WHITE PAPER

INTELLIGENT
BUSINESS
STRATEGIES

Transforming Big Data Into Insights in an IBM System z Environment

By Mike Ferguson
Intelligent Business Strategies
October 2013

Prepared for: IBM

Table of Contents

Copyright © Intelligent Business Strategies Limited, 2013, All Rights Reserved

INTRODUCTION

Customer growth and improving operational effectiveness are top of the agenda for CEOs

In a recent survey of Chief Executive Officers (CEOs)[1] 1,330 CEO respondents indicated that their top two major priorities going forward were:

- Customer retention, loyalty and growth
- Operational effectiveness

Additional customer insight is now needed to remain competitive

The impact of competition on the web has been profound

You might argue that traditional data warehouses have been at the centre of these two priorities for a long time. However when you look at competition today, there is no doubt that additional insight is now needed if these two priorities are to be realised. In fact, it is fair to say that we are at the dawn of an Information-Age, where new forms of data being created now hold the key to competitive advantage. These new forms of data contain high value insights that every company must learn how to discover if they are to survive in a market where the web is continuously changing customer behaviour and loyalty is easily forfeited in favour of a better deal.

On the web the customer is king. They can surf around comparing products, services and prices any time anywhere all from a mobile device. In addition, with new 'lighter weight' web-based companies springing up everywhere, the customer has much more choice and can easily switch loyalty with a few clicks of a mouse. The impact of the web has been profound. In some cases, on-line competition has even resulted in whole industries being "re-wired" e.g. travel, accommodation and insurance. Given that this is the case, we are now at the point where transaction data is not enough to provide comprehensive customer insight. We also need to understand on-line customer behaviour as well.

Transaction data on its own is no longer enough to provide comprehensive customer insight

Click stream and social media data are also needed

The web is also a place where the customer has a voice. Access to social networks such as Facebook and Twitter is now ubiquitous and with so many social network members, companies would be foolish not to recognize that they are a rich source of customer insight. In addition, when it comes to buying decisions, most people trust their network of friends and contacts more than anyone else and so understanding social relationships, interactions and influencers can help provide insight into who or what is influencing purchasing behaviour. Also people are quick to compare products and prices and share this information with others across these networks making potential buyers better informed. Review web sites also provide customers with ways to rate products and services giving prospective buyers ways to find reassurance before making purchasing decisions. Also, if negative sentiment is voiced on social networks it can spread like wildfire. All of this can impact on customer loyalty, retention and growth. Therefore, in addition to needing to analyse clickstream data to understand on-line behaviour, there is also a need to understand social networks and social interactions to determine sentiment, highlight influencers and to determine the overall potential value of customers' social networks to the business.

[1] PricewaterhouseCoopers 16th Annual Global CEO Survey 2013

Copyright © Intelligent Business Strategies Limited, 2013, All Rights Reserved

Integrating BI and analytics into business processes can improve operational effectiveness

With regards to operational effectiveness, it has long been the case that operations have been starved of actionable insight than can be used to improve 'on the ground' every day decision-making. Yet business executives expect business strategy execution to occur in all parts of the business including operations because strategic and tactical decision-making is not enough. Operational effectiveness can be improved by integrating existing and big data insights into operational business processes to provide timely on-demand access to actionable insight and recommendations. There is also a need to automatically analyse events to drive automated decision-making to improve responsiveness.

This paper looks at how Big Data can be used to improve customer insight as well as operational effectiveness. It looks at the data and analytical requirements needed and at problems with traditional analytical environments in dealing with them. It then defines new requirements to accommodate Big Data analytical workloads and shows how your analytical Infrastructure needs to change to accommodate these extraordinary requirements and deal with the diversity of today's data as well as the new forms that have yet to be created. Finally, it looks at how this can be done in an IBM System z environment where the vast majority of transactional data is often created and housed.

