Integrating R with Azure for High-throughput analysis

Hugh Shanahan

Department of Computer Science
Royal Holloway, University of London

hugh.shanahan@rhul.ac.uk
@HughShanahan

Applicability to other domains

- This project started out doing something very specific for the domain I work in (Computational Biology).
- I promise that there will be no Biology in this talk !!
- Realised can be extended to running high-throughput jobs in R.
- Contrast with MapReduce / R formalisms (`HadoopStreaming`, `Rhipe`, Revolution Analytics, ...) - parallelisation happens outside of individual R script.

Applicability to other domains

- This project started out doing something very specific for the domain I work in (Computational Biology).
- I promise that there will be no Biology in this talk !!
- Realised can be extended to running high-throughput jobs in R.
- Contrast with MapReduce / R formalisms (`HadoopStreaming`, `Rhipe`, Revolution Analytics, ...) - parallelisation happens outside of individual R script.

Applicability to other domains

- This project started out doing something very specific for the domain I work in (Computational Biology).
- I promise that there will be no Biology in this talk !!
- Realised can be extended to running high-throughput jobs in R.
- Contrast with MapReduce / R formalisms (`HadoopStreaming`, `Rhipe`, Revolution Analytics, ...) - parallelisation happens outside of individual R script.

Applicability to other domains

- This project started out doing something very specific for the domain I work in (Computational Biology).
- I promise that there will be no Biology in this talk !!
- Realised can be extended to running high-throughput jobs in R.
- Contrast with MapReduce / R formalisms (`HadoopStreaming`, `Rhipe`, Revolution Analytics, ...) - parallelisation happens outside of individual R script.

IaaS clouds

- We all now know what clouds are !
- Infrastructure as a Service (IaaS)
- Access Virtual Machine via the command line
- Amazon, Rackspace, OpenStack ...

IaaS clouds

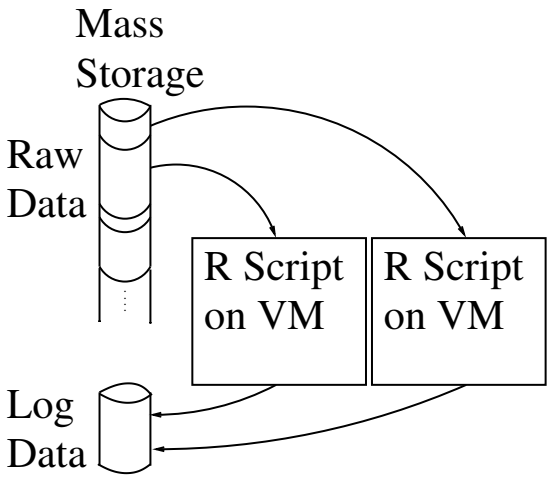
- Platform as a Service (PaaS)
- Access Virtual Machine programatically.
- Explicitly allows for batch control, more complicated workflows etc.

Microsoft Azure and Generic Worker Libraries

- Azure offers both IaaS and PaaS.
- IaaS VM's can run a variety of different flavours of Linux and Windows OS's
- PaaS (they refer to this as a Cloud Service) only runs Windows Server.
- Mass Storage (not storage associated with VM).
- Programatic access is via ASP.NET and C#
- Access mass storage via a variety of languages.
- Set of libraries which allow control of jobs running on VM's.
- Generic Worker (GW)

