Next Generation Business Intelligence Software, Areas for Growth & Opportunities for Innovation

by

Yusuf Bashir

BSc. (Hons) Computer Science
University of Leeds, 1998

SUBMITTED TO THE MIT SLOAN SCHOOL OF MANAGEMENT IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN MANAGEMENT
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2011

Signature of Author:

MIT Sloan School of Management
May 6, 2011

Certified By:

Erik Brynjolfsson
Schussel Family Professor, MIT Sloan School of Management
Director, MIT Center for Digital Business
Thesis Supervisor

Accepted By:

Stephen Sacca
Program Director, Sloan Fellows Program in Innovation & Global Leadership
MIT Sloan School of Management

[Page intentionally left blank]

Next Generation Business Intelligence Software, Areas for Growth & Opportunities for Innovation

by

Yusuf Bashir

Submitted to the MIT Sloan School of Management on May 6, 2011
in partial fulfillment of the requirements for the degree of Master of
Science in Management

ABSTRACT

In today's world, as the volume of business and consumer data continues to grow at an unprecedented pace, there is increasing desire to utilize that data in new and innovative and ways to provide insight and improve decision making.

For businesses, data is being generated from transactions, machine logs, digital media and feeds from sensors and wireless devices at a volume and velocity not seen before. When combined with data from external sources such as partners, or from the Internet from blogs, social networking sites, YouTube, Facebook and Twitter, it has the capability to provide organizations with new insight, a more holistic picture of customer and stakeholder behavior and new ways of gaining competitive advantage.

Consumers are being presented with applications of increasing analytical sophistication, leading to growing comfort in making fact-based decisions. New devices will help monitor energy usage within the home and provide insight on the optimal times to schedule devices and run household appliances.

As data volumes continue to grow, systems will need to automate the uncovering of patterns and trends in data if they are to scale. Business Intelligence (BI) software, which has traditionally been used to gain insight from data, will need to evolve and new capabilities developed to support these significant changes. Areas of growth and opportunities for new innovation within the BI software industry will be explored that will enable stakeholders to take full advantage of this new and exciting opportunity.

Thesis Supervisor:	Erik Brynjolfsson
Title:	Schussel Family Professor, MIT Sloan School of Management Director, MIT Center for Digital Business

[Page intentionally left blank]

Table of Contents

[Page intentionally left blank]

Introduction

As data growth rates sky rocket, we are currently witnessing an "information explosion." According to analyst firm IDC the amount of data created globally by the end of 2011 is expected to exceed a whopping 2 zettabyes (or 1,750 exabytes[1]) and will continue along an exponential trajectory from a relatively paltry 250 exabytes back in 2007.[2]

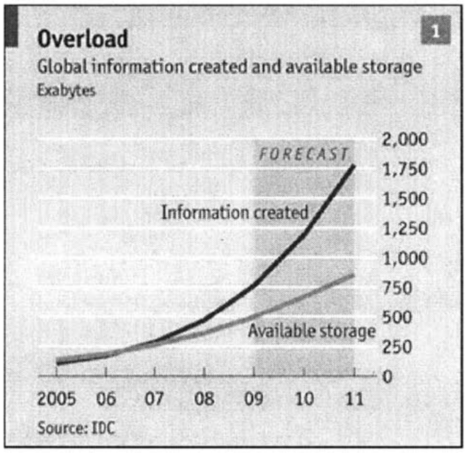


Figure 1: The growth of data, according to IDC[3]

The reason for such sustained growth lies in the greater number of systems generating data as well as increase in the granularity and detail of the data being generated and stored. Many organizations have moved beyond just using Information Technology to run mission-critical parts of their business to implementing new, non-critical systems to more effectively manage their customer relationships, sales activities, measure company performance and track employee development.

Machine-generated data has also grown tremendously. Cameras and sensors tracking everything from the assembly of products on a factory line, to the movement of components, finished goods, people, even vehicles are becoming more commonplace.

Contributing to this information explosion, individuals are also generating more data than ever before. Of the 1,200 exabytes of data generated in 2010, according to IDC approximately 300

[1] 1 exabyte (EB) = 1,000,000,000,000,000,000 B = 10^{18} bytes = 1 billion gigabytes = 1 million terabytes [Source: Wikipedia]

[2] The Digital Universe Decade – Are You Ready? IDC report, May 2010, John Gantz and David Reinsel, sponsored by EMC Corporation, http://idcdocserv.com/925

[3] Data, Data, Everywhere, The Economist special report, Feb 25 2010, http://www.economist.com/node/15557443

exabytes of it was User Generated Content, with consumers creating, capturing or replicating personal information.[4]

With the use of electronic forms of communication such as email, instant messaging and the increasing use of digital devices conversations and electronic files are being sent and stored more than ever before. Culturally, much of the world has adapted to digital forms of entertainment from their older, analogue versions. Television broadcasts, video, music and even radio have all converted. For example, in just a few decades cassette tapes and vinyl records have been replaced Compact Disks, in turn replaced by digital music files flexible enough to be bought and downloaded over the Internet and played on computers, portable music players and other devices.

With widespread adoption of digital cameras, photographs printed on paper have been replaced by digital files viewed on computers and portable devices. The average US household taking 390 digital photographs in 2009 with an average file size of 3 megabytes[5]. How the news is consumed and read has also dramatically changed from printed form to being viewed on a computer, mobile phone or the screen of a tablet device.

These new forms of information consumption have contributed to the tremendous growth of global information assets and based on the increasing adoption of digital devices and new technology is a trend set to continue. Managing all this data is spawning new technologies, innovations and approaches in data storage, management and processing capabilities.

The Growth of Business Data & the Need for Intelligence

Organizations today are drowning in data. Companies in every industry are not just storing information about this past business history but are finding new ways to collect more data from their operations. Retailers are collecting more granular Point-of-Sale information to better understand consumer behavior; retail banks are running cross-promotions between checking accounts and certain high-fee credit cards and telecommunications firms are running call behavior analysis to determine which customers are most likely to defect to a competitor. Information has gone from "scarce to superabundant"

With this information explosion we are witnessing a dramatic shift from data being seen as a natural by-product of the running of an organization to something considered as valuable as other corporate assets such as human capital, plant, machinery, equipment and intellectual property (I.P.).

Companies are increasingly willing to invest time, money and expertise in further exploiting their information assets for competitive advantage. This has resulted in the growth of several information management industries, from relational database software, analysis tools to the network infrastructure needed to transmit increasingly large volumes of data to bigger and more powerful computer hardware to store and process the data.

[4] The Digital Universe Decade – Are You Ready? IDC report, May 2010, John Gantz and David Reinsel, sponsored by EMC Corporation, http://idcdocserv.com/925

[5] http://pmanewsline.com/2010/11/15/generation-y-not-from-pma-magazine/

By reviewing data on their operations, organizations are able to view past performance through a quantifiable lens. What product lines performed well in stores, what plants are operating at maximum capacity and which employees attained their goals relative to plan? These are all questions than can now be definitely answered using historical data compared to previously when much of a company's performance was subject to interpretation.

As the business world grows increasingly competitive and complex, with greater desire to deliver differentiated service to customers whose expectations have risen dramatically, Information Technology (IT) is being seen as one of the primary ways to help businesses achieve these goals. In a globally dispersed business world, IT is enabling companies to operate in many countries as one entity, through integrated systems that help consolidate and streamline business processes, process customer orders quickly and provide a holistic view of their customers' interactions; creating what Jack Welch, former CEO of General Electric liked to call "the *boundaryless* organization".

As these IT systems proliferate, the information needed to run businesses efficiently is often locked in proprietary Enterprise Resource Planning (ERP) systems or a plethora of custom-built, add-on or customized applications with data stored in locked in different data formats, structures and file systems that cannot easily be extracted let alone integrated, creating gaps in the ability for business users to gain insight from the data in these systems, and the ability for IT to function efficiently. It is with the obscurity that comes from a myriad of information-based systems that resulted in a greater inability for decision makers within organizations to gain a holistic picture of their operations. With the proliferation of these systems came a desire to gain improved insight and this spurred the development of new tools, as well as innovative techniques and methodologies to obtain that insight – collectively known as Business Intelligence (BI).

What is BI?

According to the industry analyst firm Forrester, BI can be defined as:

> *"A set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making."*[6]

In practice, BI deployments consist of a four-layered conceptual architecture as illustrated in figure 2, consisting of:

1. Data sources
2. Data preparation and loading routines
3. Data storage, typically referred to as a data warehouse
4. Data analysis, typically done using a BI tool

[6] The Forrester Wave: Enterprise Business Intelligence Platforms, Q3 2008, Boris Evelson, Forrester, July 31 2008.

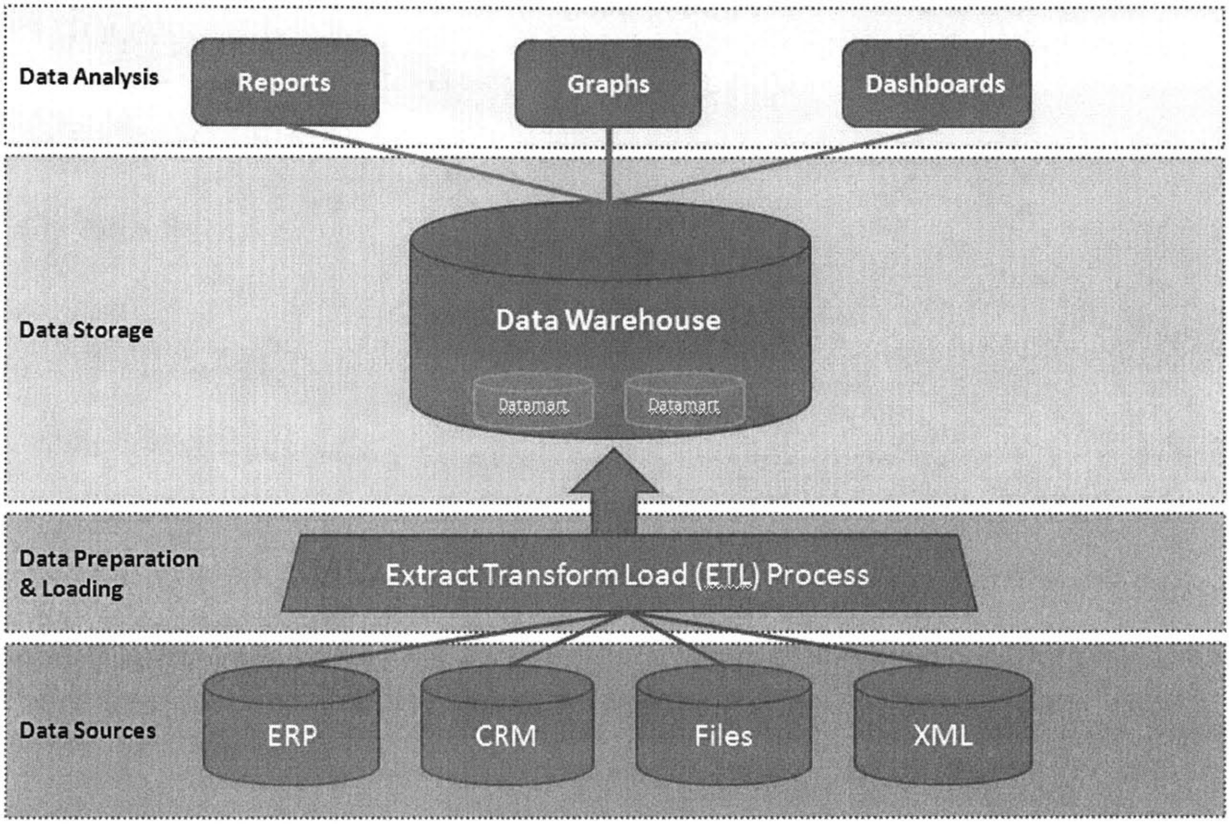


Figure 2: A typical BI architecture

In the data source layer, data can originate from any number of systems such as transactional systems, Customer Relationship Management systems, databases or files. It is almost always structured in nature, meaning that its form and composition follows a previously defined schema or metadata outline and it is stored typically in a relational database or other storage system that can be accessed using the SQL query language. The source data is typically extracted from these systems using SQL commands or purpose-built Extract, Transform and Load (ETL) tools that in addition to modifying the data and making it ready for analysis, can also schedule an ETL session, recover from data errors and keep a log of how data was moved – very useful for auditing purposes.

The data storage layer consists of one or more relational databases specifically designed to store data for analysis purposes. These databases are optimized for analysis through the use of a dimensional design where data is heavily denormalized and stored in separate tables based on its relationship to the key metrics being utilized. In contrast, transactional databases are designed for low data redundancy are highly normalized, and therefore would not store data in a dimensional schema. This can be a single enterprise-wide data warehouse or a number of smaller, departmental typically known as "data marts."

The data analysis layer consists of the BI software needed to analyze the data, build reports and distribute information to others within an organization. It includes business logic, also known a semantic logic, which acts as a translation layer between the physical data and the fields that end-users see on their screens. This layer also includes the BI application software that automatically creates SQL statements from users' commands, executes that SQL against the enterprise data warehouse and formats the results before sending it out to the users' interface of choice (e.g. web, iPad, iPhone, PDF, email, etc.)

Definitions of BI, CPM & EPM

While there are no unanimously agreed-upon definitions within the industry, analyst firms such as Gartner and Forrester Research often use the terms BI, Corporate Performance Management (CPM) and Enterprise Performance Management (EPM) synonymously, with the following subtle differences:

Corporate Performance Management (CPM) is typically used to refer to BI combined with pre-built analytical content (such as reports and dashboards) for a specific industry (e.g. retail, financial services, telecommunications, etc.) or for a specific horizontal specialization (e.g. Customer Relationship Management, Supply Chain Management, Human Resource Management, etc.)

Enterprise Performance Management (EPM) is typically used to refer to CPM combined with middleware software that is used to connect different applications along with the data integration software needed to physically move data and application logic such as metadata, over to a single reporting environment.

For simplicity, this study will define BI as software that provides the capability to analyze and report on data. As this study will clarify, it is redundant to make a distinction between business data and consumer data, as has been done in the past, due to the increasing importance of integrating consumer data into the decision making within an organization. The same data could be both "business" and "consumer" depending upon context. And it is the context of data that BI software is designed to help define.

According to a survey of 385 business technology professionals conducted by Intelligent Enterprise in 2008, the majority of respondents indicated that the use of BI has been successful in improving business performance:

Figure 3: *"How successful is your organization's use of BI in supporting improved business performance?"*[7]

Somewhat successful	56%
Very successful	19%
Less successful than expected	23%
Mostly a failure	2%

Many discussions of the BI industry often understate the importance of BI best practices, methodologies and related processes that are needed to effectively create and maintain a BI environment. For this reason, the Forrester definition for BI makes explicit reference to this. It could be argued that it is due to this limited understanding that leads to failure of many BI projects.

[7] Data obtained from InformationWeek Business Intelligence Survey of 385 business technology professionals that are using Business Intelligence tools.

Development & Growth of the BI Software Industry

With origins in the early days of Decision Support Systems (DSS) and Executive Information Systems (EIS), BI has become one of the largest (in terms of revenue) and most significant (in terms of numbers of deployments) of all the types of analytical software available today.[8]

While the definition of BI is broad, according to the analyst firm Gartner it includes enterprise reporting as well as analysis tools, but does not include similar analytical software segments such as financial budgeting, forecasting, predictive analytics or data mining[9]. In the future, as solutions become more integrated, it is possible that BI would be expanded to include these areas or be assumed into another category altogether. Only last year IBM, one of the major providers of BI software, renamed the group that develops and sells BI software to "Business Analytics." In addition, the combination of BI and financial budgeting/forecasting applications has become so commonplace that they are collectively referred to as Corporate Performance Management (CPM). In response, the #1 vendor in the space (by revenue), SAP BusinessObjects also renamed their BI business unit into "Business Analytics". It seems that this term is rapidly become an umbrella term to describe the various BI and analytics capabilities being offered in the market.

