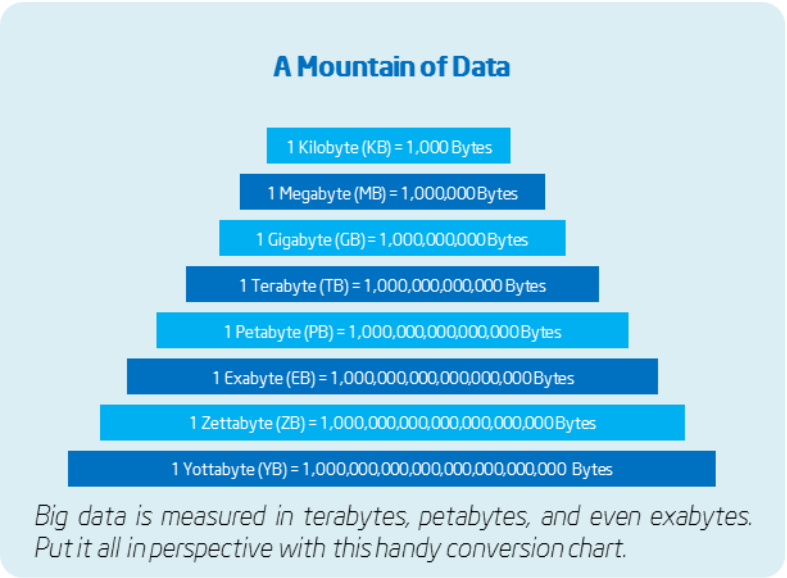
A vision for Big Data

From Intel's perspective, Big Data is an unprecedented opportunity for innovation, growth and job creation. Understanding the potential whilst identifying challenges is essential to delineate policies that encourage take-up of Big Data analytics by the public and private sectors, and protect the interests of all involved.

Big Data Fundamentals

Data has been growing in the last decade at an astonishing rate. From the dawn of civilization to 2003 mankind generated 5 exabytes[1] of data. In 2012, the digital universe of data grew to 2.72 zettabytes[2] (ZB) and will double every two years to reach 8 ZB by 2015[3]. This exponential data growth is accelerating due to the increasing digitalization and to new data generators such as the internet of things.

Data sets are larger, more varied in structure and format, and generated at a faster rate than ever before. The Big Data phenomenon creates tremendous opportunities for society to foster data-driven innovation that can enable faster and better decision-making, building a competitive advantage for our knowledge society.



Big data is measured in terabytes, petabytes, and even exabytes. Put it all in perspective with this handy conversion chart.

Big Data refers to huge data sets that are orders of magnitude larger (*volume*); more diverse, including structured, semi-structured and unstructured data (*variety*) and arriving faster (*velocity*) than any organization has had to deal with before. This flood of data is generated by connected devices –from PCs and smart phones to sensors such as RFID readers and traffic cams. Plus, it is heterogeneous and comes in many formats, including text, document, image, video, and more.

[1] 1 Exabyte (EB) = 1,000,000,000,000,000,000 Bytes
[2] 1 Zettabyte (ZB) = 1,000,000,000,000,000,000,000 Bytes
[3] "Big Data 101: Unstructured Data Analytics", Intel 2012
http://www.intel.com/content/dam/www/public/us/en/documents/solution-briefs/big-data-101-brief.pdf

© 2013, Intel Corporation. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and other countries. *Other names and brands may be claimed as the property of others.

These three Vs (Volume, Variety, Velocity) characterize what Big Data is all about, and also help define the major issues that IT needs to address:

- **Volume.** The massive scale and growth of unstructured data outstrips traditional storage and analytical solutions.
- **Variety.** Traditional data management processes can't cope with the heterogeneity of Big Data – or "shadow" or "dark data," such as access traces and Web search histories.
- **Velocity**. Data is generated in real time, with demands for usable information to be served up immediately.

This explosion in data presents significant new business opportunities. In recent market research[4] Mckinsey, the global consultants, assigned value to the Big Data. For instance, a retailer using Big Data to its full potential could increase operating margin by 60%; European governments could save more than €100 billion in operational efficiency improvements; and personal-location data could capture €485 billion in consumer surplus[5].

A Novel Technological Challenge

According to an Information Week survey in 2012[6], the top five Big Data contributors are financial transactions, e-mail, imaging data, weblogs, and Internet text and documents. This data is primarily generated by business and social sources. Intel's own research backs up these findings and adds machine generated data from sensors and devices into the top five.[7]

For decades, businesses have collected data, analysed it using a variety of Business Intelligence tools, and generated reports. But traditional reporting and data mining tools cannot handle the vast

Internet of Things

Today, traditional embedded devices - tiny, single-function computers 'embedded' in everyday items like TVs, cash machines and traffic lights - are being replaced by 'intelligent' systems that communicate and share data with other intelligent devices and people over the Internet. This is the machine-to-machine communication (M2M) which underlies the 'Internet of Things.'