Scaling up

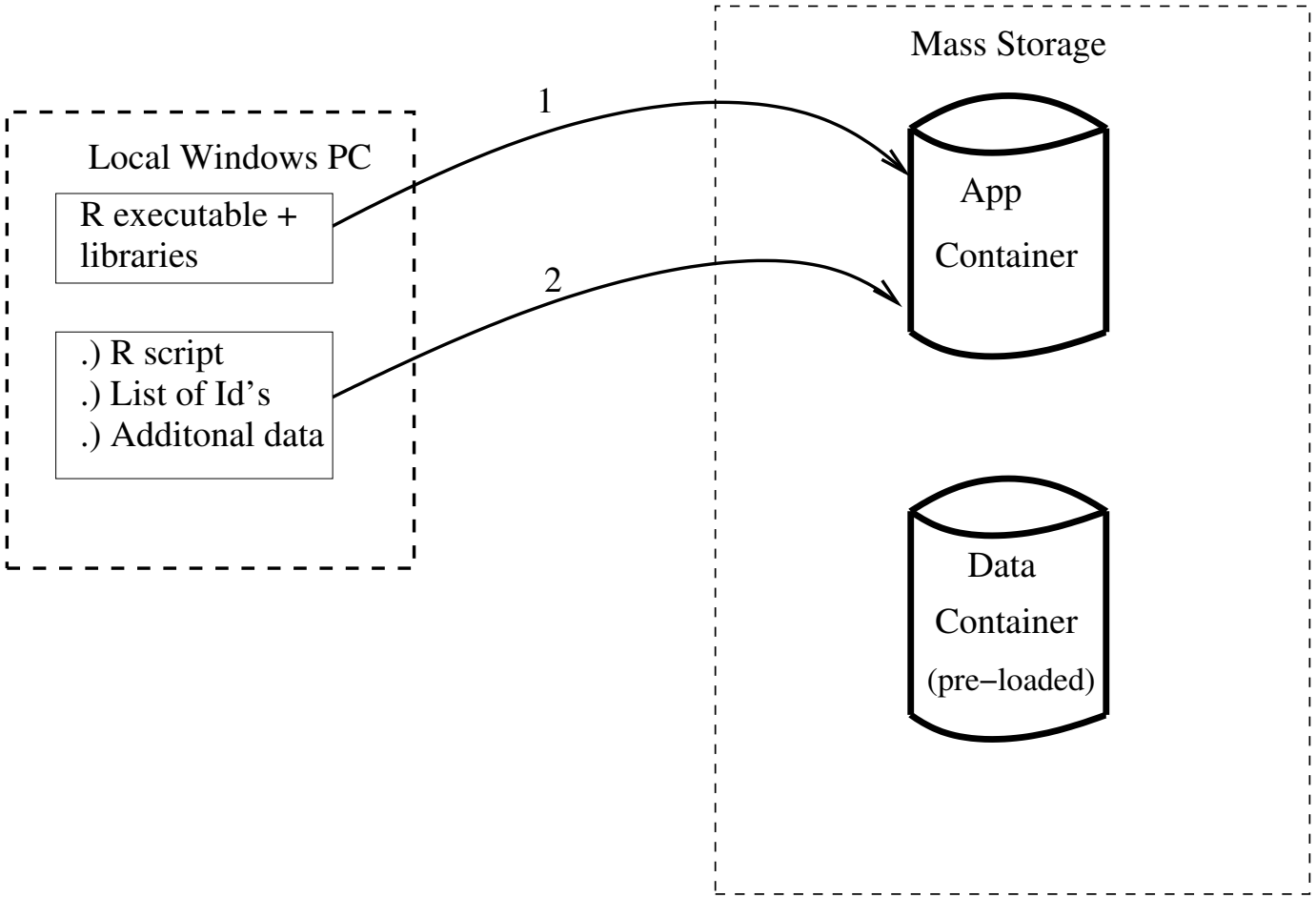
- Needed to scale up a problem based on six data sets to nearly six hundred (100 Mbyte $\rightarrow$ 1 Tbyte).
- Calculations based on an R script.
- Each data set can be analysed one at a time (batch mode).
- Individual data sets can vary by two orders of magnitude.



