WHITE PAPER

INTELLIGENT BUSINESS STRATEGIES

Architecting A Big Data Platform for Analytics

By Mike Ferguson
Intelligent Business Strategies
October 2012

Prepared for:

Table of Contents

INTRODUCTION

Organisations have been building data warehouse for many years to analyse business activity

For many years, companies have been building data warehouses to analyse business activity and produce insights for decision makers to act on to improve business performance. These *traditional* analytical systems are often based on a classic pattern where data from multiple operational systems is captured, cleaned, transformed and integrated before loading it into a data warehouse. Typically, a history of business activity is built up over a number of years allowing organisations to use business intelligence (BI) tools to analyse, compare and report on business performance over time. In addition, subsets of this data are often extracted from data warehouses into data marts that have been optimised for more detailed multi-dimensional analysis.

The BI market is mature but BI still remains at the forefront of IT investment

Today, we are over twenty years into data warehousing and BI. In that time, many companies have built up multiple data warehouses and data marts in various parts of their business. Yet despite the maturity in the market, BI remains at the forefront of IT investment. Much of this demand can be attributed to the fact that more and more data is being created. However it is also the case that businesses are moving away from running on gut feel towards running on detailed factual information. In this vibrant market, software technology continues to improve with advances in analytical relational database technology, as well as the emergence of mobile and collaborative BI.

BUSINESS DEMAND TO ANALYSE NEW DATA SOURCES

New more complex data has emerged and is being generated at rates never seen before

However, even though this traditional environment continues to evolve, many new more complex types of data have now emerged that businesses want to analyse to enrich what they already know. In addition, the rate at which much of this new data is being created and/or generated is far beyond what we have ever seen before.

Customers and prospects are creating huge amounts of new data on social networks and review web sites. In addition, on-line news items, weather data, competitor web site content, and even data marketplaces are now available as candidate data sources for business consumption.

Social network data, web logs, archived data and sensor data are all new data sources of attracting analytical attention

Within the enterprise, web logs are growing as customers switch to on-line channels as their preferred way of transacting business and interacting with companies. Archived data is also being resurrected for analysis and increasing amounts of sensor networks and machines are being deployed to instrument and optimise business operations. The result is an abundance of new data sources, rapidly increasing data volumes and a flurry of new data streams that all need to be analysed.

THE GROWTH IN WORKLOAD COMPLEXITY

Data and analytical workload complexity is growing

Looking at all these new sources of data, it is clear that complexity is growing in both the characteristics of the data itself and in the types of analyses businesses now want to perform.

THE GROWTH IN DATA COMPLEXITY

With regards to data, complexity has increased in three main ways:

- The variety of data types being captured by enterprises
- The volumes of data being captured by enterprises
- The velocity or rate at which data is being generated
- The veracity or trustworthiness of the data

Variety of Data Types

Besides the 'normal' capture of master and transactional data, new data types are now being captured by enterprises. These include:

New types of data are being captured

- Semi-structured data e.g. email, e-forms, HTML, XML
- Unstructured data e.g. document collections (text), social interactions, images, video and sound
- Sensor and machine generated data

Much of this data is un-modelled

Invenstigative analysis is needed to determine its structure before it can be brought into a data warehouse

This collection of new more complex data types is often referred to as multi-structured data. A major problem with multi-structured data is that it is often un-modelled and therefore has to be 'explored' to derive structured data from it that has business value. For this reason, investigative analysis often has to be done on multi-structured data upstream of any traditional analytical environment to identify data that could enrich what is already stored in existing data warehouses. In addition, stand-alone advanced analytical studies may also be required on this data e.g., seismic data in Oil and Gas

Data Volume

Some new sources of data are also very large in volume

In addition to the complexity caused by new data types, the rate at which companies are accumulating data is also increasing leading to much larger data volumes. Examples would include collections of documents and emails, web content, call data records (CDRs) in telecommunications, weblog data and machine generated data. These sources can run into hundred of terabytes or even into petabytes.

Velocity of Data Generation

The rate at which data is being created is also increasing

The rate at which data is being created is also increasing rapidly. Financial markets data is a good example where data is being generated and emitted at very high rates and where there is a need to analyse it immediately to respond to market changes in a timely manner. Other examples include sensor and machine generated data where the same requirement applies, or cameras requiring video and image analyses.

THE GROWTH IN ANALYTICAL COMPLEXITY

Analytical complexisty is also growing

In terms of analytical complexity, it is now the case that new algorithms and several types of analysis may be needed to produce the necessary insight

required to solve business problems. In addition, each of these analyses may need to be done on data that has different characteristics in terms of variety, volume and velocity. Retail marketing is a good example where campaign accuracy and timeliness needs to be improved in on-line channels where people are often on-line from a mobile device. This means customer insight needs to be more detailed. In this case it may require:

Several types of analysis may be needed to solve business problems

- Historical analysis and reporting of customer demographics and customer purchase transaction activity (structured data) to determine customer segmentation and purchase behaviour
- Market basket analysis to identify products that sell together to identify cross-sell opportunities for each customer
- Click-stream analysis to understand customer on-line behaviour and product viewing patterns when traversing web site content to produce accurate up-sell offers in real-time
- Analysis of user generated social network data such as profiles (e.g. Facebook, LinkedIn), product reviews, ratings, likes, dislikes, comments, customer service interactions etc.
- Real-time analysis of customer mobile phone location services (GPS) data to detect when a customer may be in the vicinity of an outlet to target them with offers that tempt them to come in

In many cases, determining the insight needed is now a process involving multiple types of analyses

The point here is that, in many cases, determining the insight needed to solve a business problem *is now a process* involving multiple analyses on different data sources where both the data and the analyses vary in complexity. Moreover, analysis of both structured and unstructured data may be needed in any single analytical process to produce the insight required. Data integration is required to merge multi-modal data to improve actionable insights.

Not all analyses in an analytical process can always be done on a single platform

Furthermore, given that some data sources may be un-modelled, the steps in an analytical process cannot all be done on a single analytical platform and require multiple underlying technologies to solve the business problem.

Despite these complexities, significant demand to analyse these new data types along with traditional, already available, data is increasing in many business areas. A popular example is analysing social network data to understand customer sentiment, social graphs and influencers, complementing existing customer profile or customer segment data.

WHAT IS BIG DATA?

The spectrum of analytical workloads is now so broad that it cannot all be dealt with in a single enterprise data warehouse

The emergence of new data sources and the need to analyse everything from live data streams in real time to huge amounts of unstructured content has made many businesses realise that they are now in an era where the spectrum of analytical workloads is so broad that it cannot all be dealt with using a single enterprise data warehouse. It goes well beyond this. While data warehouses are very much a part of the analytical landscape, business requirements are now dictating that a new more complex analytical environment is needed to support a range of analytical workloads that cannot be easily supported in traditional environments.

A new extended analytical environment is now needed

This new environment includes *multiple* underlying technology platforms *in addition to* the data warehouse, each of which is optimised for specific analytical workloads. Furthermore, it should be possible to make use of these platforms independently for specific workloads *and also together* to solve business problems. The objective now is to cater for the complete spectrum of analytical workloads. This includes traditional and new 'big data' analytical workloads.

Big Data is a term associated with new types of workloads that cannot be easily supported in traditional environments

Big Data is therefore a term associated with the new types of workloads and underlying technologies needed to solve business problems that we could not previously support due to technology limitations, prohibitive cost or both.

Big Data is therefore NOT just about data volumes

Big Data is therefore *not* just about data volumes. It may be the case that data volumes are moderate but that data variety (data and analytic complexity) are significant. Big Data analytics is about analytical workloads that are associated with some combination of data volume, data velocity and data variety that may include complex analytics and complex data types. Therefore Big Data can be associated with *both* structured and multi-structured data and not just the latter. For this reason the Big Data analytics can *include* the traditional data warehouse environment because some analytical workloads may need both traditional and workload optimised platforms to solve a business problem. The new enterprise analytical environment encompasses traditional data warehousing *and* other analytical platforms best suited to certain analytical workloads. Big Data does not replace a data warehouse. On the contrary, *the data warehouse is an integral part of the extended analytical environment.*

Big Data can be associated with both structured and multi-structured data

The data warehouse is an intregal part of the extended analytical environment

Analytical requirements and data characteristics will dictate the technology deployed

Analytical requirements and data characteristics will dictate the technology deployed in a Big Data environment. For this reason, Big Data solutions may be implemented on a range of technology platforms including stream-processing engines, relational DBMS, analytical DBMS (e.g. massively parallel Data Warehouse appliances) or on non-relational data management platforms such as a commercialised Hadoop platform or a specialised NoSQL data store e.g. a graph database. More importantly, it could be a combination of all of these that is needed to support business requirements. It is certainly not the case that relational DBMS technology cannot be used for Big Data analytics.

TYPES OF BIG DATA

Types of data frequently associated with Big Data analytical projects include web data, industry specific transaction data, machine generated/sensor data and text.

Web logs and social network interaction data

Web data includes web log data, e-commerce logs and social network interaction data e.g. Twitter streams

High volume transaction data

Industry specific transaction data examples include telecommunications call data records (CDRs) and geo-location data, retail transaction data and pharmaceutical drug test data

Sensor data

Machine generated / sensor data is one of the fastest growing areas. Today, sensors exist to monitor everything from movement, temperature, light, vibration, location (e.g. inside smart phones), airflow, liquid flow and pressure. In addition we are seeing more and more data generating electronic components going into other products all of which can be connected to the Internet to flow data back to collectors and command centres. The era of 'The Internet of Things' is now upon us.

Text

In the world of unstructured content, text is by far the most popular to analyze. Many companies are now beginning to realise that there may be significant value in text whether that be from archived documents, external content sources or customer interaction data to name a few examples. Technology limitations have prevented or limited the analysis of this kind of data in the past. However, with these barriers now lifted, sentiment analysis is one example where text has become high priority. Also companies are now collecting data to fend off future liabilities. As an example, oil and gas companies are collecting data that will span 20 to 30 years in order to capture environment data pre, during, and post operations.

Companies are now collecting data to fend off future liabilities

WHY ANALYSE BIG DATA?

