4. Big Data

4.1 Introduction

In 2004, Wal-Mart claimed to have the largest data warehouse with 500 terabytes storage (equivalent to 50 printed collections of the US Library of Congress). In 2009, eBay storage amounted to eight petabytes (think of 104 years of HD-TV video). Two years later, the Yahoo warehouse totalled 170 petabytes[1] (8.5 times of all hard disk drives created in 1995)[2]. Since the rise of digitisation, enterprises from various verticals have amassed burgeoning amounts of digital data, capturing trillions of bytes of information about their customers, suppliers and operations. Data volume is also growing exponentially due to the explosion of machine-generated data (data records, web-log files, sensor data) and from growing human engagement within the social networks.

The growth of data will never stop. According to the 2011 IDC Digital Universe Study, 130 exabytes of data were created and stored in 2005. The amount grew to 1,227 exabytes in 2010 and is projected to grow at 45.2% to 7,910 exabytes in 2015.[3] The growth of data constitutes the "Big Data" phenomenon – a technological phenomenon brought about by the rapid rate of data growth and parallel advancements in technology that have given rise to an ecosystem of software and hardware products that are enabling users to analyse this data to produce new and more granular levels of insight.

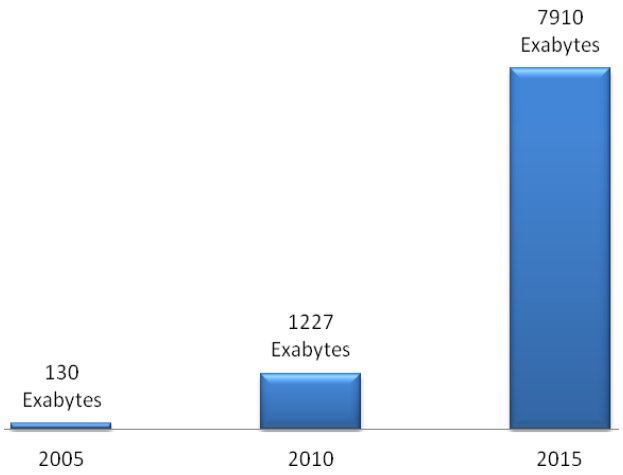


Figure 1: A decade of Digital Universe Growth: Storage in Exabytes **Error! Reference source not found.**[3]

[1] Ovum. What is Big Data: The End Game. [Online] Available from: http://ovum.com/research/what-is-big-data-the-end-game/ [Accessed 9th July 2012].

[2] IBM. Data growth and standards. [Online] Available from: http://www.ibm.com/developerworks/xml/library/x-datagrowth/index.html?ca=drs- [Accessed 9th July 2012].

[3] IDC. The 2011 Digital Universe Study: Extracting Value from Chaos. [Online] Available from: http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf [Accessed 9th July 2012].

4.1.1 What is Big Data?

According to McKinsey,[4] Big Data refers to datasets whose size are beyond the ability of typical database software tools to capture, store, manage and analyse. There is no explicit definition of how big a dataset should be in order to be considered Big Data. New technology has to be in place to manage this Big Data phenomenon. IDC defines Big Data technologies as a new generation of technologies and architectures designed to extract value economically from very large volumes of a wide variety of data by enabling high velocity capture, discovery and analysis. According to O'Reilly, "Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or does not fit the structures of existing database architectures. To gain value from these data, there must be an alternative way to process it."[5]

4.1.2 Characteristics of Big Data

Big Data is not just about the size of data but also includes data variety and data velocity. Together, these three attributes form the three Vs of Big Data.

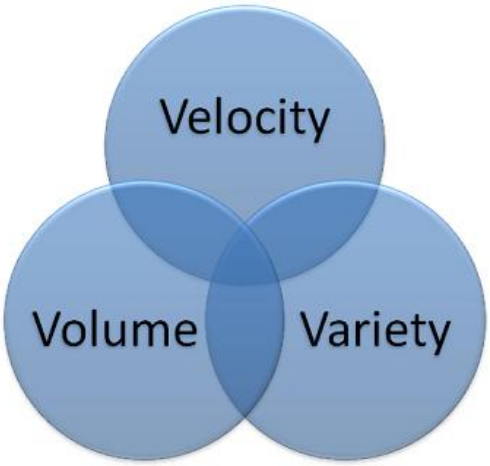


Figure 2: The 3 Vs of Big Data

Volume is synonymous with the "big" in the term, "Big Data". Volume is a relative term – some smaller-sized organisations are likely to have mere gigabytes or terabytes of data storage as opposed to the petabytes or exabytes of data that big global enterprises have. Data volume will continue to grow, regardless of the organisation's size. There is a natural tendency for companies to store data of all sorts: financial data, medical data, environmental data and so on. Many of these companies' datasets are within the terabytes range today but, soon they could reach petabytes or even exabytes.

Data can come from a *variety* of sources (typically both internal and external to an organisation) and in a variety of types. With the explosion of sensors, smart devices as well as social networking, data

[4] James Manyika, et al. Big data: The next frontier for innovation, competition, and productivity. [Online] Available from: http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation [Accessed 9th July 2012].

[5] Edd Dumbill. What is big data? [Online] Available from: http://radar.oreilly.com/2012/01/what-is-big-data.html [Accessed 9th July 2012].

in an enterprise has become complex because it includes not only structured traditional relational data, but also *semi-structured* and *unstructured* data.

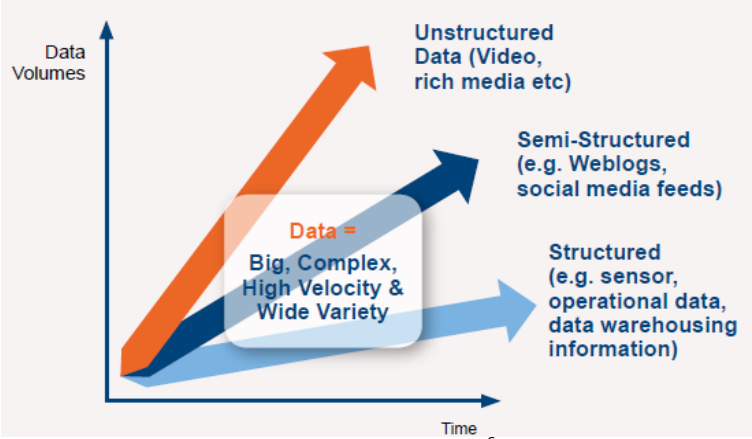


Figure 3: The three data types[6]

Structured data: This type describes data which is grouped into a relational scheme (e.g., rows and columns within a standard database). The data configuration and consistency allows it to respond to simple queries to arrive at usable information, based on an organisation's parameters and operational needs.

Semi-structured data[7]: This is a form of structured data that does not conform to an explicit and fixed schema. The data is inherently self-describing and contains tags or other markers to enforce hierarchies of records and fields within the data. Examples include weblogs and social media feeds.

Unstructured data: This type of data consists of formats which cannot easily be indexed into relational tables for analysis or querying. Examples include images, audio and video files.

The *velocity* of data in terms of the frequency of its generation and delivery is also a characteristic of big data. Conventional understanding of velocity typically considers how quickly the data arrives and is stored, and how quickly it can be retrieved. In the context of Big Data, velocity should also be applied to data in motion: the speed at which the data is flowing. The various information streams and the increase in sensor network deployment have led to a constant flow of data at a pace that has made it impossible for traditional systems to handle.

Handling the three Vs helps organisations extract the *value* of Big Data. The value comes in turning the three Vs into the three Is:

1. Informed intuition: predicting likely future occurrences and what course of actions is more likely to be successful.

2. Intelligence: looking at what is happening now in real time (or close to real time) and determining the action to take

---

6 IDC, 2012

7 Peter Buneman. Semistructured Data. [Online] Available from: http://homepages.inf.ed.ac.uk/opb/papers/PODS1997a.pdf [Accessed 9th July 2012].

3. Insight: reviewing what has happened and determining the action to take.

4.1.3 Why is Big Data important?

The convergence across business domains has ushered in a new economic system that is redefining relationships among producers, distributors, and consumers or goods and services. In an increasingly complex world, business verticals are intertwined and what happens in one vertical has a direct impact on other verticals. Within an organisation, this complexity makes it difficult for business leaders to rely solely on experience (or pure intuition) to make decisions. They need to rely on good data services for their decisions. By placing data at the heart of the business operations to provide access to new insights, organisations will then be able to compete more effectively.

Three things have come together to drive attention to Big Data:

1. The technologies to combine and interrogate Big Data have matured to a point where their deployments are practical.
2. The underlying cost of the infrastructure to power the analysis has fallen dramatically, making it economic to mine the information.
3. The competitive pressure on organisations has increased to the point where most traditional strategies are offering only marginal benefits. Big Data has the potential to provide new forms of competitive advantage for organisations.

For years, organisations have captured transactional structured data and used batch processes to place summaries of the data into traditional relational databases. The analysis of such data is retrospective and the investigations done on the datasets are on past patterns of business operations. In recent years, new technologies with lower costs have enabled improvements in data capture, data storage and data analysis. Organisations can now capture more data from many more sources and types (blogs, social media feeds, audio and video files). The options to optimally store and process the data have expanded dramatically and technologies such as MapReduce and in-memory computing (discussed in later sections) provide highly optimised capabilities for different business purposes. The analysis of data can be done in real time or close to real time, acting on full datasets rather than summarised elements. In addition, the number of options to interpret and analyse the data has also increased, with the use of various visualisation technologies. All these developments represent the context within which “Big Data” is placed.

4.1.4 Big Data and the public service sector

Big Data affects government agencies the same way it does other organisations. Big Data brings the potential to transform the work of government agencies by helping government agencies operate more efficiently, create more transparency and make more informed decisions. The data stores that various government agencies accumulate over the years offer new opportunities for agencies which can use Big Data technologies to extract insights and keep track of citizens’ specific needs. In turn, these insights could then be used to improve government services.

A Deloitte report highlights how predictive analytics – a particular form of data analytics that uses data mining to provide actionable, forward-looking intelligence – is key to business improvement

across government agencies.[8] Data analytics offers a range of new capabilities for government agencies - these include operational improvements in the areas of citizen service provision and tax fraud detection, policy development (specifically in the form of evidence-based policy making), and policy forecasting.

In 2012, the US government committed US$200 million, including US$73 million in research grants, to its Big Data R&D strategy.[9] The initiative involves the National Science Foundation, National Institutes of Health, Department of Defence, Department of Energy and United States Geological Survey (USGS). The coordinated effort represents the government's attempt to make better use of the massive data sets at its disposal and reflects the commitments from federal agencies to develop new technologies in science, national security and education.

4.2 Market Drivers

As the vendor ecosystem around Big Data matures and users begin exploring more strategic business use cases, the potential of Big Data's impact on data management and business analytics initiatives will grow significantly. According to IDC, the Big Data technology and service market was about US$4.8 billion in 2011[10]. The market is projected to grow at a compound annual growth rate (CAGR) of 37.2% between 2011 and 2015. By 2015, the market size is expected to be US$16.9 billion.

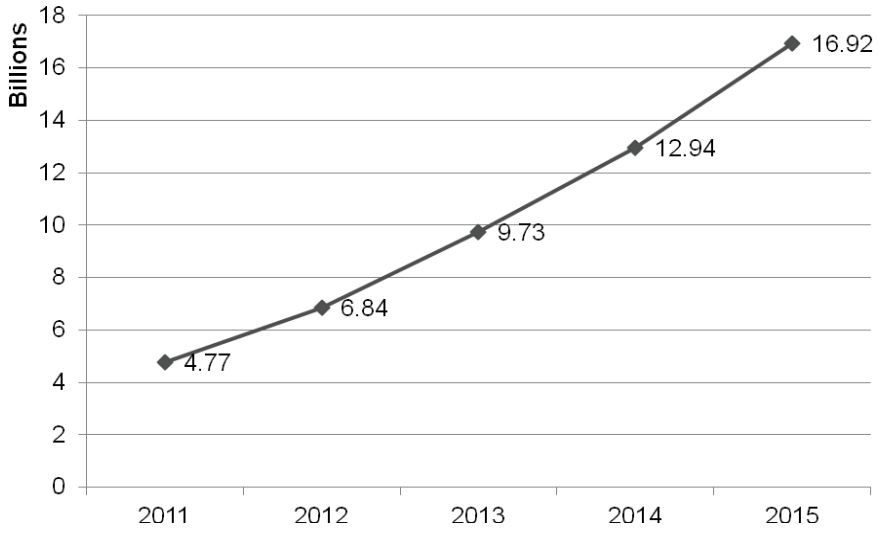


Figure 4: Global Big Data Market Projection[10]

There are several factors that will drive this market:

4.2.1 Continuous growth of digital content

[8] Deloitte. Deloitte Analytics – Insight on tap: Improving public services through data analytics. [Online] Available from: http://www.deloitte.com/assets/Dcom-UnitedKingdom/Local%20Assets/Documents/Industries/GPS/uk-gps-deloitte-analytics.pdf [Accessed 9th July 2012].

[9] J. Nicholas Hoover. White House Shares $200 Million Big Data Plan. [Online] Available from: http://www.informationweek.com/government/information-management/white-house-shares-200-million-big-data/232700522 [Accessed 9th July 2012].

[10] IDC. Worldwide Big Data Technology and Services 2012-2015 Forecast. [Online] Available from: http://www.idc.com/getdoc.jsp?containerId=233485 [Accessed 9th July 2012].

The increasing market adoption of mobile devices that are cheaper, more powerful and packed with apps and functionalities is a major driver of the continuing growth of unstructured data. Gartner estimates the 2012 smartphone shipment to reach 467.7 million units. By 2015, the expected number of smartphones in the market will reach 1.1 billion. The market adoption of tablets is also expected to increase significantly over the next few years, further contributing to the growth of data. In 2012, shipment of tablets is expected to reach 118.9 million tablets, with the number projected to rise to 369.3 million by 2015.[11] This market adoption of mobile devices and the prevalence of mobile Internet will see consumers increasingly being connected, using social media networks as their communication platform as well as their source of information.

The convergence of mobile device adoption, the mobile Internet and social networking provides an opportunity for organisations to derive competitive advantage through an efficient analysis of unstructured data. Businesses that were early adopters of Big Data technologies and that based their business on data-driven decision making were able to achieve greater productivity of up to 5% or 6% higher than the norm[12]. Big Data technology early adopters such as Facebook, Linkedln, Walmart and Amazon are good examples for companies that plan to deploy Big Data analytics.

4.2.2 Proliferation of the Internet of Things (IoT)

According to Cisco's Internet Business Solutions Group (IBSG)[13], 50 billion devices will be connected to the Web by 2020. Meanwhile, Gartner reported that more than 65 billion devices were connected to the internet by 2010. By 2020, this number will go up to 230 billion.[14] Regardless of the difference in estimation, these connected devices, ranging from smart meters to a wide range of sensors and actuators continually send out huge amounts of data that need to be stored and analysed. Companies that deploy sensor networks will have to adopt relevant Big Data technologies to process the large amount of data sent by these networks.

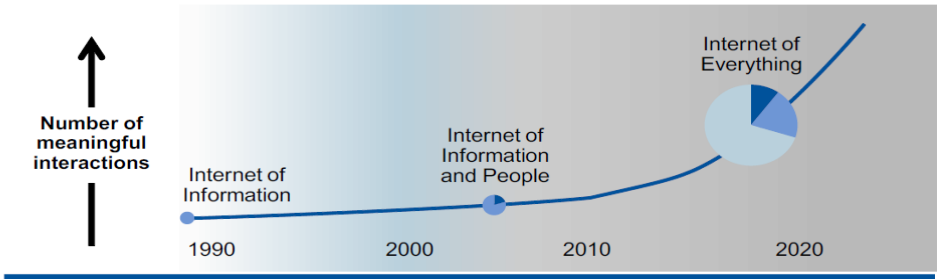
---

11 Gartner. Gartner Says Worldwide Media Tablets Sales to Reach 119 Million Units in 2012. [Online] Available from: http://www.gartner.com/it/page.jsp?id=1980115 [Accessed 9th July 2012].

12 Erik Brynjolfsson, et al. Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance. [Online] Available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486 [Accessed 9th July 2012].

13 Cisco. The Internet of Things: How the Next Evolution of the Internet is Changing Everything. [Online] Available from: http://www.cisco.com/web/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf [Accessed 9th July 2012].

14 John Mahoney, Hung LeHong. The Internet of Things is Coming. [Online] Available from: http://www.gartner.com/id=1799626 [Accessed 9th July 2012].