By using BI software, organizations are able to gain insight into their operations and learn more about areas that require attention. The content produced by BI tools is typically a report, containing the following analytical content delivered in either tabular or chart form:

- Quantitative measures such as Revenue or Cost, usually known as measures or metrics.
- Descriptive data elements such as Product Name, Region, State or City, known as dimensional attributes.
- Data filtering criteria, usually a dimensional attribute or combination of attributes, such as Time (e.g. This Month, This Quarter or Year-to-Date) or Region=West, known as filters.
- Advanced statistical analysis and measurement such as multi regression models.

In the past decade or so, dashboards have also become a common way of delivering analytical content especially to executive decision makers, for whom reports were often too detailed. Dashboards enabled users to track individual metrics over time, commonly known as Key Performance Indicators (KPIs). Perhaps the most well-known implementation of an analytical dashboard is the Balanced Scorecard developed by Harvard Business School Professor Dr. Robert Kaplan, who together with his HBS colleague, Professor David Norton, created a new visualization paradigm called strategy maps which illustrate the linkages between different metrics and their role in measuring a particular business process.[10]

In the mid-1990's a well-known retailer began to mine the Point-of-Sale (PoS) data they had been collecting to try and identify customer buying trends and discover new insight. While it was fairly

[8] The BI Verdict (previously known as the OLAP Report) http://www.bi-verdict.com/index.php?id=122

[9] Gartner http://www.gartner.com/it/summits/748720/Gartner_BI_research_note_142827.pdf

[10] The Balanced Scorecard Collaborative http://www.balancedscorecard.org/BSCResources/AbouttheBalancedScorecard/tabid/55/Default.aspx

common for the software tools of that time to allow decision makers to query data and get simple answers to straightforward questions, such as "how many blue shirts have we sold this month?" it was much harder to understand buying behavior beyond what was understood through common sense and in particular to learn something new for which the question was, until that point, completely unknown.

For this particular retailer, the investments made in data management infrastructure and querying tools suddenly (and famously) paid off when they discovered a correlation that no one would have predicted – a direct relationship between the sales of baby diapers and beer. Further investigation lead them to confirm that indeed, late in the day, many young men were coming into the store on their way home from work, and while picking up diapers would also pick up a six-pack of beer. In a few pilot stores, the retailer then brought these two products physically closer, and sales increased. They then separated out the products requiring these shoppers to pass by other aisles, and as a result sales of other products (such as chips and salsa) increased.

While certain products are clearly correlated (e.g. sales of gin often lead to sales of tonic water and sometimes limes) it was these unexpected nuggets of insight that fueled the imagination of many who began to see the value of data mining and analytics. This led many organizations (not just retailers) to start valuing the data they were generating, and start collecting as much of it as they could. And as time went by they realized the more granular this data was, the great the insight they were able to derive from it and the higher quality those insights would be.

The segment of the software industry that evolved from this was known as data warehousing, and later on was broadened to include analysis and reporting tools, and became known as Business Intelligence, or simply BI. As this story implies, BI has origins in early Executive Information Systems (EIS), also known as Decision Support, and now had become a multi-billion dollar segment of the software industry.

As data volumes increased exponentially, many organizations started to use these "data assets" in new and more creative ways to gain insight into their operations and use these new insights to improve effectiveness and profitability. BI systems thus became useful for many critical business tasks such as:

> *Measurement*: to provide a way of communicating organizational performance through Key Performance Indicators (KPIs) and metrics often displayed in a dashboard or scorecard.
>
> *Analytics*: to help deliver clearly-defined, quantifiable metrics and calculations based off transactional data that provide insight into corporate performance, rather than relying upon qualitative approaches or hunches based on gut-feeling or at best, unreliable data.
>
> *Segmentation*: to better understand the primary constituency being served by an organization (e.g. customers, suppliers, partners, etc.), how to target them more effectively with offerings that are likely to be successful.

Holistic View of the Customer: to gain a complete view of products, customers or other business dimensions and to understand all the different ways an organization influences these relationships or touch points.

Reporting: to provide historical data in table or chart form with the ability to filter out any irrelevant fields or data elements.

Distribution: to provide capabilities for the effective distribution of business insight by sending reports, dashboard or KPIs to users that need them through a wide variety of mediums such as web interfaces, wireless devices, text messages, facsimile or email.

It could be argued that insight being derived from these systems is more dependent upon the data being collected then the tools being used to analyze it. However, the reality is that both were needed. The data had to be (a) as granular as possible, because whenever new insights were found, decision makers would demand evidence and further detail, which for many organizations could only be addressed through the presentation of the actual transactional data, rather than the aggregates which were subject to interpretation and sometimes error due to miscalculation, and (b) reliable and of sufficiently high quality, which included minimal human intervention (thus reducing the chances of input errors). As the old adage stated, "*garbage in, garbage out*" - data of a sufficiently high quality was necessary as a prerequisite before any findings were to be trusted. Tools such as SQL (a structured querying language used by database programmers to answer queries of data) and Microsoft Excel could all be used to get insight from this data but what set BI tools apart was their ability to go beyond just reporting into the realm of analytics but answering more open ended questions such as "which products sold most in our top performing stores, and what other products did customers buy along with them?").

These more complex questions went beyond the capabilities of tools such as the SQL querying language or Microsoft Excel and helped users understand past behavior in unique ways, as well as provide predictive insight into what could happen in the future (e.g. "based on historical trends, which customers are most likely to defect in the next 30 days?").

While the value of BI has become much better understood and appreciated in recent years, many organizations have still been slow to adopt it. This provides opportunities for future growth for BI vendors and solution providers. In 2009, the analyst firm Gartner conducted a study into BI adoption and concluded:

> *"Because of lack of information, processes, and tools, through 2012, more than 35 percent of the top 5,000 global companies will regularly fail to make insightful decisions about significant changes in their business and markets."*[11]

Of the organizations that planned to adopt BI solutions, many of them viewed it as being strategically important to their overall corporate goals. In a recent survey of 2,000 Chief Information Officers, conducted by Gartner, the top priorities for CIOs in 2011 were found to be the following:

[11] Gartner Reveals Five Business Intelligence Predictions for 2009 and Beyond, http://www.gartner.com/it/page.jsp?id=856714

Figure 4: Top 10 Technology Priorities 2011[12]

01. Cloud computing

02. Virtualization

03. Mobile technologies

04. IT Management

05. Business Intelligence

06. Networking, voice and data communications

07. Enterprise applications

08. Collaboration technologies

09. Infrastructure

10. Web 2.0

BI Market Trends

From its early years in the early 1990's until more recently in 2006-7, the BI market was dominated by several specialist vendors, the largest of which by revenue were BusinessObjects, Cognos, Hyperion Solutions, SAS, MicroStrategy and Information Builders. These vendors were pure-play BI companies deriving almost all of their revenue from selling BI software licenses or related professional services.

This was an ideal environment for what the Delta Model framework, defined by MIT Professor Arnoldo Hax, would describe as "*customer bonding*", where small vendors treat every customer in a tailored and customized manner[13]. Each vendor in the BI industry was differentiated in one technical way or another, creating an environment where vendors rarely competed on price, but sought to serve customers in deeper and more innovative ways. Repeat business (i.e. selling into existing accounts) was common, while signing new customers (i.e. "new logos") was quite rare. Once customers were acquired, vendors worked extremely hard to keep them and maintain loyalty and customer satisfaction.

All of these vendors prided themselves in their independence from platform technologies such as operating systems, databases and the hardware that their solutions ran on. The platform technologies were required to be in place before they BI could be implemented. Due to their independence the BI vendors were able to partner with the larger platform vendors, such as

[12] Gartner EXP (January 2011) Gartner Executive Programs Survey of 2,000 CIO's, January 2011, http://www.gartner.com/it/page.jsp?id=1526414

[13] The Delta Model: Reinventing Your Business Strategy, Arnoldo Hax, Springer, Dec 14, 2009.

Microsoft, IBM, SAP and Oracle (collectively known by the acronym "MISO", named after the traditional Japanese seasoning) to provide integration between these environments and the BI software they were continually developing and improving. This resulted in close relationships being formed between the platform and BI vendors, with all parties focused on delivering a superior customer experience.

As BI software required these platform technologies to be in place before it could be used, it was critical that each BI vendor support an array of the most common platform products such as specific databases, operating systems and hardware environments from any of the MISO vendors. It was a nightmare for BI vendors to test and certify against any combination of platform technology, but one that was necessary to prove true platform independence and satisfy all types of customers.

BI customers valued this independence as it assured them that (a) their BI investments would not be effected by their choice of platform, in case they decided to switch from one MISO vendor to another, and (b) the BI technology that would become critical to their operations would not be sold to them with an ulterior motive, such as a way of up-selling or cross-selling more expensive database software or larger more powerful computer servers. The BI vendors had no hidden agenda.

According to the OLAP Report[14] (now known as the BI Verdict), a prominent BI industry report updated every year, the BI market was showing very healthy growth and this was benefiting all the major vendors. Technology was being constantly improved and the BI capabilities included in each vendors solution stack were broadening to include new capabilities such as financial planning and predictive analytics, all for the benefit of customers.

However, as the platform markets became more and more competitive, the mega vendors were continually looking for new ways to expand their solutions and increase revenue by offering customers a wider and more diverse selection of products. Their attention soon turned to complementary technologies such as application servers which powered many new Internet based businesses (which also required platform technologies to be in place), enterprise applications such as ERP (known as Enterprise Resource Planning, and was a category that included any type of software a company would use to run its operations) as well as BI which was benefiting from the growth in corporate data assets, most of which were stored on databases from one of the MISO vendors.

The Wave of Industry Consolidation and Its Effect on Innovation

No analysis of the BI market would be complete without an in-depth look at one of the most significant industry shifts that occurred, beginning in 2007 – the wave of vendor consolidation.

A useful tool in measuring the development of an industry is the Dynamic Model of Process and Product Innovation developed by MIT Professor James Utterback and William Abernathy from Harvard Business School[15]. The model provides a comprehensive method for examining an

[14] http://www.bi-verdict.com/

[15] Mastering the Dynamics of Innovation by J.M. Utterback, Boston, 1994

organization's capability for innovation. They argue that this capability is influenced primarily by environmental factors such as the market, competitive landscape and the expectation of customers. Their model assumes that an organization has minor influence on these environmental factors. Figure 5 illustrates the model, and the key concept - that an organization moves through three distinctly different phases over time; where initial high rates of *product* innovation get replaced by higher rates of *process* innovation.

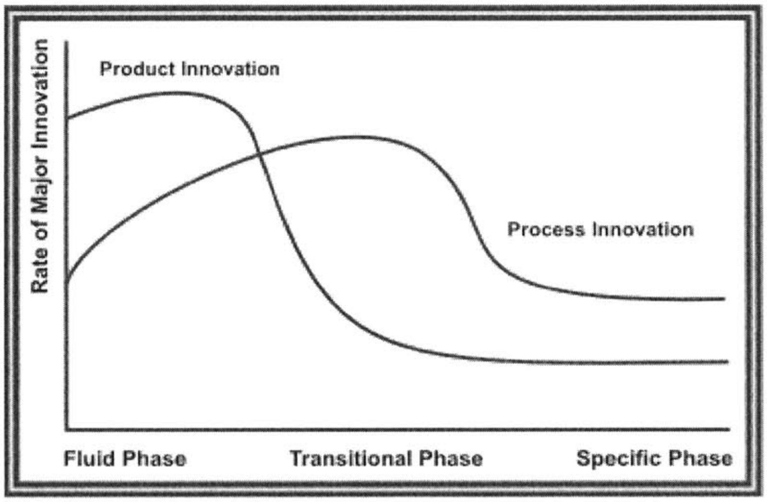


Figure 5: The Utterback Model of the technology life cycle[16]

Prior to 1999, the BI market was showing both healthy growth (~12% CAGR in North America and ~8.5% in Europe)[17] and an increasing number of vendors entering and competing in the market. 2001 saw a turning point in the market as the larger BI vendors began to acquire smaller BI companies in search of greater product differentiation, new customers, employee talent and new sources of innovation.

As 2007 began, a seismic shift occurred in the BI industry as enterprise software "mega-vendors" Microsoft, IBM, SAP and Oracle, commonly known by the euphemism "MISO soup," entered the market and began to look to BI as a way of differentiating themselves in a highly competitive (and growing) market for Enterprise Resource Planning (ERP) software.

The ERP software market by revenue was over 4 times larger than the BI market[18], and gave the MISO vendors sizable market presence as well as cash reserves to acquire even the largest BI

[16] Mastering the Dynamics of Innovation by J.M. Utterback, Boston, 1994

[17] Gartner http://www.gartner.com/press_releases/pr4feb2004.html

[18] AMR Research http://www.sap.com/usa/solutions/business-suite/erp/pdf/AMR_ERP_Market_Sizing_2006-2011.pdf

companies. For example, in 2007 the market capitalization of the largest ERP vendor was SAP (by market share, according to Gartner) at over $25 billion compared to the largest BI vendor, Business Objects at a paltry $3.9 billion.[19]

The wave of BI acquisitions by ERP vendors began in Q1 of 2007 with Oracle acquiring Hyperion Solutions, followed by Microsoft buying ProClarity, SAP acquiring the largest BI vendor BusinessObjects in Q1 2008, quickly followed by IBM acquiring the second largest BI vendor Cognos. By the end of Q1 2008, over two-thirds of the BI market lay in the hands of the MISO companies.

The wave of BI acquisition that then occurred in late 2007 and early 2008 was kicked off by Oracle, acquiring Hyperion Solutions. While Hyperion was not the largest BI vendor, it was a great fit for Oracle due to them already possessing a pure-play BI product from a previous acquisition (a product known as nQuire, which had been acquired by Siebel, which in turn was acquired by Oracle). At that point, many industry analysts and observers speculated at who would acquire the industry leader, BusinessObjects. SAP seemed distracted with their recent acquisition of Outlooksoft (a small financial planning company) but rumors were abound that the founders of BusinessObjects were willing to sell and would consider offers from any interested party.

During this time IBM, which had stated previously that it wouldn't enter the applications market was continuing along its application-agnostic strategy. Similar to the BI vendors, who were valued for their independence, since the early 1990's IBM had built a very healthy business around middleware, a software segment that included all the plumbing and connectivity technology that many large, diverse companies needed to integrate and maintain their IT systems. IBM was the largest vendor in this lucrative market and needed to maintain its relationships with all the BI vendors to ensure compatibility and interoperability.

In addition, in 2006 IBM acquired the consulting division of PricewaterhouseCoopers, and successfully built a thriving consulting practice called Business Consulting Services (BCS). In 2007, BCS had over 4000 consultants certified to install, configure and maintain BusinessObjects software (compared to only 300 certified on Cognos, and ever fewer for the other BI vendors) making IBM largest BusinessObjects consulting and implementation company in the world.

BI-related projects were a huge revenue earner for BCS and BusinessObjects projects were the biggest slice of that pie. If IBM were to acquire a BI vendor BCS would lose their neutrality in the face of clients, so vital for consultants, potentially destroying a very significant revenue stream.

It was rumored that during this time, IBM ran several simulations to determine the fallout of a potential loss of BusinessObjects-related business to BCS if the IBM Software Group decided to acquire BusinessObjects. It also seemed prudent for IBM to open up discussions with BusinessObjects and begin due diligence. But before IBM was able to find a foothold in these discussions, SAP announced their intention to buy BusinessObjects and they were willing to pay a 21% premium on a stock price that based on all the acquisition speculation, was already at a 52-week high.

[19] SeekingAlpha, http://seekingalpha.com/article/39638-is-business-objects-the-next-bi-buyout-target

While IBM BCS must have felt relief at the acquisition, after all SAP was already a significant BCS partner, IBM Software Group was worried. This new escalation would mean that customers would now start seeing BI being sold as part of a larger software solution built around Enterprise Resource Planning (ERP) software from vendors Oracle and SAP. With better-integrated solutions, IBM Software would stand to lose revenue from its middeware offerings which benefited from a fragmented software environment. In addition, IBM Software Group did not have a BI product as an independent alternative to the ERP vendors.

