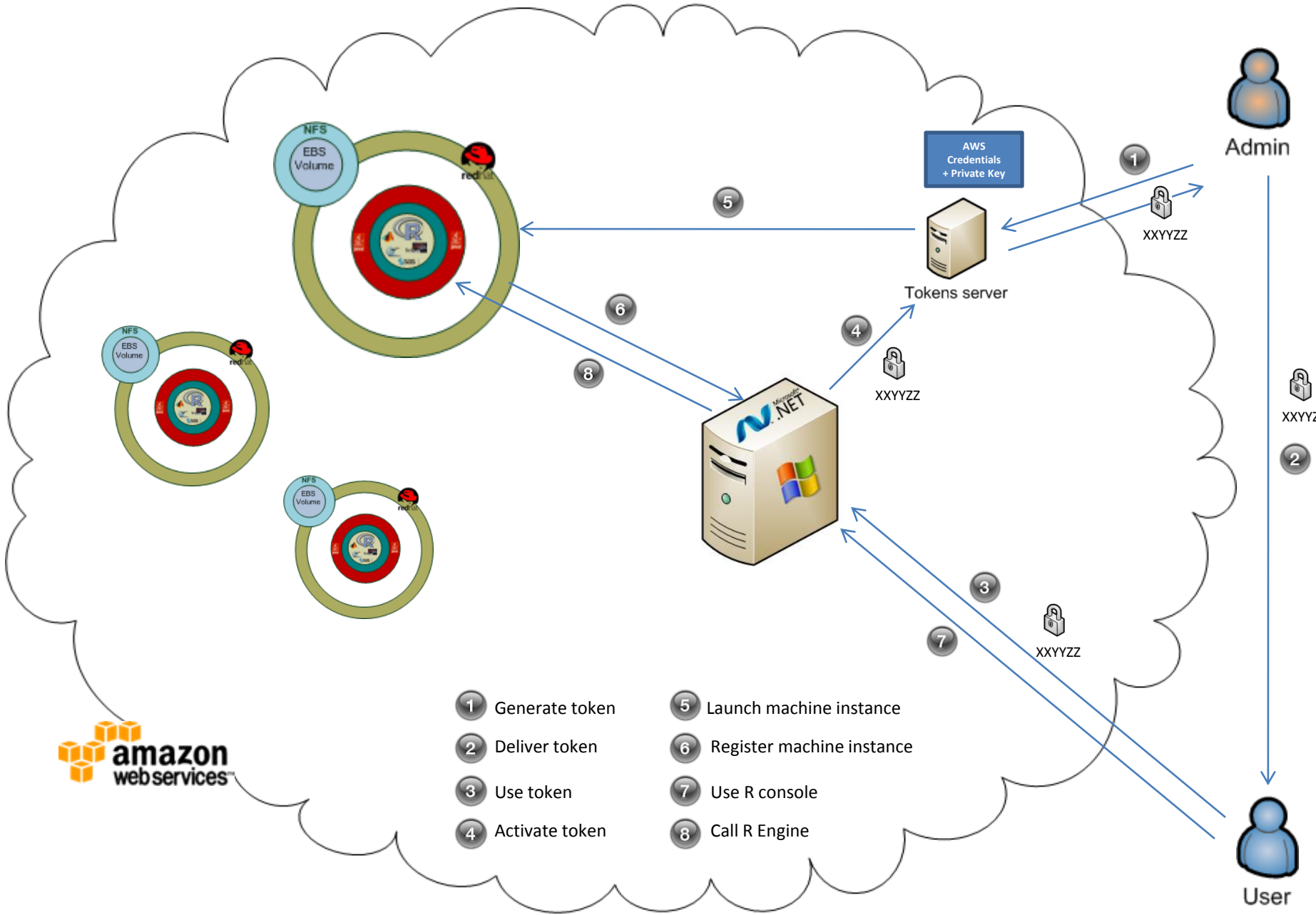
Decentralized collaboration : Elastic-R portal as an EC2 AMI



The IaaS-style cloud as a reproducible research platform



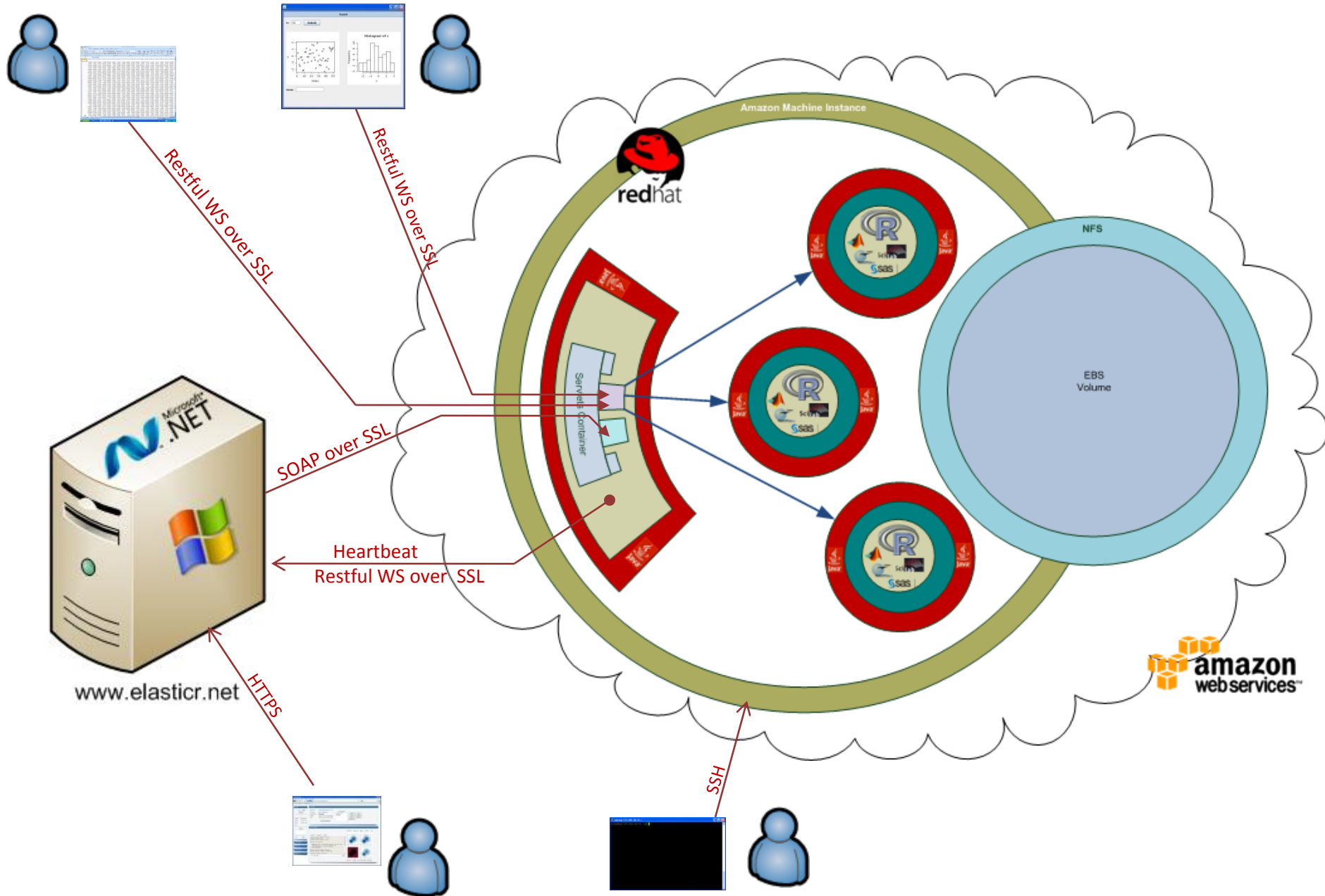
One Amazon account and many users : Elastic-R signed tokens



