NESSI White Paper, December 2012

Big Data

A New World of Opportunities

Contents

1. Executive Summary

The amount of available data has exploded in the past years because of new social behaviors, societal transformations as well as the vast spread of software systems. Big Data has become a very important driver for innovation and growth that relies on disruptive technologies such as Cloud Computing, Internet of Things and Analytics. Big Data is thus very important to foster productivity growth in Europe since it is affecting not only software-intensive industries but also public services, for example the health, administration and education sectors.

A challenge for Europe is to ensure that software and services providers are able to deliver high quality services along the lines of the fast growing number of services and users. Providers of Software and Services must be able to ensure the protection of their users' data (in terms of availability and use) while at the same time allowing the use of the information to improve the quality of the services they deliver.

Users also want more personalised and more responsive applications without unsatisfied requests. The goal is to implement the vision of full-time availability of services and resources based on data everywhere for anyone, at all time. In this perspective, the flow of Big Data is a key enabler for the provision of resources anytime, anywhere, and the adaptation to demand.

Big Data software and services generate value by supporting an innovative eco-system and by enabling completely new solutions that have not been possible before. The value lies in the applications based on advanced data-analysis on top of more general Big Data layers, semantic abstractions or network and physical objects virtualization.

In this white paper, the European Technology Platform NESSI (Networked European Software and Services Initiative) seeks to contribute to the current European Big Data discourse, based on the expertise of its member community. NESSI offers a comprehensive view on technical, legal, social and market-related aspects of Big Data that have a direct impact on applications, services and software technologies practices. This paper addresses the challenges rising from the use of Big Data and how to overcome those in order to enable new technical and business opportunities for Europe's software industry and economic sectors. The key points stressed in this paper seek to ensure that the necessary technical conditions and expertise of a skilled work force are available, together with the right business and policy environment in order for Europe to take advantage of these opportunities and regain its competitive position with respect to the rest of the world.

As a result, NESSI brings forward the following recommendations, divided into four different, but highly interlinked sections:

Technical: NESSI supports the need to direct research efforts towards developing highly scalable and autonomic data management systems associated with programming models for processing Big Data. Aspects of such systems should address challenges related to data analysis algorithms, real-time processing and visualisation, context awareness, data management and performance and scalability, correlation and causality and to some extent, distributed storage.

Business: On the EU-level we need a mechanism to foster and structure information between key actors in the European "Big Data Ecosystem" which should be considered from a data process-centric perspective. A Big Data business ecosystem can only be built with a clear understanding of policies addressing legal and regulatory issues between countries, cross-enterprise collaboration issues and the nature of data sources. EU should establish European Big Data Centres of excellence (like Hack/Reduce in Boston) and fostering Open Source Big Data Analytics to make sure the benefits stay in the EU, with European developers, users and entrepreneurs. This should include the awareness of integrating existing private data with external data to enhance existing products and services.

Legal: Within the context of the revised Data protection legal framework, NESSI suggests an in-depth privacy by design scheme, which is not imposing a 'one size fits all' rule but leaves flexibility to business in order to ensure that the economic potential of Big Data is cared for in a way that allows for evaluating and analysing data and use such data predicatively for legitimate business purposes, including identity verification and fraud detection and prevention.

Skills: NESSI encourages the creation of new education programs in data science and supports the idea of creating a European Big Data Analytics Service Network to foster exchanges between data scientists and businesses and as a way up to building advanced data-analysis products within the software industry. NESSI stresses the importance of creating resources for using commoditized and privacy preserving Big Data analytical services within SME's.

2. Introduction

2.1. Political context

In recent years, Big Data has become a major topic in the field of ICT. It is evident that Big Data means business opportunities, but also major research challenges. According to McKinsey & Co[1] Big Data is "the next frontier for innovation, competition and productivity". The impact of Big Data gives not only a huge potential for competition and growth for individual companies, but the right use of Big Data also can increase productivity, innovation, and competitiveness for entire sectors and economies. This is in line with the targets set out in the Europe 2020 Strategy aiming to foster a sustainable, smart and inclusive European economy, where the EU-flagship the Digital Agenda for Europe (DAE) has the overall aim to create a sustainable and economic European digital single market with a number of measures directed at the use of data sources in Europe.

To be able to extract the benefits of Big Data, it is crucial to know how to ensure intelligent use, management and re-use of Data Sources, including public government data, in and across Europe to build useful applications and services. The Digital Agenda emphasizes the importance of maximizing the benefits of public data, and specifically the need for opening up public data resources for re-use (Action 3 of the Digital Single Market Pillar). Public Sector Information (PSI) is the single largest source of information in Europe. Its estimated market value is €32 billion. Re-used, this public data could generate new businesses and jobs and give consumers more choice and more value for money, as the DAE points out. The European Commission has already pushed for some actions at technical level (mainly related to data formats in order to promote interoperability and re-use) but also at regulatory and policy levels, by fostering transparency and availability of data and encouraging access. The EU Open Data strategy, which is an amendment of the Public Sector Information Directive, encourages more openness and reuse of public sector data[2]. Even though the process is still in a very early stage there is still a need to analyse both sides of the problem as open data can be a threat for privacy.

In addition, the European Commission is developing a Data Value Strategy which will address the entire data life cycle. NESSI seeks to contribute to the Data Value Strategy, based on the expertise of its members. NESSI offers a comprehensive view of key technical, market-related and social aspects which forms the Big Data landscape with specific attention to the position of Europe software industry and research.

[1] McKinsey Global Institute, *Big Data: The next frontier for innovation, competition and productivity* (June 2011)

[2] Article: *"Big Data at your Service"* http://ec.europa.eu/information_society/newsroom/cf/dae/itemdetail.cfm?item_id=8337 (July 2012)

2.2. Research and Big Data

There are currently a number of initiatives aimed at adjusting the research landscape to lodge the rapid changes taking place in the processing of data under the current and seventh framework programme for Research and Innovation, FP7[3]. As an example, the Big Data research challenges under FP7, include 81 projects addressing a vast set of topics ranging from content creation & processing (including multimedia and games), "Big Data" analytics and real-time processing,.

In Horizon 2020, Big Data finds its place both in the Industrial Leadership, for example in the activity line "Content technologies and information management", and the Societal Challenges, relating to the need for structuring data in all sectors of the economy (health, climate, transport, energy, etc.). Not surprisingly, Big Data is also important to the Excellent Science priority of Horizon 2020, especially to part I, sections 4.1 and 4.2, on scientific infrastructures and development of innovative high value added services.

In addition, also other research areas will experience the growing role of the efficient use of data. For example, data from space observations have been specifically singled out as a resource to be exploited together with data from big medical studies. [4]

2.3. Purpose of the paper

The objective with this NESSI White Paper is to highlight selected aspects of Big Data that create particular opportunities, but also challenges for the software and services industry in Europe. Therefore, NESSI will address current problems and materialize opportunities associated to the use of Big Data, by applying a unifying approach, where technical aspects are aligned with business-oriented, regulatory and policy aspects.

We will examine how Big Data actors can create a better momentum for Europe in terms of global competition. Thus, the overarching goal is to identify Big Data research and innovation requirements in the context of Horizon 2020. NESSI will also put forward policy recommendations of a more general character to provide input to the Commission Data Value Strategy.