Achieving Deeper Insight Using Big Data Analytics

In order to achieve deeper insights many companies are now looking at new data sources to remain competitive. When looking at the candidate data available, it is not long before you realise that there is a treasure trove of untapped high value data available from within the enterprise and from external sources.

Data Requirements For Growth and Operational Effectiveness

Customer Loyalty, Retention and Growth

New data sources are needed to deepen customer insight

Data requirements associated with deepening customer insight to improve growth has seen a leap in demand for behaviour and social data. Behavioural and social insight takes companies beyond just having customer demographics typically seen in traditional data warehouses. In particular, inbound and outbound interaction data needs to be collected from customer touch points including those that are not owned by the business. Interaction data to be collected includes data from touch points such as:

Companies now need to collect data from every interaction point

- Direct mail
- Customer service
- Call centres
- Email
- SMS/ MMS
- Mobile devices
- Website logs
- Social networks
- Review web sites
- Blogs
- Search
- Web events
- Physical events
- In-store Point-of-Sale terminals
- Kiosks
- Online advertising sites

Companies now need to analyse strucured and multi-structured data to truly understand prospect and customer behaviour

It is not until you look at this that you realise that there is a lot of untapped data that organisations can potentially analyse over and above transaction data and demographics acquired from OLTP systems that currently finds its way into data warehouses. In particular there is very high demand to analyse click stream data to provide insight into on-line behaviour. These new data sources are critical to meeting the needs of chief marketing officers (CMOs) who need to visualise how customers and prospects are interacting with their business. Understanding the time series “interaction fingerprint” of each and every customer across all touch points as well as preferred touch points, provides the deeper insight needed to make more personalised offers and create better customer service.

Copyright © Intelligent Business Strategies Limited, 2013, All Rights Reserved

Social network data is another new data source that is high priority to analyse. Adding this kind of insight to what is already known about customers opens up the opportunity to establish personalised "systems of engagement" at every single customer touch point. Integrating these insights into operational business processes can help build customer relationships, improve customer satisfaction, build loyalty and increase retention. Deeper customer insight also means smaller more accurate customer segments and more personalised marketing campaigns to drive growth.

Operational Effectiveness

Many companies also lack insight into business operations

In addition to deepening customer insight, many companies lack insight in the area of operations mainly because:

- Analytical queries on live transaction data is often discouraged (e.g. for performance reasons)
- Insights and recommendations produced are not integrated into operational business processes
- Data is simply not being collected to monitor operational activity

Today, business requirements dictate that all of these are needed to improve operational effectiveness. Lack of data about what is happening in operations is perhaps the most urgent requirement.

Many companies are now intrumenting business operations to understand and monitor business activity

To address this, organisations are starting to:

- Instrument business operations and grids using sensor networks
- Embed sensors in so called 'smart' products (e.g. GPS sensors in smart phones, consumption sensors in smart meters)
- Monitor live transaction activity
- Monitor live market activity (e.g. financial and energy markets)
- Monitor events during business process execution

This extends data requirements to include the capturing, transforming and integrating of:

- Transaction data (which is increasingly high volume)
- Real-time streaming data from sensors, smart devices and markets
- Events occurring during operational business process execution
- Web feeds e.g. news, weather, etc.
- Rich media streams e.g. video streams

Being able to detect events and patterns means companies can improve responsiveness, reduce risk and optimise business operations

By capturing and analysing this kind of data, companies can monitor their business operations as it happens to detect patterns, predict outcomes and, based on analytical results, drive actions to improve responsiveness, reduce risk and keep business operations optimised. Examples of this include fraud detection, field service optimization and supply chain optimization. However it is not just about event-driven analysis of big data. On-demand access to the insight produced is also a critical success factor, which means that tight integration between transaction processing and big data analytical systems is critical to maximising the business value.