Key milestones	> 50% of Internet connections are things	Things create more traffic than information and people
Number of permanently Internet-connected devices	> 15 billion	> 30 billion
Number of intermittently Internet-connected devices	> 50 billion	> 200 billion
Volume of traffic from things	Minority	Majority

Source: Gartner (September 2011)

Figure 5: Exponential growth in number of connected devices[14]

4.2.3 Strong open source initiatives

Many of the technologies within the Big Data ecosystem have an open source origin, due to participation, innovation and sharing by commercial providers in open source development projects. The Hadoop framework, in conjunction with additional software components such as the open source *R language* and a range of open source Not Only Structured Query Language (NoSQL) tools such as Cassandra and Apache HBase, is the core of many Big Data discussions today. The popularity and viability of these open source tools have driven vendors to launch their own versions of such tools (e.g., Oracle's version of the NoSQL database[15]) or integrate these tools with their products (e.g., EMC's Greenplum Community edition which includes the open source Apache Hive, HBase and ZooKeeper[16]).

Some of the technology companies that are driving the technology evolution of the Big Data landscape are affiliated to the open source community in different ways. For example, Cloudera is an active contributor to various open source projects[17] while EMC's Greenplum launched its Chorus social framework as an open source tool to enable collaboration on datasets in a *Facebook-like* way.[18] Hortonworks has also formed a partnership with Talend to bring the world's most popular open source data integration platform to the Apache community.[19] The situation where open source

[15] Doug Henschen. Oracle Releases NoSQL Database, Advances Big Data Plans. [Online] Available from: http://www.informationweek.com/software/information-management/oracle-releases-nosql-database-advances/231901480 [Accessed 9th July 2012].

[16] Doug Henschen. Oracle Releases NoSQL Database, Advances Big Data Plans. [Online] Available from: http://www.informationweek.com/software/information-management/oracle-releases-nosql-database-advances/231901480 [Accessed 9th July 2012].

[17] Cloudera. Open Source. [Online] Available from: http://www.cloudera.com/company/open-source/ [Accessed 9th July 2012].

[18] Greenplum. EMC Goes Social, Open and Agile With Big Data. [Online] Available from: http://www.greenplum.com/news/press-releases/emc-goes-social-open-and-agile-with-big-data [Accessed 9th July 2012].

[19] Hortonworks. Hortonworks Announces Strategic Partnership with Talend to Bring World's Most Popular Open Source Data Integration Platform at Apache Community. [Online] Available from: http://hortonworks.com/about-us/news/hortonworks-announces-strategic-partnership-with-talend-to-bring-worlds-most-popular-open-source-data-integration-platform-to-a/ [Accessed 9th July 2012].

technologies dominate the Big Data solutions may perpetuate as the technologies are changing rapidly and the technology standards are not well established. In turn, this posts significant risk to any vendors who want to invest in proprietary Big Data technologies. Hence, the "open source" nature of the Big Data technologies will encourage bigger adoption.

4.2.4 Increasing investments in Big Data technologies

Information has always been a differentiator in the business world, allowing better business decisions to be made in an increasingly competitive landscape. Previously, market information was largely made available through traditional market research and data specialists. Today, virtually any company with a large datasets can potentially become a serious player in the new information game. The value of Big Data will become more apparent to corporate leadership as companies seek to become more "data-driven" organisations. According to O'Reilly[20], a data driven organisation is one that "acquires, processes and leverages data in a timely fashion to create efficiencies, iterate on and develop new products to navigate the competitive landscape."

A Big Data Insight Group survey[21] of 300 senior personnel from a broad range of industry sectors[22] revealed that many organisations are seeing Big Data as an important area for their organisations. Among the respondents, 50% indicated current research into, and sourcing of, Big Data solutions while another 33% acknowledged that they were implementing or had implemented some form of Big Data solutions. This survey indicates that many organisations perceive Big Data as an important development and this interest could translate into future demand for Big Data technologies.

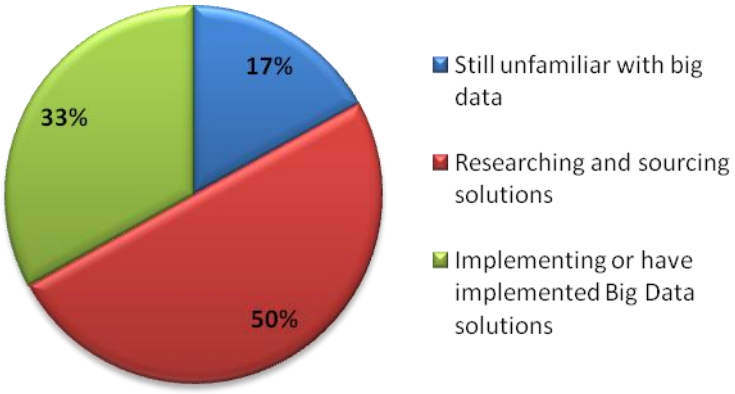


Figure 6: Current status of Big Data initiatives

4.2.5 Research and development involving high-performance computing

[20] DJ Patil. Building Data Science Teams. [Online] Available from: http://assets.en.oreilly.com/1/eventseries/23/Building-Data-Science-Teams.pdf [Accessed 9th July 2012].

[21] The survey, conducted in February and March 2012, was completed by 300 senior business, finance and IT personnel from a broad range of industry sector including financial, retail, telecommunications and the public sector. They represented companies of all sizes from SMEs to blue chip organisations

[22] Big Data Insight Group. The 1st Big Data Insight Group Industry Trends Report. [Online] Available from: http://www.thebigdatainsightgroup.com/site/article/1st-big-data-insight-group-industry-trends-report [Accessed 9th July 2012].

Research and Development (R&D) that involves data-intensive workloads, such as high-performance computing, life sciences and earth sciences, find value in Big Data technologies. For example, at CERN, the Swiss research laboratory outside Geneva, physicists studying the results of tests at the Large Hadron Collider (LHC) have to decide how to store, process and distribute the usable information accruing from the 22 petabytes of data generated annually.[23] Big Data technologies have to be in place to support such R&D efforts because they have to support the growth of digital content and enable more efficient analysis outputs. Traditional technologies such as symmetric multiprocessing (SMP) which also enables system scalability can be prohibitively expensive for many granular R&D use case scenarios. Hence, the need for cost-efficient scalable hardware and software resources to process related business logic and data volumes becomes more apparent than before.

[23] Irfan Khan. CERN, US, UK projects push big data limits. [Online] Available from: http://www.itworld.com/big-datahadoop/271154/cern-us-uk-projects-push-big-data-limits [Accessed 9th July 2012].

4.3 Diving deep into Big Data

4.3.1 Existing Big Data Technologies

There are no comprehensive Big Data technology standards in place today. The main reason is that the Big Data analytics projects companies are taking on are typically complex and diverse in nature. A proven comprehensive Big Data certification and standards are not yet in place although some vendors such as IBM and EMC have announced certification programmes centred on providing training for their Hadoop-based products.

Hadoop is almost synonymous with the term "Big Data" in the industry and is popular for handling huge volumes of unstructured data. The Hadoop Distributed File System enables a highly scalable, redundant data storage and processing environment that can be used to execute different types of large-scale computing projects. For large volume structured data processing, enterprises use analytical databases such as EMC's Greenplum and Teradata's Aster Data Systems. Many of these appliances offer connectors or plug-ins for integration with Hadoop systems.

Big Data technology can be broken down into two major components – the hardware component and the software component, as shown in the figure below. The hardware component refers to the component and infrastructure layer. The software component can be further divided into data organisation and management software, analytics and discovery software, and decision support and automation software.[24]

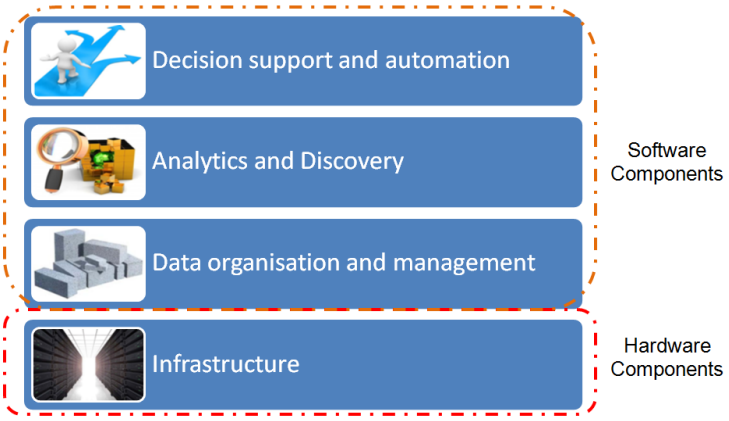


Figure 7: Big Data Technology Stack

Infrastructure

Infrastructure is the foundation of the Big Data technology stack. The main components of any data storage infrastructure - industry standard x86 servers and networking bandwidth of 10 Gbps - may be extended to a Big Data storage facility. Storage systems are also becoming more flexible and are being designed in a scale-out fashion, enabling the scaling of system performance and capacity. In-memory computing is supported by increased capabilities in system memory delivered at lower prices, making multi-gigabytes (even multi-terabytes) memory more affordable. In many instances, Not And (NAND) flash memory boards are deployed, together with traditional Dynamic Random Access Memory (DRAM), producing a more cost-effective and improved performance.

[24] IDC. IDC's Worldwide Big Data Taxonomy, 2011. [Online] Available from: http://www.idc.com/getdoc.jsp?containerId=231099 [Accessed 9th July 2012].

Data organisation and management

This layer refers to the software that processes and prepares all types of structured and unstructured data for analysis. This layer extracts, cleanses, normalises and integrates data. Two architectures – extended Relational Database Management System (RDBMS) and the NoSQL database management system – have been developed to manage the different types of data.

Extended RDBMS is optimised for scale and speed in processing *huge relational data* (i.e., structured data) sets, adopting approaches such as using columnar data stores to reduce the number of table scans (columnar database) and exploiting massively parallel processing (MPP) frameworks. On the other hand, the NoSQL database management system (NoSQL DBMS) grew out of the realisation that SQL's transactional qualities and detailed indexing are not suitable for the processing of unstructured files. (More discussion on NoSQL DBMS can be found in the segment on Technology Outlook.)

Data Analytics and Discovery

This layer comprises two data analytics software segments – software that supports offline, *ad hoc*, discovery and deep analytics, and software that supports dynamic real-time analysis and automated, rule-based transactional decision making. The tools can also be categorised by the type of data being analysed, such as text, audio and video. The tools within this layer can also be at different levels of sophistication. There are tools that allow for highly complex and predictive analysis as well as tools that simply help with basic data aggregation and trend reporting. In any case, the usage of the tools is not mutually exclusive – there can be a set of tools with different features residing in a system to enable Big Data analytics.

Decision support and automation interface

The process of data analysis usually involves a closed-loop decision making model which, at the minimum, includes steps such as track, analyse, decide and act. To support decision making and to ensure that an action is taken, based on data analysis, is not a trivial matter. From a technology perspective, additional functionalities such as decision capture and retention are required to support collaboration and risk management.

There are two decision support and automation software categories: *transactional decision* management and *project-based* decision management software. The former is automated, embedded within applications, real-time and rules-based in nature. It enables the use of outputs to prescribe or enforce rules, methods and processes. Examples include fraud detection, securities trading, airline pricing optimisation, product recommendation and network monitoring. Project-based decision management is typically standalone, *ad hoc* and exploratory in nature. It can be used for forecasting and estimation of trends. Examples include applications for customer segmentation for targeted marketing, product development and weather forecasting.

4.3.2 Big Data and Cloud Computing

Cloud services with the ability to ingest, store and analyse data have been available for some time and they enable organisations to overcome the challenges associated with Big Data. Early adopters of Big Data on the cloud would be users deploying Hadoop clusters on the highly scalable and elastic environments provided by Infrastructure-as-a-Service (IaaS) providers such as Amazon Web Services and Rackspace, for test and development, and analysis of existing datasets. These providers offer data storage and data back-up in a cost-effective manner. They deliver a low-cost and reliable environment that gives organisations the computing resources they need to store their structured and unstructured data. At the Software-as-a-Service (SaaS) level, embedded analytics engines help analyse the data stored on the cloud. The analytics output can then be provided to the end users through a graphical interface. However, the development of queries and integration to the data source on the cloud are prerequisites that organisations need to undertake before the usability can be delivered.

Cloud services are useful for organisations looking to test and develop new processes and applications before making large investments. Hence, cloud computing provides the support for Big Data deployment. However, when it comes to the transfer of data to and from these two environments, bandwidth and integration issues can become major barriers for the use of Big Data on the cloud.

4.3.3 The Data Challenge

The systematic approach toward data collection in order to enhance randomness in data sampling and reduce bias is not apparent in the collection of Big Data sets.[25] Big Data sets do not naturally eliminate data bias. The data collected can still be incomplete and distorted which, in turn, can lead to skewed conclusions. Consider the case of Twitter which is commonly scrutinised for insights about user sentiments. There is an inherent problem with using Twitter as a data source as only 40% of Twitter's active users are merely listening and not contributing.[26] This may suggest that the tweets come from a certain type of people (probably people who are more vocal and participative in social media) than from a true random sample. In addition, Twitter makes a sample of its materials available to the public through its streaming Application Programming Interfaces (APIs).[27] It is not clear how the sample of materials is derived.

Big data can also raise privacy concerns and reveal unintended information. Researchers from the University of Texas at Austin have found that *anonymised* data from a social network platform, combined with readily available data from other online sources, can render the data *de-anonymised* and reveal sensitive information about a person.[28] In a test involving Flickr (a photo-sharing site) and Twitter, the researchers were able to identify a third of the users with accounts on both sites simply by researching for recognisable patterns in *anonymised* network data. The researchers found that they could extract sensitive information about individuals using just the connections between users, even if almost all of the personally identifying information had been removed, provided they could

---

25 Danah Boyd, Kate Crawford. Six Provocations for Big Data. [Online] Available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431 [Accessed 9th July 2012].

26 Matt Rosoff. Twitter Has 100 Million Active Users – And 40% Are Just Watching. [Online] Available from: http://articles.businessinsider.com/2011-09-08/tech/30127585_1_ceo-dick-costolo-twitter-users [Accessed 9th July 2012].

27 Twitter. Public streams. [Online] Available from: https://dev.twitter.com/docs/streaming-apis/streams/public [Accessed 9th July 2012].

28 Erica Naone. Unmasking Social-Network Users. [Online] Available from: http://www.technologyreview.com/web/22593/ [Accessed 9th July 2012].

compare these patterns with those from another social network graph where some user information was accessible.

Big Data is not just about the technology; it is also about the people and business processes. Many discussions in the Big Data space have revolved around the benefits of the technologies and how they help companies gain competitive advantage. There is a danger that Big Data adopters are missing the bigger picture by excluding the discussions around the people and business processes. Before jumping onto the Big Data bandwagon, companies must first evaluate the business case and specific outcomes of their proposed Big Data initiatives. They would need to know what to ask of the data, assess how the business will react to it and be able to offer actionable operational measures. They would then need to assess the IT capability and if necessary re-architect their internal systems from a data management and business process standpoint to make the investment more strategic. Developing a roadmap of how to achieve the desired business outcomes will give the organisation the understanding of what is required and enable it to be prepared financially and organisationally.

4.3.4 Data Science and the rise of the Data Scientist

Data science is the general analysis of the creation of data. This refers to the comprehensive understanding of where data comes from, what data represents, and how to turn data into information that drives decisions. This encompasses statistics, hypothesis testing, predictive modelling and an understanding of the effects of performing computations on data, among other things. Data science pools these skills together to provide a scientific discipline for the analysis and productizing of data.[29]

Although the term "data science" has been around for years, as early as 1974 when Peter Naur first defined it,[30] the term "data scientist", in its current context, is relatively new. Gartner defines the data scientist[31] as an individual responsible for modelling complex business problems, discovering business insights and identifying opportunities through the use of statistical, algorithmic, mining and visualisation techniques. In addition to advanced analytic skills, this individual is also proficient at integrating and preparing large, varied datasets, architecting specialised database and computing environments, and communicating results. A data scientist may or may not have specialized industry knowledge to aid in modelling business problems and with understanding and preparing data.

[29] Mike Loukides. What is data science? [Online] Available from: http://radar.oreilly.com/2010/06/what-is-data-science.html [Accessed 9th July 2012].

[30] Whatsthebigdata.com. A Very Short History of Data Science. [Online] Available from: http://whatsthebigdata.com/2012/04/26/a-very-short-history-of-data-science/ [Accessed 9th July 2012].

[31] Douglas Laney, Lisa Kart. Emerging Role of the Data Scientist and the Art of Data Science. [Online] Available from: http://www.gartner.com/id=1955615 [Accessed 9th July 2012].