The analysis of multi-structured data may produce additional insight that can enrich what a company already knows

There are many reasons why companies want to analyse Big Data. Technology advances now make it possible to analyse entire data sets and not just subsets. For example, every interaction rather than every transaction can be analysed. The analysis of multi-structured data may therefore produce additional insight that can be used to enrich what a company already knows and so reveal additional opportunities that were previously unknown. This means much more accurate business insights can potentially be produced to help improve business performance. Analysing more data even to get small improvements of 0.5%, 1%, 2% or 3% in key performance indicators is considered very worthwhile in many organisations. Also, introducing the analysis of data streams can also improve responsiveness and reduce risk.

More detail improves the accuracy of business insights and reponsiveness

However there are still inhibitors to analysing Big Data. Two reasons for this are as follows:

A shortage of skills and market confusion are inhibiting adoption of Big Data technologies

1) The shortage of skilled people and

2) Confusion around what technology platform to use

The internet is awash with hype around Relational DBMS Vs Hadoop Vs NoSQL DBMSs with many not sure as to when one should be used over the other and for what kinds of analytical workloads.

BIG DATA ANALYTIC APPLICATIONS

Many analytic applications have emerged around structured and multi-structured data. The following table also shows some examples of industry use cases for Big Data analytics

Industry	Use Case
Financial Services	Improved risk decisions "Know your customer" 360º customer insight Fraud detection Programmatic trading
Insurance	Driver behaviour analysis (smart box) Broker document analysis to deepen insight on insured risks to improve risk management
Healthcare	Medical records analytics to understand why patients are being re-admitted Disease surveillance Genomics
Manufacturing	'Smart' product usage and health monitoring Improved customer service by analyzing service records Field service optimization Production and distribution optimization by relating reported service problems to detect early warnings in product quality and by analysing sensor data
Oil and Gas	Sensor data analysis in wells, rigs and in pipelines for health and safety, risk, cost management, production optimization
Telecommunications	Network analytics and optimization from device, sensor, and GPS inputs to enhance social networking and promotion opportunities
Utilities	Smart meter data analyses, grid optimisation Customer insight from social networks

A broad range of use cases exist for big data analytics

Web data, sensor data and text data have emerged as popular data sources for big data analytical projects.

Web site optimisation is achieved by analysing web logs

On-line advertising requires analysis of clickstream while users are on-line

With respect to web data, analyses of clickstream and social network content have been popular. Web log data is often analysed to understand site navigation behaviour (session analysis) and to link this with customer and/or login data. Media companies often want to analyse 'click through' on on-line advertising. This is particularly challenging, as it requires analysis of large volumes of streaming data in real-time while users are on-line to dynamically influence on-line navigation behaviour by placing ads. Social network analysis is also a growth area.

Sensors are opening up a whole new range of optimisation opportunities

Analysis of machine generated / sensor data is being adopted for supply/distribution chain optimisation, asset management, smart metering, fraud and grid health monitoring to name a few examples.

Text analysis is needed to determine customer sentiment

In the area of unstructured content, text in particular is being targeted for analysis. Case management, fault management for field service optimisation, customer sentiment analysis, research optimization, media coverage analysis and competitor analysis are just a few examples of Big Data analytic applications associated with unstructured content.

BIG DATA ANALYTICAL WORKLOADS

There are a number of Big data analytical workloads that extend beyond the traditional data warehouse environment

Given the range of analyses that could be supported in a Big Data environment, it is worth looking at the new types of big data analytical workloads that extend beyond those of the traditional data warehouse. These workloads are as follows:

- Analysis of data in motion
- Exploratory analysis of un-modeled multi-structured data
- Complex analysis of structured data
- The storage and re-processing of archived data
- Accelerating ETL and analytical processing of un-modeled data to enrich data in a data warehouse or analytical appliance

ANALYSING DATA IN MOTION FOR OPERATIONAL DECISIONS

The purpose of analysing data-in-motion is to analyse events *as they happen* to detect patterns in the data that impact (*or are predicted to impact)* on costs, revenue, budget, risk, deadlines and customer satisfaction etc. When these occur, the appropriate action can then be taken to minimise impact or maximise opportunity.

Event-stream processing is about automatically detecting, analysing and if necessary acting on events to keep the business optimised

This type of big data analytical workload is known as event stream processing and is most often used to support every day operational decisions where all kinds of events can occur throughout a working day.

Examples include a sale of shares on the financial markets, a price change, an order change, an order cancellation, a large withdrawal on a savings account, the closure of an account, a mouse click on a web site, a missed loan payment, a product or pallet movement in a distribution chain (detected via RFID tag), a tweet, a competitor announcement, CCTV video on street corners, Electrocadigram (EKG) monitors, etc. Whatever the events or streams of data, there are literally thousands or even millions of these that can occur in business operations each second. And while not all data are of business interest, many require some kind of responsive action to seize an opportunity or prevent a problem occurring or escalating. That response may need to be immediate and automatic in some cases or subject to human approval in others.

Thousands of events can occur in business operations thoughout a working day

Event stream processing requires analysis of data to take place before data is stored anywhere

Stream processing is unique because analysis of data needs to take place *before* this data is stored in a database or a file system. Given the velocity at which data is generated and the volumes of data typically involved in stream processing, it also means that human analysis is often not feasible. Analysis therefore has to be automated using a variety of analytic methods, such as predictive and statistical models or acoustic analysis to determine or predict the business impact of these events. Decision-making may also have to be automated to respond in a timely manner to keep the business optimised and on track to achieving its goals. Actions may vary from alerts to completely automated actions (e.g., invoke transactions or close a valve in an oil well). The latter is most likely when common problems occur while the former is triggered when exceptions occur that require manual intervention. Note that rules are needed to make automated decisions and to trigger automated

People cannot be expected to spot every problem

actions. Therefore a rules engine is an important component of stream processing.

In some industries the volume of event data can be significant

In some industries, the volumes of streaming data can be significant. Of course not all this data must be stored. Only if a pattern deviates from the norm is the data likely to be persisted for subsequent historical analysis to identify recurring patterns, problems and opportunities all of which may drive further tactical and strategic decisions. Nevertheless, even when filtered, the volumes of event data can be significant.

EXPLORATORY ANALYSIS OF UN-MODELLED MULTI-STRUCTURED DATA

Un-modelled multi-structured data needs to be explored to determine what subset of data is of business value

The issue with multi-structured data is that it is often un-modelled and therefore requires exploratory analysis[1] to determine what subset of data is of value to the business. Once this has been done, any data identified as being of value can be extracted and put into data structures from where further analysis can take place and new business insight produced.

Popular sources of multi-structured data include web logs and external social network interaction data.

Reputation management and 'voice of the customer are dominating analysis of text

A recent survey[2] showed that analysing and extracting data from social networks currently dominates text analytical activity in customer-facing organisations. The same survey highlighted top business applications driving analysis of text as:

- Brand/product/reputation management (39% of respondents)
- Voice of the Customer (39%)
- Search, Information access or question answering (39%)
- Research (36%)
- Competitive intelligence (33%)

Text can vary in terms of language and format

The challenge with this type of data is that it can be very large in volume and may contain content in different languages and formats. It may also contain considerable amounts of poor quality data (e.g. spelling errors or abbreviations) and obsolete content. A key requirement for successful text analytics is to 'clean' the content before analysis takes place. However, many companies often have no mechanism to do this. Pre-processing text before analysis involves extracting, parsing, correcting and detecting meaning from data (using annotators) to understand the context in which the text should be analysed. These issues highlight multi-structured data complexity.

Quality can also be a problem

Multi-structured data is hard to analyse

Multi-structured data is also hard to analyse. Take social interaction data for example. Analysing user generated social interaction data may require *multiple* analytical passes to determine the insight needed. For example:

- The first pass involves text analysis (mining) to extract structured customer sentiment and also to extract social network 'handles' embedded in interaction text that represent members of a social graph

[1] Often conducted by Data Scientists

[2] Text/Content Analytics 2011, Grimes, Alta Plana published September 2011

- The second pass is to analyse the extracted data for negative and positive sentiment
- The third pass loads the social network handles into a graph database where new advanced analytic algorithms (e.g. N-Path) can be used to navigate and analyse links to identify contacts, followers and relationships needed to piece together a social network and to identify influencers.

Multiple analytical passes may be needed to determine insights

Search based analytical tools may help with this type of workload

Predictive analytics, more sophisticated statistical analysis and new visualization tools may also be needed. Also search based analytical tools that use search indexes on multi-structured data may also help with this type of workload.

Content analytics can go beyond text to analyse sound and video

Content analytics goes beyond text analytics in that it can also handle audio, video and graphics. Digital asset content e.g. sound and video is more difficult to parse and derive business value from because the content is not text. Deriving insight from this kind of content is more heavily reliant on sophisticated analytic routines and how well the content has been tagged to describe what the content is and what it is about.

Exploratory analytics of un-modelled data is a process

Exploratory analysis of un-modelled multi-structured data is therefore a process in its own right. This big data analytical workload involves

- Obtaining the necessary un-modeled data
- Cleaning the data
- Exploring the data to identify value
- Producing a model from the exploratory analysis (structure)
- Interpreting or analysing the model to produce insight

COMPLEX ANALYSIS OF STRUCTURED DATA

Data mining is a popular example of complex analysis on structured data

This type of big data analytical workload may be on structured data taken from a data warehouse or from other data sources (e.g. operational transaction systems) for the specific purpose of doing complex analysis on that data. This may be needed so that power users can mine data to produce predictive models for use in every-day business operations. An example would be to build models that score customer risk. These models may be for use in recommendation services so that the same recommendations are used across all channels of the business that a customer can interact with e.g. website, contact centre, sales person, partner, kiosk and physical outlet (e.g. branch, store). Another example would be to support detailed statistical analysis on spend data to produce models that predict when spending is likely to exceed budget to issue early warnings that keep spending under control.

Predictive and statistical models can be built for deployment in database or in real-time operations

Some vertical industries are investing heavily in complex analysis to mitigate risk

Oil and Gas provides another example of complex analytics in the area of well integrity, which is very important in managing environmental risk, health and safety and maintaining production volumes. Disasters can be very damaging to corporate brands and so it is often the case that detailed data from multiple files associated with many individual wells are loaded into an analytical DBMS to look for possible problems and to compare readings across wells. It may also be the case that there is a need to contrast actual well data against seismic data to compare actual geology against simulated geology taken from seismic surveys. This may lead to better identification of drilling opportunities.