Data Challenges

These new big data sources bring new challenges. First, there is data variety in the form of complex data types as many of the aforementioned new data sources are:

New types of data are being captured

- Semi-structured e.g. email, e-forms, HTML, XML

- Unstructured e.g. document collections (text), social interactions (text), images, video and sound
- Machine generated e.g. web logs, sensor data

Much of this data is un-modelled, challenging to analyse and quality can be poor

These new more complex data types, often referred to as multi-structured data, are harder to analyse than structured data. Analysis is made doubly difficult because much of this data is also un-modelled. Furthermore, multi-structured social media data may also contain things like emoticons (e.g. :-) :-< :0)), twitter hash tags (e.g. #bigdata), "Yoda speak", slang, abbreviations, sarcasm, spam as well as being written in multiple languages. It may also be poor quality (e.g. speling erors).

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

These problems taken together make multi-structured data a real challenge which is why it is important to quickly identify issues with the data and work out what needs to be done to prepare it for analysis. Therefore, a key requirement is to 'clean' the content before analysis takes place. Also, exploratory analysis has to be done on multi-structured data before any high value subset of it can be extracted and analysed to produce insights that can enrich data in traditional data warehouses.

The second challenge is data volume. Social media data and machine generated data (e.g. click stream, sensor data) can be very large in size requiring companies to be able to store, process and analyse hundreds of terabytes or even petabytes.

Finally there is the challenge of having to support fast capture and analysis of data that is being generated at very high rates - so called 'high velocity' data. Examples include real-time clickstream and streaming sensor data. Sensor networks in particular can emit data at extremely high rates. As an example, 1,000 sensors emitting at 50Hz frequency produces 50,000 events per second. Organisations need scalable solutions to handle this.

Big Data Integration Requirements

In addition to dealing with the variety, volume and velocity issues that big data brings, it is often not enough to just analyse big data on its own. Analysis may need to be done in a business context. For example, to analyse social sentiment data it is highly likely that multi-structured big data may also have to be integrated with product and perhaps customer master data. This is necessary to associate positive, negative or neutral sentiment with specific products and services and with specific customers. This means that multi-structured data may need to be cleaned and integrated with structured data in preparation for analysis. It also means that sensitive data such as customer data needs to be secured and protected in big data analytical environments.

ANALYTICAL REQUIREMENTS FOR GROWTH AND OPERATIONAL EFFECTIVENESS

Big data imposes new analytical requirements

The characteristics of Big Data, create a unique set of requirements that need to be met to analyse it. These include:

- Support for the following big data analytical workloads over and above that supported in a traditional data warehouse
  - Complex analysis of structured data
  - Analysis of data in motion

New analytical workloads exist

  - Exploratory analysis of un-modeled multi-structured data
  - Graph analytics
  - Accelerating ETL and analytical processing of un-modeled data to enrich data in a data warehouse or analytical appliance
  - The long term storage and selective re-processing of archived data (including data warehouse data)

Search technology is needed to index and analyse multi-structured data

- Using search technology to index and explore un-modelled, multi-structured data as early as possible without the need to develop code so that free-form exploration of newly loaded multi-structured data can take place quickly. In this way data scientists can quickly navigate the data to understand what needs to done to prepare it for analysis and what needs to be done to derive structure and produce insight

Text analytics is needed

- Using techniques such as text analytics to extract structure from unstructured data

Analysis of click stream is needed

- The need to "sessionize" click stream data from multiple web logs to prepare it for analysis so that the full sequence of individual on-line sessions are available for analysis. This provides the best chance of producing insight into on-line behaviour

Multi-platform analytics may be required to derive insight

- If necessary, utilise multiple analytical techniques across multiple analytical platforms to derive insight from multi-structured data. For example, both text analytics and graph analytics may be needed on social network data to determine sentiment and social network relationship insights needed to enrich what an organization knows about its customers

- 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. Therefore, it should be possible to automate the analysis of streaming data using a variety of analytic methods, such as predictive and statistical models to determine or predict the impact of events on business

