Embedded systems for global e-Social Science

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

Title: Embedded systems for Global e-Social Science: Moving Computation rather than Data

Author names and affiliations

Ashley D Lloyd[1]
Terence M Sloan[2*]
Mario Antonioletti[3]
G.A. McGilvary[4]

[*]Corresponding Author

[1]Business School
The University of Edinburgh
29 Buccleuch Place
Edinburgh, EH8 9JS, UK
Email: ashley@edinburgh.ac.uk

[2]EPCC
The University of Edinburgh
James Clerk Maxwell Building
Mayfield Road
Edinburgh, EH9 3JZ, UK
Email: tms@epcc.ed.ac.uk
Tel: +44 (0) 131 650 5155
Fax: +44 (0) 131 650 6555

[3]EPCC
The University of Edinburgh
James Clerk Maxwell Building
Mayfield Road
Edinburgh, EH9 3JZ, UK
Email: mario@epcc.ed.ac.uk

[4]Edinburgh Data-Intensive Research Group,
School of Informatics,
The University of Edinburgh,
Edinburgh, EH8 9AB, UK
Email: gary.mcgilvary@ed.ac.uk

Author Contributions

Ashley Lloyd was the Principal Investigator on much of the funding that contributed to this paper. Ashley guided the direction of the reported research as well as participating fully in its activities. Ashley contributed directly to the writing of this paper.

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

Terence Sloan was the Co-Principal Investigator on some of the funding sources that contributed to this paper. Terence participated directly on the reported research in both Project Manager and technical analyst roles. Terry contributed directly to the writing of this paper.

Mario Antonioletti participated directly on the reported research with the Chinese commercial organisation both in the setting up of the required research infrastructure and in conducting data analyses. Mario contributed directly to the writing of this paper.

Gary McGilvary participated directly on the reported cloud experiments between the UK and Thailand. Gary contributed directly to the writing of this paper.

All authors have approved the final article.

Author Vitae

Ashley Lloyd has researched widely in the emerging technologies arena, from fundamental materials research to innovative ICT applications. Publications range from the IEEE Journal of Quantum Electronics to the International Journal of Innovation Management (top three articles cited >300 times). Research support received from industry and research councils in 4 continents.

Terence Sloan is a Software Development Group Manager at EPCC. He has extensive experience of managing projects in high performance computing and distributed computing with both academia and business in the UK, Europe, Asia and Australia. He has published more than 40 articles in journals and conference proceedings.

Mario Antonioletti works as a Software Architect at EPCC. Mario has a Mathematics and Physics background and a PhD in astrophysics. He has been involved in various HPC projects as well as the OGSA-DAI project (accessing and integrating databases using web services) and standards-based work at the Open Grid Forum (OGF).

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

Gary McGilvary is studying for a Ph.D. in the Edinburgh Data-Intensive Research Group within the School of Informatics at the University of Edinburgh. His research investigates how to create, deploy and manage *ad hoc* clouds easily and effectively upon an organization's non-exclusive infrastructure.

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

Role of the funding Source

This trans-national work would not have been possible without the sustained investment in building the INWA Grid over the last decade, both as a technological infrastructure capable of secure, distributed, cooperative analysis of large datasets, but also as a set of trusted working relationships that enables commercially sensitive data to be shared.

We gratefully acknowledge the support of the UK Economic and Social Research Council (award RES-149-25-0005) for the initial phase of the INWA Grid and its 'Follow-On Funding' (award RES-189-25-0039); the UK EPSRC (award EP/H006753/1 on "Building Relationships with the 'Invisible' in the Digital (Global) Economy") for supporting the reported work with collaborators in China and Thailand; the Scottish Funding Council (edikt2 grant HR04019); the Australian Research Council in partnership with Singapore Telecom for its support (awards LP0454322 and SR0567388) and the endowed SingTel Optus Chair of eBusiness held by Lloyd at Curtin University; Sun Microsystems and the Australian Academic and Research Network (AARNet) for continued support since the INWA Grid became operational in 2003, and colleagues at the Computer Network and Information Center of the Chinese Academy of Sciences who have enabled and hosted the connections within China since 2004. The Biotechnology and Biological Sciences Research Council [grant number BB/J019283/1] supported the writing of this paper.

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

Abstract

There is a wealth of digital data currently being gathered by commercial and private concerns that could supplement academic research. To unlock this data it is important to gain the trust of the companies that hold the data as well as showing them how they may benefit from this research. Part of this trust is gained through established reputation and the other through the technology used to safeguard the data. This paper discusses how different technology frameworks have been applied to safeguard the data and facilitate collaborative work between commercial concerns and academic institutions. The paper focuses on the distinctive requirements of e-Social Science: access to large-scale data on behaviour in society in environments that impose confidentiality constraints on access. These constraints arise from both privacy concerns and the commercial sensitivities of that data. In particular, the paper draws on the experiences of building an intercontinental Grid - INWA - from its first operation connecting Australia and Scotland to its subsequent extension to China across the Trans-Eurasia Information Network - the first large-scale research and education network for the Asia-Pacific region. This allowed commercial data to be analysed by experts that were geographically distributed across the globe. It also provided an entry point for a major Chinese commercial organization to approve use of a Grid solution in a new collaboration provided the centre of gravity of the data is retained within the jurisdiction of the data owner. We describe why, despite this approval, an embedded solution was eventually adopted. We find that ‘data sovereignty’ dominates any decision on whether and how to participate in e-Social Science collaborations and how this might impact on a Cloud based solution to this type of collaboration.

Keywords

Grid, Cloud, Social Science, Data Sovereignty

Highlights

- digital data gathered by commercial organisations can supplement academic research
- trusted access to the supplied, commercial data in an acceptable secure environment is crucial
- reports UK(academic)-China(industry) collaboration on millions of consumer transactions
- builds on a decade of Grid work in China but rejects Cloud/Grid for embedded system
- data sovereignty key to research collaborations with companies from different countries

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

1. Introduction

"Social science is, in its broadest sense, the study of society and the manner in which people behave and impact on the world around us."

UK Economic and Social Research Council, 2005
www.esrcsocietytoday.ac.uk

Understanding this behaviour, modeling its evolution and accurately forecasting future patterns is key to the formulation of a successful investment policy for both businesses and government, and hence key to the competitiveness of nations. For social sciences researchers accessing the quality and volume of data required to build these models however has always been difficult. This was highlighted by the Nobel Laureate for Economics Clive Granger, where, during his keynote address to the International Institute of Forecasters in 2004 [1], he noted that improvements in forecasting techniques over his career had been largely counter-balanced by a reduction in the quality of available data.

As noted by Halfpenny et al. [2] however, given the enormous amount of digital data that we generate as citizens in a digital world and the scale and variety of computing resources available to help researchers make sense of it, there are real opportunities to conduct social science that was impossible only a decade ago. Halfpenny et al. state that using routinely recorded transaction and administrative data potentially frees researchers from relying on relatively small sample surveys that ask people what they do. Instead, researchers can access enormous bodies of digital data about what people actually do, and when and where. Savage & Burrows [3] argue that only by embracing the opportunities that new forms of social data make possible can academic sociology be rescued from irrelevance, left behind by the numerous private and public sector organizations that are already exploiting this wealth of data to refine their products and services. However, in the authors' experience, to make the best use of this data, input is required from the commercial entities that produced it since knowledge of the business processes applied to the data are required to properly interpret it.

EPCC and the Business School at the University of Edinburgh have a long history (since 1994) of investigating ways to exploit commercial, operational data in social sciences research. These collaborations with commerce and industry have all been, by their very nature, data intensive with the data coming from transport [4],[5], automotive [6], financial [7] and telecommunications [8] businesses. Such collaborations have arisen due to a perceived likely mutual benefit – the researchers get access to new data sources to analyse while the companies have early access to the results of the analyses and the possibility of influencing the research focus. Originally the scope of the data involved in these collaborations lay at a UK, regional or national level but in recent years as shown in Figure 1, this has extended to an international dimension with projects spanning national and even continental boundaries [9].