2.4. Approach

NESSI wants to impact the technological future by identifying strategic research directions and proposing corresponding actions. NESSI gathers representatives from industry (large and small), academia and public administration and is an ETP active at international level. NESSI closely monitors technology and policy developments of Big Data as it has a large impact on the software and services domain. In the beginning of this year, NESSI formed a Task Force consisting of Big Data experts from member companies and universities to coordinate the work with this white paper on Big Data.

The NESSI Big Data task force further asked NESSI's members to contribute to the Big Data white paper in order to support the ideas put forward regarding of the requirements of the Software and Services sector.

[3] Article: *"Big Data at your Service"* (July 2012)
http://ec.europa.eu/information_society/newsroom/cf/dae/itemdetail.cfm?item_id=8337

[4] Answer to Parliamentary question, Neelie Kroes (23 July 012)
http://www.europarl.europa.eu/sides/getAllAnswers.do?reference=E-2012-006126&language=EN

The members were asked to answer ten questions distributed in an on-line survey. The responses from our members[5] have been integrated into this document.

2.5. Document outline

The paper firstly defines the Big Data concept and describes its origins and main characteristics (section 3).

Section 4 discusses economic and societal impact of Big Data usage for applications and services. Section 5 introduces some of the technical and scientific challenges associated to collecting, capturing, curating, searching, sharing, analysing and visualizing data faced by the software industry to exploit Big Data. Section 6 provides an overview of the main actors and market trends in the European Big Data business ecosystem. From a societal perspective, section 7 analyses opportunities and challenges related to privacy, with a special emphasis on business impact but also regulatory concerns.

In the last parts of the paper, we present a brief SWOT (Strengths, Weaknesses, Opportunities and Threats) analysis for the Software and Services Industry in Europe, within the context of Big Data. We conclude the white paper by bringing forward a number of recommendations identified by NESSI.

3. Big Data concept

3.1. Definitions

Big Data is a notion covering several aspects by one term, ranging from a technology base to a set of economic models. In this white paper, the following definition of Big Data will be applied:

> *"Big Data" is a term encompassing the use of techniques to capture, process, analyse and visualize potentially large datasets in a reasonable timeframe not accessible to standard IT technologies. By extension, the platform, tools and software used for this purpose are collectively called "Big Data technologies".*

Big Data is not a new concept, and can be seen as a moving target linked to a technology context.

The new aspect of Big Data lies within the economic cost of storing and processing such datasets; the unit cost of storage has decreased by many orders of magnitude, amplified by the Cloud business model, significantly lowering the upfront IT investment costs for all businesses. As a consequence, the "Big Data concerns" have moved from big businesses and state research centres, to a mainstream status.

[5] The total number of respondents was 34, which means that the results from the survey are not used as basis for the arguments put forward in this paper, but rather as support with additional points, to the ones already raised by the main authors in the NESSI Big Data task force.

3.2. Origins of the concept

A decade ago, data storage scalability was one of the major technical issues data owners were facing. Nevertheless, a new brand of efficient and scalable technology has been incorporated and data management and storage is no longer the problem it used to be.

In addition, data is constantly being generated, not only by use of internet, but also by companies generating big amounts of information coming from sensors, computers and automated processes. This phenomenon has recently accelerated further thanks to the increase of connected devices (which will soon become the largest source of data) and the worldwide success of the social platforms.[6]

Significant Internet players like Google, Amazon, Facebook and Twitter were the first facing these increasing data volumes "at the internet scale" and designed ad-hoc solutions to be able to cope with the situation.

Those solutions have since, partly migrated into the open source software communities and have been made publicly available. This was the starting point of the current Big Data trend as it was a relatively cheap solution for businesses confronted with similar problems.

Meanwhile, two parallel breakthroughs have further helped accelerate the adoption of solutions for handling Big Data:

- The availability of Cloud based solutions has dramatically lowered the cost of storage, amplified by the use of commodity hardware. Virtual file systems, either open source or vendor specific, helped transition from a managed infrastructure to a service based approach;

- When dealing with large volumes of data, it is necessary to distribute data and workload over many servers. New designs for databases and efficient ways to support massively parallel processing have led to a new generation of products like the so called noSQL databases and the Hadoop map-reduce platform.

The table below summarizes the main features and problems connected to handing different types of large data sets, and explains how Big Data technologies can help solve them.

Aspect	Characteristics	Challenges and Technology responses
Volume	The most visible aspect of Big Data, referring to the fact that the amount of generated data has increased tremendously the past years. However, this is the less challenging aspect in practice.	The natural expansion of internet has created an increase in the global data production. A response to this situation has been the virtualization of storage in data centres, amplified by a significant decrease of the cost of ownership through the generalization of the cloud based solutions. The noSQL database approach is a response to store and query huge volumes of data heavily distributed.

[6] McKinsey Global Institute, *Big Data: The next frontier for innovation, competition and productivity* (June 2011), p. 4 f.

Velocity	This aspect captures the growing data production rates.More and more data are produced and must be collected in shorter time frames.	The daily addition of millions of connected devices (smartphones) will increase not only volume but also velocity. Real-time data processing platforms are now considered by global companies as a requirement to get a competitive edge
Variety	With the multiplication of data sources comes the explosion of data formats, ranging from structured information to free text.	The necessity to collect and analyse non-structured or semi-structured data goes against the traditional relational data model and query languages. This reality has been a strong incentive to create new kinds of data stores able to support flexible data models
Value	This highly subjective aspect refers to the fact that until recently, large volumes of data where recorded (often for archiving or regulatory purposes) but not exploited.	Big Data technologies are now seen as enablers to create or capture value from otherwise not fully exploited data. In essence, the challenge is to find a way to transform raw data into information that has value, either internally, or for making a business out of it.

Building end-to-end platforms that can efficiently process data of all dimensions is a challenge on its own and new business opportunities appeared for:

- Traditional IT vendors who want to provide integrated, industrial grade solutions
- Small- and medium sized businesses who can offer specialized products or services around open source software
- Professional services companies who can provide training and missing skills like “data scientists”

As of today, almost all major software and service vendors have, in one way or another, jumped on to the Big Data bandwagon as it is seen as an emerging and important market.

4. Economic and societal impact of Big Data: a new world of opportunities

According to the McKinsey Global Institute report on Big Data from 2012, the most developed regions, such as Europe, have the biggest potential to create value through the use of Big Data. The enormous economic impact of Big Data is further shown in another study[7] prepared by the Centre for Economics and Business Research (CEBR), estimating the value of Big Data to the UK economy alone, being £216 billion and 58,000 jobs in the next 5 years. Big Data is further expected to add more than €250 billion a year to the European

[7] Center for Economics and Business Research: *Data Equity – Unlocking the value of big data*, p. 4 ff. (April 2012)

public sector administration.[8] Thus, the whole European Union could benefit from the cumulative financial and social impact of Big Data.

Big Data analytics has started to impact all types of organisations, as it carries the potential power to extract embedded knowledge from big amounts of data and react according to it in real time. We exemplify some of the benefits by exploring the following different scenarios.