Streaming analytics and decision management is required

- Automatically interpret the output of streaming analytics and automatically take decisions to improve operational effectiveness so that business operations remain optimized and on track to achieving its goals

- After taking action on high impact event patterns as a result of analysing streaming data, it should be possible to filter out events of interest into other analytical stores for further analysis

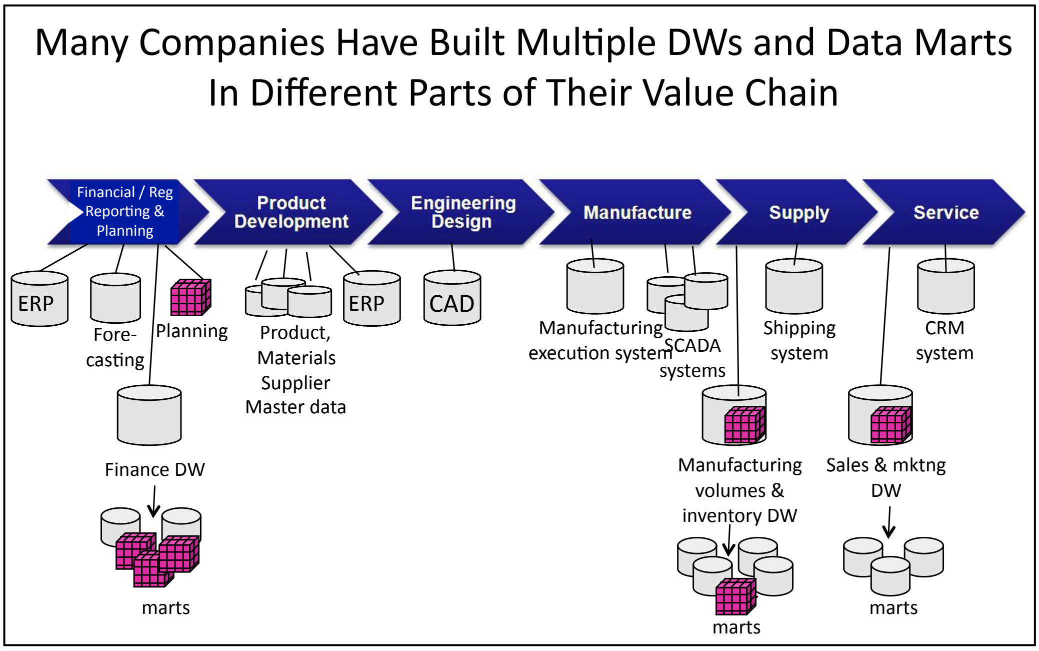
- It should be possible to offload complex analytical queries on transaction data to other analytical systems

Problems With Traditional Analytical Environments

Looking at what needs to be done to process and analyse big data the question is can the traditional data warehouse set-up cope with these new requirements? Let's take a look at the way the traditional environment has evolved to see if we can answer this question.

It is often the case that companies have built multiple data warehouses

Over the years, many people have written about a single enterprise data warehouse. However in reality, very few organisations have ended up with that kind of set-up. It is more often the case that companies have built multiple data warehouses and data marts in different parts of their value chain. This is shown in a manufacturing example in Figure 1.



Many Companies Have Built Multiple DWs and Data Marts In Different Parts of Their Value Chain

Figure 1

The siloed evolution of data warehouses has meant that business analysts now need to integrate data from multiple data stores to produce actionable insights

What Figure 1 shows is a siloed evolution of multiple data warehouses. Having historical data distributed like this makes it harder to produce management and regulatory reports that require data to be integrated to see across the value chain. Multiple copies of the same dimension data also exist and data management is often implemented separately for each data warehouse resulting in higher total cost of ownership. When you add data marts into the mix you can understand why multi-source data integration has become the norm for users of self-service BI tools. The siloed evolution of data warehouses, amid an increasingly distributed data landscape, has pushed the complexity of having to integrate data on to the business user using BI tools that were never designed to do that. This is complexity they could well do without. It also introduces the potential for each user to have to engage in personal data integration, thereby increasing the risk of data inconsistency that comes with it. Furthermore, this distributed data landscape makes it harder to integrate insights into operational business processes to improve customer loyalty, retention and growth as well as operational effectiveness – issues that are top priority at present.