- | | |
|------------------|-----------------------------|
| 1 Generate token | 5 Launch machine instance |
| 2 Deliver token | 6 Register machine instance |
| 3 Use token | 7 Use R console |
| 4 Activate token | 8 Call R Engine |



Elastic-R security



The Elastic-R AJAX Workbench I

Private virtual machines monitor

Simplified clouds management console

Virtual machines launcher

Session info

R console
+ scilab console
+ chat
+ ssh console

R Graphics
+ whiteboard
+ annotation
+ slides viewer

The screenshot shows the Elastic-R AJAX Workbench interface in a Mozilla Firefox browser window. The interface is divided into several sections:

- Portal Info:** Displays user information (User: kchine, SignOut button), Nickname (Rasalhague), Machine (173.201.20.71), Engine (Engine 1), and Time.
- Cloud Console:** Shows configuration for Amazon Elastic Cloud - EC2, including Cloud Login (karim_chine@voila.fr), Cloud Password, Machine (32-bit - R 2.10.1 - Bioconductor), and Type (32-bit - 1 core - M: 1.7 GB - D: 1). A Launch New Machine button is present.
- Private Machines:** Lists private machines with IP addresses (173.201.20.71:80, i-43871228, i-4d78ed26) and provides control buttons: Connect, Info, Putty, Winscp, Reboot, Shutdown, and Make R Links.
- Console & Graphics:** Contains an R console and a graphics area. The R console shows the following output:

```
Loading required package: lattice  
> Script sourced to R  
Loading required package: Biobase  
Welcome to Bioconductor  
Vignettes contain introductory material. To view, type  
'openVignette()'. To cite Bioconductor, see  
'citation("Biobase")' and for packages  
'citation(pkgname)'.  
Loading required package: annotate  
Loading required package: AnnotationDbi  
Attaching package: 'annotate'
```

The graphics area displays four scatter plots in a 2x2 grid, with a zoom level of 100% and a Reset Graphic button.
- Sidebar:** Contains navigation buttons for Working Directory, Graphics Tools, Workbench Layout, and Collaborators.

The Elastic-R AJAX Workbench II

Browsable contextual R help

Working directory browser
Files upload/download
to/from cloud machine instance

Collaborative console

Collaborative script editor

Working Directory

Directory: /home/gue

Filter: *

Order By: Date

Asc: No

Files: `..(dir)`, `slides(dir)`, `library(dir)`, `Image.png`, `Image.jpg`, `Image4.jpg`

Auto

File: Image Counter:

Console & Graphics

Width: 500 Height: 240

```
> kidney
ExpressionSet (storageMode: lockedEnvironment)
assayData: 8704 features, 2 samples
element names: exprs
phenoData
sampleNames: green, red
varLabels and varMetadata description:
channel: The scanner channel Cy3 or Cy5
featureData
featureNames: 1, 2, ..., 8704 (8704 total)
fvarLabels and fvarMetadata description: none
experimentData: use 'experimentData(object)'
Annotation:
> help(vsn)
```

Editor & File Browser

Script File: NewScript.R

R Documentation: vsn.old (vsn)

Variance stabilization and calibration for microarray data.

Description

Robust estimation of variance-stabilizing and calibrating transformations for microarray data. This function has been superseded by `vsn2`. The function `vsn` remains in the package for backward compatibility, but for new projects, please use `vsn2`.

Usage

```
vsn(intensities,
    lts.quantile = 0.5,
    verbose = interactive(),
    niter = 10,
    cvg.check = NULL,
    describe.preprocessing = TRUE,
    subsample,
    pstart,
    strata)
```

Arguments

intensities An object that contains intensity values from a microarray experiment. The intensities are assumed to be the raw scanner data, summarized over the spots by an image analysis program, and possibly "background subtracted". The intensities must not be logarithmically or otherwise transformed, and not thresholded or "floored". NAs are not accepted. See details.

lts.quantile Numeric. The quantile that is used for the resistant least trimmed sum of squares regression. Allowed values are

The Elastic-R AJAX Workbench III

Portal Info

Working Directory

Graphics Tools

Device: Primary

Graphics Coordinates

X: 13.409146322
Y: 12.749789451

Annotation

Mode: Circle
Color: Red
Width: 1
Style: 3
Text:
Size: 12
All Devices: No

Console & Graphics

Width: 500 Height: 240

```
featureNames: 1, 2, ..., 8704 (8704 total)
fvarLabels and fvarMetadata description: none
experimentData: use 'experimentData(object)'
Annotation:
> help(vsn)
> cells.put(exprs(kidney), 'A1', name='ssprimary')
> plot(exprs(kidney))
> plot.new();layout(matrix(1:1));par (mai=c(0.95625
,0.76875 ,0.76875, 0.39375), mar= c(5.1 ,4.1, 4.1 ,2.1) );
> nk=justvsn(kidney)
vsn2: 8704 x 2 matrix (1 stratum). Please use 'meanSdPlot'
to verify the fit.
> plot(exprs(nk))
> cells.put(exprs(nk), 'D1', name='ssprimary')
```

Spreadsheets

Name: SS Primary Range: A1:I12 Width: 100

| | A | B | C | D | E | F | G | H |
|---|--------|--------|---|----------|----------|---|---|---|
| 1 | 815.32 | 937.02 | | 9.406618 | 9.603355 | | | |
| 2 | 671.66 | 765.03 | | 9.087708 | 9.271431 | | | |
| 3 | 713.93 | 713.59 | | 9.18789 | 9.157763 | | | |
| 4 | 703.97 | 656.7 | | 9.164799 | 9.022733 | | | |
| 5 | 493.59 | 472.1 | | 8.593236 | 8.503035 | | | |
| 6 | 477.92 | 453.13 | | 8.543284 | 8.441536 | | | |
| 7 | 346.55 | 385.47 | | 8.079124 | 8.208979 | | | |
| 8 | 377.34 | 421.86 | | 8.195122 | 8.336617 | | | |

Workbench Layout

Collaborators

Graphic device selector

Graphics real coordinates

Graphic tools
Persistent collaborative
annotators + virtual laser
pointer + whiteboard + ..