New technologies produce massive streams of data in real time and space that along time can make it possible to extract patterns of how the structure and form of the city changes and the way in which citizens behave. In such "**smart cities**", data gathered by sensors integrated with transport data, financial transactions, location of users, social network interaction will provide an entirely new dimension to thinking about how cities function. The danger associated with this aspect, relates to privacy and will be elaborated further down. Managing data in an effective way opens a wide field of opportunities for cities contributing to the improvement of services for citizens, such as: "on demand" and context-sensitive transportation strategies, optimized management of energy demand, more "holistic" and "preventive" health care approaches, development of new services such as e-voting, etc.

Various branches of experimental **science** generate vast volumes of experimental data. Petabytes (PB) of data per day is not uncommon in these fields (e.g. research in particle physics produces vast amounts of experimental data within short time frames). Fulfilling the demands of science requires a new way of handling data. Optimal solutions provided by Big Data technologies to analyse and properly compare disperse and huge datasets would provide huge benefits in terms of discoveries in experimental sciences.

Big Data in **healthcare** is associated with the exploding volume of patient-specific data. A prime example is medical imaging where even small pathological features measuring just a few millimeters can be detected in magnetic resonance imaging and in CT scans.

Doctors, already under significant time and cost pressure, will find it increasingly difficult to conduct their own analyses and error-free evaluations of the growing volume of data and images. They urgently need the support of automated solutions, which are increasingly based on the application of machine learning to large sets of example images. An even more dramatic increase in data volume is expected with the introduction of modern molecular analysis to medical praxis, for example by the increasing use of next-generation DNA sequencing. Naturally, imaging data and molecular data need to be evaluated in the context of the complete available patient information, such as all clinical data but also family history. Data protection and information security are particularly sensitive issues when it comes to medical data. If there is even the slightest perception that health records could fall into the wrong hands, the corresponding services will not be adopted.

The amount of mobile data traffic is expected to grow to 10.8 Exabyte per month by 2016. This tremendous growth is driven mainly by the increased usage of smart phones and tablets[9]. Big Data technology is needed in order to realize some advanced use cases in today's **mobile networks** and will be certainly required in future networks. Big Data is important for example for managing and operating mobile networks and gaining insights into the network with the goal to improve the network quality; which includes isolation and correlation of faults within the network, support of security related detection and prevention mechanisms, traffic planning, prediction of hardware maintenance, or the calculation of drop call probability.

[8] McKinsey Global Institute, *Big Data: The next frontier for innovation, competition and productivity*, p.2. (2012)

[9] Forrester: *Mobile is the new face of engagement*, (February 2012),

The changes brought by new web **social media technologies** mainly refer to the appearance of new types of content providers and new types of content, often referred to as ‘new media’. New media is giving the power of speech to the citizens who can now very easily report, blog and send short text messages (e.g., tweets), and rapidly creating new content of huge amounts. Traditionally, in the area of news media, conventional journalism has been the main trend, operating with standard news collection and broadcasting procedures while mediating mainstream types of content (e.g., politics, sport, economy, culture, health) from authoritative sources. Since a few years however, new Internet web technologies have appeared and have disrupted this business process. Traditional news media are getting more and more overcome by the rise of web news services.

In terms of associated business and economic activity, many software and services vendors already rely on Online Analytical Programming (OLAP) systems to perform their market or sells analysis. For this usage, Big Data technologies do not provide a clear advantage, and can be at best viewed as enablers to help scale legacy systems.

However, in order to move beyond this state and access finer details, past simple statistics, different approaches that are best supported by Big Data technologies are required.

In the following section, NESSI puts forward a number of areas where such technical and scientific challenges are becoming critical.

5. Technical and scientific challenges

5.1. Big Data Analytics

Because the current technology enables us to efficiently store and query large datasets, the focus is now on techniques that make use of the complete data set, instead of sampling.

This has tremendous implications in areas like machine learning, pattern recognition and classification, to name a few. Therefore, there are a number of requirements for moving beyond standard data mining techniques:

- a solid scientific foundation to be able to select an adequate method or design
- a new algorithm (and prove its efficiency and scalability, etc.)
- a technology platform and adequate development skills to be able to implement it;
- a genuine ability to understand not only the data structure (and the usability for a given processing method), but also the business value.

As a result, building multi-disciplinary teams of “Data scientists” is often an essential means of gaining a competitive edge.

More than ever, intellectual property and patent portfolios are becoming essential assets.

One of the obstacles to widespread analytics adoption is a lack of understanding on how to use analytics to improve the business[10]. The objects to be modelled and simulated are complex and massive, and correspondingly the data is vast and distributed. At the same time, the modelling and simulation software

[10] LaValle et al: *Big Data, Analytics and the Path From Insights to Value*, (Dec 2010)

solutions are expected to be simple and general, built on the solid foundations provided by a few robust computational paradigms and naturally oriented towards distributed and parallel computing. Hence, new methodologies and tools for data visualization and simulation are required.

5.2. Context awareness

Due to the increasing mobility of users and devices, context-awareness increases in importance. A suitable and efficient content- and context-aware routing of data is needed in many cases. Facing existing infrastructures and Big Data setups many solutions focus on processing and routing all data at once. For example, in manufacturing existing data has no relation to the context about the user´s history, location, tasks, habits schedule, etc. Concepts for taking the spatial users into account are a major challenge. The goal is to take the context into account for data that is not related to a user or context and present the right data to the right people and devices.

Applying contextual awareness can thus be a suitable approach to improve the quality of existing problem solving. In the context of Big Data, contextualisation can be an attractive paradigm to combine heterogeneous data streams to improve quality of a mining process or classifier.

In the last years, context awareness has widely demonstrated its crucial role to achieve optimized management of resources, systems, and services in many application domains, from mobile and pervasive computing to dynamically adaptive service provisioning.

Context-aware Big Data solutions could exploit context awareness to focus on only some portions of data (at least in first approximation) by keeping high probability of hit for all application-relevant events, with manifest advantages in terms of cost reduction and complexity decrease.

Research on industrial applicability on how to achieve the best trade-off between limitedness of the considered portions and probability of hits still necessitate significant investigation and research efforts.

In addition, context awareness is demonstrated to be useful to reduce resource consumption by concentrating Big Data generation processes (e.g., monitoring of real-world situations via cyber-physical systems) only on the sources that are expected to be the most promising ones depending on currently applicable (and often application-specific) context. For instance, “monitoring greediness” can be context-aware, thus permitting to switch off or decrease the duty cycle of expensive (in terms of power consumption, traffic bandwidth, …) sensors; industry-oriented research along this guideline is needed, with promising opportunities of cost/complexity reduction of Big Data solutions.

5.3. Rethinking data visualization and human-computer interfaces

Data visualisation is vital if people are to consume Big Data effectively. The reports generated from the analytics can be thought of as documents. These documents frequently contain varying forms of media in addition to textual representation. Even if textual representation alone is used, the sheer amount of it in large and complex documents requires carefully designed presentation for a digital screen.