Copyright © Intelligent Business Strategies Limited, 2013, All Rights Reserved

The arrival of Big Data has and new analytical workloads has meant that a new analytical set-up is needed in modern business

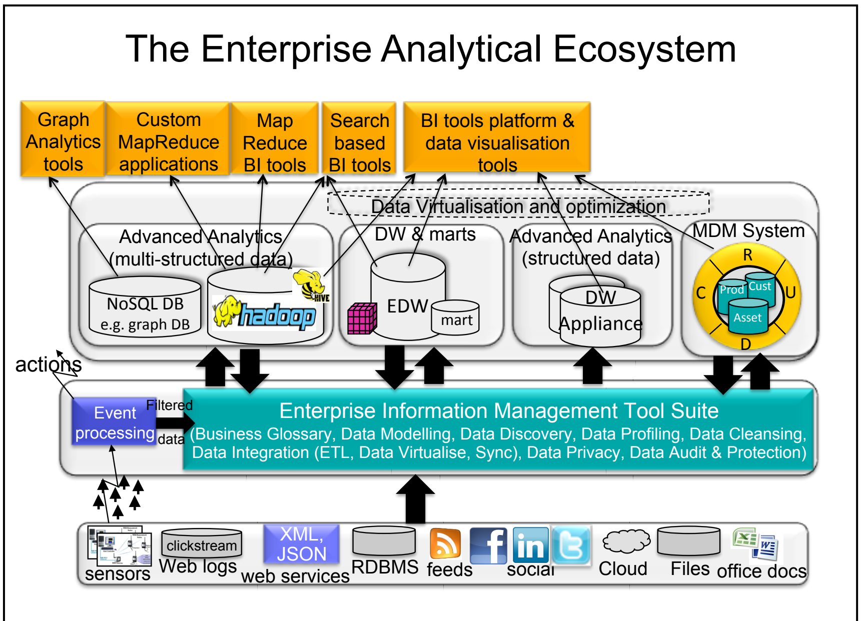
Given this typical set-up, what happens if Big Data is introduced into this landscape? How can a set-up of multiple siloed data warehouses cope with the increasing number of multi-structured data sources as well as the need to deal with data volume, data variety, and data velocity? What about the different analytical workloads that big data introduces such as stream analytics to analyse data in motion, exploratory analysis of multi-structured data, graph analytics and complex analysis of structured data? Can multiple traditional data warehouses support all the new types of data and analysis needed to manage the modern business?

The answer is undoubtedly no. A new more comprehensive and integrated set-up is now needed.

New Architectural Requirements For Competitive Advantage

Looking at the new types of data that businesses want to capture, together with the types of analytical workload they now need to implement to remain competitive, it is clear that a new architecture is needed. The architecture required[2] is shown in Figure 2. It represents an *enterprise analytical ecosystem* that supports traditional data warehouse ad hoc query processing, analysis and reporting, as well as the new big data analytical workloads now needed.

A new architecture to support new big data analytical workloads goes beyond that of the data warehouse


The Enterprise Analytical Ecosystem
Graph Analytics tools
Custom MapReduce applications
Map Reduce BI tools
Search based BI tools
BI tools platform & data visualisation tools
Data Virtualisation and optimization
Advanced Analytics (multi-structured data)
NoSQL DB e.g. graph DB
hadoop
HIVE
DW & marts
EDW
mart
Advanced Analytics (structured data)
DW Appliance
MDM System
R
C
U
D
Prod
Cust
Asset
actions
Event processing
Filtered data
Enterprise Information Management Tool Suite
(Business Glossary, Data Modelling, Data Discovery, Data Profiling, Data Cleansing, Data Integration (ETL, Data Virtualise, Sync), Data Privacy, Data Audit & Protection)
sensors
clickstream
Web logs
XML, JSON
web services
RDBMS
feeds
social
Cloud
Files
office docs