Spreadsheet selector

Server-side, R-enabled
collaborative spreadsheet

The Elastic-R Java Workbench

The screenshot shows the Elastic-R Java Workbench interface within a Mozilla Firefox browser window. The browser address bar displays `https://www.elastic.net/portal/`. The interface is organized into several panels:

- Portal Info**: A dropdown menu.
- Working Directory**: A dropdown menu.
- Graphics Tools**: A dropdown menu.
- Workbench Layout**: A panel with various checkboxes for configuring the interface, including **Console & Graphics** (checked), **Editor & Browser**, **Spreadsheets**, **Java Plugins/Dashboard**, **Graphics Comparator**, **Panel Plugins/Dashboard**, **Java Bench** (checked), **Help Frame**, **Sliding**, **Frame Radio Mode**, **Panel Radio Mode**, **Panels Resizers**, **Menu Left/Right**, **Accordion Menu**, **Canvas If Available**, and **Use Signed Applets**.
- Console & Graphics**: A panel with a text area for R code and a "Submit" button. The code includes:

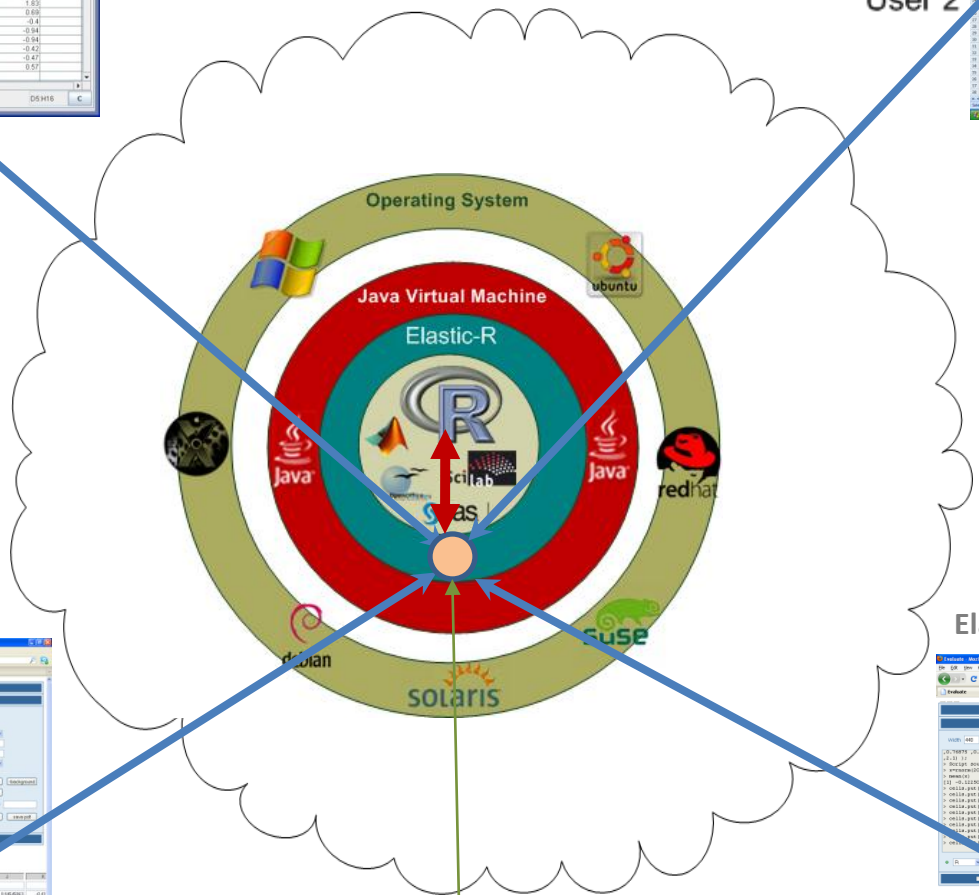
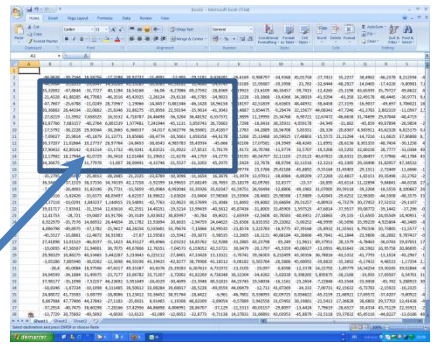
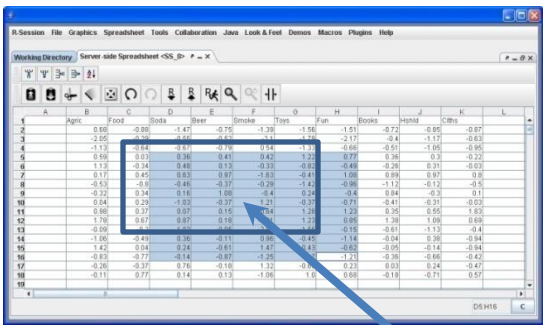
```
experimentData: use 'experimentData(object)'  
Annotation:  
> plot(exprs(kidney))  
> Script sourced to R
```
- Java Bench**: A panel containing:
 - R Console**: A text area showing the output of the R script, including details about the `primarydevice` and `featureData`.
 - Local Spreadsheet**: A table with 7 rows and 6 columns (A-F) containing numerical data.
 - Graphics**: A 3D plot titled `primarydevice` showing a terrain surface.
- Collaborators**: A dropdown menu at the bottom left.

The "Signed" status is indicated at the bottom left, and a "Show Direct URL" button is at the bottom right.