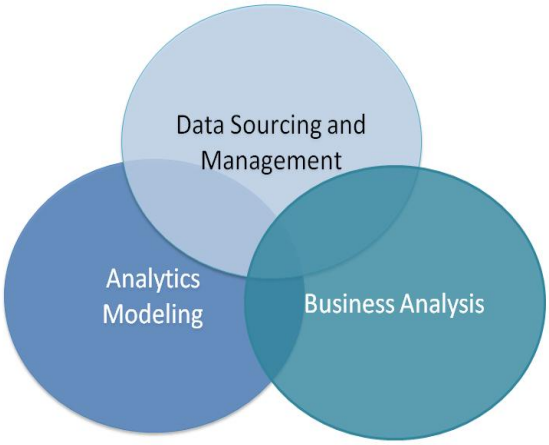


Figure 8: Core data scientist skills

Professionals having this wide range of skills are rare and this explains why data scientists are currently in short supply. By 2018, the USA alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how of Big Data analysis to make effective decisions.[4] In most organisations, rather than looking for individuals with all of these capabilities, it might be necessary to build a team of people that collectively possesses these skills[31].

Data Sourcing and Management

Knowing which data to harness and the way to harness the data are the basic requirements before any form of data analytics can be done. In addition to data that is already within the control of an organisation, a majority of the data could lie outside the organisation's ownership. This data includes social media feeds (e.g., Facebook, Twitter and LinkedIn), geo-spatial information, news, weather data, credit data and retail information. Some companies which own such data are seeing the value of this data and making it available for sale. For example, Microsoft has the Azure Marketplace Data Market which offers datasets such as employment data, demographic statistics, real estate information and weather data, among others[32].

With regard to data sources, the knowledge of data manipulation, integration and preparation is important because robust datasets are often at the core of deep analytics efforts. Data often comes from disparate locations (e.g., internal data in different repositories and data in the cloud) and in large volumes. Traditional RDBMS products tend to impose performance and flexibility restrictions on these kinds of advanced analytics activities. Significant efforts are put into databases designed for high-performance crunching of numbers, text and other types of content such as audio, video and raw data streams. An experience with some of the technologies such as NoSQL and in-memory computing is therefore necessary. Most advanced analytics involves finding relationships across datasets. Hence, data scientists must be adept at integrating data. In addition, legitimate analytics requires high-quality data and the data scientists must also be skilled at validating and cleansing data.

[32] Windows. Windows Azure Marketplace. [Online] Available from: https://datamarket.azure.com/browse/Data [Accessed 10th Oct 2012].

Analytics modelling

Analytics modelling is the process of formulating ways to explore and gain insights from data. An important functional skill set here is to be able to summarise and visualise data to understand key relationships. Depending on the applied domain and type of data, different sets of analytics algorithms will be needed. Therefore, there is a need to be able to recognise the appropriate analytics technique to use for the data and business problem. Analytics modelling skills allow a data scientist to build and diagnose models, and interpret analytic insights for business users.

Business analysis

Business analysis skills are guided by an understanding of the business context that drives the analysis and leverages the value of data. Data scientists should be able to correctly frame a problem and come up with a hypothesis. Business analysis also includes the ability to distinguish mere facts from insights that will matter to the business, and to communicate these insights to business leaders.

In response to the talent shortage of data scientists, enterprises have come up with training programmes to train interested individuals. For example, IBM has launched a partnership programme with universities in China, India, Ireland and Scotland to help universities train students adept at analytics[33]. EMC established its own Data Scientist Associate (EMCDSA) Certification which encompasses the training of various data scientist skill sets. Cloudera also has its own certification, in the form of the Cloudera Certified Developer Exam and Cloudera Certified Administrator Exam[34].

4.4 Technology Outlook

The following section discusses some of the Big Data technologies along with the projected time frame for mainstream adoption.

4.4.1 Less than three years

4.4.1.1 Hadoop MapReduce and Hadoop Distributed File System (HDFS)

Hadoop is a framework that provides open source libraries for distributed computing using MapReduce software and its own distributed file system, simply known as the Hadoop Distributed File System (HDFS). It is designed to scale out from a few computing nodes to thousands of machines, each offering local computation and storage. One of Hadoop's main value propositions is that it is designed to run on commodity hardware such as commodity servers or personal

[33] Erica Thinesen. IIBM Partners with Universities Worldwide to Promote Analytics Studies. [Online] Available from: http://www.itproportal.com/2011/12/23/ibm-partners-universities-around-worldwide-promote-analytics-studies/ [Accessed 9th July 2012].

[34] Jon Zuanich. Cloudera Training for Apache Hadoop and Certification at Hadoop World. [Online] Available from: http://www.cloudera.com/blog/2011/09/cloudera-training-for-apache-hadoop-and-certification-at-hadoop-world/ [Accessed 9th July 2012].

computers, and has high tolerance for hardware failure. In Hadoop, hardware failure is treated as a rule rather than an exception.

HDFS

The HDFS is a fault-tolerant storage system that can store huge amounts of information, scale up incrementally and survive storage failure without losing data. Hadoop clusters are built with inexpensive computers. If one computer (or node) fails, the cluster can continue to operate without losing data or interrupting work by simply re-distributing the work to the remaining machines in the cluster. HDFS manages storage on the cluster by breaking files into small blocks and storing duplicated copies of them across the pool of nodes. The figure below illustrates how a data set is typically stored across a cluster of five nodes. In this example, the entire data set will still be available even if two of the servers have failed.

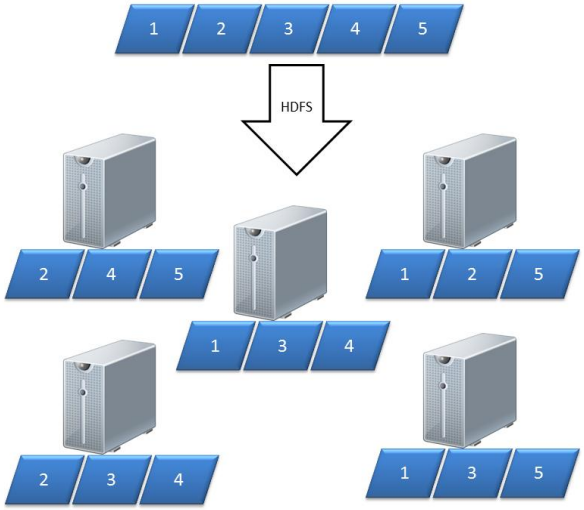


Figure 9: Illustration of distributed file storage using HDFS

Compared to other redundancy techniques, including the strategies employed by Redundant Array of Independent Disks (RAID) machines, HDFS offers two key advantages. Firstly, HDFS requires no special hardware as it can be built from common hardware. Secondly, it enables an efficient technique of data processing in the form of MapReduce.

MapReduce[35]

Most enterprise data management tools (database management systems) are designed to make simple queries run quickly. Typically, the data is indexed so that only small portions of the data need to be examined in order to answer a query. This solution, however, does not work for data that cannot be indexed, namely in semi-structured form (text files) or unstructured form (media files). To answer a query in this case, all the data has to be examined. Hadoop uses the MapReduce technique to carry out this exhaustive analysis quickly.

[35] Jeffrey Dean, Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. [Online] Available from: http://static.usenix.org/event/osdi04/tech/dean.html [Accessed 9th July 2012].

MapReduce is a data processing algorithm that uses a parallel programming implementation. In simple terms, MapReduce is a programming paradigm that involves distributing a task across multiple nodes running a "map" function. The map function takes the problem, splits it into sub-parts and sends them to different machines so that all the sub-parts can run concurrently. The results from the parallel map functions are collected and distributed to a set of servers running "reduce" functions, which then takes the results from the sub-parts and re-combines them to get the single answer.

The Hadoop eco-systems

In addition to MapReduce and HDFS, Hadoop also refers to a collection of other software projects that uses the MapReduce and HDFS framework. The following table briefly describes some of these tools.

HBase	A key-value pair database management system that runs on HDFS
Hive	A system of functions that support data summarisation and *ad hoc* query of the Hadoop MapReduce result set used for data warehousing
Pig	High-level language for managing data flow and application execution in the Hadoop environment
Mahout	Machine-learning system implemented on Hadoop
Zookeeper	Centralised service for maintaining configuration information, naming, providing distributed synchronisation and group services
Sqoop	A tool designed for transferring bulk data between Hadoop and structured data stores such as relational databases

According to IDC, the global market size of Hadoop projects in 2011 was US$77 million. The market is expected to grow almost ninefold to US$682.5 million by 2015.[36]

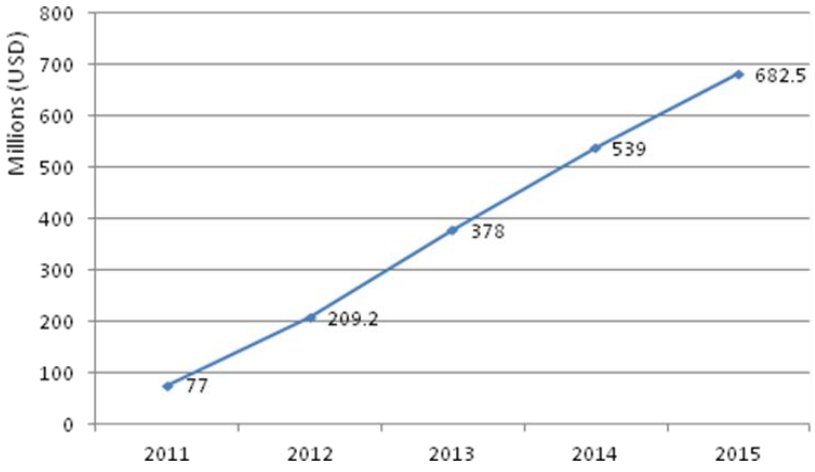


Figure 10: Worldwide Hadoop Ecosystem Software revenue forecast[36]

[36] Carl W. Olofson, Dan Vesset. Worldwide Hadoop – MapReduce Ecosystem Software 2012-2016 Forecast [Online] Available from: http://www.idc.com/getdoc.jsp?containerId=234294 [Accessed 9th July 2012].

Opportunities

Semi-structured and unstructured data sets are the two fastest growing data types in the digital universe. Analysis of these two data types will not be possible with traditional database management systems. Hadoop HDFS and MapReduce enable the analysis of these two data types, giving organisations the opportunity to extract insights from bigger datasets within a reasonable amount of processing time.

Hadoop MapReduce's parallel processing capability has increased the speed of extraction and transformation of data. Hadoop MapReduce can be used as a data integration tool by reducing large amounts of data to its representative form which can then be stored in the data warehouse.

At the current stage of development, Hadoop is not meant to be a replacement for scale-up storage and is designed more for batch processing rather than for interactive applications. It is also not optimised to work on small file sizes as the performance gains may not be considerable when compared to huge data processing.

Inhibitors

Lack of industry standards is a major inhibitor to Hadoop adoption. Currently, a number of emerging Hadoop vendors are offering their customised versions of Hadoop. HDFS is not fully Portable Operating System Interface (POSIX)-compliant, which means system administrators cannot interact with it the same way they would with a Linux or Unix system.

The scarcity of expertise capable of building a MapReduce system, managing Hadoop applications and performing deep analysis of the data will also be a challenge. At the same time, many of the Hadoop projects require customisation and there is no industry standard on the best practices for implementation.

Many enterprises may not need Hadoop on a continual basis except for specific batch-processing jobs that involve very huge data sets. Hence, the return of investment (ROI) for on-premise deployment will not be attractive. In addition, the integration of Hadoop clusters with legacy systems is complex and cumbersome, and is likely to remain a big challenge in the near term.

4.4.1.2 NoSQL Database management system (NoSQL DBMS)

NoSQL database management systems (DBMSs) are available as open source software and designed for use in high data volume applications in clustered environments. They often do not have fixed schema and are non-relational, unlike the traditional SQL database management system (also known as RDMS) in many data warehouses today. Because they do not adhere to a fixed schema, NoSQL DBMS permit more flexible usage, allowing high-speed access to semi-structured and unstructured data. However, SQL interfaces are also increasingly being used alongside the MapReduce programming paradigm.

There are several types of NoSQL DBMS:

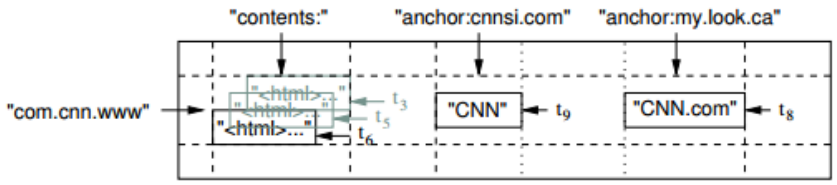
Key-value stores[37]: Key-value pair (KVP) tables are used to provide persistence management for many of the other NoSQL technologies. The concept is as follows: the table has two columns - one is the key; the other is the value. The value could be a single value or a data block containing many values, the format of which is determined by program code. KVP tables may use indexing, hash tables or sparse arrays to provide rapid retrieval and insertion capability, depending on the need for fast look-up, fast insertion or efficient storage. KVP tables are best applied to simple data structures and on the Hadoop MapReduce environment. Examples of key-value data stores are Amazon's Dynamo and Oracle's BerkeleyDB.

Keys	Value
3414	"a string of text 1"
3437	"a string of text 2"
3497	"a series of binary codes"

Figure 11: Example of Key-value pair table

Document-oriented database: A document-oriented database is a database designed for storing, retrieving and managing document-oriented or semi-structured data. The central concept of a document-oriented database is the notion of a "document" where the contents within the document are encapsulated or encoded in some standard format such as JavaScript Object Notation (JSON), Binary JavaScript Object Notation (BSON) or XML. Examples of these databases are Apache's CouchDB and 10gen's MongoDB.

BigTable database[38]: BigTable is a distributed storage system based on the proprietary Google File System for managing structured data that is designed to scale to a very large size – petabytes of data across thousands of commodity servers. Also known as Distributed Peer Data Store, this database is almost similar to relational database except that the data volume to be handled is very high and the schema does not dictate the same set of columns for all rows. Each cell has a time stamp and there can be multiple versions of a cell with different time stamps. In order to manage the huge tables, Bigtable splits tables at row boundaries and saves them as tablets. Each tablet is around 200MB, and each server saves about 100 tablets. This setup allows tablets from a single table to be spread among many machines. It also allows for load balancing and fault tolerance. An example of a BigTable database is CassandraDB.



A slice of an example table that stores Web pages. The row name is a reversed URL. The contents column family contains the page contents, and the anchor column family contains the text of any anchors that reference the page. CNN's home page is referenced by both the Sports Illustrated and the MY-look home pages, so the row contains columns named anchor:cnnsi.com and anchor:my.look.ca. Each anchor cell has one version; the contents column has three versions, at timestamps t_3, t_5, and t_6.

Figure 12: Illustration of BigTable storage[38]

[37] Carl W. Olofson. The Big Deal about Big Data. [Online] Available from: http://www.idc.com/getdoc.jsp?containerId=226904 [Accessed 9th July 2012].

[38] Google Inc. Bigtable: A Distributed Storage System for Structured Data. [Online] Available from: http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//archive/bigtable-osdi06.pdf [Accessed 10th Oct 2012].

Graph database[39]: A graph database contains nodes, edges and properties to represent and store data. In a graph database, every entity contains a direct pointer to its adjacent element and no index look-ups are required. A graph database is useful when large-scale multi-level relationship traversals are common and is best suited for processing complex many-to-many connections such as social networks. A graph may be captured by a table store that supports recursive joins such as BigTable and Cassandra. Examples of graph databases include InfiniteGraph from Objectivity and the Neo4j's open source graph database.

Opportunities

With the Big Data movement comes a series of use cases that conventional schematic DBMS is not meant to address. These cases typically include providing data processing and access environments for large-scale, compute-intensive analytics. The fact that NoSQL has been developed to handle data management problems well outside the realm of traditional databases spells new opportunities for the technology. NoSQL databases are designed to be able to scale out on commodity hardware (adopting the principle of the Hadoop framework) to manage the exploding data and transaction volumes. The result is that the cost per gigabyte or transactions per second for NoSQL can be many times less than the cost for RDBMS, allowing more data storage and processing at a lower price point. However, it is important to recognise that the NoSQL database can realistically focus on two of the three properties of Consistency, Availability and Partition Tolerance (CAP Theorem). NoSQL databases need partition tolerance in order to scale properly, so it is very likely they will have to sacrifice either availability or consistency.