Copyright © Intelligent Business Strategies, 2013

Figure 2

The enterprise analytical ecosystem now has new platforms in addition to the data warehouse

The enterprise analytical ecosystem includes a number of new analytical platforms that integrate with a more centralised data warehouse environment. In addition some technology components have been extended to support these new big data analytical platforms. The technology components required in this new architecture are as follows (in bottom up order of appearance in Figure 2):

- A broad range of data sources including multi-structured and structured data from data sources inside and outside the enterprise

An enterprise information management tools suite needs to span all platforms in the analytical ecosystem

- An enterprise information management (EIM) tool suite.
  This underpins all analytical platforms in the enterprise analytical ecosystem that includes support for a multi-domain MDM system. The tools within this integrated EIM platform include:
  - A business glossary tool and big data catalog
  - A data modelling tool

[2] For a complete set of requirements please refer to “Architecting a Big Data Platform for Analytics”, Ferguson, October 2012 http://www.ibm.bigdatahub.com

- Data exploration software to index and search on newly loaded multi-structured big data prior to analysis
- Data and metadata relationship discovery software
- Data quality profiling software
- Data cleansing and matching software
- Data integration software data (a.k.a. an Extract, Transform and Load (ETL)) tool) to integrate data and store it in a target data store

Data virtualization to simplify access to data

- Data virtualization software to simplify access to and integrate data *at run time* from multiple analytical data stores in the enterprise analytical ecosystem
- Data privacy software – to protect sensitive data within the analytical ecosystem
- Data Audit and Protection software – to audit data platform activity and to ensure that data is not accessed by unauthorised users
- All tools in the EIM suite are integrated by shared access to a common metadata repository

The EIM tool suite needs to support:

- Loading data into Hadoop HDFS or Hive, graph DBMS, analytical RDBMS, data warehouse and master data management
- Ability to generate HiveQL, Pig or JAQL to exploit the power of massively parallel Hadoop clusters in processing data

EIM tool suites now need to integrate with NoSQL analytical platforms like Hadoop and Graph databases as well as data warehouses

- Ability to run in-database and in-Hadoop analytics in information management workflows for automated analysis during data transformation and data movement
- Integration with stream processing to take filtered event data and store it in Hadoop or analytical DBMS for further analysis
- Integration with a rules engine to support automated decisions during workflow execution
- Support for seamless data flows across multiple SQL and NoSQL data stores

- Multiple analytical platforms[3] that are integrated to manage big data and traditional analytical workloads. These are listed in the following table:

Multiple analytical data stores in addition to the enterprise data warehouse

Analytical Platform	Purpose
A NoSQL Graph DBMS	Graph analytics e.g. social network influencer analysis, fraud pattern analysis
Hadoop platform	Exploratory analysis of multi-structured data Data landing zone/staging area/ data refinery for processing raw multi-structured data en route to other analytical data stores

[3] For more information on these please refer to “Architecting a Big Data Platform for Analytics”, Ferguson, October 2012 http://www.ibm.bigdatahub.com

An Analytical Relational DBMS	Complex analysis of structured data e.g. for data mining to build predictive and statistical models
A Data Warehouse	Traditional query, analysis and reporting Integrated with other analytical platforms in the analytical ecosystem e.g. Hadoop and the NoSQL graph DBMS
An Event stream processing engine	For analysing data in motion and for real-time decision management on high velocity data streams such as sensor data, markets data and multi-structured data streams

New analytical tools including search for exploratory analysis, graph analysis and custom mapreduce analytics