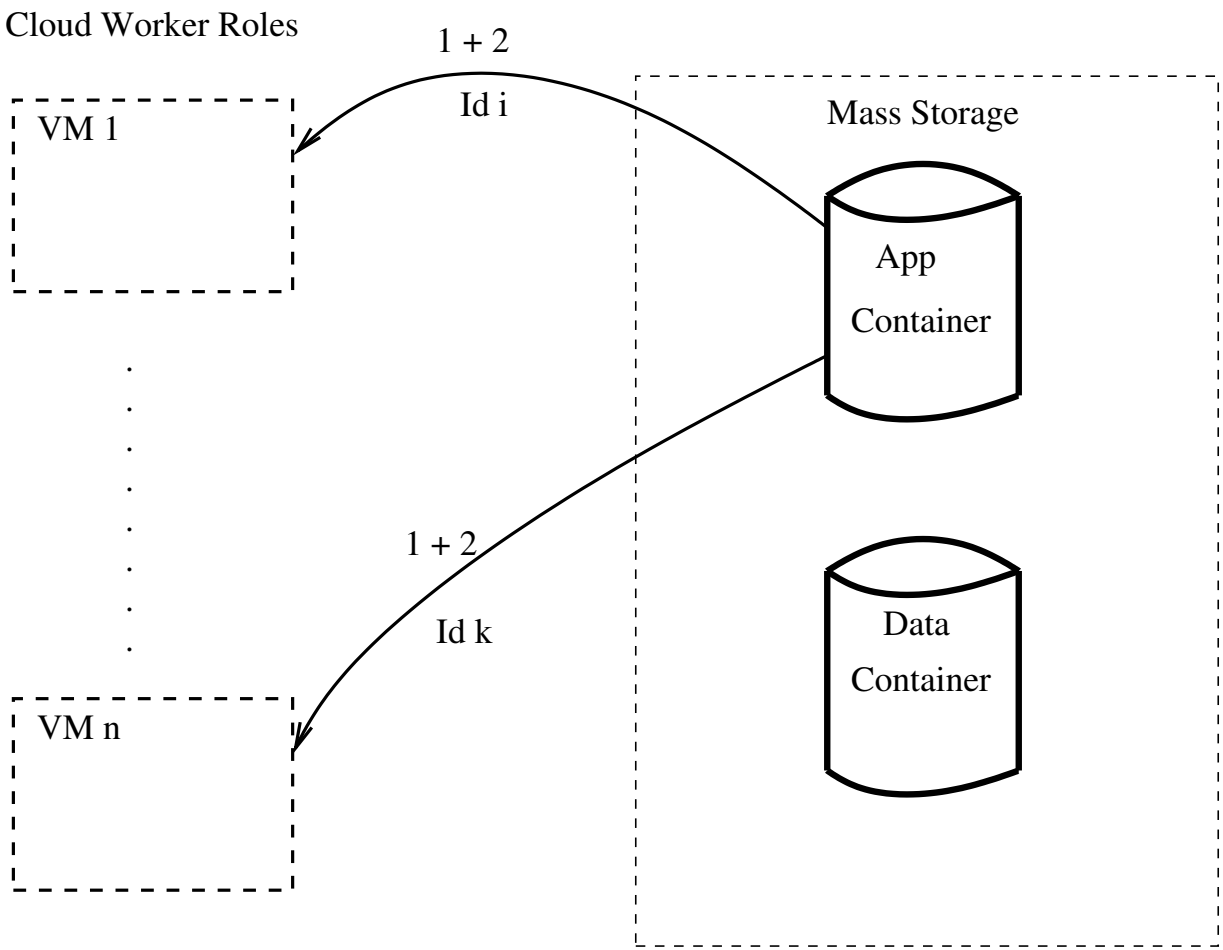
Implementation

- Made use of Azure PaaS with GW libraries.
- Written using a combination of C# and Java.
- R executables + library uploaded to mass storage.
- Data to be analysed placed in separate container of mass storage.
- R script uploaded at run time.