In undertaking these collaborations, a key aspect has been the building of computational infrastructures, modes of operations and the establishing of data exchange practices that allow research to take place within an environment that is sufficiently secure for the businesses involved to be prepared to allow their data to be analysed. Regardless, commercial sensitivities and obfuscation of data to secure commercial information and processes, in some cases, imposes quite severe constraints on the ability to interpret and analyse the data. This can make a researcher's life more difficult than it would be if they were analysing data perceived to be of less commercial significance. With the transition

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

beyond national boundaries, the issue of data sovereignty has also arisen as an additional requirement and added further complications to the existing constraints. Whilst these constraints may sometimes apply equally well to trans-national scientific collaborations it is worth highlighting the issue that is common to the analysis of commercial data about behaviour in society: that the value of the insights cannot be assessed prior to discovery and may not be of equal value to all collaborators, making goal convergence more challenging.

In this paper we describe the approaches taken in some of these collaborations, to tackle a company's concerns and provide the researchers with a suitable environment. This paper shows that as technology has progressed, it has been possible to build global infrastructures for computation and collaboration between academic and industry collaborators that make research on this quality of data easier to conduct, provided it meets the security requirements of the data owners.

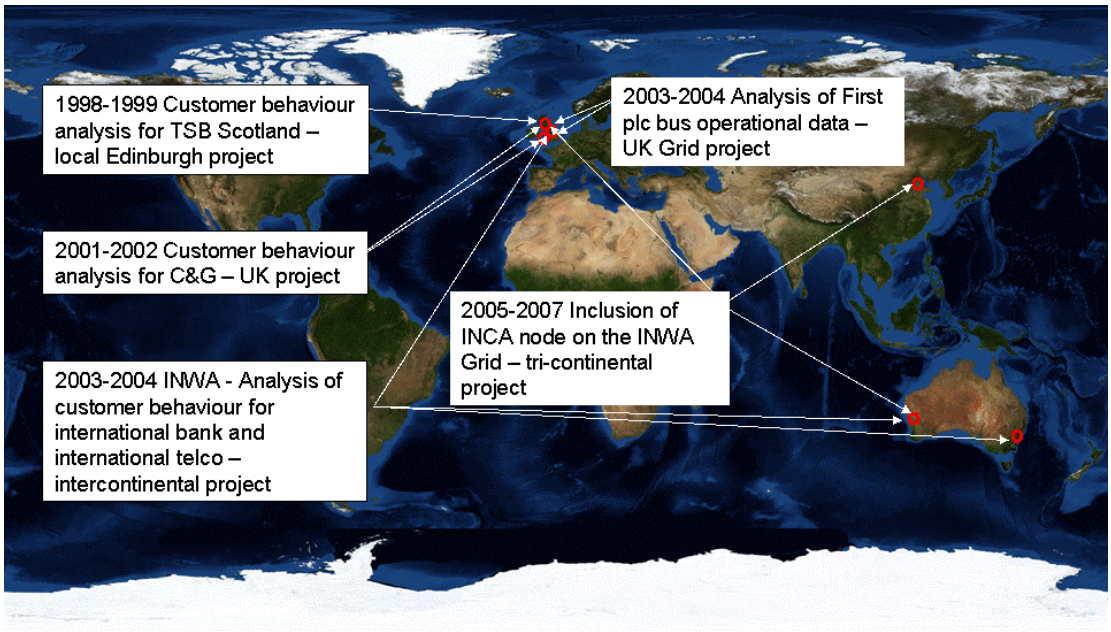

Figure 1: The location of some of the collaboration participants. Earth texture provided by NASA (visibleearth.nasa.gov)

2. Models of computation for e-Social Science

When the worth and necessary trust [10] for analysing commercial, operational data for social and scientific research has been established, the actual mechanisms for enabling researchers access to that data needs to be put in place.

For many companies, providing researchers with access to operational data on live production systems is not possible due to the potential impact on processes that are required for the company to operate. This combined with the possibility of inadvertent or even malicious interference with these production processes means that, understandably, companies are reluctant to provide live access even before other issues, such as security,

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

ethics and data protection, are taken into consideration. Moreover, production systems may not have the computational capability to undertake the types of analyses the researchers wish to do, whether that be in the form spare capacity, software licenses for the tooling required or even raw compute power.

The remaining options for providing access then generally centre around physically giving a snapshot of the data to the researcher. This allows the company to anonymise, obfuscate or remove any data that they might regard as sensitive before passing it to the researcher for them to host and analyse with their own compute platforms (i.e. moving the data to the computation) or using some form of remote access, for example by the company providing distributed access to some form of data warehouse that is under the company's control.

Figure 2 schematically represents the different operational paradigms discussed in this paper with regards to how a commercial-academic collaboration may be undertaken. These are:

- Traditional – as discussed above, where the commercial entity provides the academic researchers, or institution, with a commercially sanitised version of the data to be used by the academic parties.
- Embedded – if there are data sovereignty issues or where the data is deemed to be commercially sensitive, one possibility is to move a sufficiently large computational resource to within the company's domain where any research has to be undertaken. This allows the company to monitor what is happening to their data a little more closely and gives them ultimate access control. It however does grant an external entity (i.e. an academic researcher) access to the company's internal IT infrastructure. This requires the researcher to be isolated in some way to minimise any risk to the organisation itself.
- Grid – a virtual organisation can be created in order to share resources. In these circumstances the location of the data or computational resource can reside in either organisation but it does require the virtual organisation to be set up which is not a small undertaking.
- Cloud – the provision of third party hosting of computing power and hosting capabilities offers yet another way of establishing such an undertaking. Clearly, widening the trust circle increases the risk of data leakage. In some instances the virtualising of a resource can itself become an issue due to the increase in legislation regarding the exporting of personal data outside a region. This may impose strong requirements with regards to the localisation of the data.

The remainder of this paper sets out to describe particular instances of each of these models.