When trying to represent complex information and the associated rationale, tasks, social networks and conceptual networks on screen(s), the design issues multiply rapidly. The interface to such information needs to be humane[11], i.e., responsive to human needs and considerate of human frailties. Frailties and

[11] Raskin, J. *The Humane Interface: New Directions for Designing Interactive Systems*. Addison-Wesley, Reading, MA (2000).

needs of knowledge workers are closely linked. They need relevant information in a just-in-time manner but too much information, which they cannot search efficiently, can hide that which is most relevant. They need to understand the relevance and relatedness of information but frequently have other work commitments which stop them from striving to establish relevance and relatedness.

Other considerations are important for data visualisation like visual clutter[12] which can lead to a degradation of performance at some task. A primary measure of visual clutter is *Feature Congestion* (based on Rozenholtz' Statistical Saliency Model) which is based on the notion that the more cluttered a visual image is the more difficult it is to add contextually relevant attention-grabbing items.

5.4. Visual Analytics: how we look at data

Visual analytics aims to combine the strengths of human and electronic data processing. Visualization, whereby humans and computers cooperate through graphics, is the means through which this is achieved. Seamless and sophisticated synergies are required for analyzing Big Data, including a variety of multidimensional, spatio-temporal data, and solving spatio-temporal problems.

Visual analytics is a relatively new term; it was introduced in research agenda books published in the USA[13] and EU[14]. However, the kinds of ideas, research and approaches that are now termed visual analytics emerged much earlier. The main idea of visual analytics is to develop knowledge, methods, technologies and practice that exploit and combine the strengths of human and electronic data processing. Visualization is the means through which humans and computers cooperate using their distinct capabilities for the most effective results.

The key features of visual analytics research include:

- Emphasis on data analysis, problem solving, and/or decision making;
- Leveraging computational processing by applying automated techniques for data processing, knowledge discovery algorithms, etc.;
- Active involvement of a human in the analytical process through interactive visual interfaces;
- Support for the provenance of analytical results;
- Support for the communication of analytical results to relevant recipients

As the majority of Big Data is dynamic and temporally referenced, it is necessary to take into account the specifics of time[15]. In contrast to common data dimensions, which are usually "flat", time has an inherent semantic structure. By convention, time has a hierarchical system of granularities organized in different calendar systems.

Another key issue is supporting analysis at multiple scales. There is much to do for visual analytics in order to change the traditional practice in analysis, focusing on a single scale. Appropriate scales of analysis are not always clear in advance and single optimal solutions are unlikely to exist. Interactive visual interfaces have a great potential for facilitating the empirical search for the acceptable scales of analysis and the verification of results by modifying the scale and the means of any aggregation.

[12] Rosenholtz, R., Li, Y., Nakano, L. (2007). Measuring visual clutter. *Journal of Vision*, 7(2):17, 1-22.

[13] Thomas, J.J., and Cook, K.A., editors (2005): Illuminating the Path. The Research and Development Agenda for Visual Analytics, IEEE Computer Society, 2005

[13] (13) Daniel Keim, Jörn Kohlhammer, Geoffrey Ellis and Florian Mansmann (Eds.)
Mastering The Information Age – Solving Problems with Visual Analytics. Eurographics, 2010

[15] G.Andrienko, N.Andrienko, U.Demšar, D.Dransch, J.Dykes, S.Fabrikant, M.Jern, M.-J.Kraak, H.Schumann, C.Tominski
Space, Time, and Visual AnalyticsInternational Journal Geographical Information Science, 2010, v.24 (10), pp. 1577-1600

To realize this potential, we need to know more about appropriate visual representation of different types of data at different spatial and temporal scales.

There is a need to develop corresponding analysis supporting interaction techniques, which should enable not only easy transitions from one scale or form of aggregation to another but also comparisons among different scales and aggregations.

5.5. Data management performance and scalability

Performance and scalability are central technical issues in order to deal with the huge volume of data to be stored and processed by Big Data systems and technologies. Two primary groups of technical issues call for significant advancements and industrially applicable research results.

On the one hand, there is the need for novel effective solutions dealing with the issue of data volume per se, in order to enable the feasible, cost-effective, and scalable storage and processing of enormous quantities of data. Promising areas that call for further investigation and industrially applicable results include effective non-uniform replication, selective multi-level caching, advanced techniques for distributed indexing, and distributed parallel processing over data subsets with consistent merging of partial results, with no need of strict consistency at any time.

On the other hand, another relevant performance/scalability issue worth of significant efforts relates to the need that Big Data analysis is performed within time constraints, as required in several application domains. The possibility to define quality constraints on both Big Data storage (for instance, where, with which replication degree, and with which latency requirements) and processing (for instance, where, with which parallelism, and with which requirements on computing resources) should be carefully taken into account.

Currently, process analysis in areas such as Business process management (BPM) is disconnected from data-based analyses in areas such as Data Mining and Business Intelligence. Bridges focusing on process mining will be also essential for progression of the Big Data theme.

5.6. Correlation and Causality

The causality of events (e.g. the unusual hot weather led to increased sales of soft drinks – where the event we are interested in is the increased sales of soft drinks and the cause was the unusually hot weather) in processes is of interest to enterprises (regardless of whether these are internal or external to the enterprise). In more complex scenarios causality will usually be latent across collections of data and spread across the four `V' dimensions of big-data. For example, sudden increases of Tweets consisting of unstructured mentions of flu symptoms as well as structured contextual data such as location and demographics may shortly precede a spree of medicinal purchases and stock exhaustion in particular channels and regions. In this contrived example there is a predicted *correlation* between Tweets and medicinal stock levels but the *causal claim* is that a statistically significant number of ‘medical problem’ Tweets (the cause) results in a ‘medical solution’ prediction about stock (the event). Research is needed to discover these kinds of latent causality patterns and enable causal chain exploration supporting people to make better informed decisions in shorter time.

When enterprise data, knowledge management, decision making and so forth, are in view, the notion of cause[16,17] is equally crucial.

The challenge is manifold particularly when considering the mix of structured/unstructured data. Therefore, any Big Data strategy needs to focus on a few areas:

- Discovering and modelling causality in structured data (reasonably well understood from the Data Mining perspective)
- Discovering and modelling causality in unstructured data (not well understood but growing work in Artificial Intelligence and Machine Learning etc.)
- Integrating unstructured causality models with structured causality models (not well understood but growing work in System Dynamics, Complex Event Processing, etc.)

Causality Modelling has had most attention in Computer Science from the perspective of known, well-defined data i.e. structured data. The Data Mining community has a long history (for the Computer Science discipline) that has built on the even more historical disciplines of Statistics, Operational Research, and Probability. Such a history provides a solid start for Causality Modelling. However, a particularly difficult subset of Causality Modelling is unstructured data e.g. text, multimedia (whether user generated content or not). This is in contrast to structured data e.g. point-of-sale transactions, stock prices, sensor readings. As well as the large volume of data there are other challenges. The unstructured data must be structured or pre-processed into a form amenable for analysis. A further subset of unstructured data which is of particular interest to a broad range of businesses is text data. Text data is prevalent in practically every business domain imaginable, from Customer Relationship Management to Marketing to Product Design and so forth. Complex, valuable knowledge involving topics, products, designs, relationships, opinions, sentiment and argumentation, to name but a few, is locked in text data.