With BusinessObjects acquired, all eyes turned to the next largest vendor, Cognos. Rumors began to swirl that Cognos executives were in open discussions with Microsoft as well as discussing potential mergers with other smaller vendors such as Informatica and MicroStrategy. IBM was sitting on a large stockpile of cash, and investors began to pile on the pressure to use it to buy growth. With all the internal pressure from BCS not to acquire a BI vendor being drowned out by increasing desire from the Software Group, the new favorite child within IBM due to its spectacular revenue growth and superior margins, it was nevertheless still a surprise when the eventual announcement was made in November 2007 – IBM would be acquiring Cognos.

The BI vendor consolidation that was witnessed in 2006-07 has resulted in BI software becoming a feature of a larger solution, becoming more commonly known as "Business Analytics", and the BI component has as a result become increasingly commoditized.

While it may have been IBM's desire to see the BI vendors remain independent so that they could continue to grow their BI consulting practice (which prided itself on neutrality and derived a great deal of revenue from BusinessObjects implementations) and also grow their middleware software offerings (which benefited from customers desiring best-of-breed components from a smörgåsbord of software vendors) after the acquisition of BusinessObjects, IBM was left with no option but to acquire Cognos – the next largest BI vendor, to ensure that the IBM Software Group would not be at a disadvantage in future Business Analytics deals.

While Cognos has added significant bottom line revenue to the IBM Software Group and even allowed IBM to create and ship new more, integrated solutions combining hardware and software, this has come at a price. The Software Group itself has lost its position of neutrality and now customers are weary of buying IBM branded middleware for their environments if they do not use Cognos, for fear of IBM reducing R&D investments and support for non-IBM products in the future. And IBM BCS, which derived a great deal of revenue from BusinessObjects implementations is now under internal pressure to recommend Cognos to their clients, and risk harming their client relationships due to perceived lack of vendor neutrality.

Oracle acquires Hyperion for $3.3 billion in March 2007

Business Objects acquires Cartesis in April 2007

SAP acquires Outlooksoft Corp in June 2007

Cognos acquires Applix in Sept 2007

SAP acquires Business Objects for $6.5 billion in Q1 2008

IBM acquires Cognos for $5 billion in Q1 2008

Figure 6: Vendor Consolidation Timeline from 2007-08[20]

The BI Market Today

While the BI software industry today is dominated by the large MISO vendors, there are four main vendor types that offer a viable alternative to the MISO companies:

Independents

Perhaps the most significant category of non-MISO products, independent vendors include early BI pioneers SAS Institute, MicroStrategy, Actuate and Information Builders. All these companies have been in business for over 10 years and have a mature product line. SAS Institute, while privately held has a comprehensive product line that spans BI and related pre-built analytics, middleware and data integration capabilities. MicroStrategy was one of the first vendors to launch a web-based user interface for reporting and analysis back in 1999 and was also first with a plug-in for Microsoft Excel enabling users to seamlessly download BI content into their spreadsheets with a single click. Actute, after struggling for the past few years is finally finding its feet again, re-focusing the company on open source offerings built around its BIRT product line.

Open Source

The main vendors in the open source market for BI include Jaspersoft, a long-time pioneer in the industry now with a new management team and a revamped set of products, as well as Pentaho and Actuate (with their newer BIRT offering). While open source has its champions in the small and medium sized enterprise market, its still failing to establish itself as a viable alternative to commercially licensed products. Taking a leaf from Red Hat's business model, Jaspersoft and Pentaho have recently both launched "commercial open source" offerings that that extend the open source platform with proprietary features that companies have to license commercially, generating a similar license revenue stream to closed source vendors and ensuring that the open source projects can remain viable and commercially sustainable.

[20] BI Verdict http://www.bi-erdict.com/fileadmin/FreeAnalyses/consolidations.htm

Software-as-a-Service

BI vendors offering their products through the Software-as-a-Service delivery model are relatively new in the industry and while no dominant player has yet emerged many are finding it challenging to remain in business. All companies in this segment have taken venture capital and include PivotLink (formally Seatab Software), GoodData, Birst, Success Metrics, Oco and Panorama, with one notable failure, LucidEra which closed its doors after 4 years in business and raising $15.6MM of venture capital.[21]

In-House Systems

Still one of the largest challengers in individual sales cycles, enable companies to build software far more tailored to their business. Because of this the software can include specific features that may not exist in off-the-shelf products or be relevant to other companies or in other industries. Follow-up to any bugs or issues could be faster, depending on the internal dynamics within the company.

The option to build a solution in-house has lost its allure in the past decade as companies begin to realize the hidden costs of ongoing maintenance and management of the codebase and as vendors do a better job of illustrating the Return on Investment in buying off-the-shelf software.

Challenges to Broader Adoption

The wave of vendor consolidation that spread through the industry put BI firmly on the agenda of CIOs and executive managers of organizations of almost every size. But like other high-profile, IT projects, BI has also suffered from its fair share of negative press, overselling and visible project failures. While sometimes, external factors are to blame, many BI project failures occur due to back fit, poor consultative selling skills resulting in a poor technical fit and wider business process challenges hindering adoption.

In addition, due to the nature of BI being a decision support system rather than a mission-critical application, it is much harder for CIOs to quantify the investments made in BI. While it might be possible to calculate the results of an improved decision due to insight gleaned from BI, very few users would directly credit the system over their own decision making ability.

Due to the complexities involved in data extraction from source systems and the ETL process required to merge the data streams together, BI projects carry considerable risk of scope creep, shortage of skills and an underestimation of time required to build and load the data warehouse even before any analysis can be done.

Once the BI tool is deployed by IT and the data put in front of business users for the first time, business-specific issues are often uncovered which IT could not have foreseen leading to low rates of adoption and satisfaction.

And for even the most well-deployed and carefully crafted applications, poor data quality can destroy end-user confidence in the system and eventually kill usage altogether.

[21] http://searchbusinessanalytics.techtarget.com/news/1507062/SaaS-BI-vendor-LucidEra-to-shut-down

While the causes behind low rates of BI adoption are many, the main reasons can be summarized as follows:

- Lack of user adoption
- Complexity of integration
- Lack of sufficient data quality
- Users not willing to give up outdated spreadsheets
- Slow Implementation speed
- No integration with external data of systems
- Lack of involvement or sponsorship from the business
- Lack of political clout in decentralized IT teams requiring a company-wide effort
- Difficulty obtaining skilled talent
- Lack of sufficient budget, particularly for services and implementation work

Looking ahead, BI vendors will be continually challenged to address these concerns while around them the technology and data landscape is changing considerably. With the continued adoption of popular consumer websites and applications such as Facebook and Twitter, the expectations of the users of a superior product experience through clean, well-designed user interfaces and well-designed, useful features will continue to increase. BI vendors will have to continue to work harder to match and exceed these rising expectations.

Areas for Growth

The transition from Web 1.0 to Web 2.0 has had a profound effect on how users of the Internet viewed and interacted with information. From static, hyperlinked web pages, to the “social web” experience of Web 2.0 with user-generated content, hyper- personalization and social participation.

Similarly, we are currently witnessing a transition from Data 1.0, with its closed, silos of data, proprietary standards and poor integration across website or systems, to an era of what could be called Data 2.0, recognizable through its greater accessibility, open APIs and data platforms and third-party analytics. For many organizations, Data 2.0 will herald a new opportunity to gain external “eyes and ears” and better understand the needs, feelings and habit of their customers, as well as leverage these new data sources to gain insight that might lead to competitive advantage.

While Data 2.0 will have a profound effect on all information management functions and processes within an organization, BI as the primary interface into corporate data arguably has the most important role to play. Unless it evolves and adapts to support the new challenges that Data 2.0 brings, it will be made obsolete and replaced by newer enterprise software products that will.

According to Gartner, BI has been the fastest growing segment of the enterprise software industry with 8.1% CAGR through 2013. The primary drivers of this growth have been larger enterprise deployments and greater penetration into the small and medium sized business market.

There is little doubt that a significant factor behind the growth of existing deployments can be related to the growing data volumes that now exist within the company network as well as outside – in the cloud-based applications, in the supply chain, among channel and sales partners and from customer interactions. All this activity is driving organic growth that has been growing at a steady but not meteoric rate. This has arguably influenced vendors not to invest too heavily in R&D to fuel new innovation, but take a steady approach in-line with their peer group, impacting the rate at which the industry incumbents have been able to define and create innovative new capabilities and features.

It is not easy to identify opportunities for new innovation in any industry, let alone one as fast moving as Information Technology. However, as has been described in earlier sections of this study, some major trends are impacting the BI industry and are presenting vendors with exciting new opportunities to address these new challenges.

These trends include the ever-increasing volumes of enterprise and real-time data as well as increased granularity of that data due to the low cost of storage and improvements in capturing technology (e.g. better Point-of-Sale systems in retail stores, high-resolution sensors, RFID and other advancements). Based on these trends, vendors (both existing companies and possibly new startup ventures) could respond by expanding their offerings to source new forms of data, such as real-time data streams or build new offerings that cater for markets that have traditionally not used BI software before, such as home consumers.

Creativity will be the key. Data volumes are enabling the sandbox of opportunities to expand greatly, while new business models such as subscription pricing and pay-per-view are making previously unviable ideas now a sustainable possibility.

The Big Data Phenomenon

> *"Terms like "big data," "open data" and "linked data" are part of the new era in which the economics of data (not the economics of applications, software or hardware) will drive competitive advantage."*
>
> —David Newman, Gartner[22]

In the area of big data - a popular catch phrase being used to describe the rapidly growing stores of online and offline data, there are two main areas that are driving this trend:

1) Data being generated and collected from sensors, measurement/monitoring systems and new wireless devices and machine generated data such as web logs.
2) Data from the real-time internet, which includes sources such as Twitter, blogs and social networking web sites.

While the current obsession with big data may seem like a passing fad, the lessons learned in supporting such voluminous and fast-changing datasets will have huge ramifications to the way data is managed within corporations. In a recent Quora post many industry leaders and analysts revealed a breath of views but a consistently strong message: "*big data is going to underpin the next generation of web 2.0 and 3.0 applications*."[23]

Data is now proliferating both within the enterprise and in the cloud. As adoption of Software-as-a-Service applications such as Salesforce and Netsuite continues to grow the challenge of integrating data from across these applications together with data already within the enterprise will become even more challenging. For example, if executive managers need to gain insight into the Cost of Sale for a particular transaction, data from a Sales Force Automation (SFA) application, such as Salesforce would have to be integrated with financial data from and Enterprise Resource Planning (ERP) applications such as Quickbooks or Netsuite. The data from each system would have to be transformed to ensure that the granularity of the data matches up (e.g. transactions by day, hour or minute or sales rep by named account, city or state, etc.) and then integrated together to enable a final, combined report to be produced.

Many of these applications have an open API where the data can be accessed via another application rather than downloading a local version and then integrating the data by hand - a time consuming and error-prone task. Utilizing a separate, cloud-based integration tool such as SnapLogic

[22] How to Plan, Participate and Prosper in the Data Economy, 29 March 2011, http://my.gartner.com/portal/server.pt?open=512&objID=260&mode=2&PageID=3460702&resId=1610514&ref=QuickSearch&sthkw=big+data

[23] http://www.quora.com/Whats-the-next-big-innovation-in-web-interface-design

or Informatica Cloud would provide the necessary capabilities to perform the extract transform and integrate functions required to produce this type of report.

Sensor, Wireless Device & Machine-Generated Data

According to the IEEE a sensor, or transducer as it's more accurately known, is "a device for converting energy from one form to another for the purpose of measurement of a physical quantity or for information transfer."[24]

We are living at a time when these devices are collecting data at a volume and level of detail that is unprecedented. And as they continue to proliferate and take root quietly in almost every industry and physical setting, the data being generated is increasingly being seen as a valuable resource.

While the use of sensor technology is vast, for purposes of brevity this chapter will examine the use of sensors within the energy industry, one of the fastest growing and arguably most exciting applications of this technology. New energy management initiatives such as Smart Energy are being increasingly adopted through a combination of governmental pressure, increasing costs for suppliers and consumers, a greater concern for the environment among the general public, and the stark reality that fossil fuels are of finite supply. These factors are driving important innovations in this industry where there is no doubt that data and analytics will play a key role.

Smart Energy & Data from Home Appliances[25]

From site managers' optimizing their energy usage, to plant and line managers monitoring production lines for defects, to fleet managers being pro-actively alerted when vehicles need servicing, to logistical companies optimizing delivery routes in real-time, sensors and monitoring devices have become a critical part of operational efficiency of many organizations.

During a keynote speech to the GridWise Global Forum, the first major gathering of regulatory, governmental and commercial organizations interested in furthering smart energy and smart grid initiatives in 2010, the Chairman and CEO of IBM, Samuel Palmisano provided some significant examples of the growth of smart energy initiatives:

- Energy Australia has installed more than 14,000 new grid sensors that deliver cutting-edge monitoring and control capabilities for their 1.5 million homes and businesses.
- In an innovative program called Smart Meter Texas, Centerpoint, Oncor and American Electric Power have deployed more than 7 million advanced meters. These are enabling consumers to make more informed choices on energy use, enroll in energy supply contracts and take advantage of innovative new energy services.[26]

[24] http://www.its.bldrdoc.gov/fs-1037/dir-037/_5539.htm

[25] Information for this section was sourced from a telephone interview with John Lin, Co-Founder & CTO, Wireless Glue Networks, Berkeley, California, on Mar 31, 2011.

[26] http://www.ibm.com/smarterplanet/us/en/smart_grid/article/palmisano_gridwise_speech.html

The use of sensors and next-generation meters is not something restricted just to Smart Energy initiatives. For many years, sensors have been used on production lines to alert for defects, within dangerous environments monitoring for temperature changes and radiation, along freeways monitoring traffic flow and adjusting speed limits, and in many other industries. However, Smart Energy initiatives such as those described by Mr. Palmisano have the potential to make sensors a ubiquitous part of every home, being embedded in everything from appliances, security systems, thermostats, lawn sprayers, smoke detectors and light bulbs.

As initiatives such as these grow the data being generated, streamed, captured and stored is creating a huge new opportunity to provide analytically-based insight via a host of new applications and systems that are yet to be developed. These consumer applications will be challenged to present data in a meaningful way that could enable households to make decisions that save them significant amounts of money, keep them safe, and improve the quality of their lives.

For example, studies have shown that improved energy management of home electrical devices can reduce peak demand by 20% or more.[27] As the storage of electrical power is cost-prohibitive for utility companies, many of them charge their highest rates for power during times of highest demand. Using improved analytics delivered to consumers in their homes, that inform them of the cheapest times during the day to run appliances, consumers could make positive choices to help reduce peak time demand, reduce the need for new, high-cost power generation projects, reduce CO2 emissions, and save themselves money in the process. According to a report by the Brattle Group, even a 5% drop in peak demand in the United States would lead to approximately $35 billion in savings in power generation, transmission and distribution costs over the next 10 years.[28]

For a home device, or “gateway” such as this to be successful and adopted widely, it would have to be both unobtrusive and provided free or at little/no cost to the consumer. It is in the interests of the utility company to have consumers play a role in reducing peak demand and home gateways such as this would help educate the consumer on how they can use energy more efficiently and lower their bills.

While smart meters are steadily rolled out across the country and are providing more timely and accurate information to the utilities, they are still “dumb” and unable to perform any meaningful role in the management of home electricity usage other than conveying the amount consumed to the utility company. Instead the gateway could act as a bridge between their home devices and the grid, connecting and registering appliances in the home and tracking usage down to the device level. This gateway would provide the intelligence in the system – acting as monitor of usage, advisor on what appliances to run now and which ones to delay, and provide a far more accurate measure of energy usage down to a device-level of granularity. And in the future this gateway could provide arbitrage opportunities for the utilities through the direct control and scheduling of those devices (if consumers allow, of course).

[27] PG&E, http://www.pge.com/mybusiness/energysavingsrebates/demandresponse/peakdaypricing/

[28] The Power of Five Percent, Brattle Group report, http://sites.energetics.com/MADRI/pdfs/ArticleReport2441.pdf

For commercial customers, this gateway technology could also play an important role. Electrical energy usage is often capped by the utility company for each client and any excess is subject to overage charges. Staying within the quota limit is a key consideration when scheduling the usage of high-power equipment and in the future, an intelligent gateway device could become a key part of a the operational process, managing the energy input required any part of the manufacturing process.