THE STORAGE, RE-PROCESSING AND QUERYING OF ARCHIVED DATA

Storing and analysing archived data in a big data environment is now attacting interest

Increasingly, big data systems are being looked at as an inexpensive alternative for storing archive data. It is also the case that organisations are increasingly required to archive data. There are many reasons for this including:

Compliance, audit, e-discovery and data warehouse archive are all reasons for wanting to do this

- The need to store data for many years to remain compliant with legislation and/or industry regulations. This includes data associated with business transactions, master data, retired legacy applications, documents, emails and audio files
- The need to store archived data on-line for auditing purposes
- The need to collect and archive both structured and multi-structured data for the purposes of fending off future liabilities.
- The need to manage cost and performance of data warehouses where data volumes continue to grow from increasing transaction volumes, more data sources and from business users demanding lower levels of detail. This is done by archiving older data whose value has diminished

While storing this data in a big data environment is achievable today, the challenge comes when needing to restore that archived data to respond to legal challenges, support auditing requirements or to re-process data for specific analytical purposes. In this case, data may need to be accessed and restored in multiple ways. It may need to be queried and analysed 'as archived' for compliance purposes or restored to structures that may have changed since archives were taken. Also, data currently held in data warehouses may need to be compared to archived historical data while preserving versions of hierarchies to facilitate more accurate comparative analysis.

Integration between multi-structured data platforms and data warehouses may be needed to process and analyse data

This big data analytical workload may therefore require integration between multi-structured data platforms and structured data warehouse platforms. This may be to re-provision archived data into traditional environments using ETL processing or to run data warehouse predictive and statistical models against archived warehouse data stored elsewhere. Multi-structured archive data such as audio files or emails and attachments may also require analytical processing. All of this highlights the need to manage mixed workloads and queries across different types of archived data.

ACCELERATING ETL PROCESSING OF STRUCTURED AND UN-MODELED DATA

Finally, there is a big data analytical workload that needs to accelerate the filtering of un-modelled data to enrich data in a data warehouse or analytical database appliance. This is shown in Figure 1. With so many new sources of data now available to business and data arriving faster than business can consume it, there is a need to push analytics down into ETL processing to automatically analyse un-modelled data so that data of value can be extracted and consumed rapidly. The purpose of this is speed up the consumption of un-modelled data to enrich existing analytical systems. This improves agility and opens up the way for more timely production of business insight.

There is now a need to push analytics down into ETL processing to automatically analyse un-modelled data in order to consume data of interest more rapidly

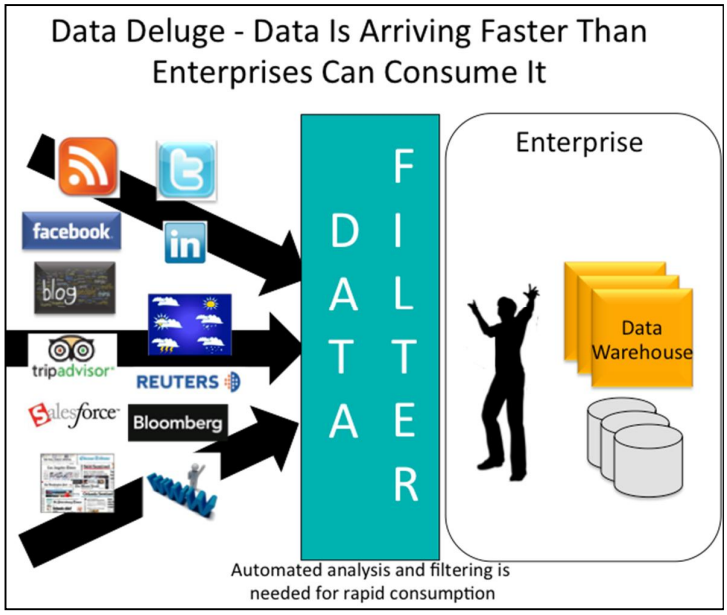


Figure 1

TECHNOLOGY OPTIONS FOR END-TO-END BIG DATA ANALYTICS

New technologies need to be added to traditional environments to support big data analytical workloads

Based on the Big Data analytical workloads defined, the next question is what additional technologies are available to support these workloads that need to be introduced into traditional data warehouse environments to extend the environment to support end-to-end big data analytics?

EVENT STREAM PROCESSING SOFTWARE FOR BIG DATA-IN-MOTION

Stream processing software supprts real-time analytical applications designed to continuously optimise business oprtations

Stream processing[3] software, is used to support the automatic analysis of data-in-motion in real-time or near real-time. Its purpose is to identify meaningful patterns in one or more data streams and trigger action to respond to them as quickly as possible. This software therefore provides the ability to build real-time analytic applications whose job it is to continuously keep different parts of a business operation optimized. These applications must be capable of automated analysis of event data streams containing either multi-structured data (e.g. Twitter streams or video streams) or structured data or both. Predictive and/or statistical models deployed in real-time analytical workflows provide this automated analytical capability in stream processing software. In addition, a rules engine is needed to automate decision-making and action taking.

The software must cope with high velocity 'event storms' where events arrive out of sequence at very high rates

One challenge with this software is to scale to handle identification of event patterns in very high velocity 'event storms'. This needs to occur even when the events in a pattern do not arrive in sequence and multiple instances of a pattern co-exist in the same time-series. Also each instance of an event pattern must be identified even when events from multiple instances of the same pattern arrive inter-mixed and out of sequence.

Another challenge is the integration of multimodal data. An example would be performing facial recognition on streaming video data from street corner cameras, and then integrating this with GPS information to inform the closest law enforcement personnel of the location of a suspected person of interest. A further example is the use of multivariate mining models such as regressions or clustering to analyse purchase transactions and integrating this with Facebook or Twitter feeds. This could be used to correlate a gasoline purchase far from home (normally triggering a suspected stolen credit card) with tweets about being away from home to correctly identify a proper purchase, and not a fraudulent purchase.

[3] Similar to complex event processing (CEP) but with emphasis on analytics, not simply rules, and the ability to analyse multi-structured data.

STORAGE OPTIONS FOR ANALYTICS ON BIG DATA AT REST

There are multiple storage options for supporting big data analytics on data at rest

Apart from data-in-motion, all of the other big data workloads discussed are associated with data at rest. In other words data needs to be stored prior to analysis taking place. Bear in mind that the aforementioned workloads are *in addition to* those running in a traditional data warehouse. In this new *extended* analytical environment, there are multiple storage options available to support big data analytics on data at rest. These options include:

- Analytical RDBMSs
- Hadoop solutions
- NoSQL DBMSs such as graph DBMSs

Also a hybrid option of Analytical RDBMS with Hadoop Map/Reduce integration may be needed.

Analytical RDBMSs Appliances

Analytical RDBMS appliances are hardware/software offerings specifically optimised for analytical processing

Analytic RDBMS platforms are relational DBMS systems that typically run on their own special purpose hardware specifically optimised for analytical processing. This combination of hardware and special purpose DBMS software is often known as an *appliance* and is a workload-optimised system. Analytical RDBMSs have been continually enhanced over the years to improve scalability, query performance and complex analyses on well understood structured data. Improvements include:

Analytical RDBMS appliances have been continually enhanced over the years

- The introduction of solid state disks (SSDs) for faster I/O processing
- Special processors that filter data close to the disk
- Columnar storage and database query processing
- Data compression
- Scan sharing so queries can 'piggy back' data brought in by others
- In-database analytics to run analytical functions closer to the data to exploit the power of the hardware and massively parallel processing
- In-memory data
- Multi-temperature data split across SSDs and spinning disk

Hadoop Solutions

The Hadoop stack enables batch analytic applications to use thousands of compute nodes to process petabytes of data stored in a distributed file system

Apache Hadoop is an open source software stack designed to support data intensive distributed applications. It enables batch analytic applications to use thousands of computers or computer nodes on petabytes of data. Besides the open source Apache version, several commercial distributions of Hadoop are available in the marketplace, many of which run on dedicated hardware appliances. The components of the Hadoop stack are:

Component	Description
Hadoop HDFS	A distributed file system that partitions large files across multiple machines for high-throughput access to data
Hadoop MapReduce	A programming framework for distributed batch processing of large data sets distributed across multiple servers
Chukwa	A platform for distributed data (log) collection and analysis

Hive is a data warehouse system for Hadoop that provides a mechanism to project structure on Hadoop data

Hive provides an interface whereby SQL can be converted into Map/Reduce programs

Mahout offers a whole library of analytics that can exploit the full power of a Hadoop cluster

Hive	A data warehouse system for Hadoop that facilitates data summarization, ad-hoc queries, and the analysis of large datasets stored in Hadoop compatible file systems. Hive provides a mechanism to project structure onto this data and query it using a SQL-like language called HiveQL. HiveQL programs are converted into Map/Reduce programs
HBase	An open-source, distributed, versioned, column-oriented store modeled after Google' Bigtable
Pig	A high-level data-flow language for expressing Map/Reduce programs for analyzing large HDFS distributed data sets
Mahout	A scalable machine learning and data mining library
Oozie	A workflow/coordination system to manage Apache Hadoop jobs
Zookeeper	A high-performance coordination service for distributed applications

Hadoop does not normally compete with RDMS technology. It expands the opportunity to work with a broader range of content. For this reason Hadoop is well suited to exploratory analysis of multi-structured data although structured data can also be analysed in this environment.

Hadoop is well suited to exploratory analysis of un-modelled multi-structured data

Typically, un-modelled data is stored in the Hadoop HDFS file system where exploratory analysis occurs to derive structure which may then be stored in Hive for further analysis. Data scientists develop batch analytic applications in languages like Java, Python and R to run in this environment using a style of programming known as MapReduce. This allows programs to be copied to thousands of compute nodes where the data is located in order to run in parallel. In addition in-Hadoop analytics in Mahout can run in parallel close to the data to exploit the full power of a Hadoop cluster. Hive is also available to SQL developers and/or tools to access data in Hadoop using the HiveQL language. In addition, Analytical RDBMS vendors are announcing external table functions and utilities to open up multi-structured data in Hadoop to the SQL community.