Software + services = applications convergence + collaboration

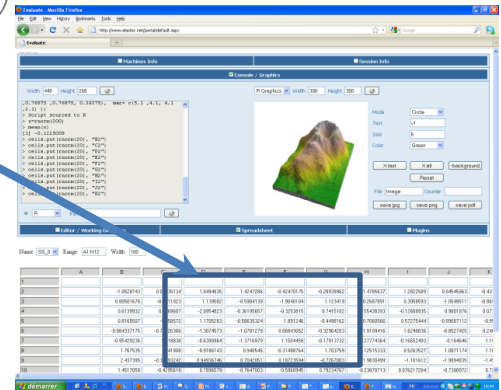
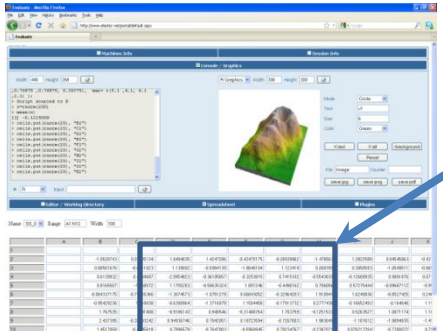
Elastic-R Java Workbench

Microsoft Excel



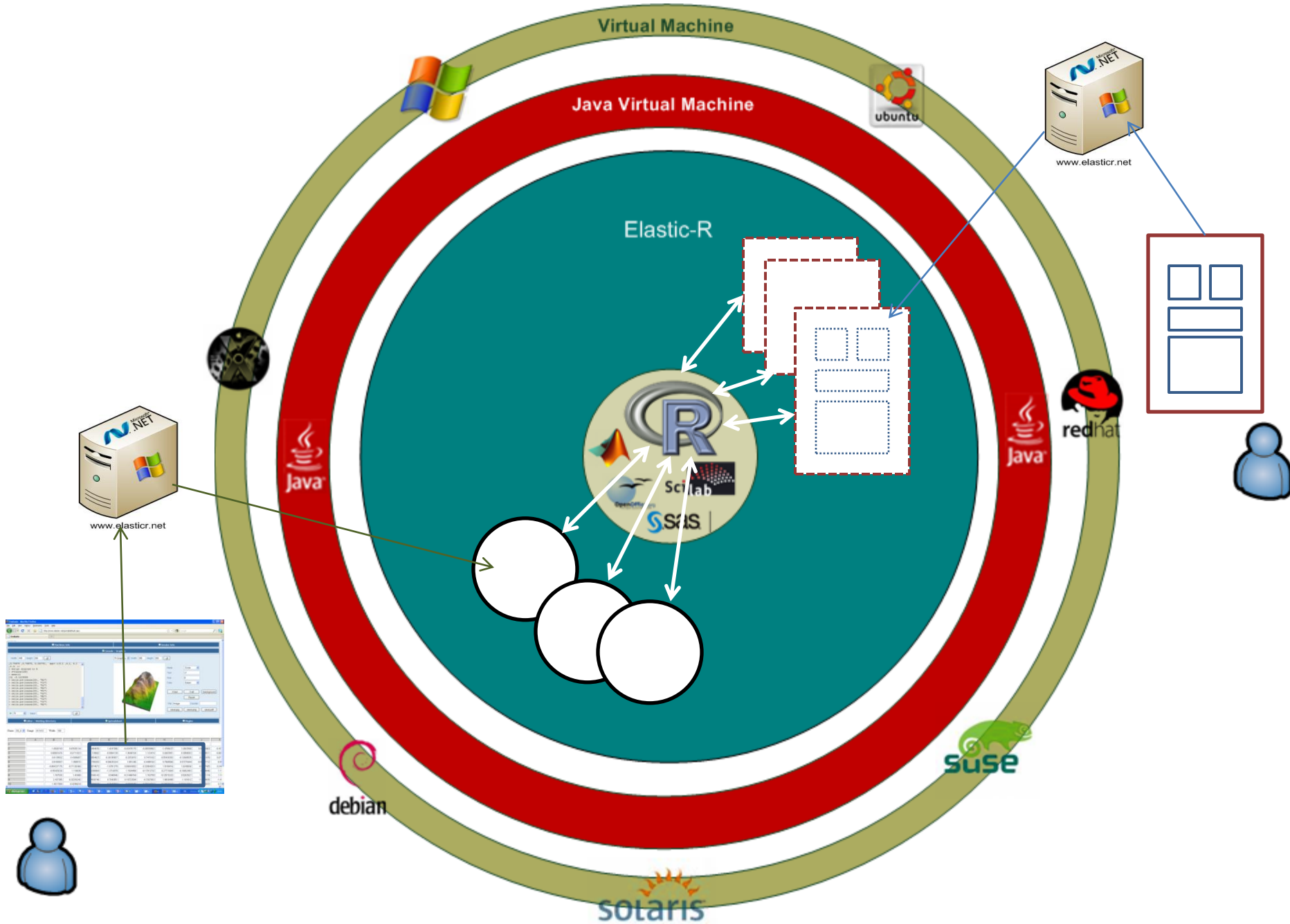
Elastic-R AJAX Workbench

Elastic-R AJAX Workbench



Elastic-R Spreadsheet model

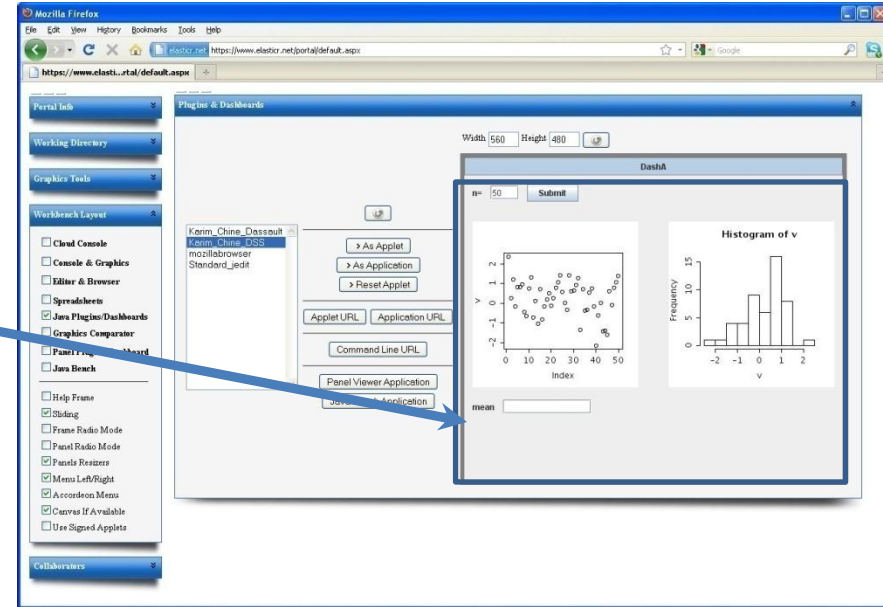
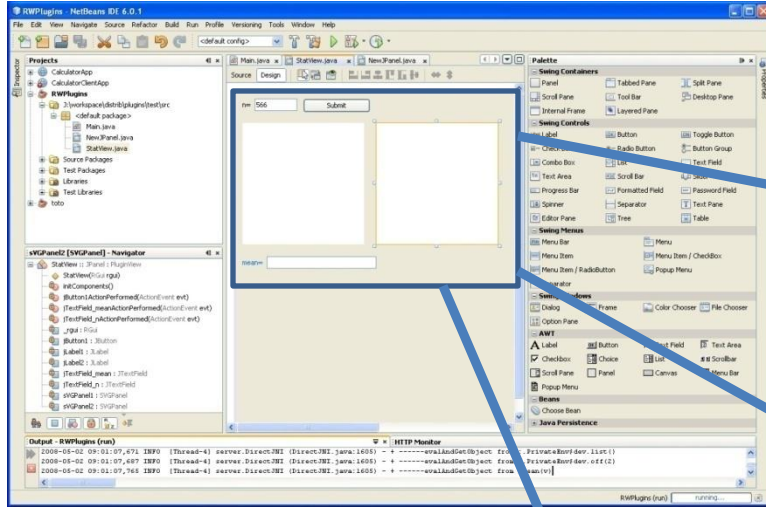