Technology startup companies have begun to exploit opportunities in this space by developing gateways such as these that enable consumers to view electrical usage data in real-time. While usage statistics are a good first step, for a solution to be truly comprehensive, it would need to include the ability to manage down to the device level and be able to communicate that to the utility company. A full solution would include the following:

1. An intuitive user interface where consumers define and register the appliances and devices they have in their homes. Appliances and devices could also be grouped together into "virtual devices" such as air conditioning which could be a grouping of thermostat control and electronic window shades, enabling decisions to be made more easily by consumers (e.g. "put my AC system into maximum economy mode during weekdays")
2. A centrally managed resource database which would enable the hub to download information on the appliance that has been selected, including manufacturer and power consumption. Appliances vendors can compete to develop appliances that are promoted due to their low power consumption or Energy Star rating.
3. Smart socket plugs with the ability to register a specific socket with an appliance or device definition in the gateway, enabling the gateway to recognize and differentiate one device from another.
4. Communication with the smart meter and one or more utility companies, to optimally manage supply and monitor costs.

With this approach consumer will be able to view their usage data down to individual devices, and see exactly where their money is being spent. The utility company will be able to provide specific advice on device management that can save consumers money and lower peak demand. Finally, the gateway or hub vendor will be able to provide analytics on the stored data and provide that insight to policymakers, device manufacturers and advertisers.

Growth of Open Data

When President Obama took office in 2009, he took significant steps to open up government data to its citizens. Within the first few weeks of taking office, he signed the Federal Funding Accountability and Transparency Act (FFATA)[29] and appointed a Chief Information Officer, the country's first. The Transparency Act mandated that federal and governmental agencies provide access to their vast stores of information which led to the data.gov website, a centralized place in which government datasets were provided to the public.

In the same spirit as open source and creative commons software movements of the 1990's and 2000's, this new "open access" philosophy has not only grown but also emulated by a number of US

[29] https://www.fsrs.gov/

state and local governments, such as San Francisco, New York, Washington, DC and Chicago. In addition, the practice has been adopted by an increasing number of academic and scientific organizations, leading to formalized licensing structure and the first ever Science Commons conference in Washington, DC in 2006 to discuss open data and how micro-protection, particular in the biotech industry was stifling the progress of research in the field.[30]

The availability of these new datasets has spurned a whole new ecosystem of companies. Over 60 companies have created application on the 100 datasets made public by the City of San Francisco alone[31] the majority of which are related to mass transit.

While many of these datasets are available, critically many of them are not in a format that can be analyzed easily. While transactional data, such as closed contracts or breakdowns of city expenditures, is increasingly being made available online, in many cases they require cleansing, transforming and aggregating before they can be useful or insightful for the average taxpayer. This requirement for data manipulation is currently beyond the capability of most tools available to the average home users such as Microsoft Access or Excel and will require knowledge of data integration processes and tools, or time before existing BI tools evolve to have some basic data integration built-in.

Some of the most well-known open data initiatives and websites include:

US Federal Government	http://data.gov
World Bank	http://databank.worldbank.org/
Space-Time Research	http://www.spacetimeresearch.com/
UK Government	http://data.gov.uk
Kno.e.sis Weather Sensor Data	http://wiki.knoesis.org/index.php/SSW_Datasets
City of Ottawa	http://www.apps4ottawa.ca/
City of Portland	http://civicapps.org/
City of San Francisco	http://datasf.org/
State of Massachusetts	https://wiki.state.ma.us/confluence/display/data/Data+Catalog
District of Columbia	http://octo.dc.gov/DC/OCTO/Data
City of New York	http://nycbigapps.com/

Figure 7: Examples of prominent Open Data initiatives

There are several reasons why governments have begun to open up data to the public: it is perceived as politically astute especially during times of economic crisis to provide the general public with information about where their taxes are being spent. As was shown in the City of San Francisco, with 3 different Apple iPhone applications that were built by the for-profit, private sector companies for

[30] Heller, M. A.; Eisenberg, R. (May 1998). "Can Patents Deter Innovation? The Anticommons in Biomedical Research," *Science* 280 (5364): 5364 http://www.sciencemag.org/content/280/5364/698

[31] http://www.datasf.org/

sale on the Apple AppStore, to help people monitor bus and train times using the open data sets provided by local government.

This alleviates the pressure on governments at state, local and federal levels to analyze and interpret data and also allows useful insight to be delivered and disseminated through communities quicker and without bureaucratic hurdles. Lastly, by opening data and enabling a platform of useful, third-party applications to be built upon it allows the providers to become an even more critical part of the information ecosystem. While governments may look at this initially as a noteworthy way to serve their constituents it could potentially become a valuable source of revenue in the future.

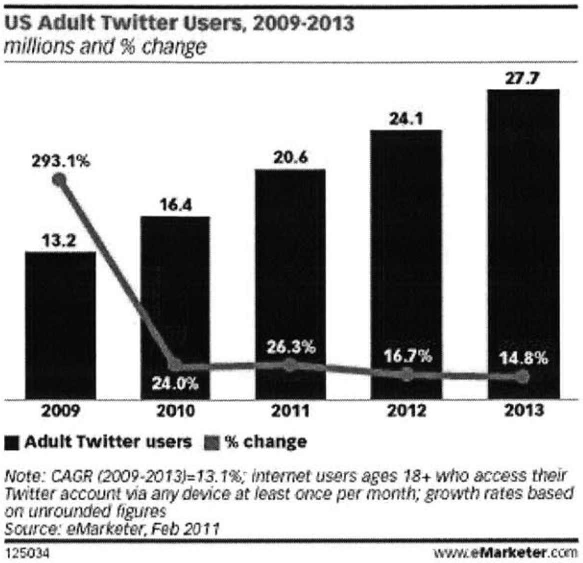
Twitter

The growing importance of the Twitter data stream was poignantly illustrated on April 14th, 2010 when the United States Library of Congress announced that it would be keeping a digital archive of every single tweet sent out since March 2006 – an average of 50 millions tweets a day!

While obtaining such a large slice of Twitter data by the US Library of Congress required special permission from executive management at Twitter, the company has publically stated that it will be opening up an Application Programming Interface (API) to the public Twitter stream to enable third-party applications, services and data brokers to access the Twitter stream programmatically. According to the company, this API, currently limited to just 150 API calls per hour, is already being used by over 50,000 applications[32]. Once the API is expanded to include unlimited API calls, programmers will have access to all Twitter data, enabling much broader analytical applications that can aggregate all Twitter streams (not just those connected to a particular user) to be built.

This data streaming API, affectionately known as "The Fire Hose" will become an even more critical part of the information ecosystem as the number of users on Twitter continues to increase. As figure 8 illustrates, according to eMarketer, the number of adult Twitter users rose from 13.2 million in 2009 to 20.6 million in 2011, and is expected to reach almost 30 million users by 2013.

[32] Twitter company blog at http://blog.twitter.com/2010/03/enabling-rush-of-innovation.html



Note: CAGR (2009-2013)=13.1%; internet users ages 18+ who access their Twitter account via any device at least once per month; growth rates based on unrounded figures
Source: eMarketer, Feb 2011

Figure 8: The Number of Adult Twitter Users Worldwide, 2009-2013

For enterprises, this API will enable them to quickly scan the *Twittersphere* to gain knowledge of trends and public sentiments that could be related to their products or services. For example, in response to broadcasting a commercial during a major sporting event, a company like General Motors could quickly measure the number of related Tweets that are sent out in the minutes following the commercial as a gauge of its success. The analytics possible using the Streaming API would enable General Motors to not just obtain a count of the times certain keywords are mentioned, but perform set analysis on that data and gain insight into the context or general sentiment of those tweets (i.e. were they positive or negative?)

As the market for sentiment analysis on social data grows, the traditional boundary separating enterprise data from the consumer internet is rapidly becoming blurred. Organizations are realizing that public data streams are a critical way of gauging reaction and sentiment among their target markets and those that want to provide high levels of customer satisfaction are even using these mediums, to engage and converse with customers.

In early 2010, the PR firm Wieden+Kennedy were retained by Procter & Gamble to try and revive sales of one of their fledgling brands, Old Spice. Quirky ads uploaded to YouTube starring a popular celebrity (Isaiah Mustafa) quickly gained notoriety and buzz. Mustafa then began to respond individually to fans via Twitter and even filmed personal video messages via YouTube[33]. The popularity and buzz surrounding the ads quickly exploded, with the Old Spice Twitter account

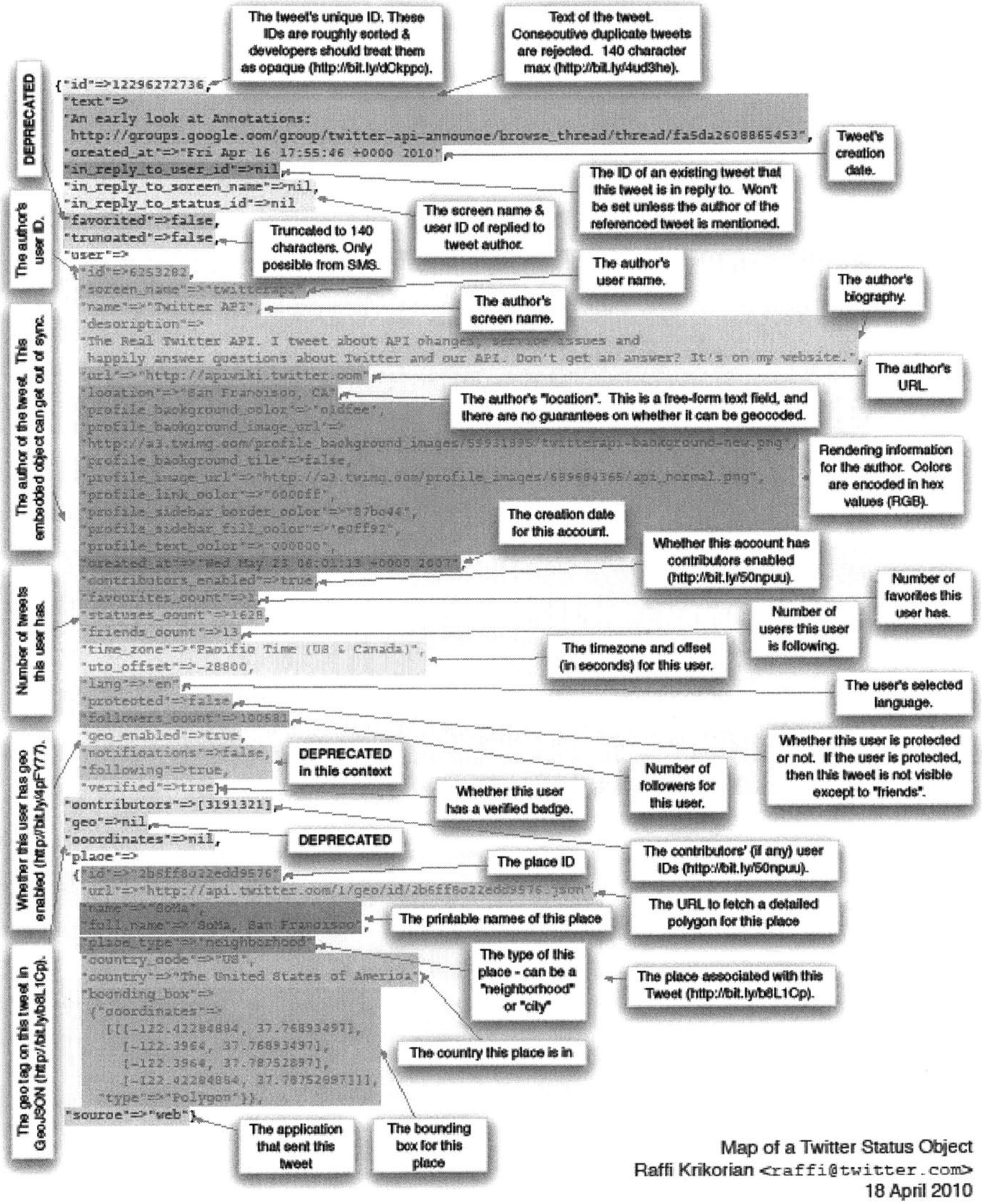
[33] Digital Buzz Blog at Digital Buzz Blog at http://www.digitalbuzzblog.com/old-spice-twitter-social-video-replies/

(@OldSpice) gaining hundreds of thousands of followers in just a few days and 186 personalized videos starring Isaiah uploaded to YouTube during the course of the campaign. [34]

While Procter & Gamble did not release sales figures as a result of the campaign, the viral success of the campaign illustrated the power and global reach of successful online campaigns that customers find engaging through the use of a social media such as Twitter and YouTube. 186 personalized videos were produced but over 180,000 people watched them and forwarded the links to others via Twitter and Facebook. Going forward, as more and more PR firms begin leveraging the power of Twitter in their campaigns, the use of the Streaming API will become even more critical to measuring the success of these campaigns and being able to fine-tune them quickly to respond to public sentiment, buzz and real-time events.

The amount of data being generated by Twitter is astounding. While every "tweet" (a message sent via the Twitter platform) is restricted to 140 characters, each tweet contains over 51 different fields, technically collectively known as a "Twitter Status Object." The breakdown of the Twitter Status Object is shown in figure 9. While the object looks overly verbose, it is to Twitters credit that much of this metadata around each tweet is already being leveraged by the Twitter ecosystem of add-on tools, products and services.

[34] YouTube http://www.youtube.com/user/OldSpice#p/c/60/xi-abeYlgxU



```
{"id"=>12296272736,
 "text"=>
 "An early look at Annotations:
  http://groups.google.com/group/twitter-api-announce/browse_thread/thread/fa5da2608865453",
 "created_at"=>"Fri Apr 16 17:55:46 +0000 2010",
 "in_reply_to_user_id"=>nil,
 "in_reply_to_screen_name"=>nil,
 "in_reply_to_status_id"=>nil
 "favorited"=>false,
 "truncated"=>false,
 "user"=>
  {"id"=>6253282,
   "screen_name"=>"twitterapi",
   "name"=>"Twitter API",
   "description"=>
   "The Real Twitter API. I tweet about API changes, service issues and
    happily answer questions about Twitter and our API. Don't get an answer? It's on my website.",
   "url"=>"http://apiwiki.twitter.com",
   "location"=>"San Francisco, CA",
   "profile_background_color"=>"c1dfee",
   "profile_background_image_url"=>
   "http://a3.twimg.com/profile_background_images/59931895/twitterapi-background-new.png",
   "profile_background_tile"=>false,
   "profile_image_url"=>"http://a3.twimg.com/profile_images/689684365/api_normal.png",
   "profile_link_color"=>"0000ff",
   "profile_sidebar_border_color"=>"87bc44",
   "profile_sidebar_fill_color"=>"e0ff92",
   "profile_text_color"=>"000000",
   "created_at"=>"Wed May 23 06:01:13 +0000 2007",
   "contributors_enabled"=>true,
   "favourites_count"=>1,
   "statuses_count"=>1628,
   "friends_count"=>13,
   "time_zone"=>"Pacific Time (US & Canada)",
   "utc_offset"=>-28800,
   "lang"=>"en",
   "protected"=>false,
   "followers_count"=>100581,
   "geo_enabled"=>true,
   "notifications"=>false,
   "following"=>true,
   "verified"=>true},
 "contributors"=>[3191321],
 "geo"=>nil,
 "coordinates"=>nil,
 "place"=>
  {"id"=>"2b6ff8c22edd9576",
   "url"=>"http://api.twitter.com/1/geo/id/2b6ff8c22edd9576.json",
   "name"=>"SoMa",
   "full_name"=>"SoMa, San Francisco",
   "place_type"=>"neighborhood",
   "country_code"=>"US",
   "country"=>"The United States of America",
   "bounding_box"=>
    {"coordinates"=>
      [[[-122.42284884, 37.76893497],
        [-122.3964, 37.76893497],
        [-122.3964, 37.78752897],
        [-122.42284884, 37.78752897]]],
     "type"=>"Polygon"}},
 "source"=>"web"}
```

DEPRECATED
The author's user ID.
The author of the tweet. This embedded object can get out of sync.
Number of tweets this user has.
Whether this user has geo enabled (http://bit.ly/4pFY77).
The geo tag on this tweet in GeoJSON (http://bit.ly/b8L1Cp).
The tweet's unique ID. These IDs are roughly sorted & developers should treat them as opaque (http://bit.ly/dCkppc).
Text of the tweet. Consecutive duplicate tweets are rejected. 140 character max (http://bit.ly/4ud3he).
Tweet's creation date.
The ID of an existing tweet that this tweet is in reply to. Won't be set unless the author of the referenced tweet is mentioned.
The screen name & user ID of replied to tweet author.
Truncated to 140 characters. Only possible from SMS.
The author's user name.
The author's screen name.
The author's biography.
The author's URL.
The author's "location". This is a free-form text field, and there are no guarantees on whether it can be geocoded.
Rendering information for the author. Colors are encoded in hex values (RGB).
The creation date for this account.
Whether this account has contributors enabled (http://bit.ly/50npuu).
Number of favorites this user has.
Number of users this user is following.
The timezone and offset (in seconds) for this user.
The user's selected language.
Whether this user is protected or not. If the user is protected, then this tweet is not visible except to "friends".
Number of followers for this user.
DEPRECATED in this context
Whether this user has a verified badge.
The contributors' (if any) user IDs (http://bit.ly/50npuu).
DEPRECATED
The place ID
The URL to fetch a detailed polygon for this place
The printable names of this place
The type of this place - can be a "neighborhood" or "city"
The place associated with this Tweet (http://bit.ly/b8L1Cp).
The country this place is in
The application that sent this tweet
The bounding box for this place

Map of a Twitter Status Object
Raffi Krikorian <raffi@twitter.com>
18 April 2010

Figure 9: Map of a Twitter Status Object by Raffi Krikorian, Twitter Senior Engineer

How Will The Growth of Big Data Be Handled?