[4] "Big Data: The next frontier for innovation, competition, and productivity", McKInsey Global Institute, May 2011
[5] "Consumer surplus, (...) in economics, the difference between the price a consumer pays for an item and the price he would be willing to pay rather than do without it." Encyclopædia Britannica http://www.britannica.com/EBchecked/topic/134573/consumer-surplus
[6] "The Big Data Management Challenge." Information Week 2012
http://reports.informationweek.com/abstract/81/8766/business-intelligence-and-information-management/research-the-big-data-management-challenge.html
[7] "Big Data Analytics: Intel's IT Manager Survey on How Organizations Are Using Big Data" Intel 2012
intel.com/content/www/us/en/big-data/data-insights-peer-research-report.html

volume of Big Data, although the variety and velocity of the data often present even greater challenges.

Big Data analytics is one of the great new frontiers of IT. Data is exploding so fast and the promise of deeper insights is so compelling that organisations and governments are highly motivated to turn Big Data into an asset they can manage and exploit.

Big Data includes three types of data—structured, semi-structured, and unstructured—and Intel's IT Manager Survey of 200 IT professionals found that four of the top five data sources for IT managers today are semi-structured or unstructured. Many businesses are simply unable to analyse these emerging forms of data, which include everything from e-mails, photos, and social media to videos, voice, and sensor data.

New technologies are emerging to make unstructured data analytics possible and cost-effective. The new approach redefines the way data is managed and analysed by leveraging the power of a distributed grid of computing resources. For instance, Big Data requires clusters of servers to support the tools that process large volumes, high velocity, and varied formats of Big Data. Clouds are already deployed on pools of servers and can scale up or down as needed for Big Data.

Apache* Hadoop is evolving as a new approach to unstructured data analytics. Hadoop is an open-source framework that uses a simple programming model to enable distributed processing of large data sets on clusters of computers. The complete technology stack includes common utilities, a distributed file system, analytics and data storage platforms, and an application layer that manages distributed processing, parallel computation, workflow, and configuration management.

Intel is committed to work with the open source community. The Intel Distribution for Apache Hadoop software is the first with hardware-enhanced performance and security capabilities.

At the technology level Intel provides technologies across the entire Big Data spectrum, from capturing data to storage, management, analysis and presentation. For example, Intel Intelligent Systems Group, the embedded systems division, has developed technologies that capture, secure and aggregate the data before it is shipped across the network to the data centre.

The policy environment

Many sectors can benefit from Big Data. Some of them have their own policy and regulatory environments (e.g. healthcare, finance, retail), and those will constrain the way Big Data can be applied and used. This is no different from what already happens today to any other kind of data. There are, however, policies that

affect Big Data as a whole, and will impact how it translates in real gains for the European economy. Examples of such policies are those related to R&D, trust, the role of public sector and skills.

Research priorities

Research in Big Data should be grounded in the quadruple helix model[8] where civil society joins with business, academia, and government sectors to drive changes far beyond the scope of what any organization can do on their own.

Managing and making sense of large volumes of data are the main challenges behind Big Data. Intel, through the Intel Science and Technology Center for Big Data (ISTC)[9] and across Intel Labs has focused its research on five major themes: databases and analytics; data math and algorithms; visualisation; architecture, and streaming. These themes address issues such as developing new software platforms for storing and processing massive amounts of data; the development of algorithms for parallel execution and for data that does not necessarily fit in memory; designing visualizations and interfaces that allow users to interact with massive data sets, on displays ranging from phones to video walls; understanding how next-generation hardware innovations – such as many-core chips, non-volatile random-access memories, and reconfigurable hardware – affect the design of data processing systems; or building data processing systems that facilitate rapid processing and which can ingest data streams to cope with the "big velocity" problem.

A New Paradigm for Life Sciences Computing

The combination of heterogeneous computing and cloud computing is emerging as a powerful new paradigm to meet the requirements for high-performance computing (HPC) and data throughput throughout the life sciences (LS) and healthcare value chains. Of course, neither cloud computing nor the use of innovative computing architectures is new, but the rise of Big Data as a defining feature of modern life sciences and the proliferation of vastly differing applications to mine the data have dramatically changed the landscape of life science computing requirements.

Both horizontal and vertical approaches to Big Data research should be combined. Big Data may foster solutions for some of the grand societal challenges of today's world. Therefore enabling research in Big Data software and hardware is critical and will have a fundamental impact in many other areas of research[10] such as health, transport or machine translation.