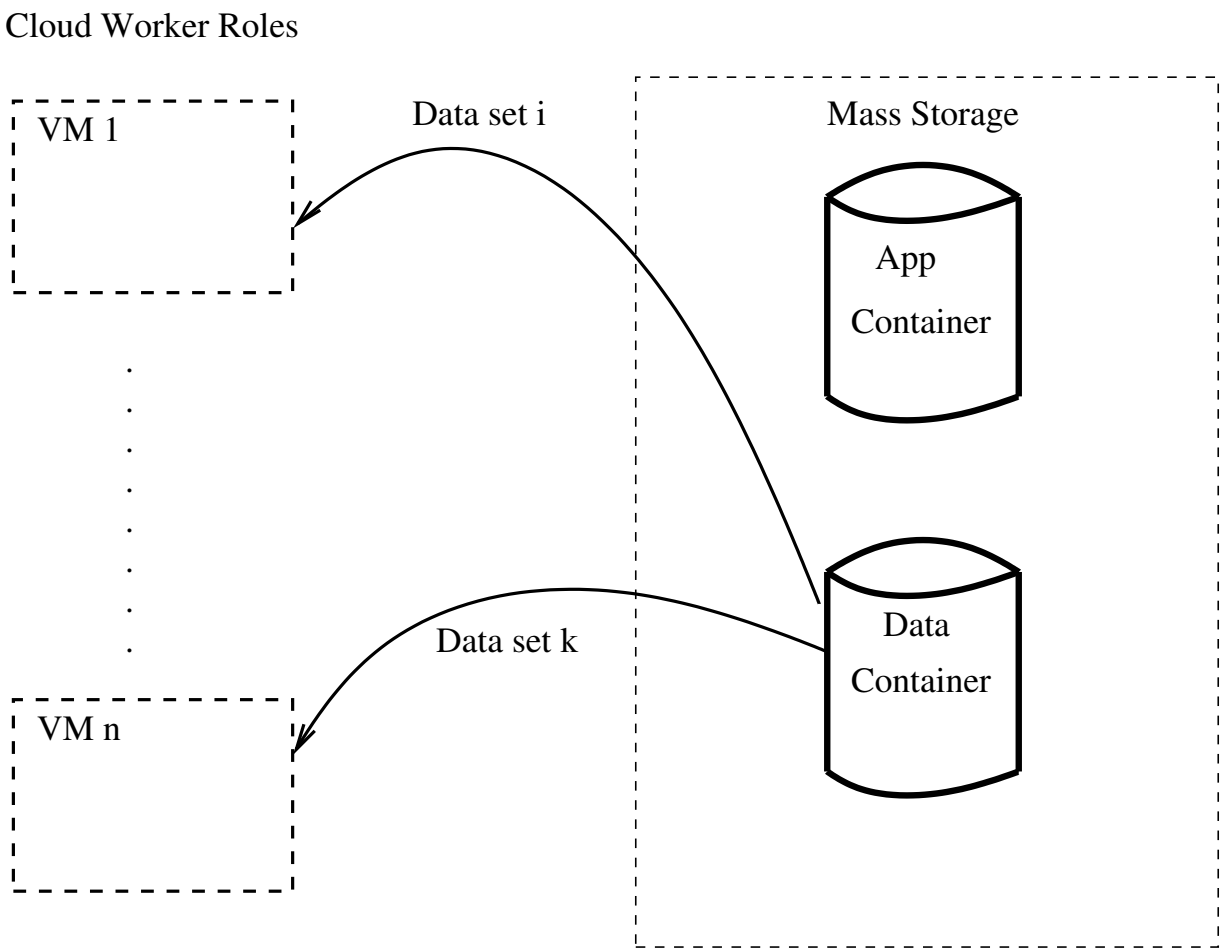
Operation


Mass Storage
Local Windows PC
R executable + libraries
.) R script
.) List of Id's
.) Additonal data
1
2
App Container
Data Container (pre−loaded)