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

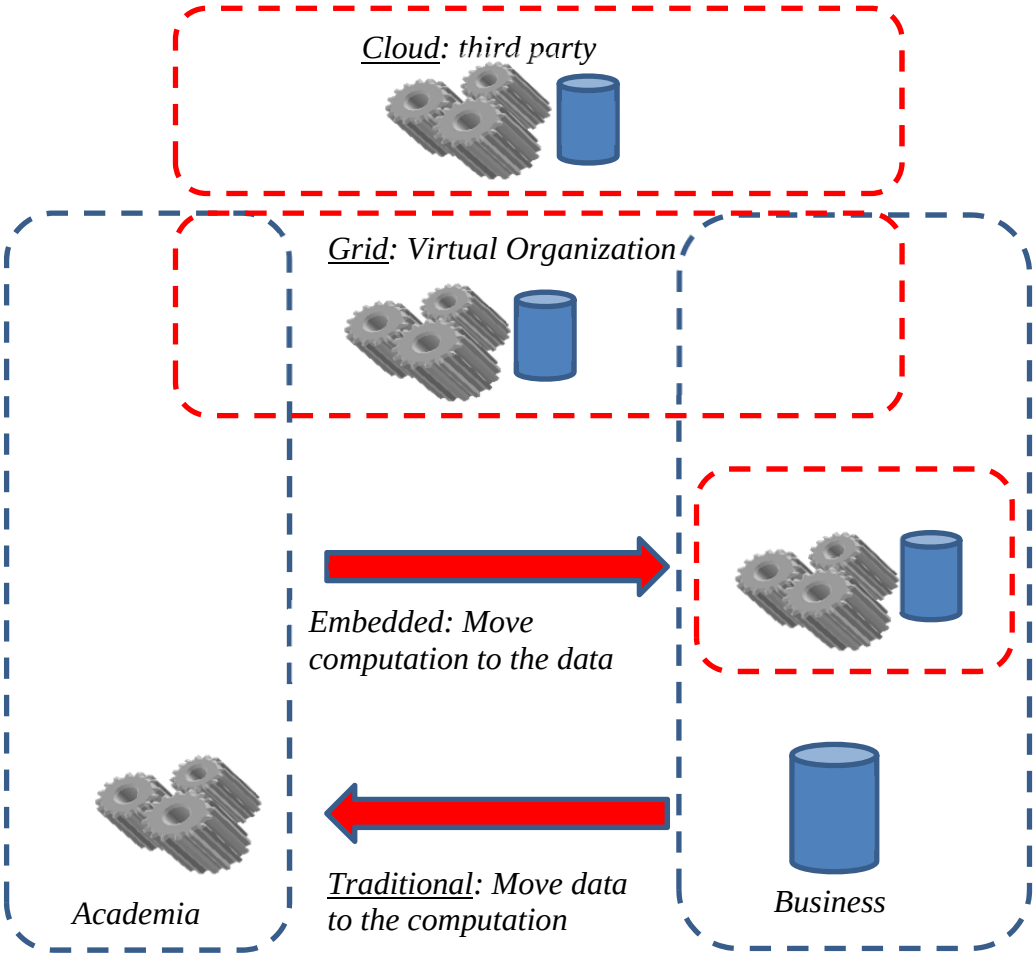

Figure 2: Operational paradigms

3. Traditional – Moving the data to the computation

In the 1960s and 70s when the procurement lifecycle of university computers with government funding was measured in decades, computing at The University of Edinburgh had strong links with industry. Corporate computing was specialised, risky and expensive and so commercial companies used the main university computing facilities for batch processing at night. Note that the principal reason for the collaboration was operational rather than strategic research and consulting. This requirement was eliminated when companies started to invest in core transaction processing systems at levels that far exceeded that of the University sector.

The collaborations that developed to enable researchers from EPCC and the University of Edinburgh Business School access to commercial, operational data were close-knit affairs. Often the researchers and the company involved were located within the same city. All the collaborations, were, however, characterised by the need to transfer operational data to the researchers' platforms. This was because enabling researchers to have direct access to on-line production systems would have interfered with operational processes. Moreover, the

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

companies did not have alternative computing platforms available to host the data for the researchers to access.

In the case of a collaboration with Kwik-Fit, the UK vehicle repair company, the research made use of the vast quantities of data keyed in daily from some Kwik-Fit centres to the main Kwik-Fit database [6]. At the time Kwik-Fit operated around 600 centres in the UK, performing vehicle repairs and supply replacement parts. Customers would drive to a centre and leave their car without a reservation, or make reservations. Customers expected fast, efficient service; generally a delay meant the customer going elsewhere. Clearly to satisfy this kind of customer expectation, centres had to be adequately staffed, otherwise business is lost. To explore this issue data was extracted from the central Kwik-Fit database and transferred to the researchers for analysis to uncover staff-related factors affecting business performance.

Towards the end of the 90s, the volume and sensitivity of the data involved increased significantly along with the relevance of computational requirements and restrictions on data exchange. In a collaboration with Lloyds TSB Bank Scotland plc, the aim was to evaluate and predict customer trends within the mortgage business [7]. As with previous collaborations, it was not possible to have direct access to the operational data on the company's systems. Instead, the company's database schemas were evaluated to determine the nature of the information it held. This involved direct cooperation with the company staff in order to understand their business processes and through this identify the appropriate and relevant data for analysis. This data was then transferred to an off-site computational platform. This meant extracting the data from the company's systems, anonymising it by removing personal information and also the removal of any information deemed commercially sensitive by the company. This whole extraction and data screening process was undertaken by the company before a tape containing the data was finally handed over to be used for research. Further processing was then necessary to deal with formatting issues resulting from the use of different computer architectures at the research and company sites. Only once this was complete could the actual analysis begin. Due to the volume of data involved these various steps required significant processing times. Even though this collaboration involved significantly increased complexity and constraints, it was still essentially a local collaboration with all the participants as well as the company data centre, located within an hour's car journey of each other thus making the transfer of data via physical media and meetings between relevant company staff and researchers fairly straightforward.

In the beginning of the 00s, these local collaborations evolved into national level collaborations such as that with the Cheltenham & Gloucester plc. Here, once again, this involved a data transfer but the interactions between relevant company staff and researchers also shifted to more use of email and telephone with less frequent face-to-face contact due to the distances involved. This style of interaction was in effect just an extension to the local interactions of the previous projects where the participants were within a few hours drive where instead of car the form of transport was a plane.

4. Distributed computing: Moving the computation to the data

The advent of the Grid as a distributed computing model offered the promise of "pervasive and inexpensive access to high-end computational capabilities" [11]. This 'pervasiveness'

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

made it attractive to think of this approach to high-end computing as an infrastructure for a highly competitive 'virtual' organisation.

Competitiveness in this required more emphasis on control and by 2000 the same infrastructure was described as: "necessarily, highly controlled, with resource providers and consumers defining clearly and carefully what is shared, who is allowed to share, and the conditions under which sharing occurs" [12].

The benefits that Grid technologies could provide is in abstracting the data and compute locations and enable interested parties (people) at geographically distributed locations to collaborate using shared resources. Moreover it potentially allowed the data provider to have more control of access to the data and so enable them to turn off access to all or just part of the data as they saw fit. It also potentially allowed them to open up access if they become more trusting of the researcher and see a particular route of enquiry. From the company' perspective this meant that they could allow researchers access without the time-consuming overhead of data preparation and transfer and enable easier access to more up to date data. The collaboration with First plc [4], [5] , the global transport provider, was concerned with investigating precisely this.

This collaboration investigated the deployment of an early implementation of the Open Grid Services Architecture Data Access and Integration services (OGSA-DAI) [13]software within First's South Yorkshire bus operational environment and answered specific business questions through a short data mining analysis using OGSA-DAI enabled data sources. Through OGSA-DAI interfaces, disparate, heterogeneous data sources and resources could be treated as a single logical resource. The OGSA-DAI services provided the basic operations for performing sophisticated operations such as data federation and distributed queries, hiding concerns such as database driver technology, data formatting techniques and delivery mechanisms. The data sources used were from the following systems within First South Yorkshire:

- Customer Contact – this recorded correspondence with customers including commendations and complaints.
- Vehicle Mileage – this recorded the daily vehicle mileage for bus services.
- Ticket Revenue – this contained the daily tickets sold and the money taken for the bus services.
- Schedule Adherence – a satellite tracking system that recorded whether a bus is arriving and departing on time from a bus stop.

These systems were located at various company sites, on differing platforms in different database systems. The databases ranged from SQL sources to ODBC sources to COBOL files.

This work established the viability of using Grid technologies to enable researchers to have remote access to commercial data. It also demonstrated the potential of enabling such analyses across different data sources. Figure 3 illustrates the combination of data from two of these sources showing the obvious relationship between customer satisfaction and vehicle reliability.