The potential outcomes of causality discovery for the business enterprise will be dependent on machine processing due to the Big Data four V's and the need for near real-time analysis to support human decision makers and more accurate simulations and predictions.

5.7. Real time analytics and stream processing

In many business scenarios it is no longer desirable to wait hours, days or weeks for the results of analytic processes. Psychologically, people expect real-time or near real-time responses from the systems they interact with. Real-time analytics is closely tied to infrastructure issues and recent move to technologies like in-memory databases is beginning to make 'real-time' look achievable in the business world and not just in the computer science laboratory.

Handling large amounts of streaming data, ranging from structured to unstructured, numerical to micro-blogs streams of data, is challenging in a Big Data context because the data, besides its volume, is very heterogeneous and highly dynamic. It also calls for scalability and high throughput, since data collection related to a disaster area can easily occupy terabytes in binary GIS formats and data streams can show bursts of gigabytes per minutes.

The capabilities of existing system to process such streaming information and answer queries in real-time and for thousands of concurrent users are limited. Approaches based on traditional solutions like Data Stream Management Systems (DSMS) and Complex Event Processors (CEP), are generally insufficient for the

[16] Pearl, J. *Causality: Models, Reasoning and Inference*. 2nd ed. Cambridge University Press. (2009)
[17] http://www.causality.inf.ethz.ch

challenges posed by stream processing in a Big Data context: the analytical tasks required by stream processing are so knowledge-intensive that automated reasoning tasks are also needed.

The problem of effective and efficient processing of streams in a Big Data context is far from being solved, even when considering the recent breakthroughs in noSQL databases and parallel processing technologies

A holistic approach is needed for developing techniques, tools, and infrastructure which spans across the areas of inductive reasoning (machine learning), deductive reasoning (inference), high performance computing (parallelization) and statistical analysis, adapted to allow continuous querying over streams (i.e., on-line processing).

One of the most open Big Data technical challenges of primary industrial interest is the proper storage/processing/management of huge volumes of data streams. Some interesting academic/industrial approaches have started to mature in the last years, e.g., based on the Map Reduce model to provide a simple and partially automated way to parallelize stream processing over cluster or data centre computing/storage resources. However, Big Data stream processing often poses hard/soft real-time requirements for the identification of significant events because their detection with a too high latency could be completely useless.

New Big Data-specific parallelization techniques and (at least partially) automated distribution of tasks over clusters are crucial elements for effective stream processing. Achieving industrial grade products will require:

- New techniques to associate quality preferences/requirements to different tasks and to their interworking relationships;
- New frameworks and open APIs for the quality-aware distribution of stream processing tasks, with minimal development effort requested by application developers and domain experts18.

5.8. Distributed Storage

It is an obvious fact that there are increasingly distributed data sets being used for Big Data analytics. Even within one organization there tends to be a number of data sets that require analysis. If this is expanded to a supply chain in one particular business domain and then to cross-domain supply chains the distributed nature of data sets increases. Including also the volume and velocity of the data and performing meaningful analytics becomes a major challenge. Then, if security and legal issues are considered, the distributed nature of data sets presents one of the most complex problems to solve.

One approach which displays promise is the use of data marts. However, data marts are not yet designed with the features of Big Data in mind. Another approach is the use of Big Data Analytics Agents (BDAA). The agents perform a specific function or set of functions on data sets. They are dispatched to appropriate locations and must be security cleared by the receiving data set. Once they have performed their analytics the data set then verifies the agents' findings before letting it return to its sending location. BDAA's could then be designed for the specific features of Big Data e.g. BDAA's that sit for some period of time on data streams or BDAA's that specifically analyse video for predefined features.

[18] In addition, the availability of such open APIs can help in reducing the entrance barriers to this potentially relevant market, by opening the field to small/medium-size companies, by reducing the initial investments needed to propose new stream processing-based applications, and by leveraging the growth of related developers/users communities

5.9. Content Validation

Validating the vast amount of information in content networks is a major challenge, since there is a very large number of different types of sources, such as blogs, social networking platforms, or news sites with social networking functionalities, and different types of content, such as articles, comments, tweets, etc. Furthermore, the complexity of human language is such that it is not easy to derive validation rules that apply to all different discussion subjects.

The vast amount of data that exists today, and the frequency with which it can be updated, render the validation task very difficult and make complex rules which require great computational resources infeasible.

Therefore, there is a need to derive simple rules for validating content, and leverage on content recommendations from other users. The recommending users must themselves be assessed on the basis of trust and reputation criteria. Further, there is a need for learning algorithms to update rules according to user feedback. Machine learning algorithms can be an efficient solution for extracting patterns or rules for Web documents semi-automatically or automatically.

6. Big Data Business Ecosystem

A business ecosystem, first defined by James F Moore[19], is “An economic community supported by a foundation of interacting organizations and individuals.”[20] In this regard, an ecosystem does not yet exist in Europe for Big Data. However, from at least one perspective a Big Data ecosystem (removal of ‘business’ is deliberate) does exist in many industries in a very simple form. For example, within an aerospace ‘ecosystem’ there will be a vast amount of data used across complex supply chains about materials, construction methods, testing, simulation, and so forth. Such data is moved from system to system in the various processes according to defined standards and within regulation. It is also analysed to establish patterns, failures, standards, and so forth.

Consider another example, fraud detection[21]. Public sector fraud for 2011-2012 cost the UK tax payer £20.3 billion and private sector fraud was £45.5 billion, of which £16.1 billion was retail fraud. In 2008-9 the National Fraud Initiative traced £215m in fraud, error and overpayments using standard analytics. How much more could be pre-empted and traced if there was a Big Data Business Ecosystem?

Moving from such specific Big Data ecosystems to a Big Data Business Ecosystem will not be a straightforward evolution. The problem for Big Data is small patterns[22] “precisely because so many data can now be generated and processed so quickly, so cheaply, and on virtually anything, the pressure is to be able to spot where the new patterns with real added value lie in their immense databases and how they can best be exploited for the creation of wealth and the advancement of knowledge. Small patterns matter because they represent the new frontier of competition, from science to business, from governance to social policies.”

[19] Moore, J.F., *Predators and Prey: A New Ecology of Competition, Harvard Business Review*, May/June 1993 (available online at: http://blogs.law.harvard.edu/jim/files/2010/04/Predators-and-Prey.pdf) (1993)

[20] Moore, J.F. *The Death of Competition: Leadership and Strategy in the Age of Business Ecosystems*, HarperBusiness. (1996)

[21] Cebr Report for SAS: *Data Equity: Unlocking the Value of Big Data*, Centre for Economics and Business Research Ltd (April 2012)

[22] Floridi, F.*, Big Data and Their Epistemological Challenge*, *Philosophy & Technology*, Nov 2012, DOI 10.1007/s13347-012-0093-4 (2012).

This gives an idea about some of the trends that are emerging and that could provide the basis for a Big Data Business Ecosystem.

