WORK PROGRAMME 2011-2012
EXTERNAL CONSULTATIONS 2009-2010

CHALLENGE 4 – INTELLIGENT INFORMATION MANAGEMENT SO4.4

1. INTRODUCTION AND BACKGROUND

External consultations for Strategic Objective 4.4 Intelligent Information Management were structured around three channels:

(1) ongoing conversations with members of the relevant constituency

(2) web questionnaire

(3) Targeted in-depth telephone interviews.

This report summarises the main issues and recommendations from actions 2) and 3) against the background of the technology and economic landscape.

2. TECHNOLOGY BACKGROUND

Strategic Objective 4.4 exists to tackle the problems and opportunities that result from the enormous growth in data volumes we are witnessing. As devices and sensors become cheaper and cheaper to produce and operate, data are being generated at faster and faster rates and greater and greater volumes. As a result, more data is being produced than we are able to make sense with using current technology. As a 25-02-2010 special report by the Economist[1] indicates, this abundance of data is at the same time a challenge and an opportunity. A challenge because new, more scalable, technologies are needed to manage such data volumes; an opportunity because much knowledge can be extracted by those who can most effectively analyse such data. Analytic abilities at the scale of Big Data are indeed the key of the technological dominance of giants like Google[2] and have been predicted to be the key of the next, impending, scientific revolution based on data intensive scientific discovery.[3]

As data volumes grow, the technology landscape is also evolving to cope with the growth. The year 2009 saw a sudden increase in the production deployment of structure storage architectures that depart from the classic relational database in order to ensure various requirements of scalability and freedom from pre-defined schemata.[4] The

[1] http://www.economist.com/specialreports/displayStory.cfm?story_id=15557443

[2] http://googleresearch.blogspot.com/2009/03/unreasonable-effectiveness-of-data.html

[3] http://research.microsoft.com/en-us/collaboration/fourthparadigm/

[4] such architectures are often collectively referred as NoSQL storage architectures: http://en.wikipedia.org/wiki/NoSQL

Semantic Web and Linked Open Data publishing philosophy that is directly endorsed by the UK government with its http://data.gov.uk initiative is indeed based on a specific type (RDF stores) of such architectures.

Similarly, the opportunity to take advantage of the multi-core architectures that are foreseen in the development roadmaps of major chip manufacturers[5] and the need to distribute computations over database partitions[6] too large to fit into a single database have sparked a renewed interest into the opportunities for distributed and parallel computing offered by functional and concurrent programming languages and exploited by frameworks such as MapReduce[7] (deployed at Google, Yahoo and many other data intensive corporations).

Our consultations were designed to elicit the broadest spectrum of opinions on the trends and developments listed in the previous paragraphs and obtain detailed feedback not only on the specific technological bottlenecks that will need to be addressed to make extremely large scale information management possible, but also on which European industrial and societal sectors might most benefit from innovation in this domain and how such innovation could be disseminated for the greatest impact.

3. THE QUESTIONNAIRE

In September 2009 we published[8] the following questionnaire, which was advertised on several channels (direct e-mail message to 3000+ contacts from the Unit; through the coordinators of projects in our portfolio; through micro-blogging venues such as Twitter).

(1) Data and data-types

- (a) What volumes of data are we dealing with today? What is the growth rate? Where can we expect to be in 2015?
- (b) What types or data can we deal with intelligently due to their inherent structure (geospatial, temporal, social or knowledge graphs, 3D, sensor streams …)?

(2) Industries, communities

- (a) Who is producing these data and why? Could they do it better? How?
- (b) Who is consuming these data and why? Could they do it better? How?
- (c) What industrial sectors in Europe could become more competitive if they became much better at managing data?

---

[5] http://techresearch.intel.com/articles/Tera-Scale/1421.htm

[6] http://en.wikipedia.org/wiki/Sharding

[7] http://en.wikipedia.org/wiki/MapReduce

[8] http://cordis.europa.eu/fp7/ict/content-knowledge/consultation_en.html

(d) Is the regulation landscape imposing constraints (privacy, compliance ...) that don't have today good tool support?

(e) What are the main practical problem identified for individuals and organizations? Please give examples and tell us about the main obstacles and barriers.

(3) Services, software stacks, protocols, standards, benchmarks

(a) What combinations of components are needed to deal with these problems?

(b) What data exchange and processing mechanisms will be needed to work across platforms and programming languages?

(c) What data environments are today so wastefully messy that they would benefit from the development of standards?

(d) What kind of performance is expected or required of these systems? Who will measure it reliably? How?