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

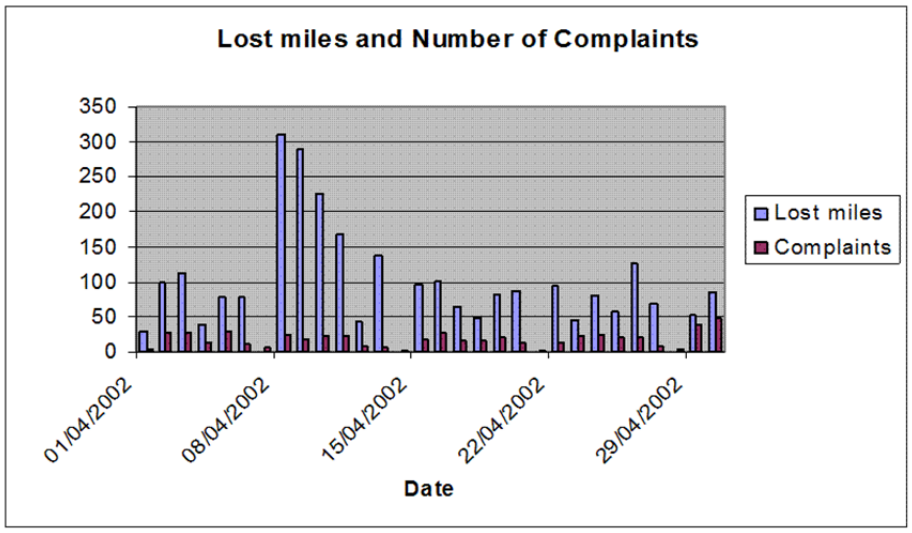

Figure 3: Lost miles and customer complaints (from [5])

Due to commercial confidentiality more detailed descriptions of the results of the various analyses undertaken were not published, however the then senior IT management at First South Yorkshire said that these provided important insights into their bus operations that would revolutionise the way they did business [5].

Though ‘control’, ‘security’ and ‘trust’ are inter-related, the distinction between access to a computational resource and a data resource was significant. Security in relation to access to data tended to be de-emphasised within the ‘Big Science’ community, as data was plentiful, openly shared and hence computational performance was the critical design objective. Securing, encrypting and protecting data would reduce this aspect of performance.

In 2003 the INWA Grid [9]went live between the UK and Australia, via the USA but in this case the data being distributed on an intercontinental scale were large national samples of the consumption of financial services and telecommunications services drawn from commercial companies. In this case the INWA team had to write the security protocols that were embedded in later releases of the Grid Middleware to overcome a significant ‘compatibility’ barrier to adoption [14].

This focus on the needs of an infrastructure for trusted collaborations between social scientists and commercial organisations also extended to considerations of how a ‘typical’ social scientist working in a commodity-computing environment would interface Grid access to their existing toolsets and hence overcome one major learning curve barrier to adoption[15].

The viability of remote access via Grid technologies was tested to the extreme on the INWA Grid [9]. This used Grid technologies to link domain experts, data analysts, data and compute resources at EPCC in The University of Edinburgh, and Curtin Business School in Western Australia. Due to the configuration of National Research and Education Networks (NRENs) at that time all network traffic generated between the UK and Australia had to be routed through the USA. The data underpinning the collaboration were large national samples of individual behaviour within global markets for telecommunications and financial services. These data were direct exports from core business transaction-processing systems and critical to the competitiveness of the data owners. The data were also subject to strict

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

privacy laws within each jurisdiction. Together these established constraints on the types of data exchanged and the types of analyses permitted. Accordingly security at all levels of the interaction was the single most critical concern, and perceptions of the intrinsic security of one model for distributed computing versus another remain critical when selecting an infrastructure for collaborations of this scale. Given the distances involved, many of the interactions between the researchers in Edinburgh and Curtin were undertaken via video-conferencing using Access Grid and with the company via telephone. There were some face-to-face meetings but these were kept to a minimum due to the costs and distances.

The partners involved in the INWA project included a bank that supplied data for 20% of the UK mortgage market and an Asian-Australian telecommunications company that supplied data on 1 million customers as well as public data providers such as the Valuer General's Office in Western Australia. The analyses of these commercial data and their combination with public data sources such as residential property valuations and classifications were undertaken by analysts located in the UK and Australia. Interactions between analysts were necessary due to their complementary skills in data analysis and business knowledge. Where time windows permitted these interactions took place via Access Grid. This ability for analysts to work in real time was crucial in ensuring that ambiguities in the data, analyses results and their interpretation could be identified and dealt with early.

Outputs from the researchers using the INWA Grid infrastructure included models of consumer behaviour with extremely high predictive accuracy (>80%) validated on unseen data, in two highly volatile global markets – telecommunications and financial services. Significant market shares in each case were analysed, allowing complex human behaviour to be modelled and predictions to be made about 'if' a certain behaviour was likely to occur and 'when'. This combination of observations provided a much richer picture of behaviour than had hitherto been available to the collaborating companies, and could even be used to identify unsatisfied, latent demand and new, emerging market segments.

The INWA Grid was successfully extended to China in 2004 with the direct help of the Director General and staff of the Computer Network and Information Center of the Chinese Academy of Sciences. It evolved from nodes in Edinburgh, Scotland and Perth, Western Australia in 2003, to a third at the Computer and Network Information Center, Chinese Academy of Sciences in Beijing in 2004 [9]. In 2007 a fourth node in Xi'An, see Figure 4, was connected to demonstrate the INWA Grid across the first large-scale research and education network in the Asia Pacific region—the Trans Eurasia Information Network (TEIN2) [16],[17]. This allowed an alternative configuration of the INWA Grid, passing data traffic across the shortest possible route between China and the UK, instead of routing the 'long way round' the world via the US. This improves the performance of distributed computing jobs and provides a network route that is faster, easier to manage, and hence potentially more secure. The Xi'An demonstration showed how public domain Grid software could carry the data, video and voice interactions needed to steer the computation, tracing each type of transfer in real time across the network, and showing how Grid security could be applied to every type of interaction managed as a single session and across the same network route [9]. A further reason for establishing the nodes in China was to find a Chinese company willing to enter into a collaboration using the INWA Grid.

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

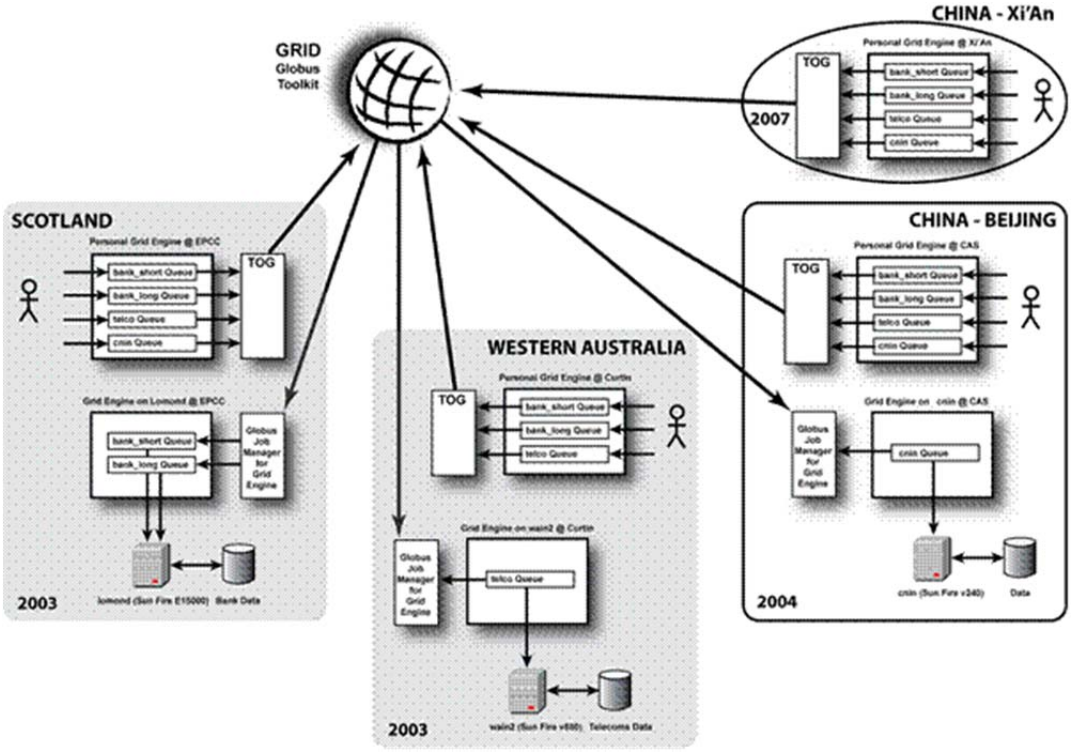

Figure 4: INWA Grid (Logical). 2003 - nodes established at EPCC (Edinburgh, Scotland) and Curtin University (Perth, Western Australia). In 2004 node added at the Chinese Academy of Sciences (Beijing). In 2007 this Grid was used to demonstrate the first connection of the Trans-Eurasian Information Network (TEIN2) into China, from Xi'An.