The Elastic-R server-side spreadsheet models / GUI widgets



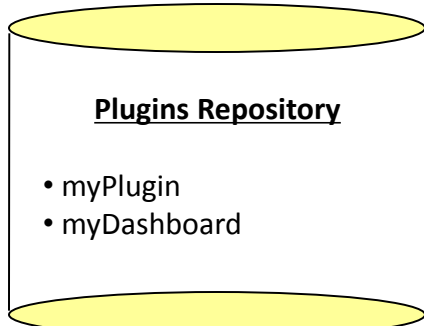
The cloud applications factory

Elastic-R AJAX Workbench

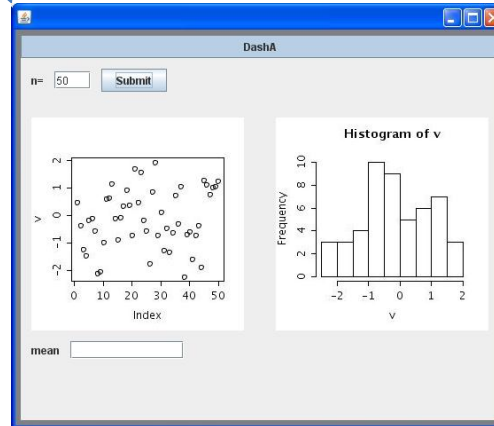
Visual Graphic User Interface Builder



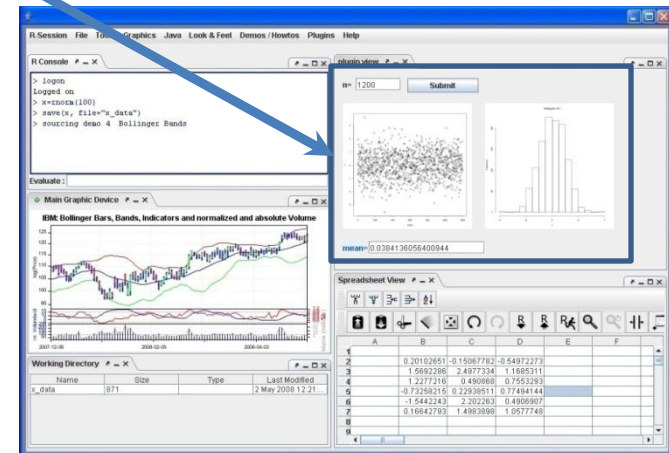
Upload plugin



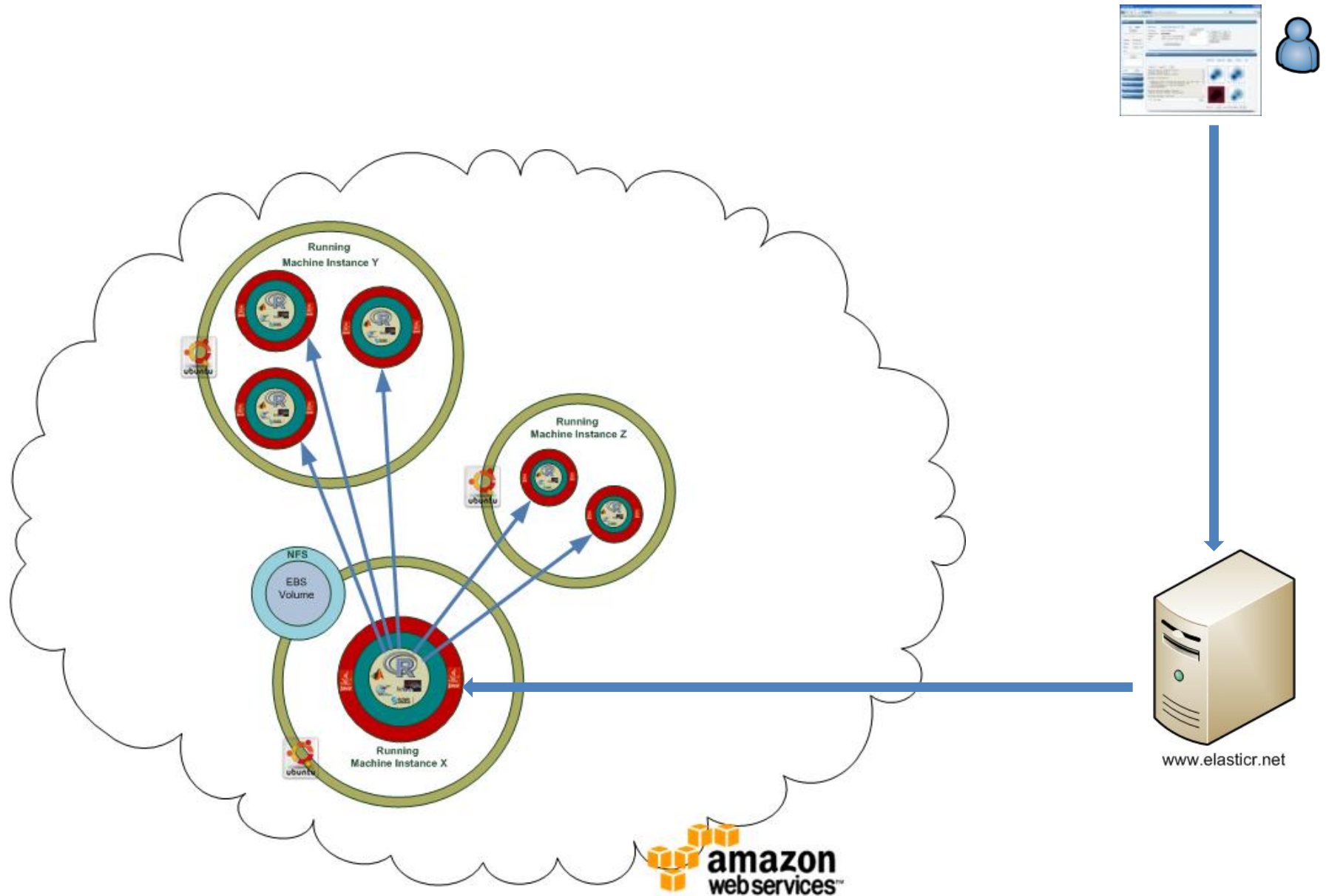
Standalone Application Accessible From a URL