[8] " Open Innovation 2.0: A New Paradigm" http://ec.europa.eu/digital-agenda/en/news/open-innovation-20-%E2%80%93-new-paradigm-and-foundation-sustainable-europe
[9] Intel Science and Technology Center for Big Data http://istc-bigdata.org
[10] "Exploring Data-Driven Innovation as a New Source of Growth" OECD 2013

The Role of the Public Sector: Unleashing Public Data

The public sector is one of the largest data-intensive sectors in the economy. The administration stores and uses an increasing volume of data. The US administration stored until 2011, 1.3 petabytes of insightful information.[11] The public sector is a crucial actor to foster data-driven innovation.

On this point Europe has been a leading data-rich region and therefore experts have outlined the importance of this resource for growth and job creation.[12] For Europe to be consolidated at the forefront of Big Data deployment and innovation, the policy responses need to foster the nascent EU Data Ecosystem by providing a balanced policy framework.

Initiatives such as the newly adopted Directive on the re-use of Public Sector Information together with the deployment of Open Data Portals are good incentives to foster the creation of an innovative data ecosystem.

The public sector can play multiple other roles apart from enabling the potential of large public datasets. The US administration announced last year the Big Data Research and Development Initiative that will foster innovative Big Data solutions in six federal departments and agencies. The public sector in itself will benefit from Big Data innovation, incentivizing at the same time the data ecosystem as a whole.

Trust: Security & Privacy

Not all Big Data is personal data. But the use of Big Data requires individuals to trust that data relating to them will be protected and used appropriately. Single data elements may not relate to an identifiable individual, but when aggregated with other data sets, organizations may then have the ability to identify an individual or a specific device. As organizations apply analytics algorithms to the data, the resulting inferences may allow for profiling of the individual or even determinations of the propensity of the individual to act in a certain way in the future. Therefore, privacy and security are fundamental to developing trust in engaging with the digital society and the use of new technologies.

EU law should aim to facilitate the trans-border flow and use of data, while also protecting individual privacy. While doing so, it will be important to focus on consistent implementation and interpretation across the different Member State legal regimes. The need for Member States to be able to customize their legislation to take into account certain unique legal, social, economic and cultural differences should be

[11] "Exploring Data-Driven Innovation as a New Source of Growth" OECD 2013
[12] "Big Data: The next frontier for innovation, competition and productivity", McKinsey Global Institute 2011

recognised. However, greater focus is required on limiting customisation to provide greater clarity of individual rights, and to simplify the implementation of operational measures.

Security

Intel looks forward to engaging with the appropriate stakeholders to think about improving the effectiveness of the information security legal framework. To this end, we value the policy efforts of the European Commission against cybercrime that threatens the potential of Big Data.

The recently tabled Cyber Security Strategy and the Network and Information Security Directive demonstrate the European Union's commitment to address and resolve these threats. Together with the recent creation of the European Cybercrime Centre (EC3) in The Hague, these proposals pave the way for a strong coordinated response against these 21st century threats. Intel welcomes the EU's commitment to address on-going and emerging challenges in cyberspace and the recognition that this requires global public and private cooperation.

Privacy

Providing privacy protection in the era of Big Data requires new thinking about how to apply the traditional Fair Information Practice Principles as described in the Organization for Economic Cooperation and Development's seminal Guidelines on the Protection of Privacy and the Transborder Flows of Personal Data (the OECD Guidelines).[13] Traditionally, laws have implemented the OECD Guidelines with focus on a Notice and Consent model. This model generally requires organizations to provide notice at the time of data collection, often obtain consent from the individual, and then only use the data in the manner described in the notice. In a Big Data environment though, much of the data relating to an individual will not come directly from that individual.

Instead, the data will be obtained from third parties (such as posts to social networks), will be observed in public spaces (such as video and sensor data) or will originate from public records (the opening up of government data sets). Privacy notices are getting much more difficult to deliver on ever shrinking devices, and research shows very few individual take the time to read these policies. While transparency and individual choice are critically important, these developments show we need to rethink how we legislatively apply these concepts so we do not artificially constrain the tremendous social and economic benefits from Big Data, while also not providing real privacy protections.

[13] "OECD Guidelines on the Protection of Privacy and the Transborder Flows of Personal Data" OECD, 1980 http://www.oecd.org/internet/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm (Revised 2013) http://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf

The best way to realize the promise of Big Data, and also engender trust, is to focus on the four concepts: Transparency, Obscurity, Appropriate Use and Accountability.

Transparency – While most individuals do not read privacy policies, these notices can still have value. Organizations describing in detail how they collect, process, store and use data is important for regulators and privacy advocates to understand how the entity is committing to protect privacy. To meet this objective, privacy policies should be comprehensive, detailed and specific.

Obscurity – While much of Big Data may not come directly from the individual, organizations should still provide reasonable rights to access, correction and deletion. Individuals need easy to use mechanisms to obscure inaccurate or disproportionate data about them that may appear in databases or on the internet.