The geographical extent of the INWA Grid is shown in Figure 5. Figure 5 also shows the principal Research and Education networks connecting the EU to East Asia including TEIN2 which gives the shortest possible network access to this region from Europe.

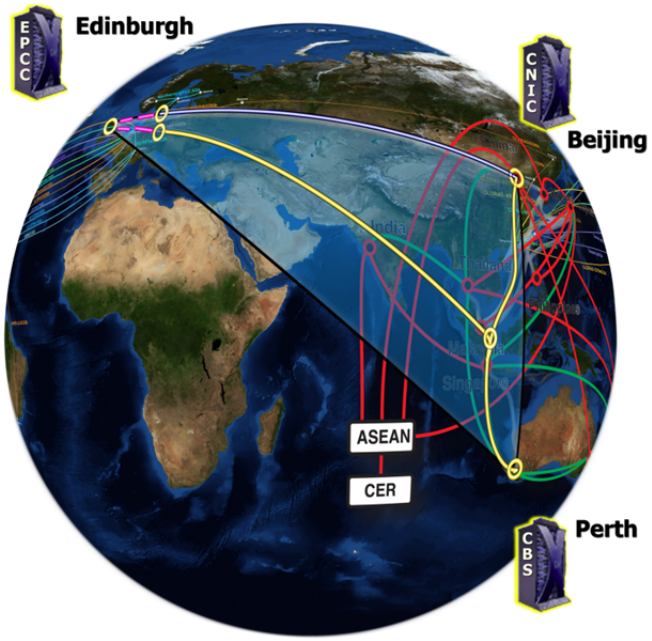

Figure 5: INWA Grid (Physical). 2003 - nodes established at EPCC (Edinburgh, Scotland) and Curtin University (Perth, Western Australia). In 2004 node added at the Chinese Academy of Sciences (Beijing).

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

In 2007 this Grid was used to demonstrate the first connection of the Trans-Eurasian Information Network (TEIN2) into China, from Xi'An. TEIN2 (Yellow), ORIENT (Blue), GÉANT2 (Purple). Earth texture provided by NASA (visibleearth.nasa.gov). Network links include abstractions from GLIF graphics (www.glif.is).

5. The Cloud – Moving the data and the computation

It has been argued that Cloud Computing has emerged from Grid Computing and represents the increasing trend towards the external deployment of IT resources, such as computational power, storage or business applications, and obtaining them as services [18]. According to [19], Cloud Computing is a model for enabling convenient, on-demand network access, to a shared pool of configurable computing resources, (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

The arrival of Cloud Computing offers the possibility to ease the potential for academic-commercial collaborations since, on the surface at least, cloud computing removes the need for either the academic or the company, to host and maintain the infrastructure for hosting the data and the computation for its analysis. This, however, requires both parties to trust a third party with the security and accessibility of the data. Establishing this trust can be a major undertaking and depends heavily on the selected deployment model, as governance of data is outsourced and delegated out of the owner's strict control[20] . In traditional architectures, establishing trust can be assisted by an efficient security policy that addresses constraints on access by external systems and adversaries including programs and access to data by people [20]. In a cloud deployment, this perception is totally obscured and in the case of public or community clouds, control is delegated to the organization owning the infrastructure [20].

Depending on the technical capabilities and software licensing on the chosen cloud platform, there may also be restrictions on the types of analyses that can be performed notwithstanding any constraints placed by the collaborating company.

Further given the global nature of cloud computing where servers are sited in locations across the world, the actual location of the data provided in the cloud infrastructure may also cause constraints on the types of analyses that can be performed. More importantly, the country of the data's origin may impose legal restrictions on where that data can be sited thus restricting the choice of the cloud platforms that can be used [21].

This is because Data Sovereignty is a major legislative hurdle to overcome for a cloud-based solution. For example in Australia, Brett Winterford in October 2010 [22] reported that the proposed Australian legislation "Privacy Principle 8 requires that any organisation storing information that identifies Australian citizens in overseas data centres must ensure that the organisation hosting that data offers the same protections as what is stated in Australia's Privacy Principles". Despite this, some companies claim to have dealt with the data sovereignty issue (see for example Wisekey [23]) and are actively promoting themselves to government bodies.

To investigate further, issues surrounding collaborations with developing countries in East Asia, the authors conducted a workshop titled 'Information Business Growth and

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

Competitiveness in the Asia and Pacific region' at the 2010 Asian Internet Engineering Conference (AINTEC) [24] in Bangkok, Thailand.

The workshop was intended to bring together industry, government and academic participants. The purpose being to explore these issues from differing social, technical and legal perspectives, structured by a critical focus on existing Eurasian collaborations carried by GEANT and TEIN3 [25] network infrastructure across a region inhabited by more than 70% of the world's population and the majority of current economic growth. The workshop attendees came from China, Malaysia, Singapore, Thailand and the UK.

Industry and academic concerns were well represented, however the active participation of government representatives was low. This was compensated for, to some degree, by the participation of Information Technology consultants and a Thai Internet Service Provider (ISP), who all cited the important role of government in regulating the competition. One example given was the unintended consequence of security interventions in the ISP market. Should any website content be found that contravenes a regulation then the servers are confiscated, creating little impact for the criminal who is highly likely to be technically competent and to have made contingency plans for this eventuality, for example simply moving a backup to another provider. This however, causes severe business continuity problems for the local business using a local web-hosting service. Such a business in these circumstances now needs to find another host to re-load their own backup since the ISP may have lost all their hardware and data. The feeling from the ISP provider was that this encourages local businesses to move 'off-shore' where the risks of operating through a distant provider are seen as lower than the business continuity risk of local sourcing. This in turn reduces job opportunities for youth in Thailand who 'want to contribute to a knowledge economy rather than simply consuming'.

Most of the challenges identified during the workshop placed Government as a major enabler and barrier to progress. Workshops such as ours were felt to be important in forming, and perhaps influencing, such priorities.

When collaborating with developing countries such as Thailand we therefore anticipate that we may experience the same data sovereignty issues. Thus any UK institution requiring access to local data may need to do so by accessing facilities in that country either on-site or remotely. However, when using local facilities to accommodate data sovereignty, the necessary compute performance is unlikely to be available for a number of reasons. Challenges specifically identified during the AINTEC 2010 workshop, which are also widely applicable included:

- Infrastructure access is limited and acceptable use policies defining access to shared national facilities are evolving rather than static. This can lead to planning difficulties.
- Licensing – the difficulty of mixing commercial and academic research in an environment where the licensing may not be appropriate. This is just as applicable in the UK.
- Sharing of both compute cycles and skills was considered more difficult locally than in other regions.
- Information goods were more expensive locally – including core application software that international partners might assume were at the same cost or cheaper. Indeed general computing costs are considered high which contributes to the illegal pirating of software in these countries.

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

- Privacy/Security concerns and regulations were seen as potentially restrictive and possibly offering a competitive differentiator if they enabled participation in specific markets for specialised analyses.
- Open source software is seen as 'Western' with limited participation from East Asia – there are limited opportunities to make this type of software reflect local needs.