Inhibitors

The history of RDBMS systems in the market indicates the level of technology maturity that gives assurance to most CIOs. For the most part, RDBMS systems are stable and richly functional. In contrast, most NoSQL alternatives are in pre-production versions with many key features yet to be implemented. There is no shortage of developers who are familiar with RDBMS concepts and programming. In contrast, there are very few who can claim to be an expert in NoSQL database concept and programming. Many of the developers dealing with NoSQL are going through a learning phase. However this situation will change over time as more become familiar with NoSQL DBMS.

4.4.1.3 Visualisation-based data discovery tool

The visualisation-based data discovery tool belongs to an emerging category of easy-to-use data analytics tools that provide business analytics capabilities suitable for workgroup or personal use. A data discovery tool provides usability, flexibility and control over how to model and build content without having to go through the IT department. The visualisation-based data discovery tool is an alternative to traditional business intelligence platforms and offers interactive graphical user interfaces built on in-memory architecture to address business user needs.

[39] Info Grid. Operations on a Graph Database. [Online] Available from: http://infogrid.org/blog/2010/03/operations-on-a-graph-database-part-4/ [Accessed 9th July 2012].

Visualisation-based data discovery tools include a proprietary data structure that stores and models data gathered from disparate sources, a built-in performance layer that makes data aggregations, summaries and pre-calculations unnecessary, and an intuitive interface that enables users to explore data without much training.

Opportunities

According to a study by the Aberdeen Group, managers who make use of visual discovery tools are 8% more likely than their peers to be able to access the right information in a timely manner that will enable them to impact decision making.[40]

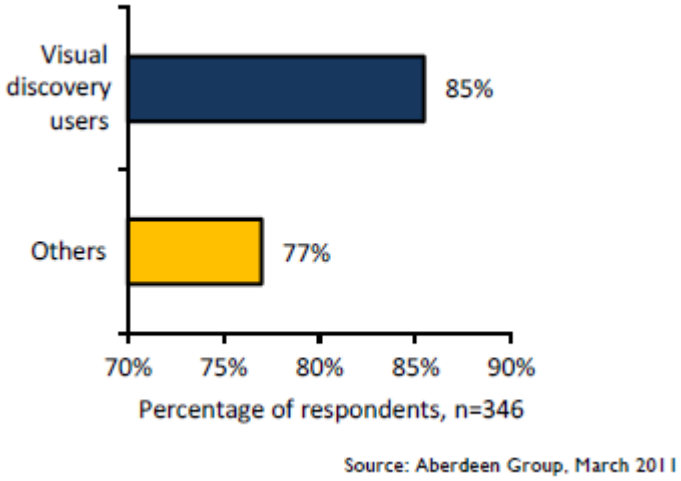


Figure 13: Visualisation-based tool provides timely access to information

Visualisation discovery also promotes self-service business analytics, enabling business users to proactively address business intelligence needs in a timely way. The self-service approach frees up the IT department to focus on other tasks while empowering end users to create personalised reports and queries. To facilitate the speedy retrieval of relevant business intelligence, the user interface must be intuitive and easy to use, the metrics and vocabulary must be comprehensible, and processes must be clear and understandable. Organisations deploying visual discovery tools can manage 65% more users with every supporting member of the IT team[40].

[40] David White. Picture this: Self-Service BI through Data Discovery & Visualization. [Online] Available from: http://aberdeen.com/Aberdeen-Library/7729/AI-business-intelligence-analytics.aspx [Accessed 9th July 2012].

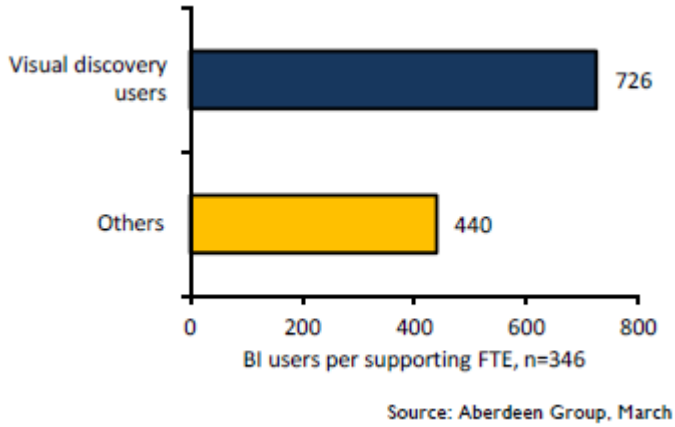


Figure 14: Visualisation based tools better leverage scarce IT resources

Inhibitors

The data discovery tool's inherent ability to work independently often results in its de-coupling from "official" cleansed and sanctioned underlying metadata, business glossaries and data stores. Data stewards may want to prevent the spread of such tools to avoid the "spread mart[41]" problem. In addition, more often than not, data regulations are present within organisations that maintain data quality and security. In this case, the self-service nature of this tool may curtail its usage.

4.4.1.4 Text analytics

Text analytics is the process of deriving information from text sources. These text sources are forms of semi-structured data that include web materials, blogs and social media postings (such as tweets). The technology within text analytics comes from fundamental fields including linguistics, statistics and machine learning. In general, modern text analytics uses statistical models, coupled with linguistic theories, to capture patterns in human languages such that machines can "understand" the meaning of texts and perform various text analytics tasks. These tasks can be as simple as entity extraction or more complex in the form of fact extraction or concept analysis.

Entity extraction: Entity extraction identifies an item, a person or any individual piece of information such as dates, companies or countries.

Fact extraction: A fact is a statement about something that exists, has happened and is generally known. It is defined by a collection of entities and fact extraction identifies a role, relationship, cause or property.

Concept extraction: Concept extraction functions identify an event, process, trend or behaviour.

Text analytics will be an increasingly important tool for organisations as the attention shifts from structured data analysis to semi-structured data analysis. One major application of text analytics would be in the field of sentiment analysis where consumer feedback can be extracted from the social media feeds and blog commentaries. The potential of text analytics in this application has

[41] Spread marts are data shadow systems in which individuals collect and massage data on an ongoing basis to support their information requirements. These shadow systems usually built on spreadsheets, exist outside of approved, IT-managed corporate data repositories (e.g., data warehouses), and contain data and logic that often conflict with corporate data.

spurred much research interest in the R&D community. In the Singapore Management University (SMU), text analytics is an important research area that includes adaptive relation extraction which is the task of finding relations between people, organisations and other entities from natural language text, unsupervised information extraction which explores the conversion of unstructured free text into structured information without human annotations and text mining on social media for the purpose of uncovering user sentiments.

Opportunities:

Combining text analytics with traditional structured data allows a more complete view of the issue, compared to analysis using traditional data mining tools. Applying text mining in the area of sentiment analysis helps organisations uncover sentiments to improve their customer relationship management. Text analytics can also be applied in the area of public security by the scrutiny of text for patterns that characterise criminal activity or terrorist conspiracy.

Inhibitors:

The text analytics solution market is still immature. The analytics engine will face challenges in dealing with non-English content and local colloquialisms. Hence, a text analytics solution developed for a particular market may not be directly applicable in another situation – a certain level of customisation will be needed. The suitability of text analytics on certain text sources, such as technical documents or documents with many domain specific terms, may be questionable as well.

Adoption of text analytics is more than just deploying the technology. Knowing the metrics to use in determining the results is also a required skill. There has to be an understanding of what to analyse and how to use the outcome of analysis to improve business. This requires a certain level of subjectivity which may not be what management desires.

4.4.1.5 In-memory analytics

In-memory analytics is an analytics layer in which detailed data (up to terabyte size) is loaded directly into the system memory from a variety of data sources, for fast query and calculation performance. In theory, this approach partly removes the need to build metadata in the form of relational aggregates and pre-calculated cubes.

The use of in-memory processing as a back-end resource for business analytics improves performance. On the traditional disk-based analytics platform, metadata has to be created before the actual analytics process takes place. The way which the metadata is modelled is dependent on the analytics requirements. Changing the way to model the metadata to fulfill new requirements requires a good level of technical knowledge. In-memory analytics removes the need to pre-model this metadata for every end user's needs. Consequently, the developer no longer needs to consider every possible avenue of analysis. The relevance of the analytics content is also improved as data can be analysed the moment it is stored in the memory. The speed that is delivered by in-memory analytics makes it possible to power interactive visualisation of data sets, making data access a more exploratory experience.

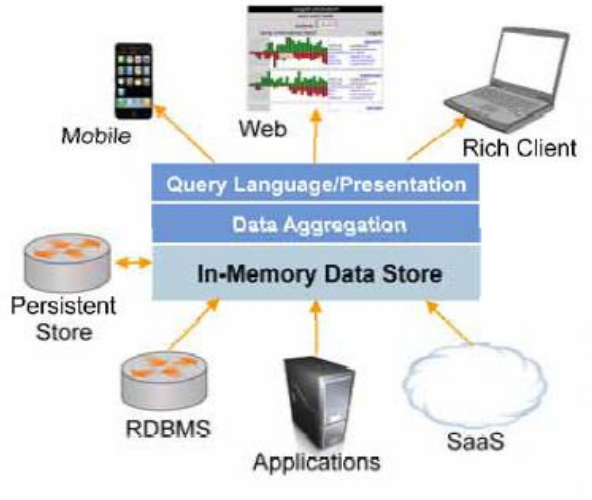


Figure 15: In-memory analytics overview[42]

In-memory analytics is enabled by a series of in-memory technologies:

In-memory data management:

I. In-memory database management system (IMDBMS): An IMDBMS stores the entire database in the computer RAM, negating the need for disk I/O instructions. This allows applications to run completely in memory.

II. In-memory data grid (IMDG): The IMDG provides a distributed in-memory store in which multiple, distributed applications can place and retrieve large volumes of data objects.

In-memory low-latency messaging:

This platform provides a mechanism for applications to exchange messages as rapidly as possible through direct memory communications.

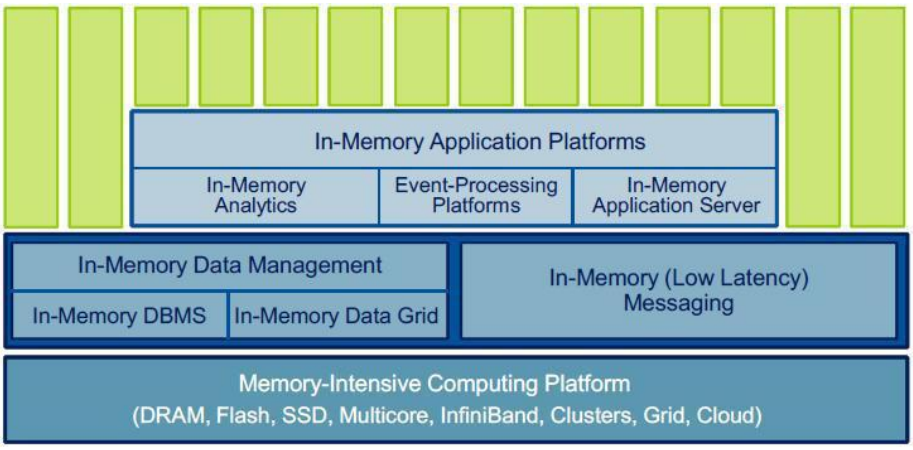


Figure 16: Taxonomy of in-memory computing technologies[43]

[42] Massimo Pezzini. Net IT Out: In-Memory Computing — Thinking the Unthinkable Applications. [Online] Available from: http://agendabuilder.gartner.com/ESC23/webpages/SessionDetail.aspx?EventSessionId=1066 [Accessed 9th July 2012].

Opportunities

In-memory analytics can work with large and complex data sets and churn output in a matter of minutes or even seconds. This way, in-memory analytics can be provided as a form of real-time or near real-time services to the users. With less dependency on the metadata, in-memory analytics will enable self-service analysis. When coupled with interactive visualisation and data discovery tools for intuitive, unencumbered and fast exploration of data, in-memory analytics can facilitate a series of "what-if" analyses with quick results. This facilitates the fine-tuning of analytics results with multiple query iterations which, in turn, enables better decision making.

In-memory analytics can be adopted to complement other Big Data solutions. One possibility is to run in-memory analytics on a Hadoop platform to enable high-speed analytics on distributed data. NoSQL databases could run on in-memory computing platform to support in-memory analytics. And finally, in-memory analytics can run on in-memory data grids that can handle and process multiple petabytes of in-memory datasets.

Inhibitors

In-memory analytics technology has the potential to subvert enterprise data integration efforts. As the technology enables standalone usage, organisations need to ensure they have the means to govern its usage and there is an unbroken chain of data lineage from the report to the original source system.

Although the average price per gigabyte of RAM has been going down, it is still much higher than the average price per gigabyte of disk storage. In the near term, memory still exacts a premium relative to the same quantity of disk storage. Currently, Amazon Web Services charges about US$0.10 per gigabyte of disk storage[44] while memory costs about 100 times as much, at about US$10.57 per gigabyte.[45] The utilisation of scaled-up in-memory analytics is likely to be restricted to the deep-pocketed and technically astute organisations.

4.4.1.6 Predictive analytics

Predictive analytics is a set of statistical and analytical techniques that are used to uncover relationships and patterns within large volumes of data that can be used to predict behaviour or events. Predictive analytics may mine information and patterns in structured and unstructured data sets as well as data streams to anticipate future outcomes. The real value of predictive analytics is to provide predictive support service that goes beyond traditional reactive break-and-fix assistance and toward a proactive support system by preventing service-impacting events from occurring.

---

[43] Massimo Pezzini, et al. Top 10 Strategic Technology Trends: In-memory Computing. [Online] Available from: http://www.gartner.com/id=1914414 [Accessed 9th July 2012].

[44] Amazon Web Services. Amazon Elastic Block Store. [Online] Available from: http://aws.amazon.com/ebs/ [Accessed 9th July 2012].

[45] Gartner. Weekly Memory Pricing Index 3Q11 Update. [Online] Available from: http://www.gartner.com/technology/core/products/research/markets/hardwareStorage.jsp [Accessed 9th July 2012].

According to Gartner[46], three methods are being evaluated within the marketplace for the prediction of technical issues within the product support space. Gartner believes though that mature predictive support services will ultimately use a combination of all three of these approaches.

Pattern-based approach: These technologies seek to compare real-time system performance and configuration data with unstructured data sources that may include known failure profiles, historic failure records and customer configuration data. Powerful correlation engines are required to seek statistical patterns within huge, multifaceted repositories to determine if the customer's current configuration and performance data indicate a likely failure.

Rules-based approach: Statistical analysis of historic performance data, previously identified failure modes and the results of stress/load testing is used to define a series of rules that real-time telemetry data is compared with. Each rule may interrogate multiple telemetry data points and other external factors such as the time of day, environmental conditions and concurrent external activities, against defined thresholds. Breaches of these rules may then be collated and escalation routines can be triggered, depending on the likely severity and impact of the resultant issues or outages.

Statistical process control-based models: Control chart theory has been a mainstay of quality manufacturing processes for more than half a century and has proven an invaluable aid to managing complex process-driven systems. The advent of retrofit-capable, real-time telemetry, improvements in data acquisition solutions and network capacity to support such large data volumes mean the statistical techniques that underpinned the quality revolution within the manufacturing space can now be used to industrialise IT service delivery. Statistical anomalies can be readily identified and used to initiate appropriate contingent or preventive action to ensure that service performance is unaffected and the business can continue functioning as normal.

Opportunities

Predictive analytics exploits patterns found in historical and transactional data to identify future risks and opportunities. Approaches are focused on helping companies acquire actionable insights, and identify and respond to new opportunities more quickly.

Predictive analytics has direct applications in many verticals:

- Law enforcement agencies can look for patterns in criminal behaviour and suspicious activity which can help them identify possible motives and suspects, leading to the more effective deployment of personnel. For example, the Edmonton Police Services uses sensors and data analytics to create maps that define high-crime zones so that additional police resources can be proactively diverted to them.[47]

- Public health authorities can monitor information from various sources, looking for elevated levels of certain symptoms that signal a potential outbreak of disease. In Singapore, the Institute of High Performance Computing (IHPC) has developed a simulation model for epidemiological study in Singapore. In case of a pandemic, the model can be used to predict the spread and

[46] Rob Addy. Emerging Technology Analysis: Predictive Support Services. [Online] Available from: http://www.gartner.com/id=1875816 [Accessed 9th July 2012].

[47] IBM. Edmonton Police Service Fights Crime with IBM Business Analytics Technology. [Online] Available from: http://www-03.ibm.com/press/us/en/pressrelease/28455.wss [Accessed 9th July 2012].

virulence of viruses, providing Singapore health authorities the basis for their public health advisories and activities.

- Tax departments can use predictive analytics to identify patterns to highlight cases that warrant further investigation. Additionally, predictive analytics can be used to understand the likely impact of policies on revenue generation.