> *"If you cannot measure it, you cannot control it. If you cannot control it, you cannot manage it. If you cannot manage it, you cannot improve it."*[35]
>
> —Dr. H. James Harrington

Two of the biggest challenges for BI systems when dealing with big data data sources are the volume and velocity of data that is being generated. As has been previously described, BI software evolved from the early days of SQL querying against structured data sources such as relational databases, where the data within the source database was fairly static and typically not updated more than once a day. Now with real-time data streams, data is being generated in greater detail and at a much faster rate to make offline storage to a database for traditional SQL querying simply unrealistic. Not only will the result set be outdated as soon as the SQL query has finished processing, but the source system itself (i.e. the relational database) will be locked up while performing inserts of the real-time data stream making querying, particularly using analytical functions, extremely difficult if not impossible.

For this reason several new approaches will be necessary to handle the volume and velocity of big data. One solution to dealing with this challenge is to create new ways of automatically detecting patterns and interconnectivity within the data based on some pre-specified, user-defined criteria. This pattern detection could occur on the real-time data stream itself, in the case of simple operations, or on a structured dataset where the real-time stream has been stored, for more complex operations.

Auto-Detecting Patterns, Trends and Interconnectivity in Data

As BI systems are built and rolled out within organizations one of the most challenging tasks for their IT departments is to know how to design the system so that it satisfies the needs of the business. Often when the system is rolled out it contains some pre-defined content to make the system immediately useful to the business as well as some level of self-service capability, ranging from the ability to create reports completely from scratch to applying simple modifications such as changing filtering criteria, such as time, product or geography, on the report data.

For the IT department it is often very challenging to build content that is useful and meaningful to the business as they are usually unfamiliar with existing business drivers, current challenges and how business performance should be measured. Without knowing the right questions, there is also no way of IT knowing whether the data being analyzed can even answer such questions.

[35] IEEE Instrument & Measurement Society, http://www.ieee-ims.org/main/index.php

Very often IT analysts, after meeting with stakeholders and potential future users on the business side, will try to take the requirements they have gathered and translate them into specifications for analytical content (i.e. a multi-dimensional data model design, reports and dashboards) that they then build and roll-out. The success of adoption critically depends upon the quality and relevance of these reports to the business which is ultimately driven by the effectiveness of this translation process.

Since the earliest days of BI software, vendors and practitioners have looked to develop ways that can more easily discover patterns and trends in data without requiring lengthy and time-consuming requirements gathering processes with cross-functional personnel. Often business users are simply left to explore the data through a trial-and-error approach where reports are built fairly randomly so that patterns and trends worthy of further investigation can be found.

Over the past several years, this requirement for improved and more guided data exploration has spurred several new startup companies some of whom have become successful and well established in the market including Qlik Technologies (NASDAQ: QLIK) and Tableau Software. Tools from both vendors help business users to quickly explore data through highly graphical user-interfaces and visually appealing charts and dashboards. While this is an improvement, users are still not given any idea which areas within the data should be explored first or are more likely to yield insights in a particular area. What is still lacking is the ability for these systems to auto-identify trends, patterns and correlations in the data based on a small set of criteria, without any analytic content having to be created, or without the need to obtain help from statisticians and predictive modeling experts.

Using Centrality to Identify Systemic Risk in the Global Financial System at the International Monetary Fund (IMF)

After the financial crisis of 2008-09 the IMF realized that they were caught by surprise at the speed and severity of the global financial crisis. In particular, they did not appreciate the extent to which some nations were exposed to high levels of risk due to their dependence upon incoming capital flows from countries that had governments, financial institutions or corporations that were perilously close to defaulting on their loans. For example, if sovereign debtors in Greece were to default what type of pressure would this place on French, Swiss and German lenders and their ability to continue to provide capital to Ireland, a country widely known to be over-exposed yet still very dependent upon these same lenders? Should this potential domino effect, force a change in the IMF's policy towards Greece? This analysis uncovered an urgent need within the IMF's Strategy & Policy Department to better understand systemic issues across the global financial system and have some way of highlighting risks and exposures related to interconnectedness between countries before it was too late.[36]

[36] Understanding Financial Interconnectedness, International Monetary Fund, Reza Moghadam and Jose Vinals, Oct 4, 2010.

Each of the 187 member countries of the IMF are mandated under Article IV of the IMF Mandate to return annual reports of their macroeconomic and financial performance[37]. One of the most critical metrics that is reported is the value of the capital flowing in and out of the country between lenders and borrowers.

During the month of January, 2011 the author conducted a 4-week project at the International Monetary Fund in Washington, DC. The objective of the project was to improve the way bilateral trade data was being presented to policy makers within the organization. While it was initially seen as a straightforward BI project, it was soon realized that the requirements of policy makers went far beyond what traditional BI tools could deliver.

During the project, annual bilateral trade data between all 187 member countries was obtained spanning back to 1962. For each year, a file of approximately 650,000 rows of data was obtained. Data from all 43 years was inputted, resulting in a total dataset of almost 28 million rows.

Astonishingly, the main way that desk economists and policy makers were analyzing this dataset was year-by-year in Microsoft Excel – a futile task because the connections between data elements were too complex and interrelated, with trends spanning across several years, to be uncovered by the human eye looking through tables and charts of yearly data.

In prior studies, centrality metrics had been successfully applied to social networks, such as by the US Federal Government agencies in identifying key protagonists within terrorist cells after the September 11th attacks in the United States. Within the finance industry, investigations had been made by the Bank of England around how network theory could be applied to financial linkages between banks and complex financial institutions.[38]

Familiarity with the success of these studies led the author to experiment in applying appropriate centrality metrics to this large bilateral trade dataset at the IMF.

With countries mapped out as nodes and the edge values representing the net capital flow between them, a network diagram was constructed allowing the application of centrality metrics to this representation of the global financial system. A sample of the dataset used to construct this network diagram is included in Appendix 1. For confidentiality reasons data from a few member countries was removed from the dataset (resulting in 179 countries being included in the analysis, rather than 187). The resulting diagram is illustrated in figure 10.

[37] The Fund's Mandate – An Overview, International Monetary Fund, http://www.imf.org/external/np/pp/eng/2010/012210a.pdf

[38] Network Models & Financial Stability, Erlend Nier, Jing Yang, Tanju Yorulmazer and Amadeo Alentorn, Bank of England Working Paper No. 346, April 2008.

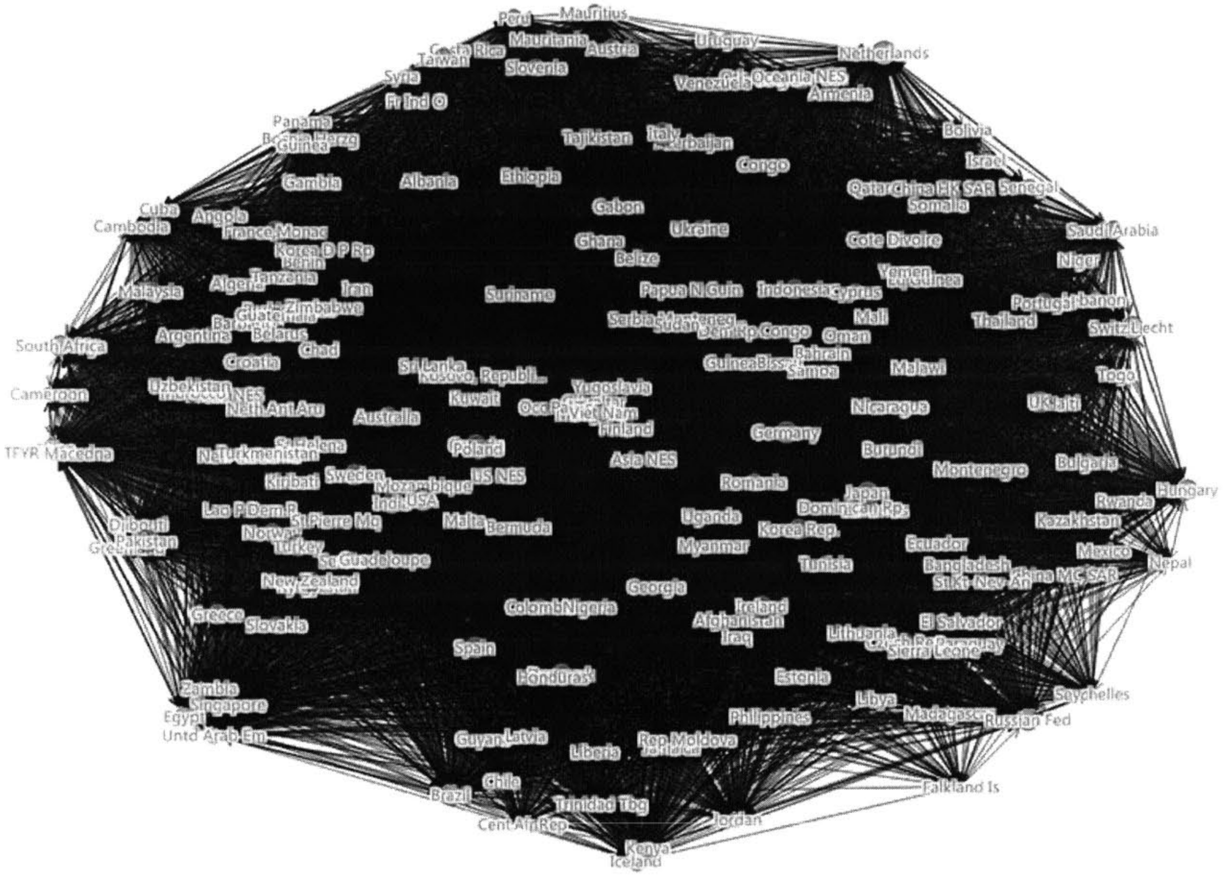


Figure 10: Network diagram of capital flows between 187 member countries of the IMF

Once the network diagram was built, an appropriate centrality metric had to be applied. There are four main centrality metrics:

(1) Degree – gives greater weight to nodes with the most connections to other nodes in the network. In a social network, degree centrality is a measure of social popularity.
(2) Closeness – as the name suggests, this metric gives greater weight to nodes that have the average shortest distance to all other nodes. In the study of the spread of infectious diseases for example, high degrees of closeness centrality indicate those likely to be exposed to infection more quickly than others.
(3) Betweeness – gives greater weight to nodes through which the largest number of interconnections between all others nodes in the network pass through. It is a measure of the level of influence exerted by a node in the overall network.
(4) Eigenvector – famously used by Google in its PageRank algorithm, this metric gives greater weight to nodes with the most connections to other nodes in the network that have the most connections. It is a measure of nodes that have the most influential connections.

Betweeness centrality was chosen as the best metric to highlight the areas of highest systemic risk (i.e. the countries with the greatest influence on the global flow of capital). The betweeness centrality calculation was executed across the complete network and then a filter applied to view just the 16 highest ranking nodes based on their betweeness centrality value. The threshold to obtain the top 16 happened to be all metric values above 140.

The resulting diagram illustrated in figure 11, showed the importance of small, European-based countries, such as Monaco and Lichtenstein, within the global financial network. Upon further investigation it was realized that they acted as important "pass-through" countries in the global flow of capital, rather than countries from which capital flows actually originated.

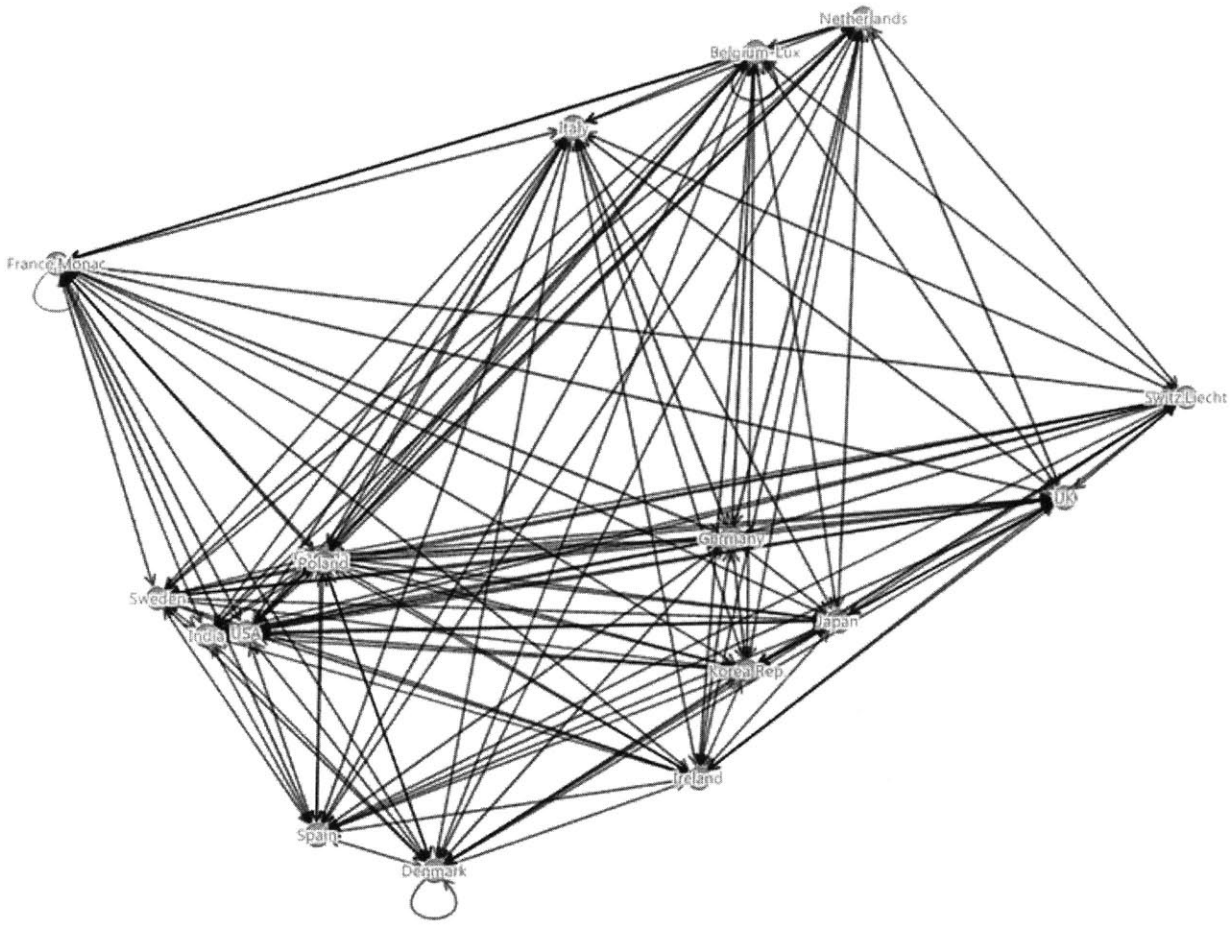


Figure 11: Betweeness centrality applied to the network and filtered to show only the top 16 nodes by metric value.