Mahout analytics can be applied to Hadoop data and the results stored in Hive

Hive provides an interface to make data available to SQL developers and tools

NoSQL DBMSs

In addition to Hadoop HDFS, HBase and Hive, there are other NoSQL DBMSs options available as an analytic data store. They include key value stores, document DBMSs, columnar DBMSs, graph databases and XML DBMSs. Some NoSQL databases are not aimed at big data analytics. Others are aimed at analysis of big data or for specific types of analyses. A good example would be graph DBMSs which are particularly suited to social graph (network) analytics. Note that there are no standards in the NoSQL market as yet.

Graph databases are one type of NoSQL data store particularly well suited to social network links analysis

Which Storage Option Is Best?

Generally speaking, the data and analytical characteristics of the big data workload will dictate which of these is the best solution. The following table shows criteria that can be used as a guideline as to where data should be stored for a big data analytical workload.

Analytical RDBMS	Hadoop / NoSQL DBMS
Data analysis and reporting or complex analysis	Data exploration followed by analysis or a very specific type of analysis for which a NoSQL DBMS is designed to excel at e.g. graph analysis in a graph database

Hadoop is suited to analysing unmodelled data or very large volumes of structured data in batch

Data is well understood	Data is NOT well understood
Schema is defined and known	Schema is not defined and variant
Batch and on-line analysis	Batch analysis with some on-line capability via Hive or Lucene
Access via BI tools that generate SQL and that can run predictive / statistical models in the database	Development of MapReduce applications in Java, R, Python, Pig etc.
Scalable to hundreds of terabytes on purpose built MPP clusters	Scalable to Petabytes on purpose built appliances or on the cloud

Analytical RDBMS is suited to complex analysis of structured data and for data warehousing systems that do not have heavy mixed workloads

Looking at the workloads for big data at rest, the following table tries to match the each workload to the appropriate data storage platform.

Big Data Analytical Workload	Big Data Storage Platform
Exploratory analysis of un-modelled multi-structured data e.g. web logs, unstructured content, filtered sensor data, email	Hadoop
Complex analysis of structured data or for data warehouses that have 'light' mixed workloads	Analytic RDBMS Appliance
Storage and re-processing of archived data	Hadoop
Accelerating ETL processing of structured and un-modelled data	Hybrid: Hadoop and Analytical DBMS
Social Graph Link analysis	NoSQL Graph DBMS

It is important to match the data characteristics and analytical workload to the tachnology to select the best platform

SCALABLE DATA MANAGEMENT OPTIONS FOR BIG DATA AT REST

A common suite of tools for information management across all analytical data stores is desireable in the extended analytical environment

In this new extended analytical environment, a critical success factor will be consistent high quality data across multiple analytical data stores including Data Warehouse RDBMSs, Analytical RDBMS Appliances, Hadoop Clusters/Appliances and NoSQL DBMSs. There are a number of options for data management that range from different data management tools for each different platform to a common suite of tools supplying data to all platforms. Ideally the latter would be better. However, it is also important to move data between platforms as part of an analytical process including:

- Moving master data from an MDM system into a data warehouse, an analytical DBMS, or Hadoop
- Moving derived structured data from Hive to a data warehouse
- Moving filtered event data into Hadoop or an analytical RDBMS
- Moving dimension data from a data warehouse to Hadoop
- Moving social graph data from Hadoop to a graph database
- Moving data from a graph database to a data warehouse

All of this is now needed and shown in Figure 2.

Information management needs to consolidate data to load analytical data stores AND also move data between data stores

Information management suites needs to integrate with Hadoop, NoSQL DBMSs, data warehouses, analytical RDBMSs and MDM

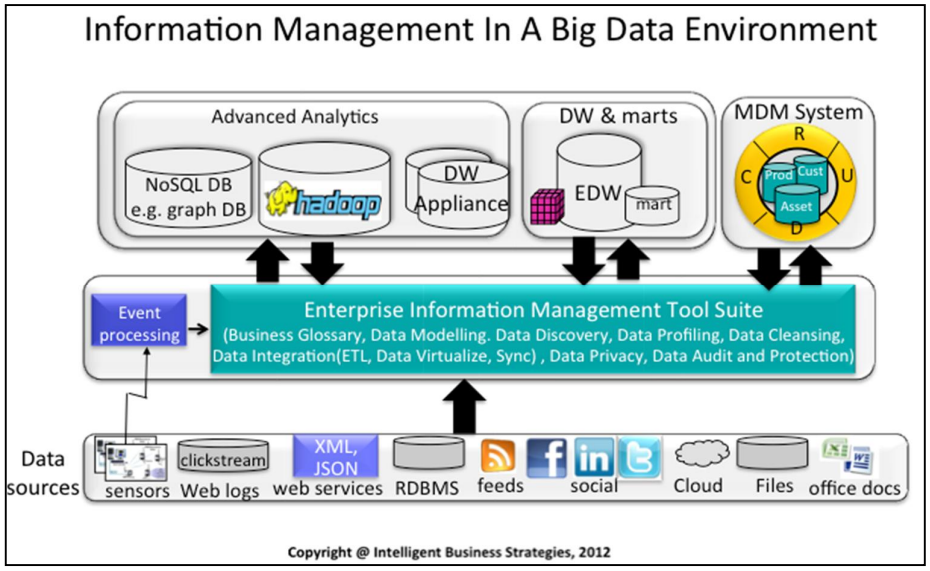


Figure 2

To make this possible, information management software needs to support ELT processing on Hadoop (multi-structured data) and/or analytical RDBMSs (structured data), interface with event processing to ingest filtered event stream data, load data into Hadoop and NoSQL DBMSs, parse data in Hadoop, clean data in Hadoop, generate HiveQL, PIG or JAQL to process multi-structured data in Hive or Hadoop HDFS, perform automated analysis on data in Hadoop and finally to extract data from Hadoop and NoSQL DBMSs. It must also support master data management.

OPTIONS FOR ANALYSING BIG DATA

In terms of analysing data in this new extended analytical environment, there are several options available depending on where that data resides. These are as follows:

A number of options are available to analyse big data at rest

- Custom Hadoop MapReduce batch analytic applications using 'in-Hadoop' custom or Mahout analytics
- MapReduce based BI tools and applications that generate MapReduce applications
- In-Database analytics on analytical DBMSs
- Traditional BI tools analysing data in Hadoop Hive and Analytical RDBMS in addition to data warehouses and cubes
- Search based BI tools on Hadoop and Analytical RDBMS
- In-flight analytics of data-in-motion in event data streams

Custom built map / reduce applications to analyse data in Hadoop

The first option is for building custom map/reduce applications to analyse multi-structured data in Hadoop HDFS. Examples would be text, clickstream data, images etc. This is likely to be an option for a Data Scientist involved in exploratory analysis and writing their own analytics in R for example or using the pre-built Mahout library of analytics from within their application.

Pre-built Mahout analytics in Hadoop

Pre-built analytic applications that use map/reduce in Hadoop

New BI tools that generate map/reduce jobs in Hadoop

There are also pre-built analytic application solutions and new BI tools available that generate MapReduce applications that exploit the parallelism in Hadoop to analyse multi-structured data such as large corpuses of content or customer interaction data.

In-database analytics in analytical RDBMSs

In-database analytics is the deployment of custom built or pre-built analytics within an Analytical RDBMS to analyse structured data. This is an example of complex analytics on structured data.

SQL-based BI tools accessing Hadoop data via Hive or accessing RDBMSs

Besides traditional BI tools accessing analytical RDBMSs and data warehouses to analyse and report on structured data, some of these tools now support a Hive interface which allows generated SQL to be converted to MapReduce applications to process multi-structured or structured data in Hadoop.

Search based BI tools and applications that use indexes to analyse data in Hadoop and/or analytical RDBMSs

Finally, given the amount of text being analysed, it is now the case that new search based BI tools (see Figure 3) are emerging to permit free form analysis of multi-structured and structured data in Hadoop and/or in data warehouse appliances. These tools can crawl structured data in analytical RDBMSs and also use MapReduce to build indexes on data in Hadoop. It then becomes possible to build analytic applications on top of these indexes to support free form exploratory analysis on multi-structured data and/or structured data. The tools may exploit the Hadoop Lucene search engine indexes or other indexes which themselves may be stored in Hadoop.

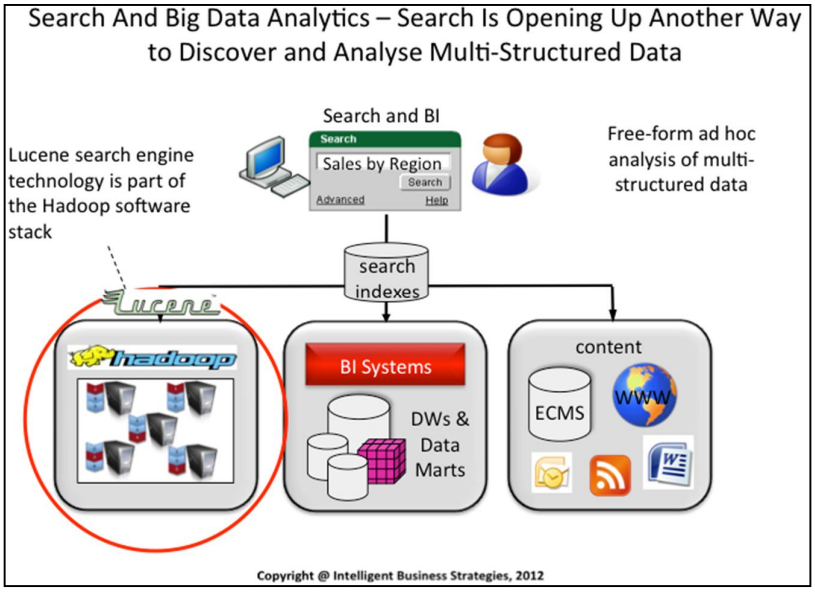


Figure 3

INTEGRATING BIG DATA INTO YOUR TRADITIONAL DW/BI ENVIRONMENT

Having looked at big data technologies, the obvious question is how do all the technology components discussed integrate with a traditional data warehouse environment to extend that environment to support all traditional and big data analytical workloads?