Inhibitors

As predictive analytics requires a multitude of data inputs, the challenge is to know which data inputs would be relevant and how these disparate sources can be integrated. In some predictive models, there is a need for user data touches on the sensitive issue of data privacy. Finally, the reliability of the prediction outcome faces scepticism, especially when the prediction outcome deviates from the decision maker's point of view. It will take time for decision makers to accept the outcomes of predictive analytics as an influencing factor in any decision. Similarly it will take time for the underlying predictive algorithms to progress and mature to more sophisticated levels.

4.4.1.7 SaaS-based business analytics

Software-as-a-Service (SaaS) is software owned, delivered and managed remotely by one or more providers. A single set of common code is provided in an application that can be used by many customers at any one time. SaaS-based business analytics enables customers to quickly deploy one or more of the prime components of business analytics without significant IT involvement or the need to deploy and maintain an on-premise solution.

The prime components:

Analytic applications: Support performance management with pre-packaged functionality for specific solutions;

Business analytics platforms: Provide the environment for development and integration, information delivery and analysis;

Information management infrastructures: Provide the data architecture and data integration infrastructure.

Opportunities

Leveraging the benefits of cloud computing, SaaS-based business analytics offers a quick, low-cost and easy-to-deploy business analytics solution. This is especially the case for enterprises that do not have the expertise to set up an in-house analytics platform nor the intention to invest in internal business analytics resources. SaaS-based business analytics may be useful for mid and small enterprises that have yet to invest in any form of on-premise business analytics solutions.

Inhibitors

There could be integration challenges for enterprises that want to export data to, and extract data from, the service provider for integration with the on-premise information infrastructure. As the data resides on the cloud, SaaS-based business analytics may not be suitable for businesses that have to worry about data privacy and security issues.

4.4.1.8 Graph Analytics

Graph analytics is the study and analysis of data that can be transformed into a graph representation consisting of nodes and links. Graph analytics is good for solving problems that do not require the processing of all available data within a data set. A typical graph analytics problem requires the graph traversal technique. Graph traversal is a process of walking through the directly connected nodes. An example of a graph analytics problem is to find out how many ways two members of a social network are linked directly and indirectly. A more contemporary example of graph analytics relates to social networks.

Different forms of graph analytics exist:

Single path analysis: The goal is to find a path through the graph, starting with a specific node. All the links and the corresponding vertices that can be reached immediately from the starting node are first evaluated. From the identified vertices, one is selected, based on a certain set of criteria and the first hop is made. After that, the process continues. The result will be a path consisting of a number of vertices and edges.

Optimal path analysis: This analysis finds the 'best' path between two vertices. The best path could be the shortest path, the cheapest path or the fastest path, depending on the properties of the vertices and the edges.

Vertex centrality analysis: This analysis identifies the centrality of a vertex based on several centrality assessment properties:

- Degree centrality: This measure indicates how many edges a vertex has. The more edges there are, the higher the degree centrality.

- Closeness centrality: This measure identifies the vertex that has the smallest number of hops to other vertices. The closeness centrality of the node refers to the proximity of the vertex in reference to other vertices. The higher the closeness centrality, the more number of vertices that require short paths to the other vertices.

- Eigenvector centrality: This measure indicates the importance of a vertex in a graph. Scores are assigned to vertices, based on the principle that connections to high-scoring vertices contribute more to the score than equal connections to low-scoring vertices.

Graph analytics can be used in many areas:

- In the finance sector, graph analytics is useful for understanding the money transfer pathways. A money transfer between bank accounts may require several intermediate bank accounts and graph analytics can be applied to determine the different relationships between different account holders. Running the graph analytics algorithm on the huge financial transaction data sets will help to alert banks to possible cases of fraudulent transactions or money laundering.

- The use of graph analytics in the logistics sector is not new. Optimal path analysis is the obvious form of graph analytics that can be used in logistics distribution and shipment environments. There are many examples of using graph analytics in this area and they include "the shortest route to deliver goods to various addresses" and the "most cost effective routes for goods delivery".

- One of the most contemporary use cases of graph analytics is in the area of social media. It can be used not just to identify relationships in the social network, but to understand them. One outcome from using graph analytics on social media is to identify the "influential" figures from each social graph. Businesses can then spend more effort in engaging this specific group of people in their marketing campaigns or customer relationship management efforts.

4.4.1.9 Master Data Management

Master data is the official, consistent set of identifiers and extended attributes that describes the core entities of the enterprise, such as customers, prospects, locations, hierarchies, assets and policies. Master data management (MDM) is a technology-enabled discipline in which businesses and IT departments work together to ensure the uniformity, accuracy, stewardship and accountability of the enterprise's official, shared master data assets.

There are four dimensions which any MDM discipline must address:[48]

Use case: This refers to how the master data will be used. There could be three MDM use case types, namely design (master data being used to build enterprise data models), operations (master data being used to process business transactions) and analytics (master data that is used to analyse business performance).

Domain: This refers to the entity with which the data has relations. The domain could refer to customer, supplier, product, material, employee and assets.

Implementation style: This is the degree to which master data is stored and governed.

Industry: This refers to the specific meaning of master data in an industry context.

Organisations use master data for consistency, simplification and uniformity of process, analysis and communication across the enterprise. Once implemented, the master data moves the organisation closer to the objective of data sharing in the application portfolio.

Opportunities

MDM helps to reconcile different data silos to create a master data "single version of the truth", either through physically consolidating data into a single data store or federation of the best source

[48] Gartner. Gartner Says Master Data Management Is One of the Fastest-Growing Software Segments in 2008. [Online] Available from: http://www.gartner.com/it/page.jsp?id=801512 [Accessed 9th July 2012].

of record data from multiple sources[49]. It is an example of an information-sharing environment that represents a key part of an enterprise information management initiative. MDM helps organisations to break down operational barriers. It can help to ensure that all master data is "clean" in a business analytical framework which can then improve the ability of users to make decisions more effectively, leading to increased performance and agility, process integrity and cost optimisation.

MDM domains have a direct impact on the business. For example, MDM for customer data creates a centralised system of records for customer data such that all applications and business processes go to the system to determine whether the right data was used and whether it is accurate. Having a single view of customer data can be useful for all processes such as marketing, and sales and service, leading to a more customer-centric customer experience and improved retention levels.

MDM ensures the entire enterprise uses one unified model for all its primary master data objects. Enterprise-wide MDM significantly reduces the costs associated with organisational integration that are needed for any transformative strategy by removing organisational barriers that inhibit information re-use.

Inhibitors

MDM requires sustained attention and any form of multi-year programme is difficult as employee roles, business conditions and workloads vary over time. In addition, functionally and geographically decentralised organisations have no organisational home for the MDM programme. Global companies present different levels of technical sophistication, and political and cultural sensitivities. This makes the execution of MDM processes challenging. MDM requires at least some central control and reconciliation, even if local customisation is allowed. Finally, breaking down the data silos is not an easy task. On the one hand, there could be no one who is willing to take on data stewardship roles and on the other, the business units may refuse to give up data ownership and control.

4.4.2 Three to five years

4.4.2.1 Complex event processing

A "complex" event is an abstraction of other "base" events and represents the collective significance of these events. Complex event processing (CEP) combines data from multiple sources to discern trends and patterns based on seemingly unrelated events. It is a style of computing that is implemented by event-driven, continuous intelligence systems. A CEP system uses algorithms and rules to process streams of event data that it receives from one or more sources to generate insights. It processes complex events, placing them in context to identify threat and opportunity situations. For example, sales managers who receive an alert message containing a complex event that says, "Today's sales volume is 30% above average", grasp the situation more quickly than if they were shown the hundreds of individual sales transactions (base events) that contributed to that complex event. This information is then used to guide the response in sense-and-respond business activities. Computation of CEP is triggered by the receipt of event data. CEP systems store large

[49] Gartner. Gartner Says Master Data Management Is One of the Fastest-Growing Software Segments in 2008. [Online] Available from: http://www.gartner.com/it/page.jsp?id=801512 [Accessed 9th July 2012].

amount of events within the memory spaces and they run continuously so that they can act immediately as the event data arrives.

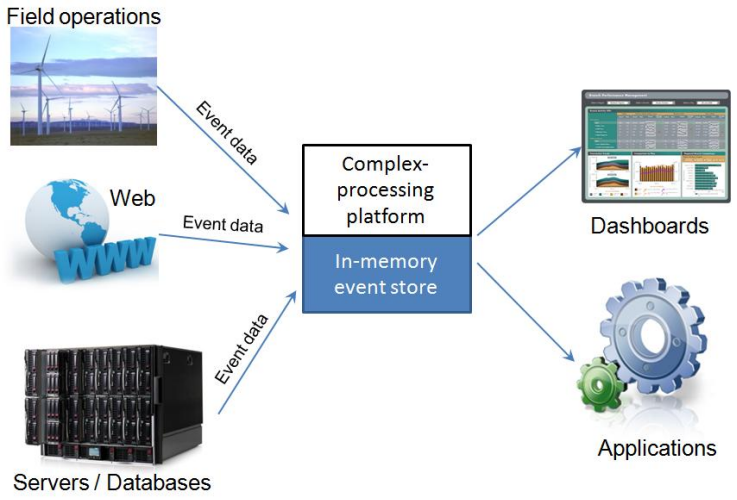


Figure 17: Event-processing platform

Opportunities

CEP is the basis for a growing number of pattern-based strategies[50], particularly those that leverage low-latency operational intelligence. CEP enables operational technology and business activity monitoring (BAM) applications. CEP improves the quality of decision making by presenting information that would otherwise be overlooked, enables faster response to threats and opportunities, helps shield business professionals from data overload by implementing a management-by-exception policy, and reduces the cost of manually processing the growing volume of event data in business.

Drivers for CEP technology will come from the sense-and-respond and situation awareness applications. CEP could be used to analyse events within a network, power grid or other large systems to determine whether they are performing optimally, experiencing problems or have become targets of an attack. It could also be used in optimisation analysis which could help activities such as inventory tracking, detecting of threats to homeland security and supply chain management. With CEP, financial trading institutions can correlate worldwide stock, commodity and other market movements to recognise potential opportunities or problems with current portfolio holdings. CEP could also examine internal corporate activities to determine if they conform to corporate policies or government regulations. With CEP, businesses can map discrete events to expected outcomes and relate them to key performance indicators (KPIs), extracting insights that will enable them to allocate resources to maximise opportunity and reduce risk.

Inhibitor

[50] Gartner defines pattern-based strategy as the discipline that enables business leaders to seek, amplify, examine and exploit new business patterns. A business pattern is a set of recurring and related elements such as business activities and events that indicate a business opportunity or threat.

At a list price ranging from US$50,000 to US$500,000, CEP platforms can be an over-investment for many event processing applications that handles modest volumes of events and simple processing rules. In most cases, event processing can still be done without adopting the CEP technology. By embedding custom CEP logic within the business applications, the outcome could be equally desirable in most cases.

The lack of awareness and understanding of CEP and what it can do is another inhibitor. On the one hand, many software developers are in fact creating applications with CEP logic without realising that they are actually implementing CEP, and hence they reinvent the wheel many times over. On the other hand, there is a lack of expertise among business analysts, system architects and project leaders in recognising how a general purpose event-processing platform can help in building CEP applications.

4.4.2.2 Mobile Business Analytics

Mobile business analytics is an emerging trend that rides on the growing popularity of mobile computing. From the mobile workforce's perspective, being able to access the latest analytics insights from back-end data stores creates a need for mobile business analytics. There are two types of mobile business analytics: passive and active.[51]

Passive mobile business analytics revolves around the "push" factor. Event-based alerts or reports can be pushed to the mobile devices after being refreshed at the back-end. Passive mobile business analytics may be a step ahead in providing accessibility convenience to the users but it is not enough to support the just-in-time analytical requirements users want. Active mobile business analytics enables users to interact with the business analytics systems on-the-fly. It can work as a combination of both "push" and "pull" techniques. An initial view of a report could constitute a "push" and further analytical operations on the report to obtain additional information could comprise the "pull" factor.

There are two approaches to develop business analytics applications for mobile devices: creating software for specific mobile operating systems such as iOS or Android or developing browser-based versions of their business analytics applications. Developing a browser-based version of the business analytics application requires only a one-time development effort as the deployment can be made across devices. On the other hand, custom-developed mobile business analytics applications can provide full interactivity with the content on the device. In addition, this approach provides periodic caching of data which can be viewed offline.

Opportunities

Mobile business analytics facilitates off-site decision making. More often than not, decision makers only need access to a few key analytics metrics churned out by the back-end data store. Having access to these metrics on a mobile device can reduce decision bottlenecks, increase business process efficiency and enable broader input into the decision at hand. In addition, device-specific

[51] Information Management. Mobile Business Intelligence for Intelligent Businesses. [Online] Available from: http://www.information-management.com/specialreports/2008_89/10001705-1.html?zkPrintable=1&nopagination=1 [Accessed 9th July 2012].

capabilities can increase the value of mobile business analytics. For instance, incorporating location awareness with a mobile business analytics query provides a location context to the query. The user can then obtain the query results that are relevant to his location.

Inhibitors

Technical and security risk concerns will inhibit the uptake of mobile business analytics, especially in the case of mission-critical deployments. Sensitive corporate data can be at risk if a tablet or smartphone device is compromised. The growth of mobile business analytics is dependent on the success of an enterprise's bring-your-own-device (BYOD) programme.

4.4.2.3 Video Analytics

Video analytics is the automatic analysis of digital video images to extract meaningful and relevant information. Also known as video content analysis (VCA), it uses computer vision algorithms and machine intelligence to interpret, learn and draw inferences from image sequences. Video analytics automates scene understanding which otherwise would have required human monitoring. It is not only able to detect motion, but also qualifies the motion as an object, understands the context around the object, and tracks the object through the scene. In theory, any behaviour that can be seen and accurately defined on a video image can be automatically identified and subsequently trigger an appropriate response.

There are two approaches to video analytics architecture, namely edge-based systems and central-based systems.

Edge-based systems

In this approach, video analytics is performed on the raw image on the video source (i.e., the camera) before any compression is applied. This means the image has the maximum amount of information content, allowing the analytics to work more effectively. Processing of locally sourced data makes the system resilient to transmission/network failures. Alerts can be responded to locally or stored for transmission once the network connection is resumed. Bandwidth usage can be controlled by reducing frame rates, lowering resolution and increasing image compression when no events or alerts are in progress. Edge-based systems must be implemented through an IP video camera or video encoder with sufficient processing power. This may result in larger units with greater power requirements compared to conventional analogue or network cameras.

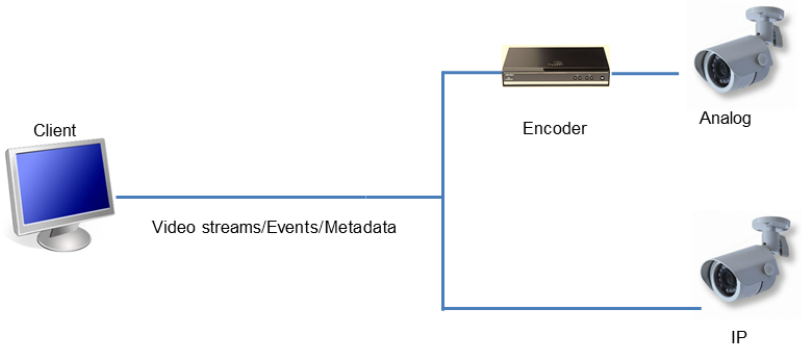


Figure 18: Edge-based system architecture

Central-based systems

In this approach, the video analytics is implemented through a dedicated server that pulls the video, analyses it and issues the alerts or analysis results. Since the analytics equipment is installed in one central location, it makes for easier maintenance and upgrading. The system can be protected from power failure by using a central uninterruptible power supply (UPS). While this independent-of-camera approach is applicable to most types of surveillance systems, large amounts of network bandwidth may be required to transmit high quality images from the image capture devices. Compression and transmission effects may impede the efficiency and accuracy of the analysis because of the loss of information content within the images.