Unlike typical analytical methods used in BI tools which required a multi-dimensional model of the data to be built and then data elements aggregated together, by using centrality analysis, which analyzes a network by examining all nodes independently, this analysis also helped shed greater light on the strength of interconnectedness between a particular country of interest and the rest of the system much more easily.

To illustrate this capability, the UK which was found to be in the top 16 in the previous analysis, was selected as the country of interest and then relationships between it and all other countries were highlighted, as shown in figure 12. This illustrated nicely the global reach and influence of this dominant country in the global financial network.

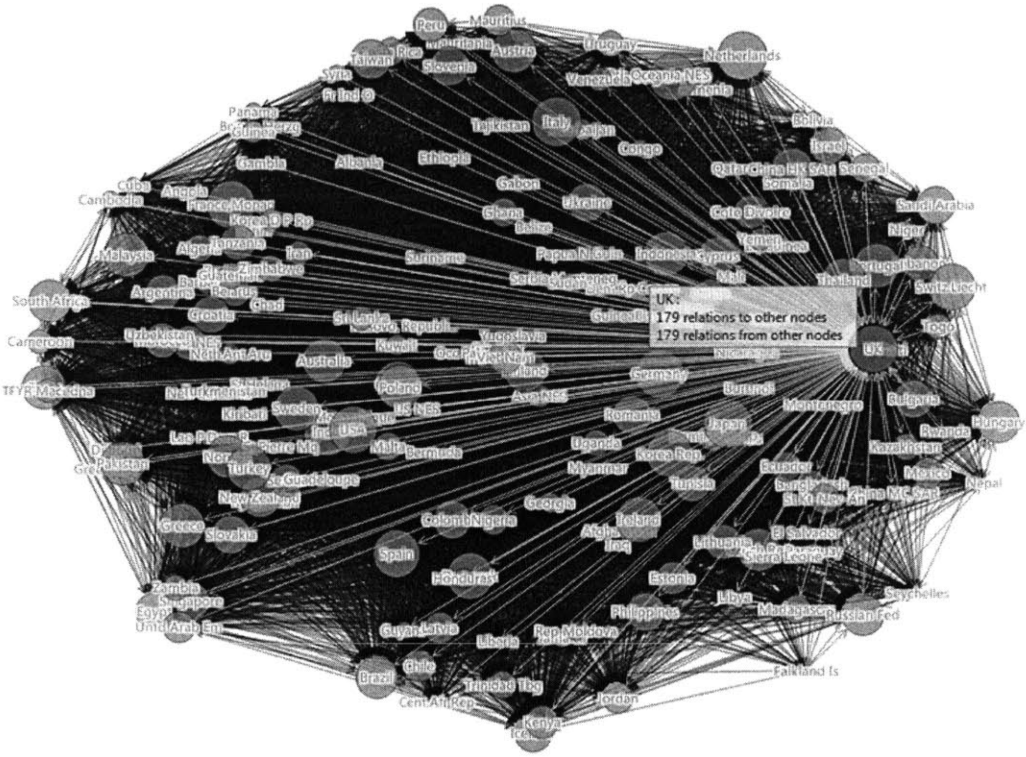


Figure 12: Network diagram of capital flows between the UK and the 179 other member countries of the IMF.

The nearest analogy to this capability in the BI world would be Market Basket Analysis in retail where the shopping baskets of customers are analyzed to determine which specific product in a basket drives the sale of most other products. Market Basket Analysis is only possible through the processing of highly granular retail data (i.e. transactions at the product SKU-level) using nested, dynamically generated, multi-pass SQL statements which currently only a handful of vendor products can perform.

To continue along this path of analysis and show the full flexibility of betweeness centrality just the nodes connected to the UK with an edge value greater than 250,00,000 (i.e. capital flows greater than $250MM) were selected, as shown in figure 13. In addition, the nodes that remained were sized according to their original betweeness centrality value within the overall global network. This diagram would help IMF desk economists working on new UK policy better understand the risk exposure that UK sovereign funds, financial institutions and corporations had with the rest of the world, and where in particular the UK's macroeconomic and financial policies should be directed.

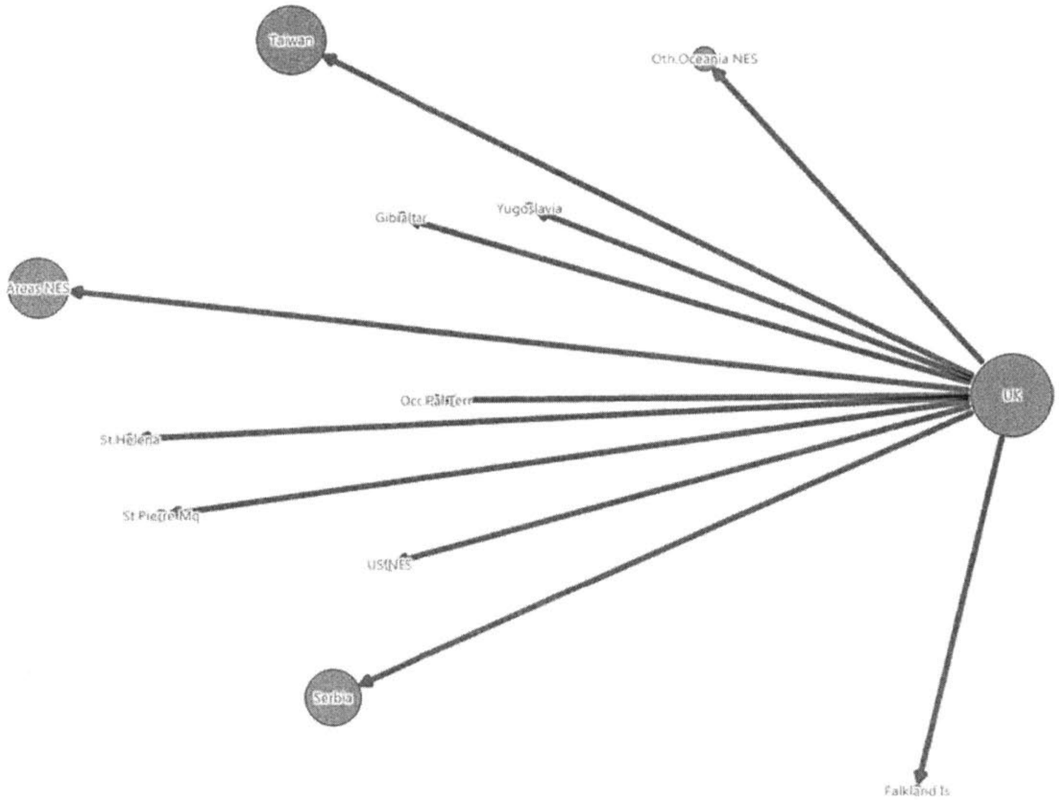


Figure 13: Network diagram of capital flows between the UK and other member countries of the IMF where capital flows are greater than $250,00,000. Size of the node is the betweeness value, where larger nodes correspond to countries with the greatest global influence.

Prior to this analysis, the Strategy & Policy Department at the IMF had not performed multi-year analysis of capital flows between all member countries at the same time due to the size and complexity of the data set. While this analysis was successful and popular among economists within the department, it required programming code and use of tools that would not be intuitive for most economists and policy makers. Instead, current BI vendors should explore ways to incorporate centrality metrics into their products and make this functionality available in an easy to use way.

Changes to the Database

Databases are evolving to satisfy the new requirements of the growing volumes of online data. Offerings from a new host of companies are being used by companies in many industries but particularly those generating large amounts of streaming data such as the media and technology industries. They are using these new databases to persist, or store the data being streamed so that it can be queried later for analysis or information retrieval purposes. For example, the BBC uses CouchBase to store all digital terabytes of high definition television content as it is being broadcast, so that it can be recalled easily later. In the future, the BBC hopes to enable analytical queries on the data, such as quickly retrieving any news content related to a particular person or event.[39]

While many of these new "noSQL" and "BigTable" databases are open source, other approaches include XML-based databases that are designed to store more complex data, such as unstructured content that may not be easy to define. For example, JetBlue Airlines uses MarkLogic to store all maintenance notes for fast and easy retrieval by engineering staff. These notes can include not just text, but images, video and even recorded speech. Information related to a particular incident or problem need to be stored together and tagged appropriately so that it can be retrieved quickly depending upon the context.[40]

Due to the very large volumes of data being stored, these new databases are typically deployed in a multi-node fashion, where the database system is spread across a number of machines, each running an independent database server and with data distributed across each node. This type of architecture ensures that massive datasets can be processed quickly and efficiently while not placing a particular node in a position of criticality. If any node fails, others can pick up its work.

Current examples of "noSQL" databases include:

- Marklogic
- MongoDB
- CouchBase
- AsterData
- Greenplum

It is unlikely that the need for local storage will ever go-away as in-memory techniques are by their very nature limited by the amount of available memory and the time the machine itself is powered on. During reboot or for redundancy purposes, a copy needs to be kept on local disk. In addition, while in-memory techniques can offer fast performance for smaller datasets, anything larger or more complex requires distribution across multiple machines, increasing overhead and latency. While this is exactly the problem that Hadoop was designed to solve, even in this scenario local persistence (i.e. storage of data into a database) helps tremendously in case any nodes go offline (the master can switch processing to another available machine) or if the master itself requires a restart, the disk copy can be easily reloaded into memory and the cluster continue processing where it left off.

Query Languages for Streaming Data

While SQL has become a de facto standard in querying relational structures such as databases, for real-time, streaming data SQL has severe limitations. The following table compares the requirements for querying relational databases versus real-time data streams:

[39] CouchBase case study on the BBC: http://www.couchbase.com/case-studies/bbc

[40] MarkLogic success story on JetBlue: http://www.marklogic.com/news-and-events/press-releases/2008/mark-logic-selected-to-power-jetblue-universitys-corporate-publications.html

Relational Database	Real-Time Data Stream
Data persisted	Data transient
One-time queries	Continuous queries
Random access to data	Sequential access to data
Well-described structure (schema)	Unpredictable form

Figure 14: Comparison of relational databases vs. real-time data streams when executing data queries[41]

Vendors such as Celequest (acquired by Cognos, which in-turn was acquired by IBM) tried to address the challenge of building analytics from real-time data sources, and did so using a proprietary system of trickle-feed databases and triggers. In 2000, the Stanford Data Stream Management project launched a more open standards-based querying language called CQL (Continuous Query Language). A CQL enabled system was able to create a loosely-defined schema for a data stream and then create queries, monitoring the data stream for the results.[42]

CQL could provide BI platforms with the capability to query real-time streams, build insightful analytics and set alerts on data objects. For example, a user may be interested in tracking several metrics pertinent to their job role. The user may have specified threshold criteria for each metric, or simply want to be alerted if data appears that contains fields used in the calculation of that metric. The system in return would track the real-time data stream and store in memory or in a relational database only those fields related to this metric. This would enable it to provide insightful and easily consumable snapshots of that data to the user at time intervals that are relevant.

Changes to the Application Layer

While this industry segment is still developing, many of the noSQL databases are designed for fast retrieval, and the requirement for more complex analytics is being satisfied by MapReduce techniques such as Hadoop. For this reason, almost all noSQL databases allow users to execute Hadoop-based MapReduce code within the database environment, and all have marketing and technology partnerships with Cloudera, a new vendor providing a commercial version of Hadoop.

Hadoop & MapReduce

Google first developed MapReduce as a way of processing very large datasets. Today, MapReduce is still core to how Google maintain indexes of sites on the Internet. As the name suggests, MapReduce consists of a two-step process. Firstly, the “map” step partitions large datasets and distributes them to servers running on other nodes. The outputs of this process are then “reduced” into a merged result set.

[41] The Stanford Data Stream Management System, Jennifer Widom, Stanford University, September 2000.

[42] http://ilpubs.stanford.edu:8090/758/

Today, MapReduce and its associated filesystem known as HFDS are available as an open source project called Hadoop which has some well-known users. Guy Harrison, writer for Database Trends and Applications compiled a few examples of how it is being used:

> *"Facebook now has more than 2 petabytes of data in Hadoop clusters, which form key components of its data warehousing solution.*
>
> *Yahoo has more than 25,000 nodes running Hadoop with data volumes of up to 1.5 petabytes. A recent Hadoop benchmark sorted 1 TB of data in just over 1 minute using 1,400 nodes.*
>
> *The well-known New York Times project that used the Amazon cloud to convert older newspaper images into PDF did so using Hadoop."*
>
> Guy Harrison, Database Trends and Applications, Sep 14 2009[43]

As data volumes continue to grow exponentially there is greater opportunity to utilize data for improved decision making. Open source initiatives such as Hadoop and Hive (analysis capabilities built on Hadoop) have been used successfully by Facebook and Yahoo to process large web logs and use the insights to radically improve user experiences. Within organizations there is a greater desire to understand customer sentiment, often requiring intelligently mining large volumes of unstructured, online blog, RSS, Twitter and Facebook data. In all these instances, the limited tools available all require hand coding and extensive software development. There is a pressing need for business-user tools that are intuitive and do not require coding; that can help decision makers more quickly analyze data and interact with it to identify actionable insight.

[43] http://www.dbta.com/Articles/Columns/Applications-Insight/MapReduce-for-Business-Intelligence-and-Analytics-56043.aspx

Opportunities for New Innovation

While the application of advanced analysis methods such as centrality to data haven't yet been fully explored nor crossed over from the realm of statistics into the mainstream of BI software, it's continued application within scenarios such as the one analyzed at the IMF will help it break eventually through. With growing data volumes, any automation around pattern detection or auto-generated insight would be much needed.

Similar to application layers, data should now be viewed at as a platform. With data originating from many systems it is through open standards and public API's that this data layer will be accessible by online applications just as easily as enterprise applications connect to their source databases.

According to the Utterback Model, which illustrates how markets mature and transition from product to process innovation, the BI market seemed to be following the typical path of Fluid to Transitional phases due to several reasons:

- Lack of new or meaningful innovation in the BI industry from the major, incumbent vendors.
- Shorter design cycles, and a dominant technology (i.e. BI became synonymous with reporting).
- Significant vendor consolidation (the larger BI vendors began to acquire the smaller ones).

However, these conclusions were shown to be premature. The characteristics shown by the BI market were altered permanently when the ERP companies entered the BI market and acquired the largest BI vendors. Far from proving that the BI market had passed through the Fluid phase, the removal of the major BI vendors created a gap in the market which was quickly filled by several emerging BI startups that began a new wave of significant BI innovation, led by inventions such as in-memory analytics, cloud-based delivery and advances in wireless and speech-based interfaces. It was the removal of the industry incumbents that enabled these significant new innovations to come to market.

BI Enhancing the Larger Decision Making Process within an Organization

Looking at how BI fits into the overall decision making process within a company, how it can be defined, captured, collaborated upon and measured still needs work. Is there an opportunity for closer integration with Business Process Management Tools? When actual outcomes and decisions (not just data) are archived and tracked over time can they be used to measure the effectiveness of the decision that was made? Answers to these questions can lead to developments where decision makers might one day be provided with post-mortem insight on the outcomes of prior decisions that they can use to tweak their current analysis.