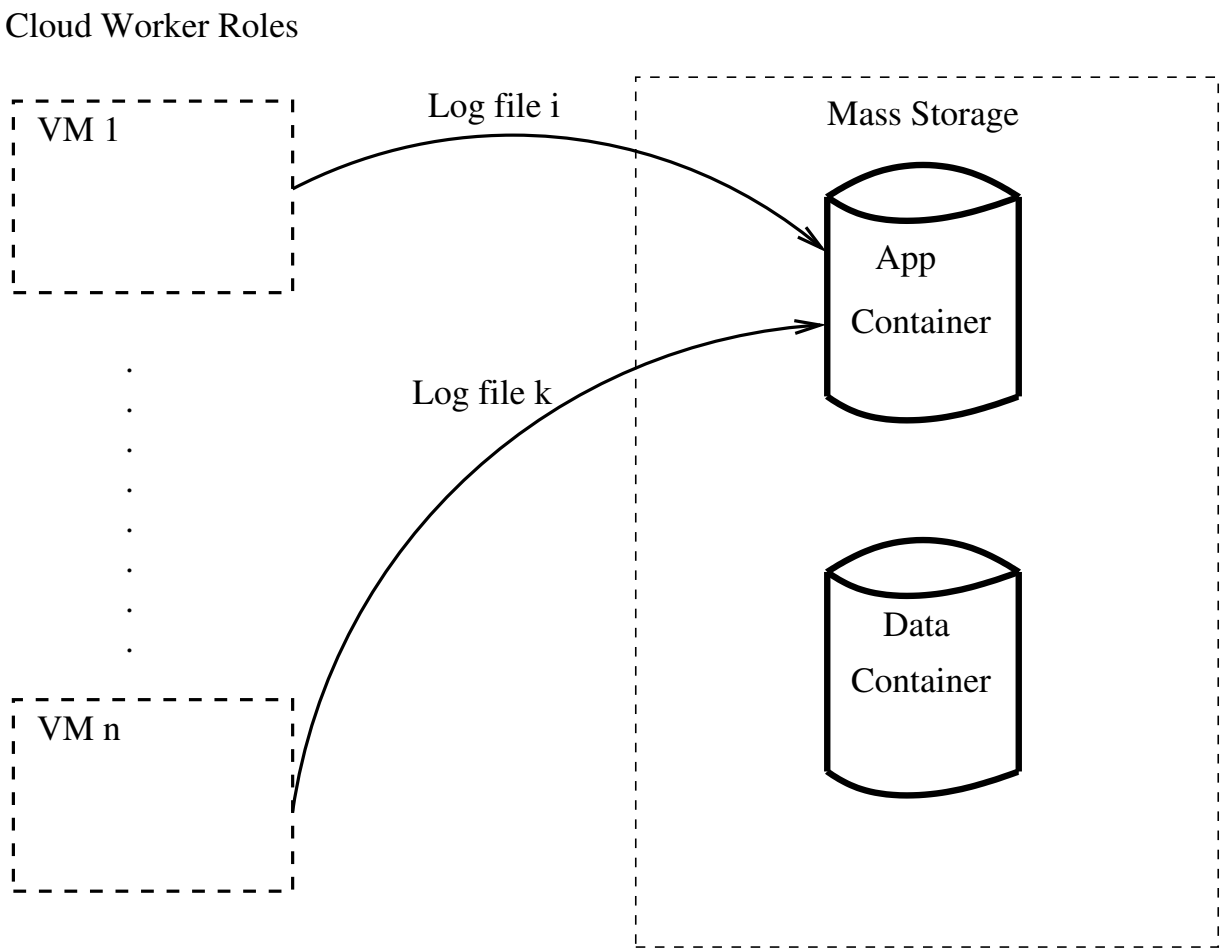
Launching


Cloud Worker Roles
1 + 2
Id i
VM 1
Mass Storage
App
Container
1 + 2
Id k
Data
Container
VM n

Running


Cloud Worker Roles
VM 1
Data set i
Mass Storage
App
Container
Data set k
Data
Container
VM n

Logging it all



In reality .....
this is less than 100 lines of C#

Extending to any R script

This can be extended to any case where

- you have data sets to be analysed by an R script,
- the data is analysed individually.
- Set of complex financial instruments
- Parameter sweeps

Extending to any R script

This can be extended to any case where

- you have data sets to be analysed by an R script,
- the data is analysed individually.
- Set of complex financial instruments
- Parameter sweeps

Key issues to fix this Summer

- Getting set up (configuration files and keys).
- Adding GUI.
- `https://github.com/hughshanahan/GWydiR`
- `https://github.com/hughshanahan/RAzureEssentials`
- Will port over to a more suitable github address for group development this Summer.

Conclusions

- C# and ASP.NET can be a learning curve for Linux users.
- Nonetheless PaaS explicitly allows control of VM's.
- Batch mode implementation for a specific problem.
- Allows analysis on Tbyte-sized data set
- Modified to run any R script in batch mode - much more general.

Shameless Plug

M.Sc. in Data Science and Analytics
M.Sc. in Machine Learning
M.Sc. in Computational Finance
All starting this year at Royal Holloway.
Please go to
`http://bit.ly/1418DOS`
for further details.

Shameless Plug

M.Sc. in Data Science and Analytics
M.Sc. in Machine Learning
M.Sc. in Computational Finance
All starting this year at Royal Holloway.
Please go to
`http://bit.ly/1418DOS`
for further details.

Acknowledgments

Andrew (Harry) Harrison

University of Essex

Anne Owen

University of Essex

Venus-C
Virtual multidisciplinary EnviroNments USing Cloud infrastructures

Funded by Venus-C EU Network

Contact `hugh.shanahan@rhul.ac.uk`
`@hughshanahan`

Thank you for your time !