THE NEW ENTERPRISE ANALYTICAL ECOSYSTEM

Traditional data warehouse environments need to be extended to support big data analytical workloads

Figure 4 shows the extended end-to-end analytical environment needed to support the big data analytical workloads discussed as well as traditional data warehouse ad hoc query processing, analysis and reporting. Some refer to this new environment as the 'enterprise analytical ecosystem' or 'logical data warehouse'. It can be seen from this architecture that event processing of data-in-motion can be done on sensor data, or indeed any other event data source like financial markets for example. When variations in event data occur, event-processing software analyses the business impact and can take action if required. Filtered events can then be picked up by information management software and loaded into Hadoop for subsequent historical analysis. If any further insight is produced using batch map/reduce analytical processing, that insight may then be fed into a data warehouse. For un-modelled multi-structured data, this data can be loaded directly into Hadoop using information management software where data scientists can conduct exploratory analysis using custom map/reduce applications, or map/reduce tools that generate HiveQL, Pig or JAQL. Alternatively search-based BI tools can be used to analyse the data using indexes built in Hadoop with map/reduce utilities. If the multi-structured data is Twitter data for example, then Twitter handles could be extracted and loaded into a NoSQL graph database for further social network link analysis. Information Management software can manage the movement of the reduced social network link data from Hadoop to the NoSQL graph DBMS for this analysis to take place. If data scientists produce any valuable insight, it can also be loaded into the data warehouse to enrich the structured data already there and so make this insight available to traditional BI tool users.

Information management has a major role in keeping this environment integrated

Complex analysis of structured data is undertaken on analytical DBMS appliances using in-database analytics. Again, if any insight is produced or any new predictive/statistical models created, then this can be moved into the data warehouse for use by information consumers in reports, dashboards and scorecards. Storage and re-processing of archived data can be managed in Hadoop with batch map/reduce applications or the aforementioned front-end tools used to analyse this data. In-Hadoop analytics (custom-built or Mahout) can be used as needed. Finally with respect to accelerating ETL processing on structured and un-modeled data, information management tools can be used to exploit Hadoop analytics and/or in-database analytics in analytical DBMS appliances (or both) for this purpose. Traditional data warehouse workloads also continue as normal.

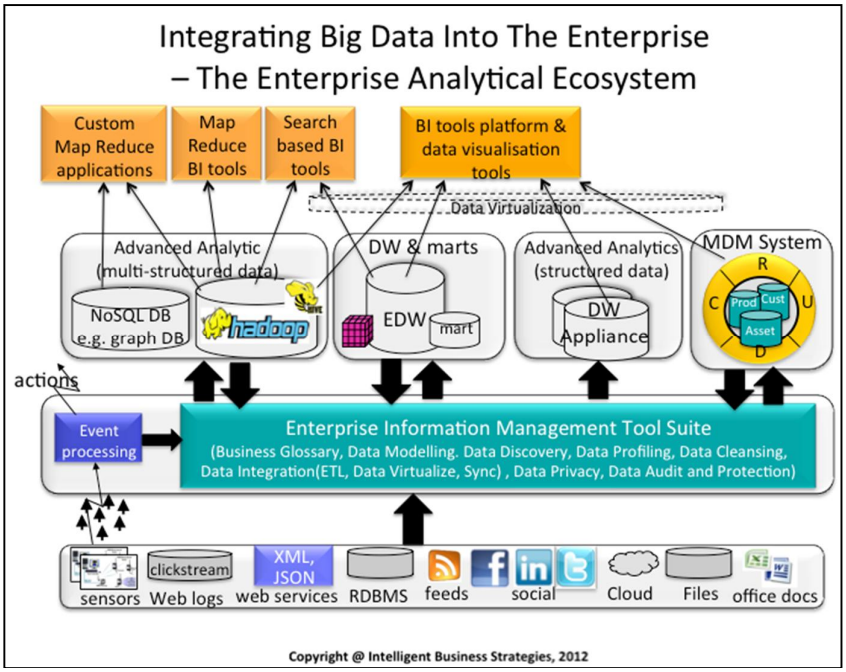


Figure 4

Data virtualization simplifies access to multiple analytical data stores

Master data management provides consistent master data to all analytical platforms

Given that this new extended analytical environment has a mix of traditional data warehouse and big data workload optimised systems, there is also a need to simplify access to these data stores for users with traditional front-end tools. This is achieved through data virtualisation software, which is designed to present data as if it is available in a single data store. Behind the scenes it uses pushdown optimization and other advanced optimization techniques to answer business queries. Also mast data is available to feed consistent dimension data to all analytical environments.

JOINED UP ANALYTICAL PROCESSING –THE POWER OF WORKFLOW

Information management workflows can be turned into analytical processes that operate across the entire analytical ecosystem

Underpinning this entire extended analytical environment is enterprise information management software. One of the major strengths of this software is its ability to define data quality and data integration transforms in graphical workflows. However, given that analytics can now be pushed down into analytical databases and into Hadoop, it transpires that analytics, rules, decisions and actions can now be added into information management workflows to create automated analytical processes. These analytical processes can then run across *multiple* platforms in the extended analytical environment. This means that workflows can be built and re-used regularly for common analytical processing of both structured and un-modelled multi-structured data to speed up the rate at which organisations can consume, analyse and act on data. All of it can be potentially automated. The power of this concept is truly awe inspiring as we transition from information management workflows to an environment where full blown analytical processes can be created across the enterprise. Furthermore, these analytical processes exploit the appropriate analytical platform best suited to the analytical workload(s) that make up the workflow. And if that was not enough, the entire workflows are service enabled making them available on-demand. So for example, social network interaction data can be loaded into Hadoop, text analytics invoked to extract sentiment data and social 'handles', master data updated with handles, sentiment associated with a customer and scored to add sentiment scores into the data warehouse. Analysis could then be done to determine unhappy *valuable* customers and action taken to retain them. In

This speeds up the rate at which organisations can consume, analyse and act on data

Powerful new analytical workflows can be used to retain customers and sharpen competitive edge

addition, social handles can be loaded into a graph database for deep analytics on social network link analysis to identify influencers and other relationships to open up cross-sell opportunities.

TECHNOLOGY REQUIREMENTS FOR THE NEW ANALYTICAL ECOSYSTEM

In order to support all this, the following requirements need to be met by any technology solution. These are not listed in any order of priority.

Multiple analytical data stores in addition to the enterprise data warehouse

- Support for multiple analytical data stores including:
  - Apache Hadoop or a commercial distribution of Hadoop for cost-effective storage and indexing of unstructured data
  - MPP Analytical DBMS offering pre-built in-database analytics and the ability to run custom built analytics written in various languages (e.g. R) in parallel to improve performance when analyzing large data volumes
  - A data warehouse RDBMS to support business-oriented, repeatable analysis and reporting
  - A graph DBMS
- Support for stream processing to analyse data in motion
- Information management tool suite support for loading Hadoop HDFS or Hive, graph DBMS, analytical RDBMS, data warehouse and master data management
- Ability for the information management tool suite to generate HiveQL, Pig or JAQL to exploit the power of Hadoop processing

Integration of information management tools with all analytical data stores and event stream processing

- Integration between stream processing and information management tools to take filtered event data and store it in Hadoop or an analytical RDBMS for further analysis
- Support for seamless data flows across multiple SQL and NoSQL data stores

Data virtualization to simplify access to data

- Data virtualisation to hide complexity of multiple analytical data stores

Best fit query optimization and in-datastore analytics

- Query re-direction to run analytical queries on the analytical system best suited to the analysis required i.e., Data Warehouse, analytical RDBMS, Hadoop platform, event stream processing engine etc.

Deployment of models in multiple analytical data stores as well as event stream processing

- Ability to develop predictive and statistical models and deploy them in one or more workload optimised systems as well as in a data warehouse e.g. in a Hadoop system, an analytical RDBMS and event stream processing workflows for real-time predictive analytics

Analytical workflows with full decision management

- Ability to run in-database and in-Hadoop analytics in information management workflows for automated analysis during data transformation and data movement
- Integration between information management workflows and a rules engine to support automated decisions during workflow execution
- Nested workflows to support multi-pass analytical query processing
- Exploitation of Hadoop parallel clusters during ETL processing
- Ability to create sandboxes for exploratory analysis across one or more underlying analytical data stores

Sandboxes for exploratory analytics by data scientists

- Ability to support data science project management and collaboration among data scientists working in a sandbox
- Ability to source, load and prepare the data for exploratory analysis of huge corpuses of content in a MPP sandbox environment by associating data sources and information management workflows with a sandbox
- Ability to integrate third party analytical tools into a specific sandbox

Governance of the entire ecosystem including sandboxes and data scientists

- Ability to control access to sandboxes
- Tools to develop MapReduce applications that can be deployed on Hadoop platforms, graph databases or analytical RDBMSs as SQL MR functions
- Parallel execution of text analytics in Hadoop clusters or in real time stream processing clusters

New tools and traditional tools working side-by-side to solve business problems

- Ability to build search indexes across all analytical data stores in the new enterprise analytical ecosystem to facilitate search based analysis of structured and multi-structured data
- Ability to connect traditional BI tools to Hadoop via Hive
- End-to-end single console systems management across the entire analytical ecosystem
- End-to-end workload management across the entire analytical ecosystem

GETTING STARTED: AN ENTERPRISE STRATEGY FOR BIG DATA ANALYTICS

BUSINESS ALIGNMENT

Big data projects need to be aligned to busines strategy

Getting started with Big Data analytics is to some extent no different than any other analytical project. There has to be a business reason for doing it so that the insights produced can help an organisation reach targets laid down in its business strategy. Alignment with business strategy has always been a critical success factor in analytics and it is no different with Big Data.