Given these challenges, provided the data sovereignty issue can be circumvented then cloud computing does offer a valid migration path. Cost of access, however, to a cloud infrastructure may be perceived as a problem relative to developed countries where costs are now low enough that many are considering it. Needless to say, cloud computing still allows cheaper, easier access to compute power compared to having to buy and operate a supercomputer [26]. The question then becomes the key concern expressed in the workshop in Thailand: what it takes to develop a nation that can participate fully in the new markets that arise as a producer 'rather than simply consuming', and here the question becomes one of breadth versus depth. It is possible to define and deploy a cloud far faster and cheaper than it is to build and operate a supercomputer and make it accessible to the same user community, but the requisite skill sets involved are different. It is widely held that: "achieving world class skills is the key to achieving economic success and social justice in the new global economy" [27] and yet "no one can accurately predict future demand for particular skill types". This remains an open question and we raise it here as a caution only.

To complement the findings of this workshop, a series of experiments were undertaken to further investigate the regional dependencies of Cloud infrastructures [26]. Due to the global nature of Cloud computing, it may be more advantageous to submit applications to the Cloud environment from one part of the world than another. This variability can have a substantial effect where, for example, the operational costs of one organisation may be greater than that of another based in a different country using the same technologies.

Regional cost differences, dependent on where an instance is deployed, are known to exist in, for example, Amazon EC2[28]. We, however, chose to test whether differences exist dependent on a user's job submission location and whether they are significant enough to make it advantageous for a researcher to submit a data analysis job to Amazon EC2 from one location than any other. To examine such differences, the same job was submitted from two distant and distinct locations in the world, the UK and Thailand.

We ran a distributed R [29] data analytics application over two Standard On-Demand Large EC2 instances (m1.large) each with 7.5 GB of main memory, and 2 cores each with 2 EC2 Compute Units. These instances were located in the US East Region within the us-east-1b Availability Zone.

The job ran the SPRINT parallel implementation ([30],[31]) of the R Pearson correlation function to process a randomly generated dataset consisting of 11,000 rows and 321 columns.

The principal focus was to create an experiment that was fair and consistent across runs, hence a collection of scripts was created to automatically instantiate instances, setup the experiment, run the experiment and teardown instances allowing the experiment to be run consistently by multiple researchers regardless of whether they were in the UK or Thailand. In collaboration with Dr Sornthep Vannarat, Head of the Large Scale Simulation Laboratory from the National Electronics and Computer Technology Center (NECTEC), the experiment could be performed in Thailand. Once the computation was complete, the cost and resource

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

usages were obtained from the Amazon EC2 invoice and resource Usage Report (a simple xml or csv file) respectively.

To ensure the results were valid, confirmation was required that Dr Vannarat's Amazon EC2 account was tied to an address in Thailand, otherwise if not, different charges could be seen. Another option, later ruled out, was to submit a job via a Thai Virtual Private Network (VPN) from the UK, however charges would still be tied to the Amazon EC2 UK account, incurring UK costs as explained in the results section later in this paper.

Table I: Difference in Resource Usage between the UK and Thailand

Location	Cost	IDT Data In	IDT Data Out	Storage	I/O Req.
UK	$2.52	0.274 GB	0.008 GB	0.151 GB	46,523
Thailand	$2.10	0.205 GB	0.007 GB	0.151 GB	84,103

From Table I we can see that the total cost of running two large instances for the same period of time, including other associated costs such as data transfer and I/O requests, is more expensive when the job is submitted from the UK than in Thailand. This is caused by the level of taxation within the two countries where the UK charges Value Added Tax (VAT) at 20%, hence an increase in $0.42, whereas Thailand charges no taxes. For small and cash-flow sensitive businesses and research institutions, this difference may have a significant impact, for example, a business spending $4,000 on cloud costs per month. For one year of use, the contribution to VAT at 20% would be $9,600; more than two months of cloud usage. The impact of VAT differentiation on market share has been studied across Europe and found to have a highly significant impact, for example a 10-15% reduction in price leading to predicted increases in demand of 70%[32]; VAT is therefore one of the many important choices a global organization must factor into account when choosing where to operate from and provides a significant incentive for adopting a cloud model to migrate data to regions where running costs are lower.

Note that total running costs must take into account regional cost differences and whether tax is charged; prices specified by Amazon are not taxed however if VAT is charged in your country, additional costs apply. In such cases, taxes are calculated based on the address of a user's account and so the service might be outsourced to a tax-free region, reducing the direct cost of the final service offered to the consumer, making it possible to price any service provided with more sensitivity to the competition.

Alongside monitoring experiment costs, the amount of resources used was also obtained. Table I~~Table I~~ shows significant differences for I/O requests and Internet Data Transfer (IDT) inwards. 70MB's of extra data transferred to the US Region from the UK was recorded, accounting for an addition of $0.01 compared to the Thailand run. This is likely caused by data retransmissions when transmitting data to the instances and installing the necessary packages for the experiment to run. This table also shows 37,580 fewer recorded I/O requests from the UK, accounting for $0.01 less than its counterpart, hence levelling the costs of both runs. Why such a phenomenon occurs is likely to result from EC2's underlying storage reading and writing mechanisms, however experimentation to uncover the exact cause of this variability is future work.

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

The structure of Amazon EC2 may also bring limitations to those with applications that require the highest performance available or those with data privacy issues. In the former case, an application that transfers large amounts of data from an external web service in the UK to run computations, would perform better if the instances were located in Amazon's European Region as opposed to the US or Asia Region, due to the locality of the data. In this case, we may see an increase in data transmission and computation speeds. Hence an organisation must perform a full analysis of their computation job to determine it characteristics - for example, whether data must be transferred long distances - allowing a decision to be made on whether lower costs or higher performance is of utmost importance.

In the latter case of data privacy, organisations may be disadvantaged by regulations from countries within their continent that do not have an Amazon Region present (e.g Africa or Australia) if no cloud or virtualisation based providers operate locally. A business or research institution with sensitive data may not be allowed to use computational services residing in other Regions, hence reducing its competitiveness with others where a cloud service is present.

A number of different researchers have performed detailed examinations into the performance of the Amazon EC2, some focused solely on performance variability. However it must not be forgotten that other cloud solutions exist, such as Google AppEngine, Microsoft Azure and Rackspace, to name a few. While Amazon EC2 may be today's most popular cloud, analysing the performance and cost variability of other providers is essential.

Iosup et al [33] examine the long-term performance variability of EC2 paying particular attention to the variability of many AWS services that are connected to EC2 such as SimpleDB and the Simple Queue Service. They also consider AppEngine's Python environment showing the completion time differences over the period of a year, as well as the variability of other related services, such as Memcache and Datastore.

Schad et al, [34] investigate the variability of EC2's CPU, I/O and network performance via micro-benchmarks. They show that small and large instances suffer from large performance variations, which may be partly caused by the underlying hardware an instance is deployed on, such as the AMD versus Intel processors. They also show variations exist between Availability Zones, both in terms of network and CPU performance. A further study would be required to determine the optimum hardware setup and Availability Zone to deploy a computation job.

6. Embedded – Moving the computation to the data

The physical networks shown in Figure 5 represent significant and sustained collaboration between multiple national agencies across all of the regions it transits. Within China, this connectivity reflects cooperation with and between different Ministries with interests in national research and education infrastructure. However, whilst the presence of such infrastructure and secure distributed computing are both necessary for intrinsically valuable data to be collaboratively analysed, they are not sufficient [35]. Since 2004, the authors have held a series of detailed discussions with organisations from the banking, telecommunications

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

and retail sectors, but it took until 2010 for a legal collaboration agreement to be formally constituted and until 2011 for data to be made available for collaborative research.