Appropriate Use – Some experts agree constraining the use of data to just those purposes specified at the time of collection will unnecessarily deprive society of tremendous innovations and value. Conversely though, allowing organizations to use data in ways that would surprise, and possibly negatively impact, the individuals to whom the data relates, is even less appealing. A solution is to focus on allowing organizations to use data in ways deemed "appropriate" through analysis of the benefits and harms to both the individual and society.

Accountability - The more organizations would like to use data in new and innovative ways, the more they should demonstrate they have invested in processes and procedures to act as responsible data stewards.

Digging into the 'circuitry' of cancer

One of the best cases for the value of Big Data is in its use for research. Big Data analytics, simulations and visualizations are enhancing research particularly in life sciences.

"Scientists can now gather billions of data points on how a specific patient's cells are malfunctioning. Genetic abnormalities that cause these tumors manifest differently in each of us. What's more, even a healthy human body creates millions of these mutations. So it's an enormous scientific challenge to determine, for each individual, which mutations are relevant in creating "my" disease. But that's where medicine must go."

Eric Dishman, Intel Fellow and GM

The innovations and benefits that Big Data analytics will bring to society may change our relationship to the disclosure of our personal data. That is why, a strong but balanced data protection framework, awareness, and education, are key components to shape any relation with data. Education has always been essential for the understanding and protection of fundamental rights. Several supervisory authorities have led by example in working together with other stakeholders in raising such awareness about the risks from privacy invasions, and how to protect oneself,

exercise rights, or lodge claims. However, more consistent investment in coordinated programs is required, as well as additional focus on educating individuals and those processing personal data. This could be accomplished by providing support for civil society, NGOs, professional associations and other organisations that have privacy awareness and education as their primary mission, and regularly report on their progress to the general public.

Education & Skills

A new kind of professional is helping organizations make sense of the massive streams of digital information: the data scientist.

Data scientists are responsible for modelling complex business problems, discovering business insights, and identifying opportunities. They bring to the job:

- Skills for integrating and preparing large, varied data sets
- Advanced analytics and modelling skills to reveal and understand hidden relationships
- Business knowledge to apply context
- Communication skills to present results

Data science is an emerging field. Demand is high, and finding skilled personnel is one of the major challenges associated with Big Data analytics. A data scientist may reside in IT or business- but either way, he or she is essential for planning and implementing Big Data analytics projects.

The OECD has warned about the upcoming shortage in data scientists and Big Data related skills in its annual OECD Skills Strategy in 2012[14]. Mismatches between supply and demand are regrettably common in the ICT sector but data

Supporting data scientist education in France

Intel with GENCI[1] combined forces aiming at developing high performance simulation solutions usage models in order to increase the competitiveness of small and medium enterprises.

Based on this goal Intel and GENCI will work together on developing education curricula with 10 partner universities in France on Big Data analytics and High Performance Computing simulation from bachelor up to master level, enhancing the student curricula.

Through this type of academic-private collaboration the government tries to stimulate local excellence in education by providing computing capabilities in order to support market competitiveness.

1 GENCI http://www.genci.fr/ the French representative of PRACE http://www.prace-ri.eu/

[14] "Better Skills, better Jobs, Better Lives: A Strategic approach to Skills Policies", OECD, 2012

scientists demand will outnumber supply by 140.000 to 190.000 only in United States. The Big Data implications in employment will be broader if we take into consideration 1.5 million potential managers and analysts only for the US by 2018.[15]

KEY ELEMENTS FOR A BIG DATA POLICY		
Area to address	End goal	Recommendations
Research priorities	Use research funding to incentivize breakthrough innovation in Big Data	• Ensure research in societal challenges that can be addressed by Big Data solutions • Focus research on hardware and software that enables Big Data processing (e.g. HPC, data centres, analytics)
Role of the Public Sector	Leading role of the Public Sector in the Data Economy	• Enable the Public Sector Information for re-use • Adopt Big Data solutions for evidence-based policy-making
Trust: Security & Privacy	Ensure citizens trust in Big Data solutions protecting privacy rights	• Clarify the distinction between personal data and non-personal data • Build a strong but balanced data protection framework to enable citizen trust • Raise network security to enable citizen trust • Research in anonymization technologies • Develop comprehensive and special regimes for scientific processing of data
Education & skills	Ensure the supply of data scientists and data analysts	• Commit the Grand Coalition for Digital Skills to address data skills • Engage with academia and the ICT sector to develop data scientist curricula • Develop a network of Centers of Excellence for Big Data

In the event that you have any questions relating to Intel's views on Big Data, please contact Adrian Martinez Pacin: adrian.martinez.pacin@intel.com

[15] "Big Data: The next frontier for innovation, competition and productivity", McKinsey Global Institute , 2011