(4) Usability and Training

(a) How difficult will it be for a developer of average competence to deploy components whose core is based on rather deep computer science? Do we all need to understand Monads and Continuations? What can be done to make it ever easier?

(b) How is a developer of average skills going to learn about these new advanced tools? How can we plan for excellent documentation and training, community mentoring, exchange of good practices, etc... across all EU countries?

(5) Challenges

(a) What should be, in this domain, the equivalent of the Netflix challenge, Ansari X Prize, Google Lunar X Prize, etc. ... ?

(b) What should one do to set up such a challenge, administer, and monitor it?

A study of the responses to the questionnaire reveals a number of important themes:

- Data abundance will create a scarcity of trust: there will be so much data available that the value will come from knowing which specific date source to trust and why. For this reason it will be important to establish data curators as intermediates and maintainers of quality and trust. Such intermediaries will likely also solve the problem that information creators today find it difficult to be compensated for their efforts. Intermediaries can establish mutually beneficial institutions where data offer meets data demand (including quality assurance). This would in turn improve data reuse. The respondents also point out that in such scenarios the management of data privacy and compliance with regulations will be of extreme importance.

- Multimedia (including 3D) and sensor data were listed as the data types most likely to experience the fastest growth and thus must in need of scalable analytics and management infrastructure. Among the most significant data producers the following sectors were mentioned: the financial industry, geospatial imagery (in particular satellite), eHealth (in particular data from wearable personal sensors), Aerospace, Big Pharma and, more generally, any experimental domain where automation is the norm (genome sequencing, physics…). This last point is completely consistent with the observation made in section 1 above on large datasets as the foundation for a scientific revolution.

- As a special case of data that is expected to play an important role in the near future are data collection that are made available by city, regional, national and EU administrations or public bodies. Although not gigantic in size they are expected to be of high value and densely interconnected, thus affording excellent opportunities for reuse, reasoning and value extraction.

- Several of the respondents highlighted the connections between a few important ideas. Since data will be produced more and more often by devices and sensors, static data collections will be in many environments replaced by streams of data. This means will require a shift in many information management practices and infrastructures. Moreover, these streams will be analysed for the occurrence of events of interest (simple events or complex events that depend on the detection of simpler ones) and this will require (particularly in enterprise settings) a shift from process based to event driven practices. Finally, since organisations will need to react efficiently to events not completely under their own control, it will be extremely important that the detection of such events be carried out in as close as possible to real time (the paradigmatic example being the financial domain).

- Two specific technical requirements were identified as being of particular importance: the need to parallelise information management operations and algorithms as much as possible and the need to retarget data intensive algorithms to emerging computing architectures. Parallelisation will be important not only in order to fully exploit the trend towards CPU with larger and larger number of cores but also because extremely large data sets make it more efficient to bring computation to the data than to load data into a central computing unit. The ability to retarget data intensive algorithms to novel hardware architectures such as Field Programmable Gate Arrays or Graphic Processing Unit will further expand the ability to exploit computational resources. Benchmarking and challenges will be particularly useful in order to measure progress along these two technical dimensions.

- Visual Analytics has been identified as a functional requirement of extreme importance. Technical issues to be solved are on one side the problem of scaling current visualisation patterns to datasets containing billions of objects and on the other the problem of supporting exploratory data analysis, supporting decision makers in the task of identifying data patterns that cannot be precisely defined a priori.

4. IN-DEPTH INTERVIEWS

In depth interviews (conducted by e-mail and telephone) were arranged with members of two communities in particular:

(1) practitioners of non-traditional computing paradigms

(2) developers of non-traditional data store architectures

It is to be noted that in both cases the individuals interviewed were targeted not for their very strong academic credentials as much as for their front-line involvement in the deployment of practical, robust and scalable information management systems.

The issues discussed in the two cases were quite distinct.

On the computing paradigms front it was pointed out that while the EU has been at the forefront of programming language design (the Erlang language was developed by Ericsson in Sweden, the Haskell language in Glasgow) what has been missing has been appropriate training and documentation to ease what is admittedly a rather steep learning curve. The experts consulted highlighted the need to develop usable data oriented abstractions so as to allow a software engineer of average skill to take advantage of these paradigms.

As far as non-traditional storage architectures are concerned, the main issue discussed was the need to engineer extremely fluid data management architectures as a requirement for managing datasets of unpredictable variety and with arbitrarily complex relationships. It was also pointed out that a consequence of that requirement is that we could potentially see the development of extremely diverse solutions, whose performance and functionality will need to be compared in a meaningful way to generate confidence and ultimately adoption. To this end rigorous benchmarking exercises will be key.