This extended period was necessary to formulate legal definitions of distributed computing in which the locus of data was explicit. It was also necessary to allow the adaptations required to domesticate a 'proven' technology, the INWA Grid, so that it aligned with existing policies, and where possible with existing business processes and procedures or explicitly established new ones. Examples of steps taken in this case include security enhancements, taking care not to impact what Silverstone [36] cautions as "the potential for real change and real engagements" of the overall system, and ISO certification as a practical step towards Leonard-Barton's [37] "implementation is innovation". This was in effect a "co-production of the social and the technical" [38].

Following the successful demonstration in 2007 of the INWA node in establishing that inter-continental research on such commercial data sources was viable, a subsequent collaboration with a Chinese commercial company started in 2010. This collaboration worked with anonymised, data on tens of millions of consumer transactions provided by the company in 2011. The original intention was to extend the INWA Grid to a node at the participating company with a further extension to the Chinese Academy of Sciences (CAS) in Beijing should extra compute and storage be required. To alleviate the company's concern over security and data sovereignty, this option required the use of a leased line between the company and CAS. Given the costs of such a leased line it was decided to purchase and locate a new machine directly in the company's data centre that would be of sufficient power and capacity for the collaboration. The intention being that this machine would therefore form the company's node on the INWA Grid with the machine sitting directly on the company's network. Subsequently, the company decided they wanted to allow access to this node via Virtual Private Network rather than Grid technologies. This final configuration therefore echoes Gray's [39] conclusion that the economic solution is usually to physically place computational resource near the data since 'it is fine to send a GB over the network if it saves years of computation - but it is not economic to send a kilobyte question if the answer could be computed locally in a second'.

It is this final configuration that this collaboration has allowed, for the first time, e-Social Sciences researchers at a UK university access to consumer data from a Chinese commercial company. Whilst it might appear that the collaboration could have taken place ten years earlier using a similar infrastructure, the reality is that politically, economically and technically this collaboration between a Chinese company and UK researchers would not have been possible. Most obviously connection speeds alone would have been prohibitive for day-to-day access to the data node thus making any interactive data analysis impossible and the building and activation of batch jobs difficult to say the least. The introduction of the TEIN2 (in January 2006 [40]) and subsequently TEIN3 (extended to South Asia in December 2009 [25][~~22~~]) networks has vastly improved connection speeds and are one of the reasons for making this collaboration possible.. There are numerous, further reasons, around even fundamentals such as power supply, but these are beyond the scope of this current paper.

7. Conclusions

Physically embedding systems with an organisation's Data Center is a challenging task for e-Social Science researchers, and at first sight a much less attractive option than moving the

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

organisation's data to a suitable computing facility. As reported here, for Edinburgh researchers collaborating with a Chinese company the need to maintain the centre of gravity of the data within a single legal jurisdiction and the long distance between that location and the INWA Grid facilities at the Chinese Academy of Sciences and The University of Edinburgh meant that cost-performance considerations made it more economic to invest in local compute power than to distribute the computation.

Alternative Grid and Cloud computing models have the same exposure to communications cost at an architectural level. We found that a Grid model was acceptable to our industry partners, even when distributed over thousands of kilometres but, the security questions that were in circulation at that time about Cloud computing prevented any exploration of that as an alternative.

When we tested these types of engagement – data sharing and analysis using third party distributed computing facilities - with business, academic and government stakeholders in Thailand we found similar concerns about systems that can be differentiated in terms of data sovereignty. However, the concerns raised were far more profoundly linked to development of these services as a business within Thailand that serves both domestic and export markets. In all cases the role of government was considered critical to successful capability development. By running the same application from two distinct parts of the world, Thailand and the UK, we showed that contrasting costs emerge due to the local taxes employed by the country, adding a new dimension to the flexibility provided by a cloud solution. This flexibility however may be constrained due data privacy laws restricting where data can be sent for data processing.

Whilst the INWA Grid vision has achieved a move from a demonstrator of UK-China e-Social Science capability to a direct industry engagement, the use of an earlier computing model for this first engagement should not be viewed as a rejection of the Grid model. Rather this is a stage in the development required because of the trust-building process.

Indeed it is because of the time taken in this trust-building process that an embedded solution has another advantage. Years in trust building and the drafting of suitable legal agreements, means that an embedded system allows trials and testing for Quality Assurance prior to release of the data, and has the potential of more currency in data terms compared to an export of large data archives that had been rationalized for record-keeping requirements and requires unpacking. With an embedded system there is the potential for developing a routine export whilst the data is still live and the data context can be inspected. If the organizational data is intended to provide a window on behaviour for social science analysis, then an embedded system has the possibility of establishing a much closer and longer-term relationship than any other computing model whilst keeping the migration path open.

Acknowledgements

This trans-national work would not have been possible without the sustained investment in building the INWA Grid over the last decade, both as a technological infrastructure capable of secure, distributed, cooperative analysis of large datasets, but also as a set of trusted working relationships that enables commercially sensitive data to be shared.

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

We gratefully acknowledge the support of the UK Economic and Social Research Council (award RES-149-25-0005) for the initial phase of the INWA Grid and its 'Follow-On Funding' (award RES-189-25-0039); the UK EPSRC (award EP/H006753/1 on Building Relationships with the 'Invisible' in the Digital (Global) Economy) for supporting the reported work with collaborators in China and Thailand; the Scottish Funding Council (edikt2 grant HR04019); the Australian Research Council in partnership with Singapore Telecom for its support (awards LP0454322 and SR0567388) and the endowed SingTel Optus Chair of eBusiness held by Lloyd at Curtin University; Sun Microsystems and the Australian Academic and Research Network (AARNet) for continued support since the INWA Grid became operational in 2003, and colleagues at the Computer Network and Information Center of the Chinese Academy of Sciences who have enabled and hosted the connections within China since 2004. The Biotechnology and Biological Sciences Research Council [award BB/J019283/1] supported the writing of this paper.

References

[1] C.W.J. Granger, Some aspects of the future of Forecasting, Keynote, 24th International Symposium on Forecasting, Sydney, July 2004. http://forecasters.org/isf/pdfs/ISF2004_program.pdf. Accessed 13th July 2012.

[2] P. Halfpenny, R. Procter, The e-Social Science research agenda, Phil. Trans. R. Soc. A 368 (2010) 3761-3778, doi: 10.1098/rsta.2010.0154.

[3] M. Savage, R. Burrows, The coming crisis of empirical sociology, Sociology 41 (2007) 885–899, doi:10.1177/0038038507080443.

[4] T.M. Sloan, A. Carter, P.J. Graham, D. Unwin , I. Gregory, First Data Investigation on the Grid: FirstDIG, in: S.Cox (Ed.) Proceedings of the 2nd UK e-Science All Hands Meeting, 200, pp. 287-289.

[5] P.J. Graham, T. Sloan, A.C. Carter, I. Gregory, FirstDIG: Data Investigations using OGSA-DAI, in: S. Cox (Ed.) Proceedings of the UK e-Science All Hands Meeting 2004, Nottingham, UK, pp 39-54.

[6] K. Smyllie, Data mining and simulation applied to a staff scheduling problem, in: P. Sloot, M. Bubak, A. Hoekstra, B. Hertzberger (Eds.), Proceedings of the 7th International Conference on High-Performance Computing and Networking Europe, Lecture Notes in Computer Science, Volume 1593, Springer 1999, pp. 1190-1193, doi: 10.1007/BFb0100687.

[7] T.M. Sloan, P.J. Graham, K. Smyllie, A.D. Lloyd, Extracting business benefit from operational data, in: M. Bubak, H. Afsarmanesh, R. Williams, L.O. Hertzberger (Eds.), Proceedings of the 8th International Conference on High Performance Computing and Networking Europe, Lecture Notes in Computer Science Volume 1823, Springer 2000 pp. 477-486.