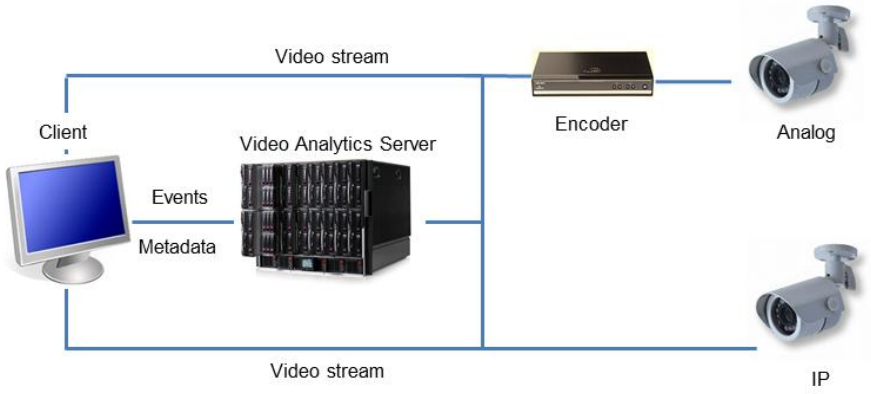


Figure 19: Central-based system architecture

Opportunities

Video analytics can complement business intelligence with security awareness. Automated surveillance cameras with analytics can detect the presence of people and vehicles, and interpret their activities. Suspicious activities such as loitering or moving into an unauthorised area are automatically flagged and forwarded to security personnel. In the retail sector, video analytics can enhance the customer experience and drive revenue generation through an improved understanding of the customer experience. With video analytics, retailers are able to analyse customer time in the store to evaluate the effectiveness of key store promotional areas. Retailers will also be able to determine the demographics of store visitors and recognise the buying patterns by correlating sales receipts to customer demographics.

Inhibitors

Despite technological advancements that lead to improved video capabilities such as scene and object recognition, the current analytics is still focused on intrusion detection and crowd management where events happen randomly and often appeal to very specific verticals that require customisation[52]. In addition, video analytics may still be plagued by accuracy and reliability issues, including problems caused by shadows or foliage and areas of high contrast in light. The video analytics system will have the issue of false alert (or false positive), largely due to the sophistication level of the analytics algorithms. False Alert Rate (FAR) refers to the number of times the system

[52] David Reinsel, Christopher Chute. ISCW12 and Video Surveillance: To the Edge and Beyond? [Online] Available from: http://www.idc.com/research/viewtoc.jsp?containerId=234255 [Accessed 9th July 2012].

creates an alert when nothing of interest is happening. FAR of one per camera per day may sound acceptable but it may be magnified after multiplying across hundreds of cameras. For example, in a network of 100 cameras, that will equate to a false alert every 14 minutes.

Hardware limitation is a concern. While video analysis can take place at the edge of devices, the deployment cost or hardware requirement for each camera, with computing capability installed in each camera to run complex video analytics algorithms, will be high. On the other hand, video analytics that takes place at the central server will require high bandwidth while the image feeds for video analysis may suffer from attenuation and loss of details. This may impede the accuracy of the video analysis. There is a rise of IP-based cameras but a majority of the current infrastructure continues to employ analogue-based cameras. The delineation of both technologies will also result in an increase in cost of adoption of video analytics[52].

4.4.2.4 Data Federation

Data federation technology enables the creation of virtual in-memory views of data integrated from one or many data sources so they can be used for business intelligence or other analyses. In effect, data federation provides the ability to create integrated views of data that appear to the applications as if the data resides in one place, when in fact it may be distributed. Data federation technology provides the ability to create abstract interfaces to data. The virtual views against one or many data sources can be presented in a way which removes the applications from needing to know anything about the physical location and structure of the data. These virtually integrated views of data from multiple distributed sources and abstracted interfaces to data are useful in supporting a variety of different data integration requirements.

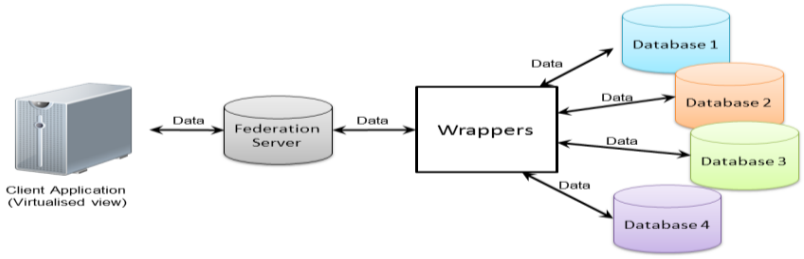


Figure 20: Overview of data federation

Data federation technology enables the creation of rapid prototypes of integrated views of data. This helps to achieve the flexibility for change and yet maintain pervasive and consistent data access with reduced costs because there is less need to create physically integrated data structures. The end result is greater agility from the organisation's data assets. The value of data federation will increase in importance as companies seek to expand their capabilities for accessing data types on top of structured databases, including documents, Web-based data and other data of a less structured variety. However, current data federation tools face performance challenges in the light of the significant and rapidly growing volumes of data managed by organisations. While data federation tools increasingly offer stronger capabilities for data transformation, the ability to deal with the complex data quality issues remains limited. As a result, deployments of data federation technology on broad mission-critical and large-scale systems remain rare.

4.4.2.5 Phase change memory (PC-RAM)

Advancements in memory technology, particularly through the form of non-volatile memory, have changed the way data is being accessed. Non-volatile memory speeds up data access and pulls the storage closer to the processor. Currently, the NAND flash memory technology with persistent storage and an access speed that is much faster than disk drives, is a new form of data storage. Beyond Flash, there are other non-volatile memory storage technologies. One of them is phase change memory.

PC-RAM is touted as a fast, low-power, random-access, non-volatile and potentially low-cost memory technology. It uses a chalcogenide-based alloy that is formerly used in rewritable optical media such as compact discs (CDs)[53] and has several significant advantages over current NAND flash memory. Firstly, it is more robust than NAND – it has an average endurance of 100 million write cycles compared to 100 thousand for enterprise-class single-level cell (SLC) NAND.[54] This also translates to the possibility of creating simpler wear levelling algorithms[55] which will require less software and memory overheads. Secondly, PC RAM is byte-addressable and this means that it is much more efficient than NAND at accessing smaller chunks of data.

PC-RAM will likely gain a foothold as replacement for low-density NAND flash or serve as a bridge memory, filling the gap between NAND and DRAM. This technology will probably have the most impact on mobile phones and embedded applications where it has the potential to gradually cannibalise the market for flash memory during the next few years. The access speed and the non-volatile nature of this technology will be very useful when it comes to high-speed data analytics (e.g., for example, by enabling in-memory analytics or complex event processing).

4.4.2.6 Audio analytics

Audio analytics uses a technique commonly referred to as audio mining where large volumes of audio data are searched for specific audio characteristics. When applied in the area of speech recognition, audio analytics identifies spoken words in the audio and puts them in a search file. The two most common approaches to audio mining are text-based indexing and phoneme-based indexing:

Large-vocabulary continuous speech recognition: Large-vocabulary continuous speech recognition (LVCSR) converts speech to text and then uses a dictionary to understand what is being said. The dictionary typically contains up to several hundred thousand entries which include generic words as well as industry and company specific terms. Using the dictionary, the analytics engine processes the speech content of the audio to generate a searchable index file. The index file contains information about the words it understood in the audio data and can be quickly searched for key words and phrases to bring out the relevant conversations that contain the search terms.

Phonetic recognition: Phonetic recognition does not require any conversion from speech to text but instead works only with sounds. The analytics engine first analyses and identifies sounds in the audio

[53] Gartner, "Emerging Technology Analysis: The Future and Opportunities for Next-Generation Memory", 2011

[54] HPC Wire, "Researchers Challenge NAND Flash with Phase Change Memory", 2011

[55] A flash cell can only be written on a finite number of times before it fails or wears out. Wear levelling algorithms make sure that the write load is spread evenly across all the flash cells in the storage environment so that a situation where a few cells in the environment receive a majority of the write load and wear out before the other flash cells does not occur.

content to create a phonetic-based index. It then uses a dictionary of several dozen phonemes[56] to convert a search term to the correct phoneme string. The system then looks for the search terms in the index.

One difference between the phonetic recognition and LVCSR relates to which stage (indexing or searching) is the most computationally intensive. With phonetic recognition, the rate at which the audio content can be indexed is many times faster than with LVCSR techniques. During the search stage however, the computational burden is larger for phonetic search systems than for LVCSR approaches, where the search pass is typically a simple operation. Phonetic recognition does not require the use of complex language models as it can be run effectively without knowledge of which words were previously recognised. In contrast, knowledge of which words were previously recognised is vital for achieving good recognition accuracy in the LVCSR system. LVCSR approaches must therefore use sophisticated language models, leading to a much greater computation load at the indexing stage and resulting in significantly slower indexing speeds.

An advantage of the phonetic recognition approach is that an open vocabulary is maintained, which means that searches for personal or company names can be performed without the need to re-process the audio. With LVCSR systems, any word that was not known by the system at the time the speech was indexed can never be found. For example, a previously unknown term called "consumerisation" is now popular. This word is not already in the dictionary of words used by an LVCSR system and this means the analytics engine can never pick out this word in the audio file that was processed by the system. In order to find matches for this new word, the LVCSR system has to be updated with a new dictionary that contains the word "consumerisation" and all the audio has to be pre-processed again. This is a time-consuming process. This problem does not occur with phonetic audio mining systems because they work at the level of phonemes, not words. As long as a phonetic pronunciation for the word can be generated at search time, it will be able to find matches for the word with no re-processing of audio required.

Opportunities

Audio analytics used in the form of speech analytics enables organisations with a call centre to obtain market intelligence. The call records that a call centre maintains typically represents thousands of hours of the “voice of the customer” which represents customer insights that the organisations will want to extract and mine for market intelligence. Speech analytics can be applied across all the call records for a company to gather data from the broadest sources possible. More importantly, the data is gathered from actual customer interactions rather than recollections of the interactions at a later date. As a result, the data is more accurately placed in the actual context.

Government agencies and regulated organisations have to handle large amounts of recorded conversations for security and intelligence purposes. Manual transcription and analysis are not only slow; they risk either late or missed intelligence. Audio analytics can be applied for real-time audio streams and users can move straight to the specific point in the conversation recording where the word or phrase of interest is used, improving the monitoring processes.

Inhibitors

[56] Phonemes are the smallest unit of speech that distinguishes one utterance from another. For example, “ai”, “eigh” and “ey” are the long “a” phoneme. Each language has a finite set of phonemes and all words are sets of phonemes.

There are limitations to the accuracy and reliability of audio analytics. It is difficult to have 100% accuracy in the identification of audio streams. In the case of speech analytics, the system may not be able to handle accented words. Moreover, even if the analytics algorithms were sophisticated enough to identify words to a high level of accuracy, understanding the contextual meaning of the words would still be a challenge.

4.4.3 Five years or more

4.4.3.1 Quantum Computing

Modern computing may be considered powerful by today's standards. However, to increase the power of conventional computers, it is necessary to pack more transistors – the basic components of a computer – within a single computer. This is getting increasingly difficult. The maximum number of transistors that can possibly be packed will soon be reached and computers will have to be revolutionised to meet the computing demands of the future.

Quantum computing, a convergence between quantum physics and computers, represents this revolutionary form of computing where the data is represented by qubits. Unlike conventional computers which contain many small transistors that can either be turned on or off to represent 0 or 1 to represent data, each qubit in quantum computing can be in the state of 0 or 1, or a mixed state where it represents 0 and 1 at the same time. This is a property known as superposition in quantum mechanics and it provides quantum computers the ability to compute at a speed *exponentially* faster than the conventional computers. For certain problems like searching databases or factoring very large numbers – the basis of today's encryption techniques – quantum computers could produce an answer in days whereas the fastest conventional computer would take longer than 13.7 billion years of computation[57].

In quantum computing, data value held in a qubit has a strong correlation with other qubits even when they are physically separated. This phenomenon, known as entanglement, allows scientists to dictate the value of one qubic just by knowing the state of another qubic. Qubics are highly susceptible to the effects of noise from the external environment. In order to ensure accuracy in quantum computing, qubics must be linked and placed in an enclosed quantum environment, shielding them from noise. Areas which require huge computational power, e.g., security and digital image processing, will witness quantum computing transforming the speed at which these processes are carried out.

To date, quantum computing has not been demonstrated in a verifiable way as the technology is still in the early stage of research. However, quantum computing continues to attract significant funding and research is being carried out on both the hardware and algorithm design. Some significant advances made during recent years may pave the way to the development of a quantum computer.

[57] Kenneth Chang. I.B.M. Researchers Inch Toward Quantum Computing. [Online] Available from: http://www.nytimes.com/2012/02/28/technology/ibm-inch-closer-on-quantum-computer.html?_r=1 [Accessed 9th July 2012].

- Google has been using a quantum computing device created by D-Wave since 2009 to research on a highly efficient way to search for images based on an improved quantum algorithm discovered by researchers at MIT[58]

- IBM researchers built on a technique developed by Robert J Schoelkopf, a physics professor at Yale, and derived a qubit which lasted as long as one-10,000th of a second in 2012[57]. This helped to lengthen the time in which error correction algorithms can detect and fix mistakes. Otherwise, generating reliable results from quantum computing is impossible as the error rate is too high.

< 3 years	3-5 years	> 5 years
• Hadoop MapReduce and HDFS • NoSQL DBMS • Text Analytics • Visualisation-based data discovery tool • In-Memory Analytics • Predictive Analytics • SaaS-based Business Analytics • Master Data Management	• Data Federation • Audio Analytics • Video Analytics (consumer marketing) • Complex Event Processing • Mobile Business Analytics • Non-Volatile Memory: PC-RAM • Improved Analytics Algorithms	• Quantum Computing

Figure 21: Big Data technology radar

4.5 Technology Outlook

Big Data can be used to create value across sectors of the economy, bringing with it a wave of innovation and productivity gains. The discussion on the impact of Big Data focuses very much on the application of Big Data analytics rather than on the middleware or the infrastructure. Therefore, the adoption of Big Data technologies always comes from the analytics perspective which in turn drives the adoption of the underlying supporting technologies. According to McKinsey[4], there are five ways to leverage Big Data's potential to create value:

Creating transparency: Making Big Data more accessible to relevant stakeholders across disparate departments/units in a timely manner can create value by sharply reducing the data search and processing time.

Enabling experimentation: As organisations create and store more transactional data in digital form, they can collect more accurate and detailed performance data on many relevant aspects, such as product inventories and staff movements. This data can be used to analyse variability in performance, identify root causes and discover needs or opportunities.

[58] Paul Marks. Google demonstrates quantum computer image search. [Online] Available from: http://www.newscientist.com/article/dn18272-google-demonstrates-quantum-computer-image-search.html [Accessed 9th July 2012].

Segmenting populations: Organisations can leverage Big Data technologies to create highly specific customer segmentations and to customise products and services that meet those needs. Though this functionality may be well-known in the field of marketing and risk management, its use in other sectors, particularly in the public sector, is not common and to a certain extent, may be considered revolutionary.

Replacing/supporting human decision making: Sophisticated analytics with automated algorithms can unearth valuable insights that would otherwise remain hidden. These insights can then be used to minimise decision risks and improve decision making. In some cases, decisions may not be completely automated but only augmented by the analysis of huge data sets using Big Data techniques rather than small data sets and samples that individual decision makers can handle and understand.

Innovating new business models, products and services: Using emerging Big Data technologies, companies can enhance and create new products and services.

The application of Big Data varies across verticals because of the different challenges that bring about the different use cases. The common driver of Big Data analytics across the verticals is to create meaningful insights which translate to new economic value. Adoption of Big Data and analytics can be seen in the following verticals:

4.5.1 Healthcare

Big Data has a huge application in healthcare, particularly in areas where analysis of large data sets is a necessary pre-condition for creating value. Possible adoption of Big Data analytics could be done in a few specific areas. One of them is comparative effectiveness research (CER). CER is designed to inform healthcare decisions by providing evidence on the effectiveness, benefits and harm of different treatment options. The evidence is generated from research studies that compare drugs, medical devices, tests, surgeries, or ways to deliver healthcare[59]. By analysing large data sets that include patient genome characteristics, and the cost and outcomes of all related treatments, healthcare services can identify the most clinically effective and cost-effective treatments. However, before any analytical techniques can be used, comprehensive and consistent clinical and claims data sets must be captured, integrated and made available to researchers. In this area, issues of data standards and interoperability, as well as patient privacy, conflict with the provision of sufficiently detailed data to allow effective analysis.

The other application of Big Data analytics can be in the area of a Clinical Decision Support System (CDSS). A CDSS is a computer application that assists clinicians in improved decision making by providing evidence-based knowledge with respect to patient data[60] .Such systems analyse physician entries on patient data and compare them against medical guidelines to alert for potential errors such as adverse drug reactions, in the process reducing adverse reactions and resulting in lower treatment error rates that arise from clinical mistakes. The system can be extended to include the

[59] U.S. Department of Health & Human Services. What is Comparative Effectiveness Research. [Online] Available from: http://www.effectivehealthcare.ahrq.gov/index.cfm/what-is-comparative-effectiveness-research1/ [Accessed 9th July 2012].