Identify candidate Big data projects and prioritise them based on business benefit

To that end, organisations should create a list of candidate Big Data analytic applications encompassing structured and/or multi-structured data-in-motion and data at rest. Obvious examples are analysis of web logs, text, social network data, financial markets and sensor data. Other signs are when current technology is limiting the ability to analyse data or when data or analytic complexity has prevented it from being done before. Just like in traditional data warehousing, business priority then needs to be assigned to these candidate applications based on return on investment.

WORKLOAD ALIGNMENT WITH ANALYTICAL PLATFORM

Match the analytical workload with the analytical platform best suited for the job

Once this has been decided, organisations should seek to match the analytical workload with the platform best suited for the job. For example, exploratory analysis of un-modelled multi-structured data, such as social network interaction data, would be a candidate for a Hadoop platform whereas complex analysis of structured data would be a candidate for an analytical RDBMS. Organisations should match the workload to the appropriate platform rather than expect all data to come into an enterprise data warehouse.

SKILL SETS

Data Scientists are new people that need to be recruited

With respect to skills, new skill sets are emerging. The new skill set is that of a data scientist. Data scientists are needed for investigative analytical projects while traditional data warehouse developers and business analysts are needed for data warehousing and BI. All need to work together in this new analytical environment.

Data Scientists are self-motivated analytically inquisitive people with a strong mathmatical background and a thirst for data

Data scientists tend to have strong backgrounds in statistical and mathematical modelling techniques and are capable of using programming languages to explore, analyse and visualise multi-structured data. For example, in Oil and Gas they may be seismic engineers who are used to mining, filtering, analysing and visualising large amounts of complex data. They need the freedom to analyse un-modelled data and to produce insights that can be used to enrich data in traditional environments.

Traditional ETL developers and business analysts need to broaden their skills to embrace big data platforms as well as data warehouses

Traditional data warehouse ETL developers need to broaden their use of information management tools to load data into Hadoop environments for data scientists to analyse and to take data from Hadoop into data warehouses that data scientists have signalled as being of business value. In addition, business analysts should broaden their use of BI tools to exploit the Hive interface to access Hadoop data as well data in data warehouses.

CREATE AN ENVIRONMENT FOR DATA SCIENCE AND EXPLORATION

Governed sandboxes are needed by data scientists wishing to conduct investigative analysis on big data

Data scientists need a governed environment where they can explore un-modelled data and/or conduct complex analyses on large amounts of structured data. Creating a project environment where small teams of data scientists can work and collaborate in 'sandboxes' on Hadoop and/or analytical RDBMS appliances is an important step. Sandbox creation and access needs to be controlled. Also data scientists need the ability to search for content to analyse and quickly highlight the data sources needed. Data going into and coming out of sandboxes then needs to be governed by using information management tools to load data into sandboxes and to extract data from sandboxes. Also tools need to be made available to build custom MapReduce applications and/or mine data to discover insights and to develop analytical models. These can then be deployed in either stream processing to analyse data in motion or in analytical RDBMSs, Hadoop, and data warehouses to analyse data at rest.

DEFINE ANALYTICAL PATTERNS AND WORKFLOWS

Event stream processing and Hadoop based analytics are often upstream from data warehouses

Use big data Insights to enrich data in a data warehouse

Patterns need to be defined to maximise business benefits of this new analytical ecosystem. Obvious ones include positioning event stream processing and Hadoop as analytical systems that are 'upstream' from data warehouses. This means that filtered and analyzed streaming data then needs to flow from stream processing systems into Hadoop, NoSQL and/or data warehouses for further analysis over time. Also insights produced from exploratory analysis in Hadoop or NoSQL graph databases needs to be brought into data warehouses to enrich what is already known.

INTEGRATE TECHNOLOGY TO TRANSITION TO THE BIG DATA ENTERPRISE

Technology needs to be extended and integrated to allow your organisation to create an extended analytical ecosystem that encompasses Big Data and traditional DW/BI analytical workloads. These are listed below:

- Add new analytical platform(s) and stream processing that support big data analytical workloads to your traditional environment as and when there is a business need to do so
- Extend the use of information management tools beyond that of data warehouses and data marts to:
  - Supply data to Hadoop, NoSQL graph databases, analytical RDBMS appliances, data warehouses and MDM systems
  - Move data between Hadoop and data warehouses and/or analytical RDBMS appliances
  - Move data from MDM systems to Hadoop, data warehouses and/or analytical RDBMS appliances
  - Integrate with stream processing to load filtered or analyzed streaming data into Hadoop and/or analytical RDBMSs
- Integrate predictive analytics with Hadoop, DW appliances and streaming engines to deploy models in workload optimized systems
- If possible integrate existing BI tools with Hadoop Hive
- Integrate search based BI tools with Hadoop and data warehouses
- Introduce data virtualization on top of all analytical systems
- Integrate front end tools into portals to create a single user interface

Technologies need to be added to and integrated with traditional data warehouse environments to create a new enterprise analytical exosystem that caters for all analytical workloads

VENDOR EXAMPLE: IBM'S END-TO-END SOLUTION FOR BIG DATA

Having defined what big data is, looked at big data analytical workloads and looked at technologies and requirements needed for the new extended enterprise analytical ecosystem, this section looks how one vendor, IBM, is stepping up to the challenge of delivering the technologies and integration needed to make this possible. In other words to enable organisations to support traditional, operational and big data analytical workloads for end-to-end analytical processing.

IBM provides a range of technology components for end-to-end analytics on data in motion and data at rest

IBM provides a number of integrated technology components for end-to-end analytics on data in motion and data at rest. These components include:

- A stream processing engine for real-time analysis of data in motion
- A data warehouse platform supporting traditional analysis and reporting on structured data at rest

This includes a data warehouse and a range of analytical appliances

- A range of analytical appliances optimised for specific advanced analytical workloads on big data
- An appliance for accelerating operational analytic query processing
- An integrated suite of self-service BI tools for ad hoc analysis and reporting including support for mobile BI
- Search based technology for building analytic applications offering freeform exploratory analysis of multi-structured and structured data
- Predictive analytics for model development and decision management
- Applications and tools for content analytics
- Pre-built templates to quick start analytical processing of popular big data sources

Information management tools to goven and manage data

- A suite of integrated information management tools to govern and manage data in this new extended analytical environment

All of these components are included in the IBM Big Data Platform

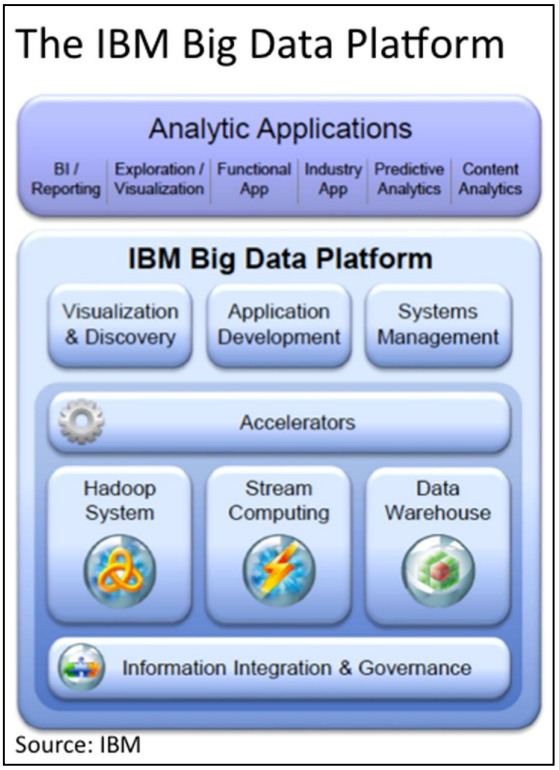
Together, this set of technologies constitutes the IBM Big Data Platform as shown below. This platform includes three analytical engines to support the broad spectrum of traditional and big data analytical workloads needed by most enterprises. These are:

Three analytical engines in the IBM Big Data Platform

- Stream computing
- A Hadoop System
- Data Warehouse (could be one or more data stores)

The platform is also extensible and can support additional third party analytical data stores e.g. non-IBM analytical RDBMSs and NoSQL data stores.

The IBM Big Data Platform is IBM's name for the enterprise analytical ecosystem



IBM INFOSPHERE STREAMS – ANALYSING BIG DATA IN MOTION

IBM InfoSphere Streams offers continuous real-time analysis of data-in-motion

IBM InfoSphere Streams is the IBM Big Data Platform technology component for building and deploying continuous real-time analytic applications that analyse data in motion. These applications continuously look for patterns in data streams. When detected, their impact is analysed and instant real-time decisions made for competitive advantage. Examples include analysis of financial market trading behaviour, analysis of RFID data for supply and distribution chain optimisation, monitoring sensor data for manufacturing process control, neonatal ICU monitoring, real-time fraud prevention and real-time multi-modal surveillance in law enforcement. IBM InfoSphere Streams can simultaneously monitor multiple streams of external and internal events whether they are machine generated or human generated. High volume structured and unstructured streaming data sources are supported including text, images, audio, voice, VoIP, video, web traffic, email, geospatial, GPS data, financial transaction data, satellite data, sensors, and any other type of digital information.

IBM InfoSphere Streams offers ships with pre-built toolkits and connectors to expedite developmemt of real-time analytic apliactuions

To help expedite real-time analytic application development, IBM also ships with pre-built analytical toolkits and connectors for popular data sources. Third party analytic libraries are also available from IBM partners. In addition, an Eclipse based integrated development environment (IDE) is included to allow organisations to build their own custom built real-time analytic applications for stream processing. It is also possible to embed IBM SPSS predictive models or analytic decision management models in InfoSphere Streams analytic application workflows to predict business impact of event patterns.

Scalability is provided by deploying InfoSphere Streams applications on multi-core, multi-processor hardware clusters that are optimised for real-time analytics. Events of interest to the business can also be filtered out and pumped to other analytical data stores in the IBM Big Data platform for further

IBM InfoSphere Streams can be used to continually ingest data into IBM BigInsights Hadoop system for further analysis

analysis and/or replay. InfoSphere Streams can therefore be used to continually ingest data of interest into IBM BigInsights to analyse. It is also possible to summarize high volume data streams and route these to IBM Cognos Real-Time Monitoring for visualization in a dashboard for further human analysis. .