Elastic-R Java Workbench



Elastic-R : user-friendly distributed computing platform



Demo

Useful links

Elastic-R Portal :

www.elasticr.net

Platform Web Site:

www.elasticr.net/platform

Articles :

Karim Chine, "Open Science in the Cloud: Towards a Universal Platform for Scientific and Statistical Computing", Chapter 19 in "Handbook of Cloud Computing", Springer, 2010 (in press)

Karim Chine, "Scientific Computing Environments in the age of virtualization, toward a universal platform for the Cloud" pp. 44-48, 2009 IEEE International Workshop on Opensource Software for Scientific Computation (OSSC), 2009

Karim Chine, "Biocep, Towards a Federative, Collaborative, User-Centric, Grid-Enabled and Cloud-Ready Computational Open Platform," *esience*, pp.321-322, 2008 Fourth IEEE International Conference on eScience, 2008

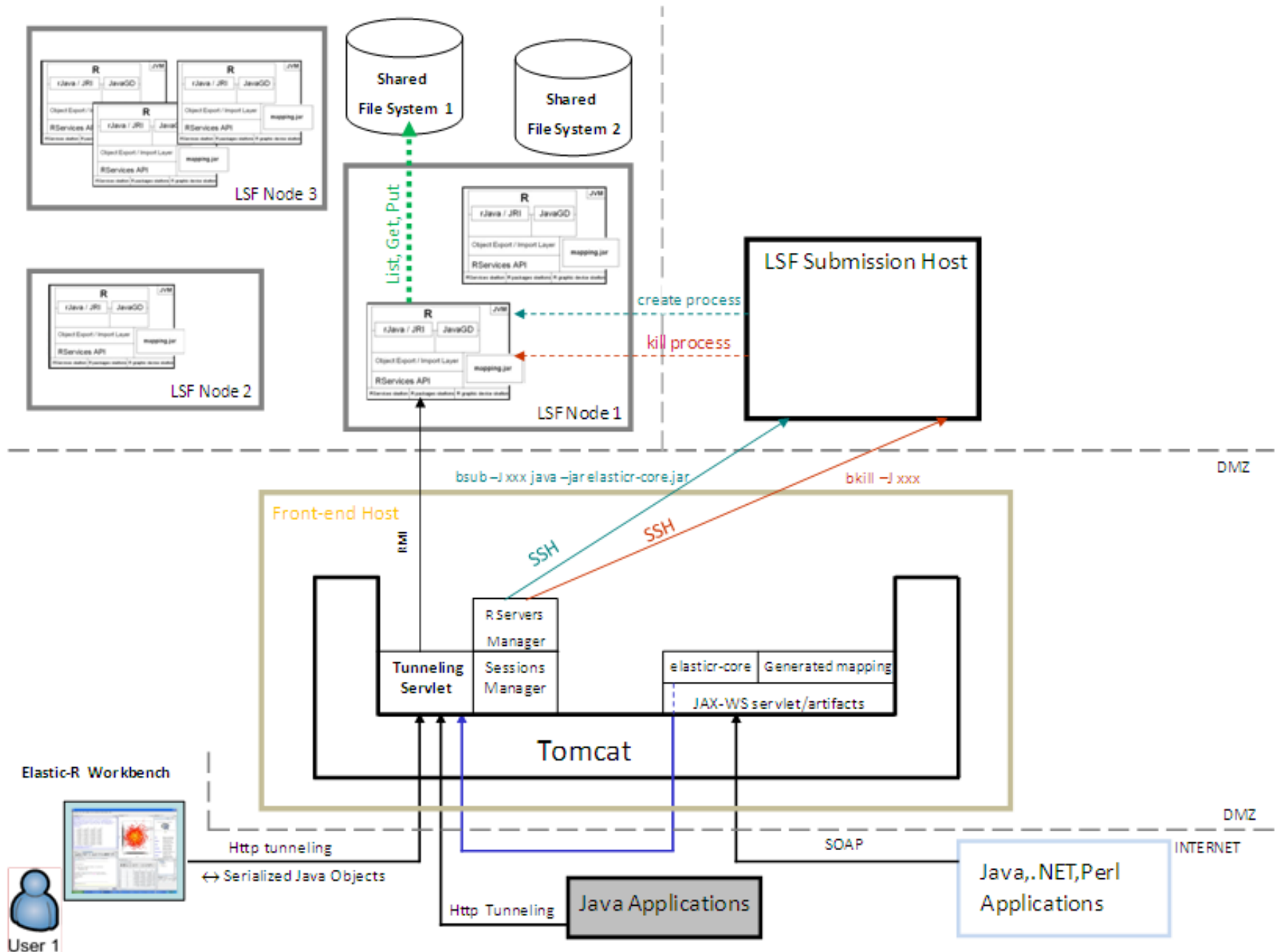
Linkedin Group:

<http://www.linkedin.com/groups?home=&gid=2345405>

Acknowledgments

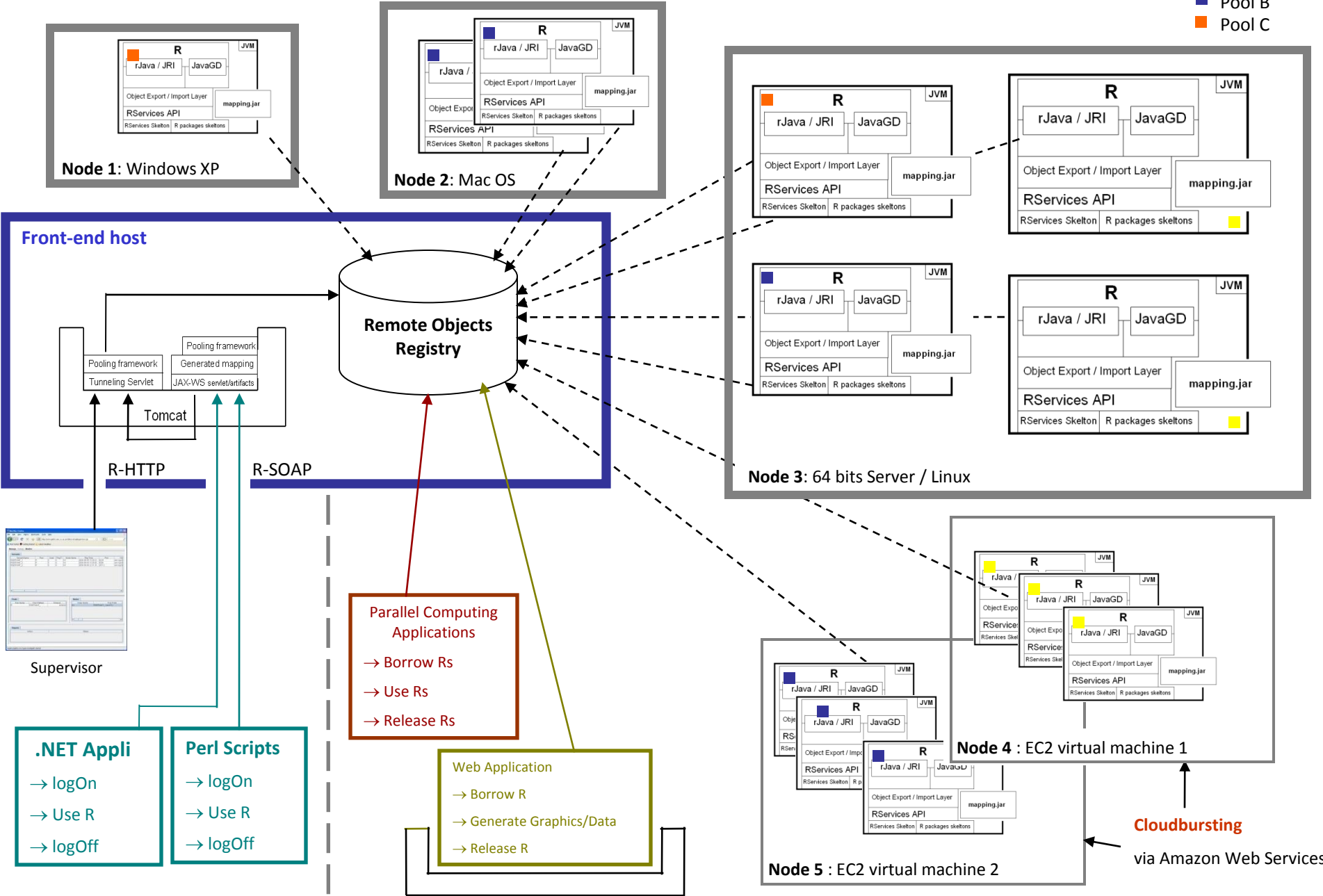
ACS: Madi Nassiri **Amazon:** Simone Brunozzi, Deepak Singh **AT&T Research Labs:** Simon Urbanek **ATUGE:** Imen Essafi, Béchir Tourki, Ilyes Gouja, HatemHachicha, Amine Elleuch **Auckland Centre for eResearch:** Nick Jones **Banca d'Italia:** Giuseppe Bruno **Bio-IT World:** Kevin Davies **BNP Paribas:** Ousseynou Nakoulima **Cambridge Healthtech Institute:** Cindy Crowninshield **City University of New York:** Mario Morales, Makram Talih **Columbia University:** Omar Besbes **Dassault Systèmes:** Omri Ben Ayoun, Patrick Johnson **Dataspora:** Michael E. Driscoll **EDF:** Alejandro Ribes **EBI:** Alvis Brazma, Wolfgang Huber, Kimmo Kallio, Misha Kapushesky, Michael Kleen, Alberto Labarga, Philippe Rocca-Serra, Ugis Sarkans, Kirsten Williams, Eamonn Maguire **EPFL:** Darlene Goldstein **ESPRIT:** Farouk Kammoun, Tahar. Benlakhdar **e-Taalim:** Nadhir Douma **ETH Zürich:** Yohan Chalabi, Diethelm Würtz, Martin Mächler **European Commission:** Konstantinos Glinos, Enric Mitjana, Monika Kacic, Ioannis Sagias **FHCRC:** Martin Morgan, Nianhua Li, Seth Falcon **Google:** Olivier Bosquet **FVG LLC:** Lisa Wood **Harvard University:** Tim Clark, Sudeshna Das, Douglas Burke, Paolo Ciccarese **IBM:** Jean-Louis Bernaudin, Pascal Sempe, Loic Simon, Lea A Deleris, Alex Fleischer, Alain Chabrier **Imperial College London:** Asif Akram, Vasa Curcin, John Darlington, Brian Fuchs **Indiana University:** Michael Grobe **INRIA:** David Monteau, Christian Saguez, Claude Gomez, Sylvestre Ledru **JISC:** John Wood, David Flanders **Johnson & Johnson - Janssen Pharmaceutica:** Patrick Marichal **KXEN:** Eric Marcade **Lancaster University:** Robert Crouchley, Daniel Grose **Leibniz Universität Hannover:** Kornelius Rohmeier **LIAMA:** Baogang Hue, Kang Cai **Limagrain:** Zivan Karaman **Mekentosj:** Alexander Griekspoor, Matt Wood **Microsoft:** Eric Le Marois, Tony Hey **Mubadala:** Ghazi Ben Amor **Nature Publishing Group:** Ian Mulvany, Steve Scott **NCeSS:** Peter Halfpenny, Rob Procter, Marzieh Asgari-Targhi, Alex Voss, YuWei Lin, Mercedes Argüello Casteleiro, Wei Jie, Meik Poschen, Katy Middlebrough, Pascal Ekin, June Finch, Farzana Latif, Elisa Pieri, Frank O'Donnell **New York Java User Group:** Frank D Greco **OeRC:** Dimitrina Spencer, Matteo Turilli, David Wallom, Steven Young **OMII-UK:** Neil Chue Hong, Steve Brewer **OpenAnalytics:** Tobias Verbeke **Oracle:** Dominique van Deth, Andrew Bond **OSS Watch:** Ross Gardler **Platform Computing:** Christopher Smith **Royal Society:** James Wilsdon **San Diego Supercomputer Center:** Nancy R. Wilkins-Diehr **Sanger Institute:** Lars Jorgensen, Phil Butcher **Shell:** Wayne.W.Jones, Nigel Smith **Société Générale:** Anis Maktouf **Stanford University:** John Chambers, Balasubramanian Narasimhan, Gunter Walther **SYSTEM@TIC:** Karim Azoum **Technische Universität Dortmund:** Uwe Ligges, Bernd Bischl **Technoforge:** Pierre-Antoine Durgeat **Tekiano:** Samy Ben Naceur **Télécom-ParisTech:** Isabelle Demeure, Georges Hebrail, Nesrine Gabsi **The Generations Network:** Jim Porzak **Total:** Yannick Perigois **Tunisian Ministry of Communication Technologies:** Naceur Ammar, Lamia Chaffai-Sghaier, Mohamed Saïd Ouerghi, Syrine Tlili **Tunisian Ecole Polytechnique:** Riadh Robbana **UC Berkeley:** Nouredine El Karoui, Terry Speed **UC Davis:** Rudy Beran, Debashis Paul, Duncan Temple Lang **UCL:** Daniel Jeffares **UCLA:** Ivo Dinov, Jeroen Ooms **UC San Diego:** Anthony Gamst **UCSF:** Tena Sakai **Université Catholique de Louvain:** Christian Ritter **University of Cambridge:** Ian Roberts, Robert MacInnis Peter Murray-Rust, Jim Downing **University of Manchester:** Carole Goble, Len Gill, Simon Peters, Richard D Pearson, Iain Buchan, John Ainsworth **University of Plymouth:** Paul Hewson **University of Split:** Ivica Puljak **UTK:** Ajay Ohri **World Bank Group-IFC:** Oualid Ammar **Yahoo:** Laurent Mirguet, Rob Weltman **Independant:** Charles Dallas, Romain François