[8] A.C. Hume, A.D. Lloyd, T.M. Sloan, and A.C. Carter, Applying Grid Technologies to Distributed Data Mining, in: S.Cox (Ed.) Proceedings of the UK e-Science All Hands Meeting 2004, Nottingham, UK, pp. 944-947.

[9] A. Lloyd and T. Sloan, Intercontinental Grids: An Infrastructure for Demand-Driven Innovation, Journal of Grid Computing 9 (2011) 2 185-200.

[10] A. Haugen Gausdal and J. Moss Hildrum, Facilitating Trust Building in Networks: A Study from the Water Technology Industry, Systemic Practice and Action Research 25 (2012) 1 15-38, doi: 10.1007/s11213-011-9199-3.

[11] I. Foster, C. Kesselman, The Grid: Blueprint for a New Computing Infrastructure. Morgan-Kaufmann 1998.

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

[12] I. Foster, C. Kesselman, T. Tuecke, The Anatomy of the Grid:Enabling Scalable Virtual Organizations. International Journal of High Performance Computing Applications. 15 (2001) 3 200-222.
[13] M. Antonioletti, M.P. Atkinson, R. Baxter, A. Borley, N.P. Chue Hong, B. Collins, N. Hardman, A. Hume, A. Knox, M. Jackson, A. Krause, S. Laws, J. Magowan, N.W. Paton, D. Pearson, T. Sugden, P. Watson, M. Westhead, The Design and Implementation of Grid Database Services in OGSA-DAI , Concurrency and Computation: Practice and Experience 17 (2005) 2-4 357-376.
[14] A.C. Hume, A.D. Lloyd, T.M. Sloan, A.C. Carter, Applying Grid Technologies to Distributed Data Mining, in: H. Jin, Y. Pan, N. Xiao, J. Sun (Eds.) Proceedings of Grid and Cooperative Computing – GCC 2004, Springer-Verlag Heidelberg, Lecture Notes in Computer Science Volume 3251 (2004) pp. 696-703.
[15] M.J. Jackson, A.D. Lloyd, T.M Sloan, Enabling Access to Federated Grid Databases: An OGSA-DAI ODBC Driver, in: S.J. Cox, D.W. Walker (eds.) Proceedings of Fourth UK e-Science All Hands Meeting, Nottingham, 2005. pp. 951-957.
[16] A.D. Lloyd, T.M. Sloan, B. Yan, Global competitiveness and regional innovation—using the Grid to "close the gap between Business, Research and Resources"? One World, One Network— 24th Asia Pacific Advanced Network Meeting, Xi'An, China, 27–31 Aug 2007. Available at http://www.apan.net/meetings/xian2007/proposals/global.html. Accessed 13th July 2012.
[17] A.D. Lloyd, Case study: Understanding global market dynamics requires global network reach. Delivering of Advanced Network Technology to Europe – DANTE (2008). Available: http://www.tein3.net/upload/pdf/INWA_final_web.pdf. Accessed 13th July 2012.
[18] K. Stanoevska-Slabeva, T. Wozniak Grid and Cloud Computing-A Business Perspective on Technology and Applications, Springer-Verlag, Berlin, Heidelberg (2010)
[19] National Institute of Standards and Technology, The NIST Definition of Cloud Computing, Information Technology Laboratory, 2009
[20] D. Zissis, D. Lekkas, Addressing cloud computing security issues, Future Generation Computer Systems, Volume 28, Issue 3, March 2012, Pages 583-592, ISSN 0167-739X, 10.1016/j.future.2010.12.006.

[21] W.K. Hon, C. Millard, Data Export in Cloud Computing - How Can Personal Data Be Transferred Outside the EEA? The Cloud of Unknowing, Part 4 (April 4, 2012). Queen Mary School of Law Legal Studies Research Paper No. 85/2011. Available at SSRN: http://ssrn.com/abstract=2034286 or http://dx.doi.org/10.2139/ssrn.1925066. Accessed 13th July 2012.
[22] B. Winterford, Privacy revisions present risk for offshore clouds, itnews for Australian business (2010). Available at http://www.itnews.com.au/News/234695,privacy-revisions-present-risk-for-offshore-clouds.aspx. Accessed 18th June 2012.
[23] Cloud Computing Data Sovereignty by WISeKey. Available at http://www.wisekey.com/en/solutions/DataSovereignty/Pages/default.aspx. Accessed 18th June 2012.
[24] Information Business Growth and Competitiveness in the Asia and Pacific region, Asian Internet Engineering Conference 2010. Available at http://www.interlab.ait.ac.th/aintec2010/appworks.php. Accessed 25th June 2012.
[25] TEIN 3 – The Research and Education Network for Asia-Pacific. http://www.tein3.net/upload/pdf/1075_TEIN3_Brochure_2.pdf, Accessed 13th July 2012.

Future Generation Computer Systems
Volume 29, Issue 5, July 2013, Pages 1120–1129
http://dx.doi.org/10.1016/j.future.2012.12.013

[26] M. Piotrowski, G. McGilvary, T. Sloan, M. Mewissen, A. Lloyd, T. Forster, L. Mitchell, P. Ghazal, J. Hill, Exploiting Parallel R in the Cloud with SPRINT, Methods of Information in Medicine 2012, (in press).
[27] Prosperity for all in the global economy - world class skills, Leitch Review of Skills – Final Report, December 2006. http://www.hm-treasury.gov.uk/d/leitch_finalreport051206.pdf. Accessed 13th July 2012.
[28] Amazon Elastic Cloud (EC2) http:// aws.amazon.com/ec2/. Accessed 27th October 2011.
[29] The R Project for Statistical Computing.http://www.r-project.org/
[30] Hill J, Hambley M, Forster T, Mewissen M, Sloan TM, Scharinger F, Trew A, Ghazal P. SPRINT: a new parallel framework for R. BMC Bioinformatics 2008; 9:558.
[31] Petrou S. SPRINTing with HECToR, 2010. http://www.hector.ac.uk/cse/distributedcse/reports/sprint/sprint.pdf. Accessed 09/05/2011
[32] Oosterhuis F, Dodoková A, Gerdes H, Greño P, Jantzen J, Mudgal S, Neubauer A, Rayment M, Stocker A, Tinetti B, van der Woerd Varma A. The use of differential VAT rates to promote changes in consumption and innovation, Final Report under DG Environment, Contract 070307/2007/482673/G1, 2008
[33] Iosup A, Yigitbasi N, Epema D, On the performance variability of production cloud services. Proceedings of the IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing 2011
[34] Schad J, Dittrich J, Quian e-Ruiz J. Runtime measurements in the cloud: observing, analyzing, and reducing variance. Proc. VLDB Endow., 3:460–471, September 2010.

[35] A.D. Lloyd, BRIDGE – Building relationships with the invisible, EPCC News, December 2011.
[36] R. Silverstone, E. Hirsch, D. Morley, Information and communication technologies and the moral economy of the household. In R. Silverstone & E. Hirsch (Eds.), Consuming technologies: Media and information in domestic spaces. London: Routledge 1992.
[37] D.A. Leonard-Barton, Implementation as mutual adaptation of technology and organisation. Research Policy, 17 (1988) 5 251−267.
[38] Sørensen, K. H. Domestication: The enactment of technology. In T. Berker, M. Hartmann, Y. Punie, & K. Ward (Eds.), Domestication of media and technology. Maidenhead: Open University Press 2006.
[39] J. Gray, Distributed computing economics, ACM Queue 6 (2008) 3 63-68.
[40] TEIN 2 - Powering research and Education in Asia-Pacific. http://www.sigmanet.lv/uploads/raksti/tein2.pdf. Accessed 14th July 2012.