IBM APPLIANCES FOR ANALYSING DATA AT REST

In addition to IBM InfoSphere Streams, the IBM Big Data Platform also includes a traditional data warehouse, appliances optimised for big data analytics, advanced analytics on structured data and for accelerating analytical query processing on operational transaction data. These are as follows:

IBM InfoSphere BigInsights

IBM InfoSphere BigInsights is IBM's commercial distribution of Hadoop

IBM InfoSphere BigInsights is IBM's commercial distribution of the Apache Hadoop system. It has been designed for exploratory analysis of large volumes of multi-structured data to gain insights that were not previously possible. IBM InfoSphere BigInsights ships with standard Apache Hadoop software. However IBM has strengthened this by adding:

A lot has been done to enhance Hadoop to make it more robust

- An enterprise scalable, Posix compliant file system GPFS-SNC [4]
- JSON Query Language (JAQL) to support easy manipulation and analysis of semi-structured JSON data
- Data compression
- Map/reduce based text and machine learning analytics
- Storage security and cluster management
- Support for Cloudera's distribution of Hadoop in addition to IBM's
- Connectors to IBM DB2, IBM's PureData Systems for Operational Analytics (DB2) and for Analytics (Netezza) to access structured data during big data analyses from JAQL based MapReduce applications
- Job scheduling and workflow management
- BigIndex – a MapReduce facility that leverages the power of Hadoop to build indexes for search based analytic applications

IBM InfoSphere BigInsights can support 3rd party Hadoop distributions as well as IBM's own

BigInsights is available in two editions:

- IBM BigInsights Basic Edition – a pre-configured free download trial version of BigInsights
- IBM BigInsights Enterprise Edition

BigInsights Enterprise Edition is available on a pre-configured IBM System x and PowerLinux reference architecture.

The analytic systems in the IBM Big Data Platform are part of a new family of systems called PureSystem. The PureSystem family of systems which debuted earlier this year, are built on a set of core principles that focus on integrated expertise, simplified management and optimized performance. In October 2012, IBM extended this family to include systems that exclusively focus on data, called PureData Systems. Two systems were introduced – PureData System for Analytics and PureData System for Operational Analytics discussed further below.

---

[4] SNC = Share Nothing Cluster

IBM PureData System for Analytics (powered by Netezza technology)

IBM PureData System for Analytics powered by Netezza technology is the next generation Netezza Appliance optimised for advanced analytical workloads on structured data.

IBM PureData System for Analytics is optimised for anvanced analytics on structured data and for some data warehouse workloads

IBM PureData System for Analytics is a compact, low cost data warehouse and analytic hardware appliance. It is scalable from 100 gigabytes to 10 terabytes of user data capacity and is fully compatible with the IBM Netezza 1000 and IBM Netezza High Capacity Appliance.

The PureData System for Analytics is a purpose-built, standards-based data warehouse appliance that integrates database, server, storage and advanced analytic capabilities into a single system. It scales from 1 TB to 1.5 petabytes includes special processors to filter data as it comes off disk so that only data relevant to a query is processed in the RDBMS. The IBM Netezza Analytic RDBMS known as Netezza Platform Software (NPS) requires no indexing or tuning which makes it easier to manage. It is designed to interface with traditional BI tools including IBM Cognos BI platform and also runs IBM SPSS developed advanced analytical models deployed in the database on large volumes of data.

IBM Netezza Analytics provides in-database analytics capabilities and it comes free in every IBM Netezza 1000 or PureData System for Analytics allowing you to create and apply complex and sophisticated analtyics right inside the appliance.

Complementing the IBM PureData System for Analytics, is IBM Netezza Analytics, an advanced analytics framework. In addition to providing a large library of parallelized advanced and predictive algorithms, it allows creation of custom analytics created in a number of different programming languages (including C, C++, Java, Perl, Python, Lua, R*, and even Fortran) and it allows integration of leading third party analytic software offerings from companies like SAS, SPSS, Revolution Analytics, Fuzzy Logix, and Zementis.

** With Revolution R Enterprise software from Revolution Analytics*

IBM PureData System for Analytics allows you to create, test and apply models to score data right inside the IBM Netezza appliance, eliminating the need to move data and giving you access to more of the data and more attributes than you might otherwise be able to use if you needed to extract the data to a laptop or other small computer.

IBM PureData System for Operational Analytics

IBM PureData System for Operational Analytics is a modular pre-integrated platform optimized for operational analytic data workloads

The IBM PureData System for Operational Analytics was introduced in October and is based on IBM Power System and the IBM Smart Analytics System. The IBM Smart Analytics System is a modular, pre-integrated real-time Enterprise Data Warehouse optimized for operational analytic data workloads available on IBM System x, IBM Power System or IBM System z servers. These systems are pre-integrated and optimized to ensure quick deployment and deliver fast time to value. Both the IBM PureData System for Operational Analytics and the IBM Smart Analytics System family include IBM InfoSphere Warehouse 10 software running on DB2 Enterprise Server Edition 10. InfoSphere Warehouse includes data modeling, data movement and transformation, OLAP functions, in-database data mining, text analytics and integrated workload management. DB2 10 includes improved workload management, table level, page level and archive log compression, new index scanning and pre-fetching, temporal data access and a new NoSQL Graph store. Notable improvements have also been made in automated optimized data placement capabilities leveraging Solid State Disk (SSD) inside the new PureData System solution. IBM Cognos Business Intelligence is also available.

DB2 10 includes a NoSQL Graph store

IBM Big Data Platform Accelerators

IBM Big Data Accelerators are designed to speed up development on the IBM Big Data Platform

In order to expedite and simplify development on the IBM Big Data Platform, IBM has built a number of accelerators. These include over 100 sample applications, user defined toolkits, standard toolkits, industry accelerators and analytic accelerators. Examples include:

- Data mining analytics
- Real-time optimization and streaming analytics
- Video analytics
- Accelerators for banking, insurance, retail, telco and public transport
- Pre-built Industry Data Models
- Social Media Analytics
- Sentiment Analytics

IBM DB2 Analytic Accelerator (IDAA)

IBM DB2 Analytics Accelerator offloads complex analytical queries from OLTP systems running DB2 mixed workloads on IBM System z

IBM DB2 Analytics Accelerator is an IBM Netezza 1000™ and/or PureData System for Analytics appliance specifically designed to offload complex analytical queries from DB2 mixed workloads on IBM System z. This is done by re-creating DB2 tables on IDAA using pre-defined administrative DB2 stored procedures and then loading the data from DB2 into IDAA. If necessary, DB2 tables can be locked to prevent update during IDAA loading. Also queries can be routed and processed by IDAA while loading occurs. No change is required to any applications or tools accessing DB2. This is because it is the DB2 optimizer that decides which dynamic SQL queries to re-route to the IBM DB2 Analytics Accelerator for parallel query processing. To all intents and purposes, IDAA is therefore "invisible" to the applications and reporting tools querying the DB2 DBMS on IBM System z. In addition, the overhead in terms of database administration is minimal given that Netezza technology does not have any indexes and that all administrative activity is via pre-built DB2 stored procedures. The result is that capacity upgrades can be avoided and service levels are improved.

There are currently three IDAA offerings available supporting 8, 16 and 32 Terabytes of user data. This can be increased with data compression.

IBM Information Management for the Big Data Enterprise

IBM InfoSphere Information Server and Foundation Tools provide end-to-end data management across all data stores in the IBM Big Data Platform

The IBM Information Integration and Governance platform provides an integrated suite of tools for integrating, governing and managing data. It includes tools for ensuring high quality information, mastering data into a single view, governing data throughout its lifecycle, protecting and securing information, integrating all data into a common view, and ensuring a single understanding and set of knowledge. IBM InfoSphere Information Server supports connectivity to IBM InfoSphere BigInsights, IBM PureData System for Operational Analytics and IBM PureData System for Analytics data warehouse appliances, and IBM DB2 Analytics Accelerator. IT also integrates with IBM InfoSphere Streams to pump filtered event data into IBM InfoSphereBigInsights for further analysis.

IBM uses InfoSphere Blueprint Director to create smart workflows that govern data cleansing, data integration, data privacy and data movement

IBM has used InfoSphere Information Server as the foundation for *Smart Consolidation* on the IBM Big Data Platform. Smart Consolidation uses IBM InfoSphere Blueprint Director to build and run workflows that leverage services on the InfoSphere Information Server to clean, integrate, protect and distribute data to the appropriate analytical data store best suited for an analytical workload. The purpose of Smart Consolidation is to increase agility, move and integrate data both in batch and in real-time, and to create a framework for integrated data management to govern and manage data across all analytical

Data virtualization is also included in IBM's Information Integration and Governance platform

data stores in the IBM Big Data Platform. This hides complexity, increases automation and opens up the way for workload routing and on-demand dynamic workload optimisation whereby data is moved in real-time as part of an optimisation plan in response to in-bound queries on the IBM Big Data Platform.

IBM Analytical Tools For The Big Data Enterprise

IBM BigSheets

BigSheets enables business users to analyse data in Hadoop

BigSheets can import data into BigInsights Hadoop from internal and external data sources

IBM BigSheets is a web based user interface facility based on the spreadsheet paradigm that allows line of business users to make use of Hadoop to analyse structured and unstructured data without the need of skilled IT professionals. It allows users to import data into BigInsights by crawling websites, selecting data from internal servers and desktops or by using custom importers to pull selected data from specific data sources like Twitter for example. Data imported into BigInsights can be visualised though the BigSheets spreadsheet user interface from where the user can filter and analyse the data using pre-built and custom built analytics that are based on Hadoop MapReduce. Results of analyses can be fed into other worksheets to zoom in on insights. A good example would be the analysis of tweets to determine lead opportunities based on derived positive sentiment. Given that spreadsheets often make it difficult to visualise insights, IBM has integrated Many Eyes into BigSheets to provide users with the ability to graphically visualise insights. Third part visualisations can also be supported via BigSheets plug-in capability.