[60] Runki Basu, Norm Archer, Basudeb Mukherjee. Intelligent decision support in healthcare. [Online] Available from: http://www.analytics-magazine.org/januaryfebruary-2012/507-intelligent-decision-support-in-healthcare [Accessed 9th July 2012].

mining of medical literature to create a medical expert database capable of suggesting treatment options to physicians based on patient records. The system can also be extended to include image analytics to analyse medical images such as X-rays and CT scans for pre-diagnosis.

Healthcare service quality can be improved using Big Data analytics. The US Centres for Medicare and Medicaid Services (CMS) announced a new data and information initiative to be administered by the Office of Information Products and Data Analytics (OIPDA). The initiative will guide the agency's evolution from a "fee-for-service" based payer to a "value-based purchaser of care" and aims to link payments to the quality and efficiency of care rather than the sheer volume of services[61]. To achieve this objective, there is a need to analyse data collected from service providers such as hospitals and physicians across the country to measure the service quality so as to establish and promote standards in the healthcare services sector. Big Data solutions may effectively meet the challenges faced by CMS and similar organisations around the world.

Finally, predictive analytics can be applied on patient profiles to identify individuals who are liable to contract a particular disease. Proactive treatments can be administered in an effort to prevent the illness or limit its severity so as to control healthcare costs. The ability to identify patients most in need of services has implications for improved treatment quality and financial savings.

4.5.2 Retail

The retail sector is built on an understanding of the consumers' retail habits. Top retailers are mining customer data and using Big Data technologies to help make decisions about their marketing campaigns, merchandising and supply chain management. Retailers are using more advanced methods in analysing the data they collect from multiple sales channels and interactions. The use of increasingly granular customer data gives retailers access to more detailed analytics insights which, in turn, can improve the effectiveness of their marketing and merchandising efforts. The clearer insights will also provide retailers greater accuracy in forecasting stock movements, thereby improving supply chain management.

Marketing

There are many ways which Big Data and analytics can be applied in retail marketing. One of these applications is to enable cross-selling which uses all the data that can be known about a customer, including the customer's demographics, purchase history and preferences, to increase the average purchase amount. Online retailer, Amazon, is a good example. With its "You may also like" recommendation engine, Amazon is able to provide product recommendations relevant to each of its customers, based on their browsing, shopping habits and other online profiles. According to the company, 30% of its sales are generated by its recommendation engine[62].

Another area would be for the purpose of analysing customers' in-store behaviour so as to help retailers create a more conducive shopping environment and increase the customers' propensity to purchase. A possibility here is to adopt video analytics to analyse customers' in-store traffic patterns and behaviour to help improve store layout, product mix and shelf positioning. Thirdly, analytics can

[61] U.S. Department of Health & Human Services. HHS hardness the power of health data to improve health. [Online] Available from: http://www.hhs.gov/news/press/2012pres/06/20120605a.htm [Accessed 1st October 2012]

[62] The Economist. Building with big data. [Online] Available from: http://www.economist.com/node/18741392 [Accessed 9th July 2012].

be used to harness consumer sentiments. Sentiment analysis leverages the streams of data generated by consumers on various social media platforms to help fine-tune a variety of marketing decisions. Finally, Big Data analytics helps retailers create micro-segments of their customer pool. Along with the increase in customer data, the increasing sophistication in analytics tools has enabled more granular customer profiling, helping retailers to provide more relevant, if not personalised, product and service offerings.

Merchandising

Retailers have to consider the optimisation of their product assortments. Deciding the right product mix to carry in the stores based on analysis of local demographics, buyer perception and purchasing habits can increase sales. Secondly, retailers can use the increasing granular pricing and sales data together with the analytics engine to optimise their pricing strategy. Complex demand and elasticity analytics models can be used to examine historical sales data to derive insights into product pricing at the stock keeping unit (SKU) level. Retailers can also use the data and consider future promotion events and identify causes that may drive sales.

Supply chain management

Inventory management is crucial in any successful retail strategy. With the details offered by Big Data analytics from the multiple huge data sets, retailers can have a clear picture of their stock movements and improve on their inventory management. A well-executed inventory management strategy allows retailers to maintain low stock levels because the suppliers respond more accurately to consumer demands. Retailers would want to optimise their stock levels so as to reduce inventory storage costs while minimising the lost sales due to merchandise stock-outs.

4.5.3 Education

In the education sector, learners are creating information at the same time as they are consuming knowledge. Students are faced with increasingly demanding curricula where they are no longer expected to regurgitate facts from hard memorising but are required to learn the subjects with deep understanding. At the same time, the onus is also on the educators as high expectations are placed on them to provide personalised teaching and mentoring. The challenge for the educators is to be able to have a clear profile of each of the students under their charge. Creating a profile for each of the students would require disparate sets of information and this is where the opportunity lies for Big Data analytics.

As more emphasis is being placed on learning that is adaptive, personal and flexible, there is the need to mine unstructured data such as student interactions and any form of student generated content. Learning analytics, such as Social Networks Adapting Pedagogical Practice (SNAPP)[63], can be deployed to analyse this data. SNAPP is a software tool that allows users to visualise the network of interactions resulting from discussion forum posts and replies. The visualisation provides an opportunity for teachers to rapidly identify patterns of user behaviour at any stage of course progression. This information provides quick identification of the levels of engagement and network

[63] Dawson, S. (2010). 'Seeing' the learning community: An exploration of the development of a resource for monitoring online student networking. British Journal of Educational Technology, 41 (5), 736-752.

density emerging from any online learning activities. From there, disengaged and low performing students can be identified.

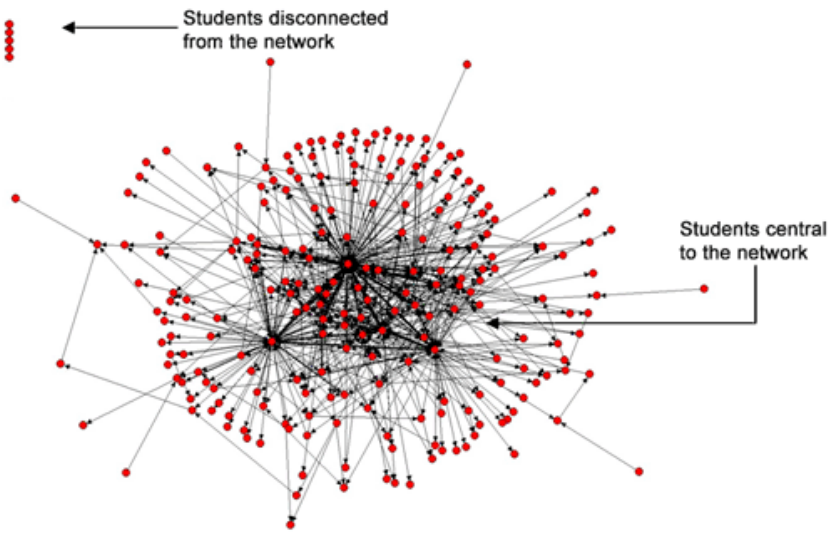


Figure 22: An example of a network visualisation diagram

Other opportunities for Big Data analytics can be exemplified by Civitas Learning, a digital education platform that uses predictive analytics to help guide educational decision making[64]. Civitas takes data such as demographic, behavioural and academic data provided by its partner institutions, and anonymises and combines them. The data is then analysed to identify trends and provide insights, such as identifying the courses and tracks that are most beneficial, based on the student profiles and the instructional approaches that tend to be most effective at ensuring good educational outcomes. These insights are than translated into real-time recommendations for students, instructors, and administrators through a customised platform.

4.5.4 Transport

For a land-scarce country like Singapore, there is a limit to the number of infrastructure developments that can be implemented to provide better public transport services. Big Data analytics offers the opportunity for public transport service operators to obtain critical insights on passenger demand trends so as to implement more effective measures in their service provisions.

Personalised real-time information for travel options

The prevalence of smartphones makes it possible for individuals to have information tools that identify and provide a variety of travel options, enabling commuters to make the best decision in terms of time and costs. By entering a destination, a commuter could be given the estimated time of arrival for different travel options that include buses, trains or private car transport. This will require

[64] Megan Garber. Can Better Data Keep Students From Dropping Out of College? [Online] Available from: http://www.theatlantic.com/technology/archive/2012/05/can-better-data-keep-students-from-dropping-out-of-college/257520/ [Accessed 9th July 2012].

feeding data from sensors placed on board the public transport vehicles as well as along the roads into computation models that predict traffic patterns and travel times at different times of the day. The model can also be built to take into account other factors, such as the likelihood of a crash on a particular roadway at a given time of day, weather conditions and even the anticipated fuel consumption and cost. These models can enable predictions of train and bus arrival times, and compare these mass transit approaches with different routes the commuter could take by car, and make recommendations of how to travel to a destination in the most effective way.

Real-time driver assistance

Analytics of the data from the sensor networks can provide information which could help drivers navigate the road. This includes incorporation of real-time road condition information to enable active routing. Drivers will always be advised in real time to use the least busy roads. As the sophistication of the algorithm increases, re-routing of vehicles will include the avoidance of road closures and roadworks. This approach can optimise road usage in a more balanced manner, thereby minimising congestion and decreasing the overall travel time without necessitating expensive new transport networks.

Scheduling of mass transit systems

Analysis of passenger boarding data, bus and train location data, and a series of sensor networks data could allow for accurate counts of the number of people currently using the systems and the number of people waiting at each stop. This information could be used to dynamically manage the vehicles, based on actual demand.

Preventive maintenance

Preventive maintenance is a major cost saver for organisations with significant infrastructure assets and regular maintenance schedules. A regular maintenance schedule is the basic requirement for any transport service provider to ensure a serviceable transport fleet (e.g., taxi or bus). However, despite the best maintenance effort, there will still be cases of vehicle break-downs which cause disruptions to the transport services. The maintenance effort is not universal across all vehicles in the transport service's fleets – certain vehicles would need more maintenance efforts compared to others. It is difficult to identify the specific vehicles that would need more maintenance efforts. However, by applying data analytics, transport service operators will be able to predict issues ahead of outages and improve client satisfaction. By capturing the log data from the vehicles onto an analytics platform, algorithms can be executed to report issues and forecast events related to each vehicle in the fleet. Preventive maintenance, rather than regular maintenance, is more effective in ensuring the transport fleet's serviceability.

Improved urban design

Part of the enhancements to the transport system includes improvements to urban design. Urban planners can more accurately plan for road and mass-transit construction and the mitigation of traffic congestion by collecting and analysing data on local traffic patterns and population densities. With the help of sensors and personal location data, urban developers can have access to information about peak and off-peak traffic hotspots, and volumes and patterns of transit use. The

information can help urban planners make more accurate decisions on the placing and sequencing of traffic lights, as well as the likely need for parking spaces.

4.5.5 Finance

Big Data plays a significant role in the finance sector, especially with regard to fraud detection with the application of Complex Event Processing (CEP)[65]. By relating seemingly unrelated events, CEP aims to give companies the ability to identify and anticipate opportunities and threats. CEP is typically done by aggregating data from distributed systems in real time and applying rules to discern patterns and trends that would otherwise go unnoticed. It is with the analysis of these huge data sets that fraud activities can be more easily detected – for example, unusual spending patterns such as buying French train tickets online from a US IP address minutes after paying for a restaurant bill in China. Nonetheless, this is only possible when IT systems of large financial institutions are able to automatically collect and process large volumes of data from an array of sources including Currency Transaction Reports (CTRs), Suspicious Activity Reports (SARs), Negotiable Instrument Logs (NILs), and Internet-based activity and transactions. It would be ideal to include the entire history of profile changes and transaction records to best determine the rate of risk for each of the accounts, customers, counter parties, and legal entities, at various levels of aggregation and hierarchy. While this was traditionally impossible due to constraints in processing power and cost of storage, HDFS has made it possible to incorporate all the detailed data points to calculate such risk profiles and have the results sent to the CEP engine to establish the basis for the risk model. A database management system will capture and store low latency and large volumes of data from various sources, as well as real-time data integration with the CEP engine, to enable automatic alerts and trigger business processes to take appropriate actions against potential fraud.

Companies in the finance sector also apply Big Data analytics to understand customer behaviour in order to increase customer intimacy and predict the right product to introduce at the appropriate time. This involves understanding customers and competitors, and using computational algorithms to make sense of the world. High-performance analytics can help to reach customers at precisely the right time and place, and with the right message, so banks can acquire and grow a profitable customer base. Many companies such as British Pearl, a technology-based financial solutions company, make use of unstructured data to find the best ways of attracting and retaining customers. Since it costs much more to acquire a customer than maintain one, another large financial services firm in the USA uses data from 17 million customers and 19 million daily transactions as an early warning system to detect customer disengagement[66]. Certain interactions and transactions trigger alerts to front-line staff who immediately contact the customer whenever there is an indication that the relationship needs to be nurtured. Not only do finance companies use their own data sets, they also work with partners operating in and out of the financial sector to get a far more comprehensive and accurate view of the market. By obtaining data about potential customers and their online presence, products and services can be tailored more accurately to specific individuals, and customers can be more effectively retained.

Working with various data sources can help financial institutions identify patterns and trends which measure people's likelihood to be fraudulent in order to reduce the bank's home loan lending risk.

---

[65] Oracle. Oracle Information Architecture: An Architect's Guide to Big Data. [Online] Available from: http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf [Accessed 9th July 2012].

[66] SAS. Banking on Analytics: How High-Performance Analytics Tackle Big Data Challenges in Banking. [Online] Available from: http://www.sas.com/resources/whitepaper/wp_42594.pdf [Accessed 9th July 2012].

Getting insights of consumers' major life events reduces the bank's risk exposure across the customers' lifecycles. To assess the credit worthiness and conduct risk evaluation, banks can apply high-performance analytical techniques to detect subtle anomaly throughout large loan portfolios. Banks can use these techniques to reduce the time needed to identify problem loans from more than 100 hours to a few minutes[66], resulting in significant savings.

4.5.6 Data privacy and regulations

The harvesting of large data sets and the use of analytics clearly involve data privacy and security concerns. The tasks of ensuring data security and protecting privacy become harder as the information can easily transcend geographical boundaries. Personal data such as health and financial data can help to bring about significant benefits such as helping to identify the right medical treatment or the appropriate financial services or products. Likewise individual shopping and location information can help to bring about better shopping experiences by informing customers of the products and services that are of greater relevance.

Organisations may have tried to use various methods of de-identification to distance data from the real identities and allow analysis to proceed while at the same time containing privacy concerns. However, researchers have proven that anonymised data can be re-identified and attributed to specific individuals[28,67]. These research outcomes disrupt the privacy policy landscape by undermining the faith that can be placed in anonymisation. The implications for government and businesses can be stark, given that de-identification has become a key component of numerous business models, more notably in the contexts of targeted marketing and health data. The challenge is for the individuals and organisations to be thoughtful and find a comfortable trade-off between data privacy and service utility.

4.5.7 Big Data related initiatives in Singapore

IDA's Internet of Knowledge Working ("IOK") Group

IDA has put forth a strong initiative in data and business analytics with its Internet of Knowledge ("IOK") effort. This integrated approach involves catalysing demand and seeding early adoption of analytics in key industry sectors, developing infocomm industry and manpower capabilities, establishing scalable and secure data exchange platforms, formulating the appropriate data policies, and developing "data hubs" that will deliver innovative data services and applications. Collectively, these will create a vibrant data and analytics ecosystem in Singapore and position Singapore strategically as an international Data & Analytics Hub. Enterprises can leverage Singapore's Data & Analytics capabilities to strategically apply analytics capabilities to guide business strategy, assist planning and optimise day-to-day business processes.

This initiative comprises three main thrusts:

Thrust 1: Developing Singapore's data and analytics capabilities to position Singapore as an international Data & Analytics Hub

[67] Paul Ohm. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. [Online] Available from: http://epic.org/privacy/reidentification/ohm_article.pdf [Accessed 9th July 2012].

The objective is to build up the local industry and manpower capabilities to serve the data analytics needs of enterprises and users around the world. From the industry development perspective, the plan is to anchor leading data and analytics vendors and users, and build platforms to focus regional demand in Singapore. This can be achieved by two measures. First, catalyse the development of innovative analytics and cloud computing products and services by anchoring intellectual property commercialisation activities into Singapore. Second, drive regional business development out of Singapore by building market channels and developing new business models to capture regional analytics users and partners.