The first one refers to **data science and that associated skills are in high demand**. The education challenge is not only to teach students fundamentals skills such as statistics and machine learning, but also appropriate programming ability and visual design skills. Individuals who have a little imagination, some basic computer science and enough business acumen will be able to do more with data either for their own business or as part of a pool of talented “data scientists”. Organisations – and governments - able to achieve this multi-disciplinary cross-fertilisation will have a tremendous advantage as, ultimately, this could result in a self-feeding cycle where more people start small and are able to work their way up to building advanced data-analysis products and techniques used as the foundation for the next generation of data products built by the next generation of data scientists. A **European Big Data Analytics Service Network** to improve skills capacities within the software industry could be highly beneficial for achieving this aim.

Secondly, the **generalisation of machine learning** is crucial for creating a viable Big Data Business Ecosystem. Consumers - and advertisers - want more personalisation and more responsive applications: the prospect of writing models that discover patterns in near real-time is so appealing that it’s difficult to imagine a company not considering it as a new revenue stream. While machine learning has been recently promoted to the front-line, it does not mean it’s easy to do. To the contrary, effective machine learning requires a strong academic background and the ability to transform a mathematical model into a marketable product.

As a third point, the **commoditisation of Big Data platforms** plays an important role for any Big Data Business Ecosystem. The technological improvements in infrastructure, database, telecoms, and so forth are nothing without applications that can take advantage of them, and two approaches are driving the market. The first one is to make Big Data accessible to developers by making easy to create applications out of pre-packaged modules tied to a Big Data backend. This is in essence the promise of delivering “Data as a Service” or “Algorithms as a Service” in a Cloud environment. The second approach is to find a convincing use case for Big Data like face recognition, user behaviour analysis or network security, and turn it into a commercial product that companies can start using off the shelf.

But there are further trends that have appeared in the context of creating a Big Data Business Ecosystem, such as[23]:

Cloud services expand the state of the art architectural design and implementation pattern, into business relationships between service providers and consumers. IT personnel need to develop new skills that come close to those required from data scientists.

Increasing Cross-Enterprise Collaboration is a business necessity in many instances, but it requires sharing, exchanging, and managing data across enterprise walls. Erbes *et al* state that such cross-enterprise collaboration “will create new ecosystems that rely on effectively and selectively enabling access to the required systems and services.”

It is also helpful to understand that Big Data has a social life[24]. It moves through processes often in a transformed state and can have different importance at different stages of a process. Davenport *et al*[25] state that organisations capitalising on Big Data differ from traditional data analysis in three ways:

[23] Erbes, J., Motahari-Nezhad, H.R. and Graupner, S. *The Future of Enterprise IT in the Cloud*, *IEEE Computer*, Vol. 45(5), pp. 66-72 (2012)

[24] Brown, J.S. and Duguid, P, *The Social Life of Information*. Harvard Business School Press (2000).

1. They pay attention to data flows as opposed to stocks.
2. They rely on data scientists and product and process developers rather than data analysts.
3. They are moving analytics away from IT function and into core business, operational and production functions.

We propose that, in light of the above, a Big Data Business Ecosystem should be considered from a process-centric perspective. Data exist, whether *at rest* or *in motion,* within processes that span traditional LoB's and sectors. Such processes will have a variety of resources, from people to facilities, which relate to the data. Delineating the processes and resources is a vital step in evolving immature Big Data Business Ecosystem to more mature levels.

Consider the following example. An automotive production line makes use of several different tools. These tools have an uptime of 99.8%. The production line process is owned by the automotive manufacturer and they monitor the tool data to ensure maximum efficiency. The tool manufacturer has a cross cutting process related to tool maintenance and receives constant data streams from the tools which they analyse for potential performance issues and service events. The production line tooling is producing data that are relevant to both processes. Some of the data will be more relevant to one or the other process.

In the above contrived example it may be straightforward to see the security and regulatory issues. However, real life scenarios will be much more complicated. Business processes will span countries which frequently have different regulations regarding data, they will span organisations and they will span various data sources. We propose that a Big Data Business Ecosystem can be built with a clear understanding of and policies addressing, the following:

- Legal and regulatory issues between countries.
- Cross-enterprise collaboration issues.
- Nature of data sources and the basic translations to be applied to them.

6.1. An EU Big Data Business Ecosystem

At the highest level an EU Big Data business ecosystem must accommodate two scenarios, the first scenario indicating **unhindered innovation supported by open data access**. Open access to Big Data should not necessarily result from a clear and well understood roadmap. Opening access to certain Big Data sources enabling innovation will help drive an ecosystem of SMEs.

The second scenario would support **innovation by principled collaboration between business process owners**. An EU Big Data business ecosystem is an important factor for commercialisation and commoditisation of Big Data services. Especially for enterprises, an ecosystem requires several roles such as data integrators, Big Data service providers, real-time data vendors, and so forth. The segments for typical Big Data vendors are mainly split into the four: applications, analytic tools, data management, and infrastructure.

Within an international ecosystem several roles have to be applied. The service provider/vendor is the first

[25] Davenport, T.H., Barth, P. and Bean, R. *How 'Big Data' is Different. MIT Sloan Management Review*, (July 2012) (available online at: http://sloanreview.mit.edu/the-magazine/2012-fall/54104/how-big-data-is-different/)

actor and basically the owner of the data. The provider can be split into the private sector, public sector and individuals. Of course, each of them has specific challenges and motivations for joining the ecosystem. A catalyst is required to serve legitimacy, which could be the role of, e.g., the government. For example, business analytics based on trade date would become easier for the EU by joining multiple geographic groups. Also, consumers with a clear benefit are additional actors. Finally, data scientists and experts are necessary due to specific technical skills and domain knowledge.

For building a working ecosystem the benefit for all actors must be clear. For example, companies must profit from offering and/or consuming services. Consequently, a challenge will become the orchestration of such services. Furthermore, the EU has to deal with special challenges due to the variety of public and private instances across country boarders. The major challenge for a broader ecosystem will be privacy and security. This is a crucial success factor for all actors to take part. Especially for the EU the instrumentation of standards across the borders is an additional gap to be fulfilled.

6.2. Big Data Market Places

A further perspective in conceiving of a Big Data Business Ecosystem is that of Big Data Market Places. Despite the value that is hidden in Big Data sets, there are certain challenges associated with the cost and complexity of publishing such data as well as the cost and complexity of consuming and utilising the data. Many data providers, who have stockpiled vast amounts of interesting data, struggle with the problem of finding ideas for creating novel services using their data, identifying what makes their data relevant for potential consumers, and deploying solutions for rapid integration of data for loosely defined services. The lack of a sustainable data marketplace ecosystem for Big Data, where producers of data can effectively disseminate their data and consumers can interact with the providers in new ways, enabling efficient delivery of new applications and services, hinders the development of novel data-driven business models in the Big Data domain.

Current data markets face various problems such as data discovery, curation, linking, synchronization and distribution, business modelling, sales and marketing. This situation calls for new technologies, infrastructure, approaches, and methodologies to create and sustain a Big Data marketplace ecosystem. This will require, amongst others, improving current practices of publishing and consuming Big Data; tool-supported methodologies for efficient publication, dissemination and consumption of data in data marketplaces; scalable data dissemination and communication approaches between data providers and consumers on the data marketplaces.