IBM Cognos 10

Many Eyes is also included to improve visualisation

IBM has integrated its Cognos BI tool suite with BigInsights via Hive and also with IBM Netezza and IBM Smart Analytics System now a part of the PureData Systems family

IBM Cognos 10 is IBM's flagship BI platform. It is an integrated system for ad hoc reporting and analysis, dashboard creation, scorecarding, production reporting, multi-dimensional analysis, budgeting, planning and forecasting. It is used to access structured data housed in IBM and non-IBM data warehouses and data marts, including IBM PureData System for Analytics (powered by Netezza) and IBM PureData System for Operational Analytics as well as non-IBM data warehouse platforms.

The IBM Cognos family of products provide starting points for an organization in an integrated, scalable approach:

- IBM Cognos Insight (personal, desktop analytics)
- IBM Cognos Express (workgroup BI)
- IBM Cognos Enterprise

IBM Cognos can be extended to mobile environments as a native application.

IBM Cognos RTM can analyse filtered event data fed to it by InfoSphere Streams for real-time exception monitoring

With respect to Big Data, the IBM Cognos BI Platform has been extended to enable reporting on MapReduce relational databases such as EMC GreenPlum, Teradata Aster and others via ODBC and on IBM InfoSphere BigInsights via a Hive adaptor.

IBM Cognos also offers **IBM Cognos Real-time Monitoring** (RTM) to monitor and visualize business events, in real-time, in order for business users to make informed decisions when immediate action is required.

IBM Cognos RTM fits into the Big Data story as it can monitor filtered events routed to it from IBM InfoSphere Streams.

Real-time business insights can be integrated into operational and managerial dashboards alongside historical and predictive intelligence

IBM Cognos RTM and IBM Cognos Enterprise integration provides real-time awareness to operational and managerial dashboards on personalised BI workspaces. This integration enables historical, real-time and predictive information to be seen at a glance, in a single user interface. Watch points and thresholds can also be defined by users when they want to be alerted in real-time.

IBM Cognos Consumer Insight (CCI)

IBM Cognos Consumer Insight is a purpose built social media analytic application that uses IBM's BigIndex to analyse consumer and customer interaction data

IBM Cognos Consumer Insight is a purpose-built social media analytic application that analyses large volumes of content collected from public websites and customer interactions stored in internal databases. This application leverages the search and text analytics capabilities of IBM InfoSphere BigInsights to analyze unstructured social media data. The purpose is to derive sentiment, word affinity and other insights about social behaviour. IBM Cognos Consumer Insight includes pre-built reports that integrate with IBM Cognos BI platform. The analytics produced by IBM Cognos Consumer Insight can also be incorporated into IBM SPSS predictive models to determine what action(s) to take in response to specific customer behaviour and intentions.

IBM SPSS

IBM SPSS is used to build and deploy advanced analytics in the IBM Big Data Platform

IBM SPSS is IBM's tool suite for building and deploying advanced analytics and for developing automated decision management applications. Using IBM SPSS, power users can design and build predictive and statistical models that automatically analyse data. These models can be deployed in

- IBM InfoSphere Streams applications to analyse big data in motion
- IBM PureData System for Analytics (powered by Netezza technology) and the Netezza Platform Software for in-database advanced analytics on structured data at rest
- IBM InfoSphere Warehouse DB2 DBMS on the IBM PureData System for Operational Analytics for in-database advanced analytics and operational BI.

IBM is extending SPSS to exploit map/reduce to scale analytics on BigInsights

IBM's direction will make SPSS developed predictive analytics available in IBM InfoSphere BigInsights alongside the Hadoop Mahout library of advanced analytics, to automatically analyse large volumes of multi-structured data in Hadoop HDFS and Hive.

IBM Vivisimo

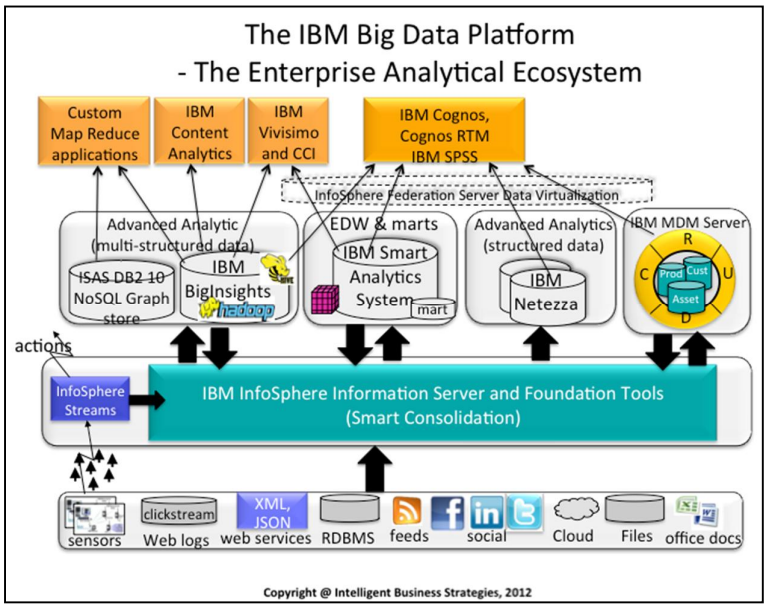
IBM Vivisimo is a search based platform for developing free-form analytic applications that analyse data on the Big Data Platform

The latest addition to the IBM Big Data Platform is IBM Vivisimo. This product speeds up time to value by allowing organisations to federate exploration and navigation of large amounts of structured and unstructured data while leaving the data where it is in multiple data sources. This is achieved by crawling structured and unstructured big data sources to build search indexes. IBM Vivisimo also provides tooling to develop applications on top of these indexes so that organisations can conduct exploratory analysis and facet-based navigation to discover valuable insights. It is particularly useful when the solution to a problem is not clear and when there is a need to explore and navigate multi-structured data freely without restriction. It is also useful for ad-hoc analysis in business operations when a response to unique problem is needed quickly and no reports or data structures have as yet been developed.

How They Fit Together For End-to-end Business Insight

The following chart shows how all these components fit together.

All of these come together to extend traditional data warehouse environments to create an enterprise analytical ecosystem that processes traditional and big data workloads



CONCLUSION

Business is now demanding more analytical power to analyse new sources of structured and multi-structured data

Many businesses are now at the point where there is an insatiable appetite among business users to want to analyse new sources of structured and multi-structured data to produce new insights for competitive advantage. The data itself is increasing in complexity in terms of the rate that it is being created, the data volumes involved and the new types of data that need to be analysed. Analytical complexity has also increased and spawned the need for new analytical techniques and tools. Also data quality is becoming an issue on new data sources like social interaction data.

New technologies have emerged to support specific analytical workloads

Traditional data warehouse environments now need to be extended to accommodate these new big data analytical workloads

To respond to this added complexity, vendors are building new analytical platforms specifically designed for certain analytical workloads. Most organisations recognise that bringing all this new data into an enterprise data warehouse is not the right approach. New big data workloads need to be identified and the appropriate analytical platforms selected. Traditional BI environments then need to be extended to accommodate an enterprise data warehouse *as well as* these new workload optimised analytical systems while still providing consistent data to all analytical data stores. In addition there is a need to shield users from this more complex analytical environment by making it easy to access data in multiple analytical systems. Queries need to be automatically directed to the appropriate system best suited for the analytical workload and data easily moved between analytical systems to process it in the best place. In addition, organisations also need to integrate user interfaces to make it easy to consume insight and see it all at a glance.

The challenge now is to find a technology partner that can deliver a new analytical ecosystem that supports the entire spectrum of traditional and big data analytical workloads now needed by business users. In addition, it all has to be done while keeping data governed and everything integrated.

The IBM Big Data Platform rises to the challenge to create this new analytical environment

Looking at the marketplace, IBM has already identified this requirement. Its Big Data Platform is a comprehensive offering that includes support for the entire spectrum of analytical workloads. It includes a strong IBM InfoSphere BigInsights Hadoop offering, the IBM Smart Analytic System or IBM PureData System for Operational Analytics running the InfoSphere Warehouse with a built-in NoSQL graph store, IBM PureData System for Analytics powered by Netezza technology for complex analytics on structured data, InfoSphere Streams for real-time analytics on data in motion and IBM DB2 Analytics Accelerator (IDAA) for accelerating operational analytics. In addition the InfoSphere information management tools suite underpins the entire IBM Big Data Platform providing consistent trusted data to all analytical data stores *and* IBM MDM Server. It also supports data virtualisation to simplify access to data. InfoSphere Blueprint Director has the ability to move data between all platforms and invoke in-database analytics during workflow execution. InfoSphere Streams can push events to IBM Cognos RTM which itself can integrate with IBM Cognos 10 to see real-time insights. IBM Cognos also connects to BigInsights and all relational analytical data stores. IBM SPSS can push analytics down into the PureData System for Analytics and the PureData System for Operational Analytics and InfoSphere Streams for big data performance and IBM Vivisimo supports free-form search-based analytics. All of this makes IBM a serious contender to be on any shortlist in any Big Data competitive environment.

The IBM Big Data Platform makes IBM a serious contender to support end-to-end analytical workloads

About Intelligent Business Strategies

Intelligent Business Strategies is a research and consulting company whose goal is to help companies understand and exploit new developments in business intelligence, analytical processing, data management and enterprise business integration. Together, these technologies help an organisation become an *intelligent business*.

Author

Mike Ferguson is Managing Director of Intelligent Business Strategies Limited. As an analyst and consultant he specialises in business intelligence and enterprise business integration. With over 31 years of IT experience, Mike has consulted for dozens of companies on business intelligence strategy, big data, data governance, master data management, enterprise architecture, and SOA. He has spoken at events all over the world and written numerous articles. He has written many articles, and blogs providing insights on the industry. Formerly he was a principal and co-founder of Codd and Date Europe Limited – the inventors of the Relational Model, a Chief Architect at Teradata on the Teradata DBMS and European Managing Director of Database Associates, an independent analyst organisation. He teaches popular master classes in Big Data Analytics, New Technologies for Business Intelligence and Data Warehousing, Enterprise Data Governance, Master Data Management, and Enterprise Business Integration.

Water Lane, Wilmslow
Cheshire, SK9 5BG
England
Telephone: (+44)1625 520700
Internet URL: www.intelligentbusiness.biz
E-mail: info@intelligentbusiness.biz

Architecting a Big Data Platform for Analytics