The manpower development perspective focuses on three strategies. The first strategy is to enable professional conversion by collaborating with leading analytics players and Institutes of Higher Learning (IHLs) to enable conversion opportunities for professionals through workshops and on-the-job training. The second strategy is to create and deliver business analytics and cloud computing training capacities by collaborating with the Singapore Workforce Development Agency (WDA) and industry players to establish business analytics and cloud computing job roles and competency standards under the National Infocomm Competency Framework (NICF). The third strategy is to build business analytics and cloud computing talent for the industry by collaborating with universities and polytechnics to roll out business analytics and cloud computing diploma and undergraduate degree programmes.

Thrust 2: Drive sectoral economic competitiveness with the innovative application of data and analytics

IDA seeks to drive the adoption of data and analytics in key sectors of the economy to enhance the competitiveness of these sectors, and work with key players and sector champions to increase the level and sophistication of data and analytics adoption. There are two main initiatives under this thrust.

The first initiative is the development of the Business Analytics Shared Services for the retail and wholesale sectors. IDA and its industry partners will be investing a total of S$5.3 million to develop a suite of Business Analytics shared services for the retail and wholesale sectors. These services, to be deployed on shared infrastructure, will enhance enterprises' capabilities in Customer & Marketing Analytics, Inventory Optimisation and Operations Analytics, starting mid-2013. Through this initiative, IDA hopes to lower enterprises' cost of adoption, shorten implementation time and make analytics services available to the retail and wholesale sectors. Applied to the vast amount of transaction data that retailers have accumulated over time, these analytics services will allow them to derive greater insights into their customers' needs, improve service level, minimise inventory holding and increase overall operational efficiency. Besides the retail and wholesale sectors, IDA's initial efforts in data and analytics adoption will also cover other sectors such as healthcare, insurance, and food and beverage.

The second initiative is the Government Business Analytics Programme. The objective of the programme is to build public sector capabilities in analytics and drive the adoption of analytics by Government agencies. The programme will focus on implementing shared business analytics services for Government to implement analytics applications, developing public sector analytics capabilities, creating awareness of analytics, and promoting a data-driven and evidence-based decision making culture.

Thrust 3: Develop supporting platforms and enablers for data and analytics

The objective of this thrust is to establish the enablers that are required to support the development of Singapore's data and analytics capabilities, and drive productivity through enterprise adoption of data and analytics. One initiative under this thrust is the SaaS Enablement Programme. which provides funding for SaaS enablement projects in order to lower the barriers to SaaS adoption, expedite the SaaS enablement process and upgrade the capabilities of Independent Software Vendors (ISVs) in SaaS enablement.

Another initiative under this thrust is the proposal for a new regulatory framework on data protection to increase consumer trust and strengthen Singapore's position as a trusted global data hub. A data protection bill is set to be tabled in parliament by Q3 2012[68]. The proposed bill will strike a balance between the need of organisations to collect, use or disclose personal data for reasonable purposes and the right of individuals to have their personal data protected. The data protection bill will impose obligations on organisations to act responsibly in the collection, use and disclosure of an individual's personal data.

SMU Living Analytics Research Centre (LARC)[69]

With research grants of S$26 million from the National Research Foundation (NRF), the Singapore Management University (SMU) has set up a Living Analytics Research Centre (LARC) with Carnegie Mellon University. The Centre's objective is to advance the national effort of developing business analytics as a strategic growth area for Singapore, and develop new techniques to acquire and analyse data on consumer and social behaviour so as to develop applications and methods that will benefit consumers, businesses and society.

The Living Analytics research programme is unique in that it combines the key technologies of Big Data (large-scale data mining, statistical machine learning and computational tools for the analysis of dynamic social networks) with analytics focused on consumer behaviour and social media. LARC will substantially strengthen the ability to execute consumer and social insight trials that will benefit from any combination of the ability to observe how large numbers of users actually behave through automatic observation of their digital traces, and understand and predict the ways in which the preferences and behaviour of individuals influence those of their associated groups. This will contribute toward positioning Singapore as a hub for companies to analyse consumer insight data originating from multiple countries across Asia. LARC's analytics and experimental methods, technology platforms, project outputs, and working relationships with key regional and global consumer and lifestyle firms will create capability that can serve as an important resource to the Economic Development Board (EDB)'s Institute of Asian Consumer Insight.

data.gov.sg

The Singapore government has opened up public data as an operating principle of public services delivered by Government agencies as well as private providers. Significant social and economic

[68] Ellyne Phneah. S'pore data protection bill to be tabled by Q3. [Online] Available from: http://www.zdnet.com/spore-data-protection-bill-to-be-tabled-by-q3-2062304224/ [Accessed 9th July 2012].

[69] Living Analytics Research Centre. LARC and Singapore's National Analytics Initiative. [Online] Available from: http://www.larc.smu.edu.sg/larc-singapore-national-analytics-initiative.html [Accessed 9th July 2012].

benefits could be realised through access to aggregated, non-personal databases such as transport and hospital statistics, collected routinely by public services.

Many of these data sets are already provided through a common Web portal – data.gov.sg. Launched in June 2011, the data.gov.sg portal brings together over 5,000 datasets from 50 Government ministries and agencies. The portal aims to provide convenient access to publicly available data published by the Government, create value by catalysing application development, as well as facilitate analysis and research. Besides Government data and metadata, data.gov.sg also offers a listing of applications developed, using Government data, as well as a resource page for developers.

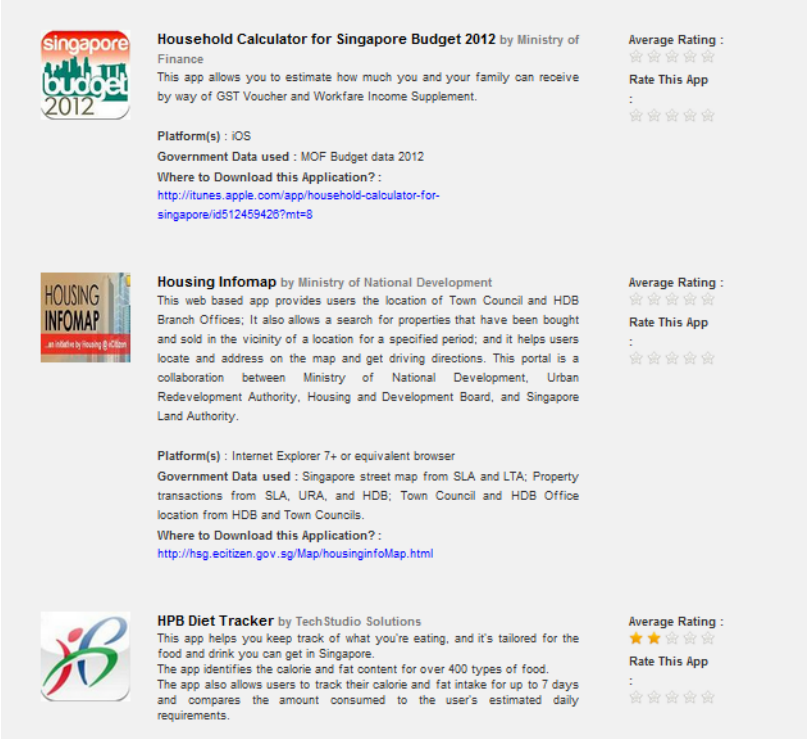


Figure 23: Listing of applications available on data.gov.sg

Local research efforts

Big Data has been a major focus for the local research institutes. In I^2R, the Data Analytics Department focuses on the effective mining of complex patterns from large data sets, both structured and unstructured. The department, which is one of Singapore's largest group of analytics researchers (with around 30 data scientists and 20 research engineers), develops machine learning methods for data mining, and focuses on teaching the computer to learn to identify complex patterns and make statistically intelligent decisions based on the data. Compared to traditional analytics methodologies such as statistical analysis and signal processing, machine learning technologies are devised to handle a diversity of data from numbers and text to graphs and Global Positioning System (GPS) trajectories. The department also develops text mining technologies to enable the computer to automatically extract information from large collections of documents and Web pages. For knowledge representation and annotation, the department develops semantic technologies and makes use of ontologies. The department has significant strength and innovation in privacy-preserving technologies for data collection, sharing and warehousing, so that privacy concerns are taken into consideration when doing data mining.

The Data Analytics department has received considerable recognition through international awards and winning international data mining benchmarking competitions. The department has active and successful industry collaborations in the following areas:

- Analysis of brand loyalty for customer relationship management;
- Intelligent data-driven diagnostics and prognostics on multi-sensor data for aircraft manufacture;
- Analysis of fluid dynamics in terms of the spatial and temporal evolution of vortices from very large-scale simulations of sensor data;
- Detection of fraudulent merchants by mining credit card transaction data;
- Leveraging of data mining to enhance financial modelling;
- Information extraction from large collections of news articles and reports;
- Patient profiling for individualised medical treatments;
- Identification of rising stars in social interaction networks;
- Social media analysis for sentiment detection;
- Mobile user analytics for understanding mobile user profiles and activities from phone-based sensors;
- Traffic prediction and discovery of supply and demand patterns from taxi GPS trajectories;
- Energy analytics for disaggregating smart meter data to understand appliance level usage without sensors/appliance monitors.

In the Institute of High Performance Computing (IHPC), the Computer Science department has related capabilities in cross-disciplinary, data-intensive analytics (CDA), distributed computing and high performance computing (HPC). This department adopts a multifaceted approach by integrating the capabilities within these three capability groups to tackle the large-scale data analytics challenges. The HPC group develops highly efficient relevant algorithms while the CDA group provides insights into the data and the Distributed Computing Group leverages the cloud infrastructure to deliver large-scale data analytics capabilities. Some of the Computing Science Department's R&D activities and achievements include:

1. A Computational Engineering Laboratory, jointly established with Rolls-Royce Singapore Pte Ltd, to conduct ongoing projects on using data-driven approaches (e.g., machine learning) to identify inter-relationships and patterns from large-scale simulation data, as well as analysing real-time machine data for effective condition monitoring.

2. The successfully application of analytics techniques to generate insights quickly and accurately to answer real domain problems - These large-scale diverse domain applications include:
    - Log analysis and fault detection at large supercomputing centres;
    - Performance tuning of supercomputers;
    - Epidemic modelling and simulation of infectious diseases;
    - Modelling the spread of dengue through correlation analysis between weather variables and dengue incidence;
    - Environmental modelling using climate data;
    - Traffic analysis based on the harmonisation of multi-domain data (road condition, traffic news, weather data).

3. The design and development of a workflow-enabled data computation platform for large-scale, data-intensive applications - This workflow management platform includes several

Web-based tools which allow end users to more effectively manage big, distributed data for data analytics workflow applications in hybrid clouds. This hybrid cloud model allows complex analytics workflow to be integrated seamlessly across private and public clouds. It also allows for dynamic resource allocation for better system utilisation. In addition, the workflow for complex applications can support multi-objective optimisation, where end users can decide what to deploy based on the cost (budget) or performance desired. The model is able to scale resources adaptively and efficiently with the integrative data management component to handle a wide variety of data, including both structured databases and a large number of data files. This platform has been successfully employed for several applications including weather data processing, traffic analysis, DNA synthesis, and optimisation algorithms (up to 56 times faster) for the logistics sector.

4. The design and development of a number of generic tools that automatically optimise, parallelise or map applications to execute efficiently on hybrid multicore and multi-Graphics Processing Units (GPUs) - For example, one of the tools developed by IHPC can enable the processing of large-scale data that comes in streaming form, such as sensor data or social network data. IHPC scientists have demonstrated that the applications derived from using the tool can achieve tremendous speed improvements of between 18 and 248 times – verified by a series of typical benchmarks – and outperform the current state-of-the-art applications. Another benefit of using this tool is that end users will need to write only in a higher-level programming language -- there is *no need to write GPU codes* in order to enjoy the performance enhancement. Besides tools, the team has also successfully developed efficient algorithms/methods:

    - Fast feature selection techniques (up to 130 times faster) for high-dimensional data;
    - Adaptive computing infrastructure for network-based simulation such as epidemic modelling (up to 21 times faster);
    - Acceleration of *key kernels* of diverse applications: cell migration (25 times faster), gene synthesis (6 times) and cancer diagnosis (725 times).

The Data Storage Institute (DSI) focuses on data storage technology research. One of its research thrusts is in non-volatile memory (NVM) technology, consisting of PC RAM, Memristor and Spin Transfer Torque Magnetic Random Access Memory (STT-MRAM). DSI collaborated with Micron to jointly conduct research in high-density STT-MRAM devices, a next generation non-volatile memory technology[70]. Another research thrust is in data centre technologies – storage system and architecture for future data centres. DSI is currently working on a Large Scale Object Storage System that can manage data in the order of billions of files, provide low latency, high bandwidth and has efficient search and retrieval functions. The clustered storage system, shown below, will be a tiered architecture consisting of a hybrid back-end storage medium based on next generation NVM technologies such as PCRAM, several solid state devices and hard disk drives (or magnetic media).

[70] PRWeb. Micron Collaborates with A*STAR on Next-generation Memory Technology. [Online] Available from: http://news.yahoo.com/micron-collaborates-star-next-generation-memory-technology-130234304.html [Accessed 9th July 2012].

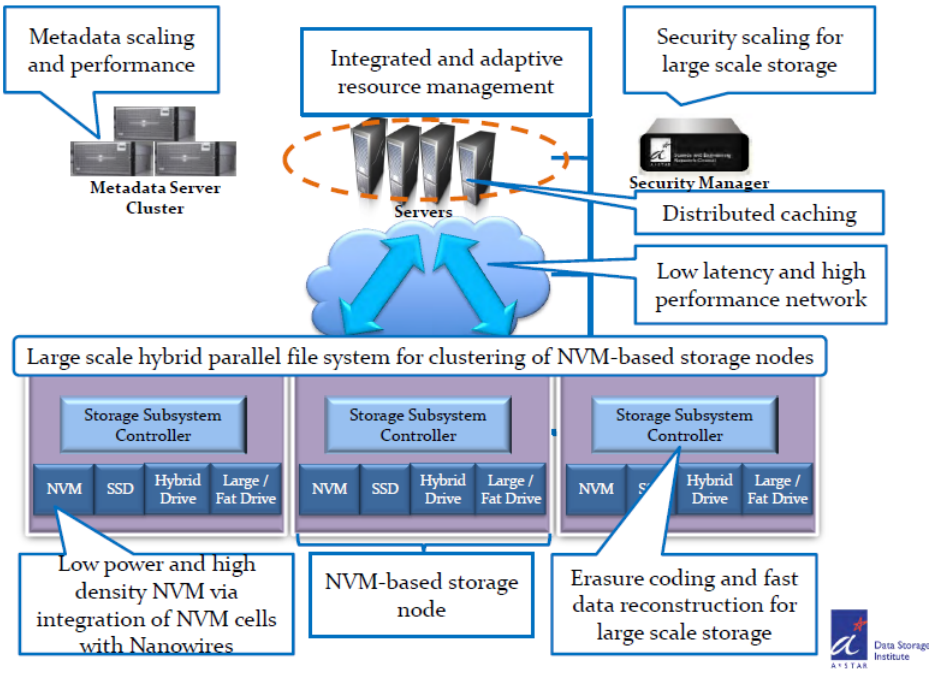


Figure 24: Storage Network Architecture for the Data Centre of the future

Finally, DSI has also developed a real-time, context-aware Big Data software that enables travel companies, hotels and meetings, incentives, conferencing, exhibitions (MICE) organisers to tap the growth potential of online travel bookings and compete for customers by harnessing the multitude of data exchanged with potential holidaymakers on the Internet[71]. A technology development study conducted by DSI researchers found that a significant amount of transactional data exchanges on travel booking and reservation information were duplicates. This is known as "look-to-book" in the industry, which is a ratio defining the number of "looks" or online searches that lead to cash paid or credit card confirmation for reservations. Based on this finding, DSI researchers designed the software to reduce the "look-to-book" ratio, which operates on in-memory compute grid technology, smart cache mechanisms and commodity servers. In the face of increasing business costs and demands on processing capabilities, this high performance product may help companies in various verticals reduce data volumes transacted by managing the influx of real-time data. Tapping on just two server nodes, the software is capable of processing real-time travel data to lower operating expenses, achieving 500 transactions per second in speed and only 200 milliseconds in latency. This also alleviates the need for expensive new equipment to handle more data.

[71] Data Storage Institute. Hoping to Get a Good Deal Online for your Next Holiday? A*STAR Data Storage Institute Unveils Smart Big Data Software for Tourism Sector. [Online] Available from: http://www.dsi.a-star.edu.sg/news-events/news-releases-speeches/Pages/HopingtoGetaGoodDealOnlineforyourNextHolidayASTARDataStorageInstituteUnveilsSmartBigDataSoftwareforTourismSector.aspx [Accessed 9th July 2012].