- New analytical tools and techniques that have also been added to cater for new analytical workloads requirements. These include:
  - Custom analytic applications written to exploit the Hadoop MapReduce framework to analyse multi-structured data in batch
  - BI tools that generate MapReduce code to retrieve and analyse data typically stored in Hadoop
  - Search based BI tools that index data typically from Hadoop in support of exploratory analysis of multi-structured data
  - Graph analysis tools that visualise data from NoSQL Graph databases in support of exploratory analysis

Existing BI platform tools with access to Hadoop and NoSQL DBMSs

- Existing BI platform tools that can access both SQL-based and NoSQL based analytical platforms (e.g. access Hadoop data via Hive) in support of different types of analytical and reporting needs. This includes:
  - Ad hoc query and reporting
  - Advanced discovery and visualisation
  - Dashboards
  - On-line analytical processing (OLAP)
  - Data Mining – statistical and predictive model development

In-Hadoop as well as in-database analytics

It should also be possible to develop predictive and statistical models and deploy them in a Hadoop system, an analytical RDBMS and event stream processing workflows for real-time predictive analytics

Cross platform analysis across the to solve business problems

Within this new analytical ecosystem there is also an integration requirement to facilitate cross platform analysis. For example, when variations in streaming data occur, stream-processing software analyses the business impact and can take action if required. However it should also be able to filter events of interest to be picked up by the EIM software and loaded into Hadoop for subsequent historical analysis. If any further insight is produced using batch map/reduce analytical processing on Hadoop, that insight may then be fed into a data warehouse to enrich what is already known. For un-modelled multi-structured data, this data can be loaded directly into Hadoop where it can be cleaned, transformed and integrated using EIM data integration software on the Hadoop platform in preparation for exploratory analysis by data scientists. They can then analyse the data 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 big data via search indexes built in Hadoop with map/reduce parallel processing. If the multi-structured data to be analysed is Twitter data for example, then not only could sentiment be determined but Twitter handles could also be extracted and loaded into a NoSQL graph database for further social network link analysis. EIM software can manage the movement of the extracted Twitter handles 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 new insight available to traditional BI tool users.

Complex analysis of structured data on analytical DBMSs can be used to produce predictive/statistical models for use elsewhere in the analytical ecosystem

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 Hadoop's pre-built Mahout analytics) 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.

In terms of multi-platform integration within the enterprise analytical ecosystem, it should be possible to access Hadoop from relational DBMSs and vice versa. There should also be tight integration between operational and analytical systems to maximise value from new insights and to leverage insight on-demand.

BIG DATA ANALYTICS IN AN IBM SYSTEM Z ENVIRONMENT

Having looked at big data, looked at the requirements for big data analytics and also defined a new big data enterprise analytical ecosystem, this section of the paper looks at how one vendor, IBM, has implemented this ecosystem in an IBM System z environment to enable organisations to transform Big Data into insights that can be used to drive competitive advantage.

IBM SYSTEM Z – THE HOME OF CORE TRANSACTIONAL DATA SYSTEMS

A considerable amount of transaction data is stored and maintained on IBM System z

IBM System z has long been at the centre of many company's mission critical transaction processing systems. Its reputation for very high availability, reliability and security have meant that IBM System z is a rich source of data needed in analytical workloads. IBM IMS and DB2 for z/OS in particular hold enormous volumes of transaction data and click stream web logs, which together are two of the most important sources of data in big data analytics.

THE IBM BIG DATA PLATFORM IN A SYSTEM Z ENVIRONMENT

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

This includes a DB2 z/OS based data warehouse that integrates with an analytic accelerator and zEnterprise big data platforms

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

Information management tools to goven and manage data

Together, this set of technologies constitutes the IBM Big Data Platform[4] as shown below. This platform includes three analytical engines to support the broad spectrum of traditional and big data analytical workloads. These are:

[4] For more information on IBM Big Data Platform please refer to the paper "Architecting a Big Data Platform for Analytics", Ferguson, September 2012