7. Privacy and Big Data

7.1. “Without appropriate privacy no benefits from using Big Data technology”

With the advent of Big Data, systematic collection, storage and analysis of personal data has dramatically increased. From internet logs, user information can be extracted that is accessible for surveillance and marketing purposes; identity management tools are now used on the Internet to track the identity of users; in the physical world cameras are used for surveillance; mobile phones send location information to the network providers; debit and credit card payment systems reveal the amounts spent and stores visited. Store loyalty cards allow analysing consumer behaviour; and social media allow user-to-user contact and access to pictures, videos and movies.

An increasing number of companies, established global players , SME’s and start-ups, built their business models on using and selling user profiles generated from these data sources. Data mining tools sift through the data to find patterns in large collections of personal data, to identify individuals and to predict preferences and interests. These patterns and predictions are stored in company databases and combined with new data. Also governments have increasing use for analysing and exchanging information about their citizens. Overall, collection, storage, analysis and usage of personal data are now part of our everyday life at all levels of society.

NESSI members do not believe that realizing the promise of analytics requires the sacrifice of personal privacy. A truly modernized legal framework should provide for a high level of data protection while leaving sufficient flexibility to business to innovation, create jobs and growth and remaining competitive at a global level.

7.2. A New Legislation

The Data Protection Directive dates from 1995, back to a time when the internet use was not widespread. So far, the 27 EU Member States have implemented the 1995 rules differently. This results in divergences in enforcement. In January 2012, the European Commission proposed a comprehensive reform of the EU's 1995 data protection rules to strengthen online privacy rights and boost Europe's digital economy.[26] A single law, so it is hoped, will do away with the current fragmentation and costly administrative burdens,

However, the benefits of greater harmonisation are at risk of being outweighed by the administrative burden and subsequent costs the draft regulation would impose on business without actually enhancing the protection of the user; the proposed text is adding administrative burden and not cutting red tape;

As currently drafted, the regulation represents a step backwards for competitiveness in the European Union at a time of economic crisis when business is facing many other challenges. Newly proposed obligations are too complex to be properly understood, coupled with constraints on implementation that limits flexibility needed by business to efficiently utilize technologies – including big data usages --to remain competitive in a global economy.

Many of the proposed provisions were drafted with a specific context in mind without considering the consequences for the wider industry. Overly prescriptive requirements inhibit the goal of the Regulation to be technology neutral. We and the upcoming generation need to be able to use new technologies to address Europe’s most difficult societal and economic problems. With a, all too rigid, legal framework we will not be able to live up to this promise.

In this context we are especially concerned about the ‘one-size fits all’ approach for consent and the wording of Article 20 of the draft which is putting even the legitimate use of analytics/big data (or here profiling) at risk. The need to ensure that the provisions relating to “profiling” do not prevent businesses from being able to evaluate and analyse data and use such data predicatively for legitimate business purposes, including identity verification and fraud detection and prevention.

An in-depth privacy by design scheme, which is not imposing a ‘one size fits all’ rule but leaves flexibility to business could go a long way in delivering both the promise that Big Data applications can bring while providing strong safeguards for data protection. This must be coupled with strong enforcement and severe fines for stakeholders abusing the scheme and being deliberately and notoriously non-compliant with data protection rules.

[26] European Commission Press release: *Commission proposes a comprehensive reform of data protection rules to increase users' control of their data and to cut costs for businesses* (January 2012) [available online: http://europa.eu/rapid/press-release_IP-12-46_en.htm?locale=en]

7.3. Privacy Preserving Data Mining.

There is a research direction called privacy-preserving data mining that aims to reconcile the tension between Big Data and privacy. The upcoming legislation, and the sentiments of at least parts of the public, makes privacy-preserving methods a topic of strong interest. It can be categorized as follows. In **privacy-preserving data publishing** the challenge is to publish the data with appropriate suppression, generalisation, distortion, and/or decomposition such that privacy of individuals is not compromised and yet the disclosed data is useful for analytical purposes. A large body of work inspired by k-anonymity has flourished in this area. In **privacy-preserving data mining**, the main assumption is that private data is collected for the purpose of consolidating the data and analysing it with one or more data mining algorithms. The collector is not a trusted party, so data is subjected to a random disconcertion as it is collected. In **privacy-preserving pattern publishing**, the central question addressed is how to publish frequent patterns without revealing sensitive information about the underlying data. In **pattern hiding** the main concern is how to transform the data in a way that certain patterns cannot be derived (via mining), while others can. In addition, **secure multiparty mining** over distributed datasets, indicates that data on which mining is to be performed is partitioned, horizontally or vertically, and distributed among several parties. The partitioned data cannot be shared and must remain private but the results of mining on the "union" of the data are shared among the participants.

Although a number of interesting approaches have been developed, these methods are not generally known and not in widespread use in industry. Attempts are needed to make them better known and to bring them to practice.

8. SWOT analysis of the EU Software and Services industry in the context of Big Data

Using Big Data as a new asset is very a promising idea for the European economy, especially for the large base of SME's. The following analysis underlines strengths, weaknesses, opportunities and threats (SWOT) with respect to the European software industry and how they could use big data analytics to improve their products and services. This section especially addresses the SME initiative on analytics aiming to help European SMEs acquire the competences and resources they need to develop innovative content and data analytics services[27].

8.1. Strengths

- Europe has a large group of commercially successful SME's focussing on market driven innovation within their niche and with a focus on export. In Germany, they are known as "hidden champions" but they exist in other European countries as well. Because these SME's operate in very narrow niches they have to focus on global markets to work on an economic scale. Those companies benefit from exclusive insights about their global customers to build market oriented products and services.

- There is deep knowledge about local markets and local customer problems and ability to develop customized products such as language dependent products, legislation dependent products.

[27] FP7, ICT Work Programme 2013, Objective 4.3, p. 53.

- Effective (although fragmented) research and development networks between universities, research centres and SME's are established and has been referred to as a "role model for decentralized R&D".
- There is a well-established understanding in export oriented companies to deliver high-value products and services (e.g. "premium cars") by using innovation networks.
- Growing interest from SME's using cloud computing and software services

8.2. Weaknesses

- While US-based companies like Yahoo, Google, or Twitter are widely recognized for their activities in Big Data, very few research organizations, including SMEs, are known for their activities and initiatives in this field in Europe.
- SME's lag behind larger enterprises in taking up the Big Data challenges, especially compared to the US
- In most SME's there is not enough understanding on how to gain new insights by using data analytics concerning their customers, products and services.
- The capabilites to develop an individual data strategy, to select and integrate the right data sources and to use effectively big data analytics for leveraging exclusive insights for product and business development are crititcally underdeveloped.
- No wide spread knowledge about freely available data ("getting new insights by merging different data sources")
- Analytical Big Data services for SME's within Europe are currently non-existing

8.3. Opportunities

- Rising customer demands in Europe for smarter products, higher individualization, and mass customization.
- Europe's SME's can benefit by enhancing their products and services with big data analytics and privacy by design, e.g. offering preventive maintenance services in the utilities industry, usage-based analytics for distant product development, self-learning behaviors for energy optimization (e.g. Nest thermostate) or new business models thanks to better usage insights (like Rolls-Royce and their power by the hour performance based contracting).
- Europe's SME's could benefit enormously by reducing copyright infringements with smart networked products and services.
- Better use of freely available data.
- Levelling the playing field by giving access to formerly very demanding analytical tools through commoditization.
- Developing new products and services enhanced with Big Data analytics and privacy by design, developing products adapted to European privacy standards

- Established innovation networks could reduce skill shortages by spreading knowledge with online trainings and generating hands-on expertise based on commoditized analytical services.