Elastic-R for clusters/grids

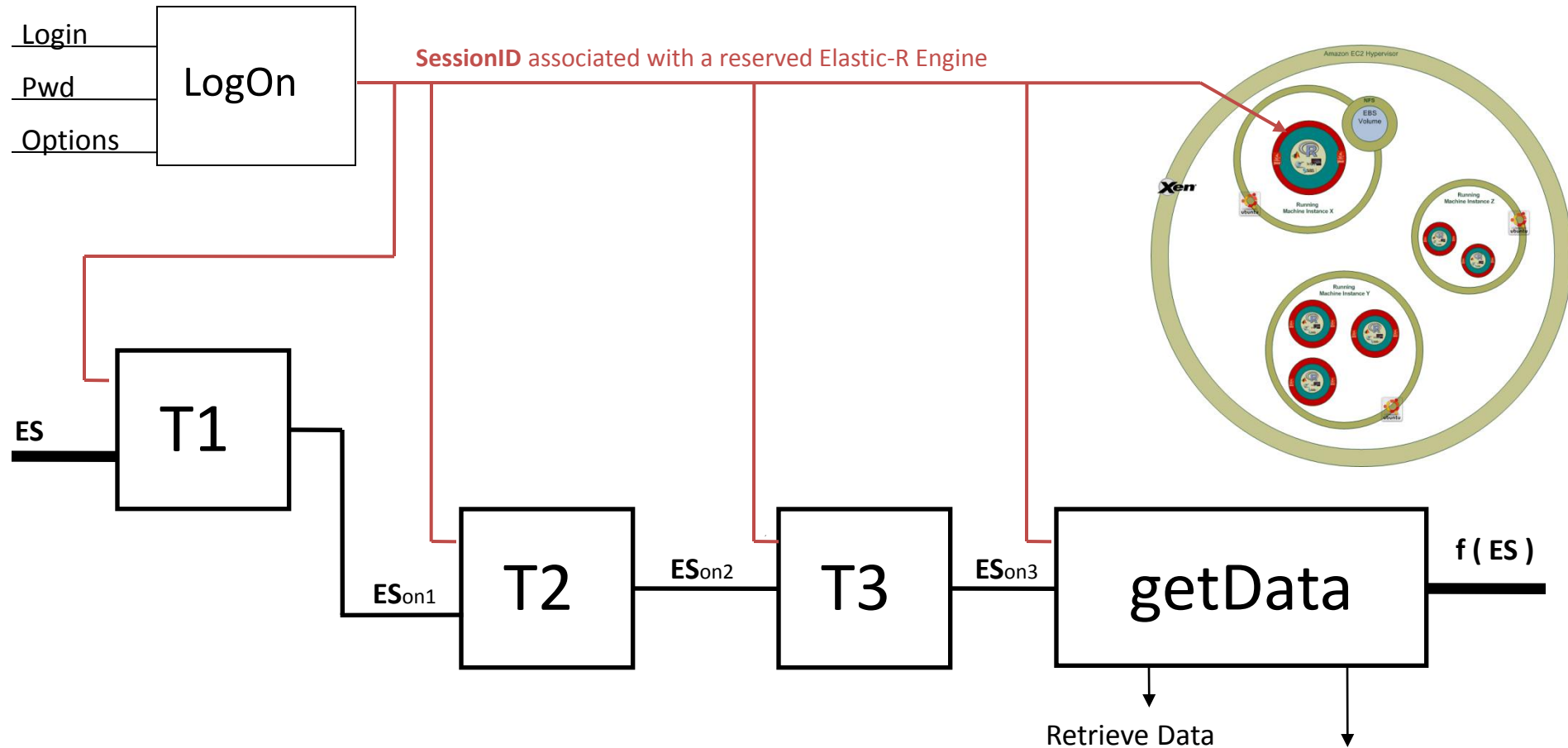


Elastic-R SOA platform

■ Pool A
■ Pool B
■ Pool C



Stateful generated Web Services : Elastic-R for workflow workbenches



T1,T2,T3 : Generated Stateful Web Services for R functions T1,T2 & T3

LogOn, getData : R-SOAP methods

ES : ExpressionSet

ESon1, ESon2, ESon3 : ExpressionSet Object Names

$f = T3 \circ T2 \circ T1$

- remove **ESonx**
- « Clean » Elastic-R Engine
- Put Elastic-R Engine back in the Pool
- kill Elastic-R Engine