BI Utilizing Data to Become a Information-as-a-Service Layer

Areas for new innovation do not have to limited to technology progress. How BI software is currently sold can also be improved. The current business models of perpetual licenses or Software-as-a-Service based subscription pricing, is starting to be seen as more and more outdated as corporate software users begin to compare what they get for hundreds of thousands of dollars versus the value they get for free with consumer websites like Facebook and Google. How these Web 2.0 companies make money is through indirect services, such as advertising or data arbitrage, where the data that gets built up over time through regular usage becomes extremely valuable. While in the B-to-B world advertising revenue doesn't make much sense, BI vendors especially those offering cloud-based BI services, could offer reports off the data they are hosting, acting as "data brokers" that no individual company could possible build alone. The healthcare industry for example, is dominated by large but fragmented healthcare systems/providers and payers. By offering each organization cloud-based reporting for free, the resulting data could provide population-level analytics that would be invaluable to the whole industry, particularly to epidemiologists and those working on healthcare policy.

BI Software Performing Scenario Analysis Using Data

Almost all of the new technology being developed to support big data is being focused in the areas of data storage (e.g. MogileFS, HDFS, Bigtable, etc) and data compute (e.g. Hadoop, Greenplum, Hive, Amazon Web Services) very little attention is being paid to the new front-end requirements for tools that users will be interacting with. While the data storage and compute innovations will enable much faster querying of massive datasets, users performing ad-hoc analysis to glean insight from data need new capabilities to maximize the benefit of having such a large and detailed dataset.

One of the challenges when making business decisions about the future is challenge around trying to predict how the market will respond to such behavior. Often companies perform simplistic scenario analysis (if any) to determine what would resonate most strongly with customers based on past behavior.

Due to the very nature of BI being a tool with which to perform analysis on historical data, BI has become synonymous with post-mortem analysis. Often, what an analyst is looking for within the tool is an area of interest that has already been identified as needing attention. But how to new problems get discovered? How can BI software be used for pre-mortem analysis?

If the vast datasets being collected, and the new techniques with which they can be queried and computed, can be better leveraged to perform scenario analysis, organizations looking to make a new decision will be able to more effectively test different hypotheses and even simulate what might happen using historical data to predict the future. Currently this type of predictive analytics has been relegated to the desk of statisticians and analysts with advanced degrees. If some simple scenario analysis could be provided to a business decision maker, through a BI tool that is easy to use yet connected to a live, massive dataset, it could become a very powerful next generation, decision support system.

BI Software That Can Auto-Detect Patterns in Data

During this project several, non-BI software applications that support centrality metrics were examined. Examples include Pajek (the Slovene word for spider) available free for non-commercial use yet lacking a usable user-interface and not suitable for business users, and Tibco Spotfire Network Analytics, a much nicer, commercially-available data mining product that comes with an appropriately large price tag.

Several products have also been developed in academia, such as Comment Flow from the MIT Media Lab (illustrated in figure 15) and SemaSpace from Ludwig Boltzmann Institute for Media, Art & Research in Austria (illustrated in figure 16).

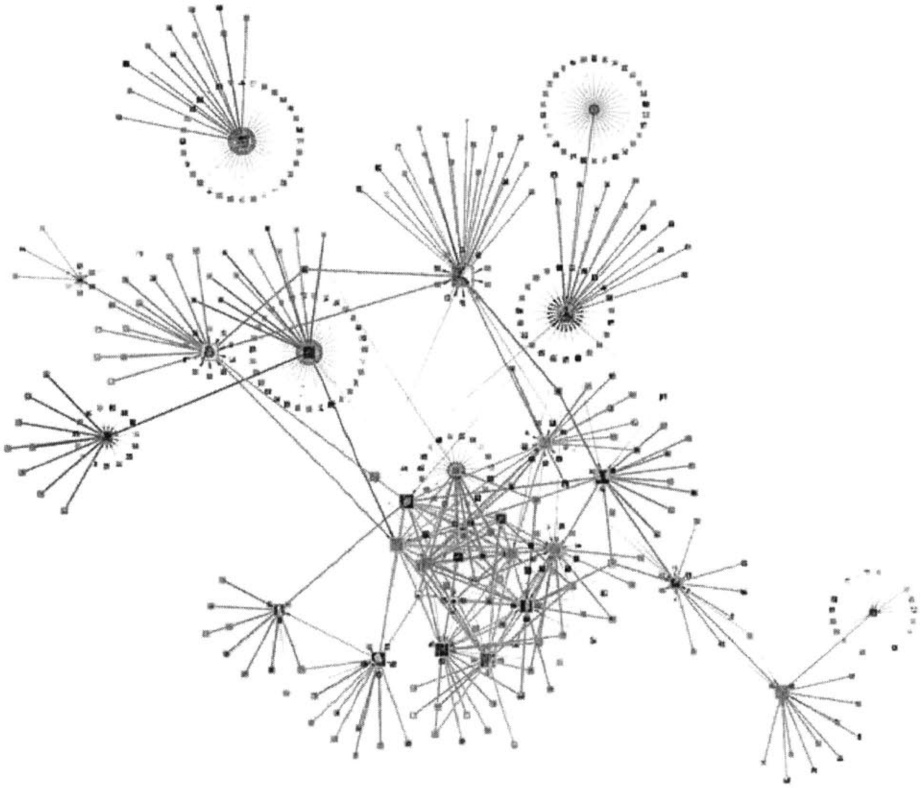


Figure 15: Comment Flow, developed at the Media Lab at MIT.

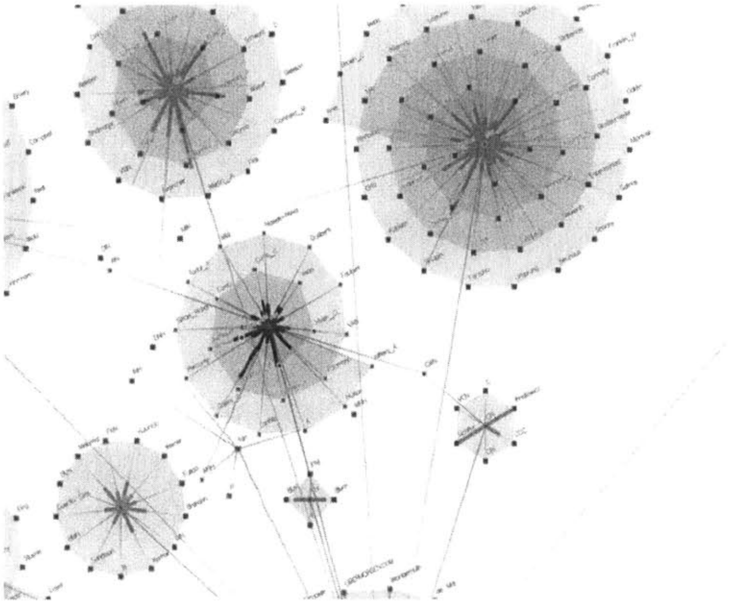


Figure 16: SemaSpace, developed at the Ludwig Boltzmann Institute in Austria.

The Growing Importance of the Software-as-a-Service Delivery Model[44]

As the SaaS industry grew and reached maturity with vendors such as Salesforce and Netsuite gaining broad acceptance and market share, several startup companies tried to address the latent need for BI in the cloud. While some of these vendors, such as Cloud9 Analytics and LucidEra described themselves as SaaS BI exclusively for data from SaaS applications others, such as PivotLink, GoodData and Birst, took a more general approach and claimed that the SaaS BI approach was superior even for data originating in on-premise enterprise applications.

To-date there are no publically traded SaaS BI companies, so the information for this section has come from several personal interviews (many off the record) with business executives, industry analyst and investors who are familiar with the SaaS BI industry.

After several years and lots of raised venture capital, it is still uncertain whether SaaS BI has been a success or not. While some vendors such as GoodData are continuing to grow by adding new customers and trying out new business models, others have lost key management personnel (e.g. PivotLink) and some going out of business completely (e.g. LucidEra).

Similar to on-premise BI applications, SaaS BI offers users the ability to report across multiple data systems and gain visibility into operations and company performance, something not possible using a single system.

For customers for whom most of their data resides on-premise and not in a cloud-based application, a data integration architecture needs to be setup that transfers a constant stream of data between the SaaS BI system and their source applications which reside within their corporate network. Many customers object to the exposure to risk that comes from this need to constantly transferring data outside of their company network to keep their SaaS BI systems up to date. Opening up a firewall, or transferring data via FTP no matter how secure, leaves CIO's feeling vulnerable and due to the sensitive nature of most of this data, often related to transactions and revenue, many organizations simply cannot risk the negative press that could result from one of these files being intercepted or lost. This has been one of the main hurdles to adoption of SaaS BI.

For those companies that use SaaS applications exclusively, or only need to report off their SaaS applications and do not have to transfer data out of the company network, SaaS BI is a better fit. However, as more and more SaaS applications start building their own reporting and analytics capability into their applications, as Salesforce, Netsuite, Successfactors and many others are beginning to do, it is arguable how big the addressable market will be for a SaaS BI solution that only makes sense for company using more than one SaaS application. Even for the companies falling into this category their strategy may be to eventually consolidate to one SaaS application vendor for cost or simplicity reasons, further eroding the addressable market.

[44] Information for this section was sourced from in-person interviews with GoodData CEO & President, Roman Stanek in San Francisco, California, on Wednesday Dec 22, 2010 and Nabil Elsheshai, Senior Analyst, Pacific Crest Securities, Portland, Oregon, on Thursday Mar 31, 2011.

BI Software That Can Enable Users to Perform Basic Data Manipulation Tasks

With the growth of open data sources, many of the datasets that are now available through data.gov are outputs from systems of record, such as ERP systems, and have not designed for public consumption let alone for analysis.

Traditionally, these data outputs need to be extracted, transformed and then loaded (ETL) into a separate system from their source. This ETL process is typically performed by a dedicated ETL tool from vendors such as Informatica, IBM, Snaplogic, Cast Iron or others, and can cost hundreds of thousands of dollars. For the average home user, cheaper, less complex and more user-friendly tools are available, such as Microsoft SQL Service Integration Services (SSIS) but this tool still requires familiarity with SQL code.

For this reason new BI software will need to include the capability for users to manipulate data sets to make them more appropriate for analysis, or better still, be able to take a typical output from a system of record and be able to automatically manipulate it and make it ready for analysis, with minimal input from the user.

A New Technical Architecture for Big Data

Many of the new innovations that have been developed to address the challenges of handling big data have focused on specific ways to query extremely large data sets and can now do this very well as illustrated by successful implementations at companies like Facebook, Google and Yahoo.

However, for big data technology to be successful within an enterprise context, it needs to fit into an existing analytical infrastructure. Figure 17 is illustrates how this might be done. Creating a unified logical layer is critical, for end-users to be able to gain a seamless experience, querying internal and external, relational and streaming data sources. The unified logical layer ensures consistent definitions for metric calculations as well as consistent views of business dimensions and attributes.

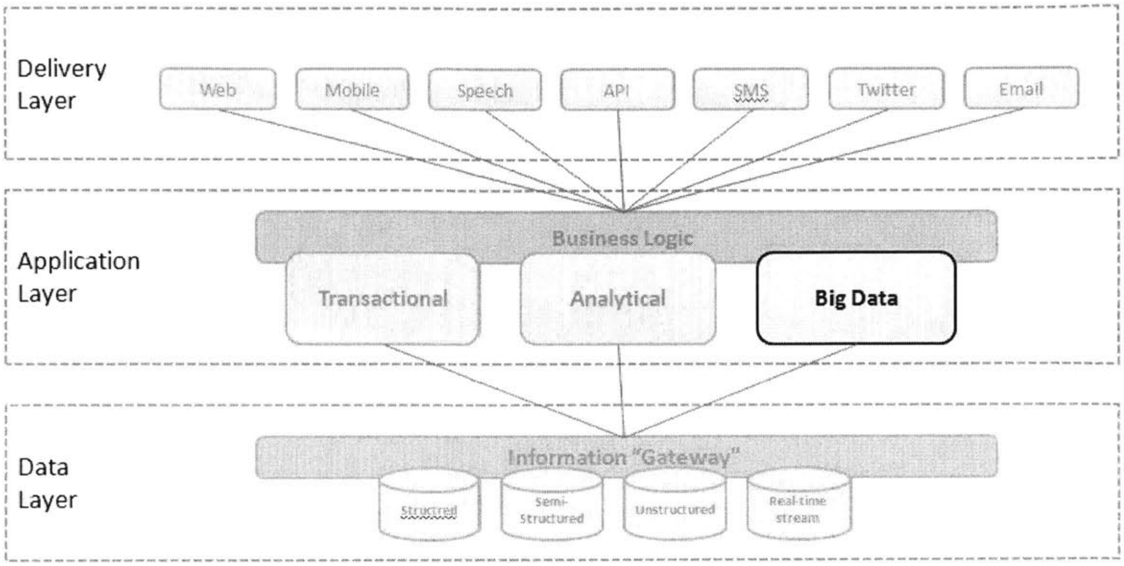


Figure 17: A Suggested Integration of Big Data & Existing Enterprise Analytical Infrastructure

The addition of the big data component within the application layer, as opposed to the data layer, is critical to ensuring that result sets from real-time data streams are identified separately to those from transactional and analytical applications. This will ensure that end users are not confused about where report data might be originating from and over time will come to understand the limitations and trade-offs of reporting against transactional, analytical and real-time data systems.

Today, the architecture of the big data component of the application layer consists of the following:

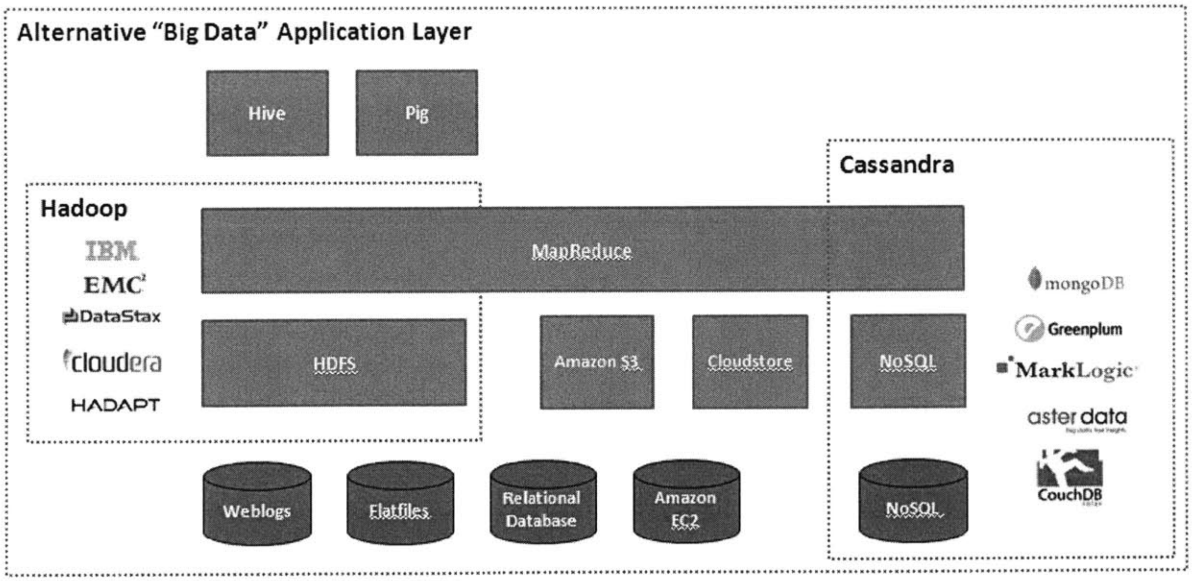


Figure 18: Suggested Architecture of the Big Data Application Layer Component

As figure 18 suggests, today there are several different ways that real-time data streams and big data can be supported. While the combination of HDFS and Hadoop is arguably the most popular, for some organizations, a packaged solution from a vendor such as Teradata Aster Data or EMC Greenplum may be a better solution.

Privacy Issues

As data volumes grow, and the use of online applications and websites continues to increase the level of risk and exposure to personal information will continue to be a growing concern for users. According to IDC, "by 2020 almost 50% of all global information will require a level of IT-based security beyond a baseline level of virus protection and virus protection[45]."