8.4. Threats

- Globally there is a fast growing knowledge about using data as an asset class for leveraging the industrial competitiveness, for example in China, India.

- There is a massive venture-capital driven development in the US to commoditize formerly complex big data analytics for the mainstream market, www.bigml.com is an example of the commoditization efforts around Hadoop. SME's in Europe have no alternatives to choose comparable services from European vendors.

- Insufficient capabilities of European companies to scale to the world market.

- There are serious concerns among European businesses about using big data analytics neglecting European privacy standards and expectations. In the meantime, foreign competitors take their market shares by pushing their own de-facto standards in respect to privacy.

9. Recommendations

In summary, there are several aspects to consider in order for Big Data to provide a beneficial impact on society and to create growth and jobs for Europe. There is especially added value in the application layer of Big Data which is why, it is important to bring together the domain knowledge with the analytical expertise needed to identify meaningful correlations of data. In extension, there is a need to develop predictive use cases, and to support real-time processing and analysis. Another important aspect is the protection of privacy related data. In the section below, NESSI aims to bring forward a number of important aspects which need to be addressed in any research programme or policy initiative on Big Data. The recommendations have been divided into four highly interlinked sections.

9.1. Technical aspects

NESSI supports the need to direct research efforts towards developing highly scalable and autonomic data management systems associated with programming models for processing Big Data. Aspects of such systems should address challenges related to real-time processing, context awareness, data management and performance and scalability, correlation and causality and to some extent, distributed storage. In detail, NESSI suggest the following directions for future research efforts.

1. Align the ability of **data** analysis algorithms towards the new constraints generated by the treatment of massive data collections (heterogeneity, multimodality, size etc.)

2. Concerning context awareness, research on industrial applicability on how to achieve the best trade-off between limitedness of the considered portions and probability of hits is highly important and still necessitate significant investigation.

3. With regards to visual analytics, it is important to - on a general scale - promote the need for understanding the relevance and relatedness of information.

4. There is a need to know more about appropriate visual representation of different types of data at different spatial and temporal scales. It is important to develop corresponding analysis tools

supporting interaction techniques, which should enable not only easy transitions from one scale or form of aggregation to another, but also comparisons among different scales and aggregations.

5. Areas that call for further investigation and industrially applicable results within the context of management performance and scalability include effective non-uniform replication, selective multi-level caching, advanced techniques for distributed indexing, and distributed parallel processing over data subsets with consistent merging of partial results, with no need of strict consistency at any time.

6. The possibility to define quality constraints on both Big Data storage (for instance, where, with which replication degree, and with which latency requirements) and processing (for instance, where, with which parallelism, and with which requirements on computing resources) should be carefully taken into account in future research activities.

7. In relation to modelling causality, research efforts would be directed towards discovering and modelling causality in structured data (reasonably well understood from the Data Mining perspective), discovering and modelling causality in unstructured data (not well understood but growing work in Artificial Intelligence and Machine Learning etc.) and finally, integrating unstructured causality models with structured causality models (not well understood but growing work in System Dynamics, Complex Event Processing, etc.).

8. New Big Data-specific parallelization techniques and (at least partially) automated distribution of tasks over clusters are crucial elements for effective real-time stream processing. Achieving industrial grade products will require new techniques to associate quality preferences/requirements to different tasks and to their interworking relationships together with new frameworks and open APIs for the quality-aware distribution of stream processing tasks, with minimal development effort requested by application developers and domain experts28.

9. Regarding validation, it is necessary to derive simple rules for validating content, and leverage on content recommendations from other users. Machine learning algorithms can be an efficient solution for extracting patterns or rules for Web documents semi-automatically or automatically.

9.2. Business aspects

10. NESSI calls for a mechanism on EU-level which will foster and structure information and business exchange between key actors in the European "Big Data Ecosystem". This Big Data Business Ecosystem should be considered from a process-centric perspective. NESSI underlines the importance of fostering industrial knowledge transfer by establishing European Big Data Centres of excellence (like Hack/Reduce in Boston).

11. NESSI stresses the importance of fostering Open Source Big Data Analytics to make sure the benefits stay in the EU, with European developers, users and entrepreneurs.

12. NESSI suggests raising awareness about the possibilities of integrating existing private data with external data to enhance existing products and services. NESSI further advocates that a Big Data

[28] In addition, the availability of such open APIs can help in reducing the entrance barriers to this potentially relevant market, by opening the field to small/medium-size companies, by reducing the initial investments needed to propose new stream processing-based applications, and by leveraging the growth of related developers/users communities.

business ecosystem can only be built with a clear understanding of, and policies addressing legal and regulatory issues between countries, cross-enterprise collaboration issues and the nature of data sources and the basic translations to be applied to them.

13. NESSI believes that the European Commission should move beyond the Open Data Initiative, towards an interoperable data scheme, including protocols that help data processing from heterogeneous sources. This could serve as an incentive for a normalized way to manage data related to people, etc.

9.3. Legal aspects

It is apparent that the Big Data topic raises legal concerns related to the fair and liable use of data. To address these issues, NESSI puts forward the following recommendations.

14. There is a strong need to ensure that the provisions relating to "profiling" do not prevent businesses from being able to evaluate and analyse data and use such data predicatively for legitimate business purposes, including identity verification and fraud detection and prevention. Within the context of the revised Data protection legal framework, NESSI suggests an in-depth privacy by design scheme, which is not imposing a 'one size fits all' rule but leaves flexibility to business.

15. NESSI supports the idea that any Big Data analytics commoditization should be developed with respect to European privacy regulations (privacy by design) as well as Cloud Computing strategies. NESSI recommends developing such privacy by design regulations within already existing standardisation bodies.

16. Although a number of interesting approaches have been developed with regards to privacy preserving data mining, these methods are not generally known and not in widespread use in industry. NESSI advocates for making them better known and to bring them to practice.

9.4. Skills

As mentioned in the chapter about a Big Data eco-system, skills for data science and associated skills are in high demand.

17. NESSI encourages the set up of education programs on data science. The education challenge is not only to teach students fundamentals skills such as statistics and machine learning, but also appropriate programming ability and visual design skills.

18. NESSI supports the idea of creating of a European Big Data Analytics Service Network (EBDAS) to increase exchanges between data scientists and businesses in order to improve skills capacities within the software industry which will lead to the next generation of data-analysis products and applications.

19. NESSI stresses the importance of creating resources for using commoditized and privacy preserving Big Data analytical services within SME's. This includes the creation of Big Data Centres, offering hands-on trainings, generating hands-on expertise for SME's using Big Data for their products and services, and leveraging access to large-scale compute clusters, online-trainings, specialized consulting.