For example, a social security number might be entered onto a form on a doctor's website or a bank website for authentication purposes, but it gets stored, compared, looked-up and backed-up countless times. After several months of use, this one field may end up existing on several different systems and if compromised by hackers could exist on many more. All without the knowledge (or consent) of the user who thought they were simply logging into a website service.

As users of social networking websites such as Facebook feel increasingly more comfortable sharing personal information such as employer names, birthdays, photographs of family members and lists of friends, the debate about privacy will rage on. For Facebook, the sharing of information by users is central to the enjoyment they will obtain from the service. But concern and suspicion about who might view that data will stop users from wanting their data shared, especially those that are not as technology-savvy and become exposed after being unaware of the specific security settings they needed to modify to gain an acceptable level of control over their privacy. If these users were aware of how available and accessible their information could be to others would they participate less? The answer today seems to yes, but this may change over time as culture and societal norms evolve.

As free online services such as Google and Twitter continue to be used, it is unclear who owns the usage data that gets generated from these sites. For example, search queries in Google can be mined and analyzed allowing anyone to see what keywords are being searched on by geography. The data is detailed enough that Google could combine search queries from a particular I.P. address with data from Gmail and Google Docs accounts accessed from that same address to improve the quality of advertising displayed to that user. While Google may argue that this is improving the users online experience, to the user this may fuel suspicion and worry about how their online habits are being tracked and who else might have access to that data.

As data volumes grow and interconnectedness between systems continues to increase, these concerns around privacy will become even more important. They will require delicate and sensitive handling by existing application vendors, third part service providers and especially new entrepreneurs who, while under pressure to deliver success may extend their creative license and change their business modes and focus and start using user data in ways that those users would not have agreed to initially. And as BI systems gain access to more big data the risk of personal

[45] The Digital Universe Decade – Are You Ready? IDC report, May 2010, John Gantz and David Reinsel, sponsored by EMC Corporation, http://idcdocserv.com/925

information “leaking” from one system to another is much greater. For organizations that care about their relationship with customers, such as retailers and organizations in healthcare, financial services and other service-related industries, maintaining this level of trust and confidence with customers is critical. One breach could have dire consequences.

Conclusion

In 1964 Michael S. Scott Morton, a student at Carnegie-Mellon University and Harvard Business School researcher coined the phrase “decision support system”. While he went on to build a successful career as a Professor of Management at Massachusetts Institute of Technology Sloan School of Management, his ideas helped lay the foundation for a multi-billion dollar software market for BI and related decision support software.[46]

From humble origins, the BI market has grown tremendously in both size and influence and has quickly become a must-have for organizations of all sizes. According to Lillian Sullivan, Director of Enterprise Information & Analysis at LexisNexis Group, Business Intelligence “is the lifeblood of the business. It is central to everything our internal customers do.”[47]

Industry analyst firm Pacific Crest securities, outlines the reasons for the continued growth and relevance of BI:[48]

- Market validation from mega vendors, particularly IBM and Oracle, is driving broader awareness and interest
  - Recession drove need for greater forward looking analytics
  - Increased levels of economic volatility (currency, commodity, consumer) is driving the need for more forward looking analytics
  - History- and batch-based BI of limited value
- Emerging in-memory platforms and mobile adoption are driving broader usage of BI
- Compliance and risk issues driving greater need for integrated reference data such as an integrated view of customer
- SaaS and Big Data trends are driving a new source of Data

According to Gartner, the BI market is still growing rapidly. This will encourage new entrants from more mature technology markets to enter the BI space and will also encourage incumbents to seek out new opportunities for innovation such as the ones described in this study. If data within context is understood as information, and information itself is increasingly being seen as a valuable asset, then the ever-increasing volumes of data being generated globally, present new opportunities for business innovation.

Arguably, this wave of new innovation has already begun with the BI industry having seen several new entrants establish their brands in the market over the past few years, including QlikTech, PivotLink, Occo, GoodData and LogiXML as well as continued product innovation from the

[46] Taken from an interview with Michael S. Scott Morton conducted by DSSReview.com, Sept 18th, 2007, http://dssresources.com/reflections/scottmorton/scottmorton9282007.html

[47] Business Intelligence is Valuable, but Falls Short of Its Potential, Allan Alter, CFO Insight Magazine, Oct 5, 2010.

[48] BI and Analytics Update: Exceptionally Strong Growth Trends, presentation by Nabil Elsheshai, Senior Analyst, Pacific Crest Securities, Feb 18, 2011.

incumbents such as IBM, Oracle and SAP. The main areas of innovation that have developed include the following:

- Use of in-memory technology, to accelerate query times and spur a new breed of analysis capabilities.
- Cloud-based BI services, which have offered BI capabilities through a significantly improved business model and cost-structure and to far more end-users than was previously economically feasible.
- Major new developments around wireless (e.g. Apple iPad) and speech-recognition interfaces.

It is ideas like this that will fuel the next major wave of BI software development.

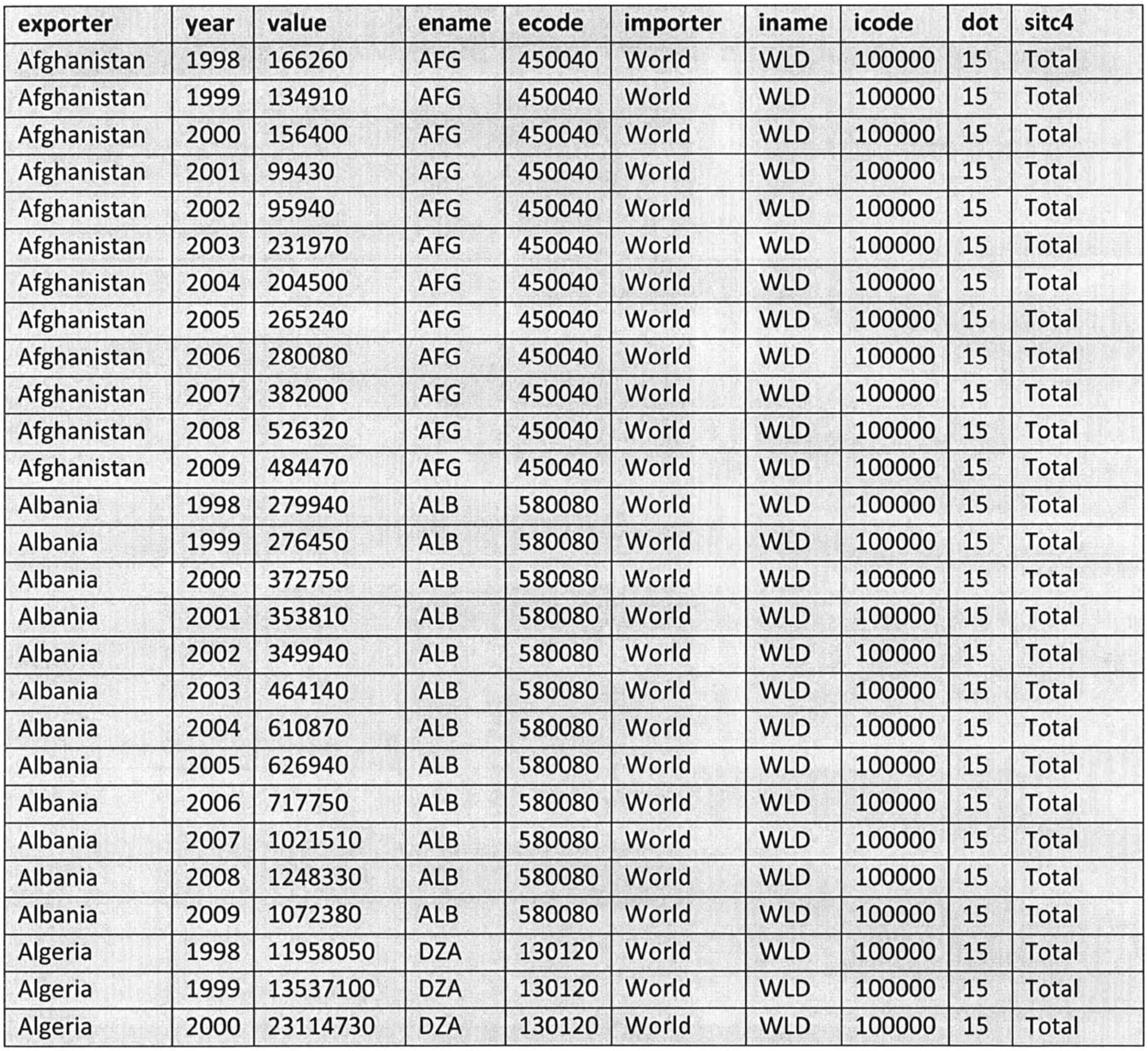
While technical innovations are pushing the BI industry forward, the data landscape is radically changing. Volumes of data being collected both within and outside the corporate world are increasing exponentially. While decision makers at all levels see value in data to make better informed, “data driven” decisions, the size and complexity of the available data is presenting new challenges to Chief Information Officers and the IT department. It’s not just a data volume problem. The velocity with which data is being generated, with greater detail, variety and all in real-time is creating unique challenges. These extreme information management issues will determine the success or failure of the next generation of BI software and therefore it is imperative that BI software vendors adapt to these changes if they are to remain relevant and survive.

Within organizations, being able to handle big data will become a greater and more critical business need as the larger environment, customers, stakeholders and even employees adapt to the new, data-rich world of the future. For those organizations that invest in new analytical infrastructure and can adapt their business models and decision making processes to leverage this new insight, significant competitive advantage can be attained.

As data volumes continue to grow, the goal to provide insight from that data will become more critical but also technically more challenging. Even in the past when data warehouses greater than 1 terabyte were uncommon, gaining actionable insight from that data was challenging. Today, as those data volumes have grown, and as BI software has improved, the same challenges still exist. Going forward it won’t be about the volume of data as much as the relevance of the insight that can be gleaned from it. This is the future challenge for BI software.

The opportunities of tomorrow will be centered on establishing BI as a foundational layer behind any information management strategy, both in the workplace and at home - to vastly improve the capability of human decision-making with the aid of machines. This dream of Scott Morton continues and with the right amount of technical innovation, business savvy and luck, we might just get there.

Appendix: Sample Data from the IMF Bilateral Trade Dataset Used for Centrality Analysis



exporter	year	value	ename	ecode	importer	iname	icode	dot	sitc4
Afghanistan	1998	166260	AFG	450040	World	WLD	100000	15	Total
Afghanistan	1999	134910	AFG	450040	World	WLD	100000	15	Total
Afghanistan	2000	156400	AFG	450040	World	WLD	100000	15	Total
Afghanistan	2001	99430	AFG	450040	World	WLD	100000	15	Total
Afghanistan	2002	95940	AFG	450040	World	WLD	100000	15	Total
Afghanistan	2003	231970	AFG	450040	World	WLD	100000	15	Total
Afghanistan	2004	204500	AFG	450040	World	WLD	100000	15	Total
Afghanistan	2005	265240	AFG	450040	World	WLD	100000	15	Total
Afghanistan	2006	280080	AFG	450040	World	WLD	100000	15	Total
Afghanistan	2007	382000	AFG	450040	World	WLD	100000	15	Total
Afghanistan	2008	526320	AFG	450040	World	WLD	100000	15	Total
Afghanistan	2009	484470	AFG	450040	World	WLD	100000	15	Total
Albania	1998	279940	ALB	580080	World	WLD	100000	15	Total
Albania	1999	276450	ALB	580080	World	WLD	100000	15	Total
Albania	2000	372750	ALB	580080	World	WLD	100000	15	Total
Albania	2001	353810	ALB	580080	World	WLD	100000	15	Total
Albania	2002	349940	ALB	580080	World	WLD	100000	15	Total
Albania	2003	464140	ALB	580080	World	WLD	100000	15	Total
Albania	2004	610870	ALB	580080	World	WLD	100000	15	Total
Albania	2005	626940	ALB	580080	World	WLD	100000	15	Total
Albania	2006	717750	ALB	580080	World	WLD	100000	15	Total
Albania	2007	1021510	ALB	580080	World	WLD	100000	15	Total
Albania	2008	1248330	ALB	580080	World	WLD	100000	15	Total
Albania	2009	1072380	ALB	580080	World	WLD	100000	15	Total
Algeria	1998	11958050	DZA	130120	World	WLD	100000	15	Total
Algeria	1999	13537100	DZA	130120	World	WLD	100000	15	Total
Algeria	2000	23114730	DZA	130120	World	WLD	100000	15	Total

References

Reza Moghadam and José Viñals, Understanding Financial Interconnectedness, October 2010, Strategy, Policy, and Review Department, International Monetary Fund http://www.imf.org/external/np/pp/eng/2010/100410.pdf

Prasanna Gai and Sujit Kapadia, Contagion in financial networks, March 2010, Bank of England Working Paper No. 383, http://www.bankofengland.co.uk/publications/workingpapers/wp383.pdf

The IMF- FSB Early Warning Exercise Design and Methodological Toolkit, September 2010 http://www.scribd.com/doc/38348537/The-IMF-FSB-Early-Warning-Exercise-Design-and-Methodological-Toolkit-September-2010

Kristin Glass, Richard Colbaugh, Toward Emerging Topic Detection for Business Intelligence: Predictive Analysis of 'Meme' Dynamics http://arxiv4.library.cornell.edu/ftp/arxiv/papers/1012/1012.5994.pdf

Emilio J. Castilla, Hokyu Hwang. Mark Granovetter and Ellen Granovetter. 2000. "Social Networks in Silicon Valley."

J. Ignacio Alvarez-Hamelin, Luca Dall'Asta, Alain Barrat, Alessandro Vespignan, Analysis and visualization of large scale, networks using the *k*-core decomposition, European Conference on Complex Systems, ECCS'05 :: Paris :: 14-18 November 2005

Eckerson [1], W.W. (2007). Beyond the basics: Acceleration BI maturity. Retrieved August 05, 2009, from http://download.101com.com/pub/tdwi/Files/SAP_monograph_0407.pdf

Eckerson [2], W.W. (2007). The myth of self-service BI. In What Works, Vol 24, 2007. Retrieved May 01, 2009, from http://www.tdwi.org/Publications/WhatWorks/display.aspx?id=8638

Gartner (2006). Gartner says business intelligence software market to reach $3 billion in 2009. Retrieved Jan 23, 2009, from http://www.gartner.com/it/page.jsp?id=492239

Golfarelli. M., Rizzi, S., & Cella, I. (2004). Beyond Data Warehousing: What's Next, retrieved August 08, 2009, from http://www.aberdeen.com/launch/report/benchmark/5873-RA-pervasive- business-intelligence.asp

IBM Corporation (2006). Information Lifecycle Management. Retrieved July 26, 2009, from Information Lifecycle Management http://mediazone.brighttalk.com/comm/sitedata/f2217062e9a397a1dca429e7d70bc6ca/download/977_Evan's%

Kanaracus, C. (2008). IDC: Oracle maintains lead in database market. Retrieved August 05, 2009, http://www.pcworld.com/businesscenter/article/147684/idc_oracle_maintains_lead_in_database_market.html

Scott Morton, Michael S. (ed), The Corporation of the 1990s: Information Technology and Organizational Transformation, Oxford University Press, 1991, ISBN 0-19-506358-9

Davenport, Thomas; Harris, Jeanne; Competing on Analytics, the New Science of Winning, Harvard Business School Press, March 6, 2007

Data, Data, Everywhere, The Economist magazine, February 27th, 2010

Evelson, Boris; The Forrester Wave™: Enterprise Business Intelligence Platforms, Q3 2008, Forrester, July 31 2008.

Gartner EXP (January 2011) Gartner Executive Programs Survey of 2,000 CIO's, January 2011 http://www.gartner.com/it/page.jsp?id=1526414

Business Intelligence Competency Centers: A Team Approach to Maximizing Competitive Advantage" by Gloria J. Miller et. Al.

TDWI Best Practices Report 2008: "Pervasive Business Intelligence: Techniques and Technologies to Deploy BI on an Enterprise Scale"