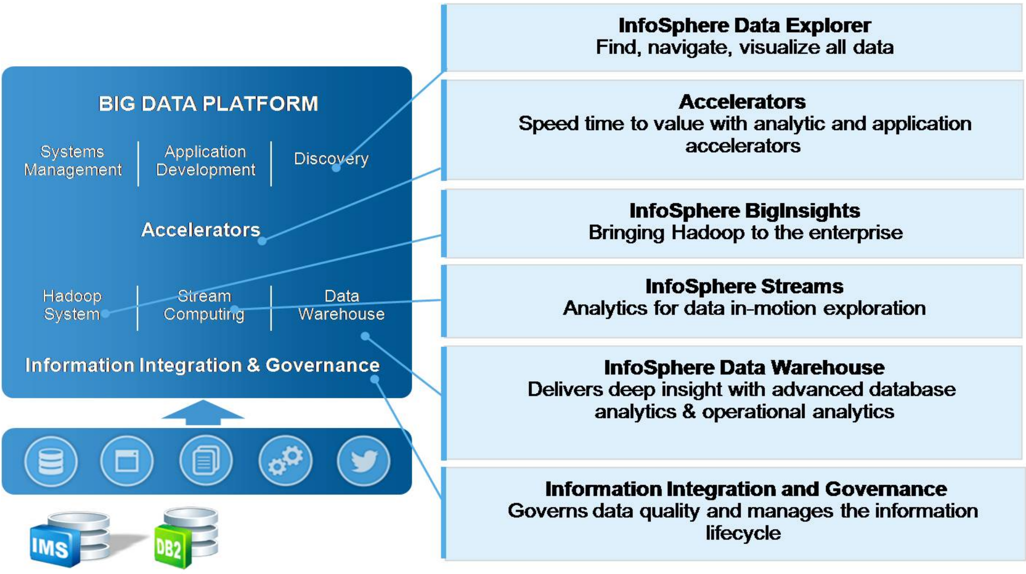
Three analytical engines in the IBM Big Data Platform

- IBM InfoSphere Streams for continuous analytics of data-in-motion
- IBM BigInsights - A Hadoop System
- Data Warehouse

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

The IBM Big Data Platform



InfoSphere Data Explorer
Find, navigate, visualize all data

Accelerators
Speed time to value with analytic and application accelerators

InfoSphere BigInsights
Bringing Hadoop to the enterprise

InfoSphere Streams
Analytics for data in-motion exploration

InfoSphere Data Warehouse
Delivers deep insight with advanced database analytics & operational analytics

Information Integration and Governance
Governs data quality and manages the information lifecycle

Source: IBM

IBM InfoSphere Streams

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.

It also ships with pre-built toolkits and connectors to expedite development of real-time analytic applications

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.

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

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 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 BI Real-Time Monitoring for visualization in a dashboard for further human analysis.

IBM InfoSphere BigInsights on IBM System zEnterprise

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

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

- An enterprise scalable, Posix compliant file system GPFS-FPO [5] with distributed metadata to eliminate single points of failure
- Hadoop HDFS NameNode high availability via NameNode failure detection and automatic failover to a standby NameNode
- 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
- Big SQL to provide a SQL interface and query engine for accessing data stored in BigInsights via the use of JDBC and ODBC drivers
- Connectors to IBM DB2 and other IBM analytical platform 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

BigInsights is available in two editions:

- IBM BigInsights QuickStart Edition – a pre-configured free download, non-production version of BigInsights
- IBM BigInsights Enterprise Edition

IBM InfoSphere BigInsights can run on IBM zEnterprise

Although, InfoSphere BigInsights does not run on Linux on z, an Apache distribution of Hadoop for IBM Linux for System z is available for download from the Apache website. Also IBM business partner Veristorm provides Vstorm Enterprise, a platform that collects z/OS data sources for processing and *zDoop* for analytics through Hadoop on Linux. Retaining your data on System z preserves physical and operational security, delivering quick access to the data and processing needed for secure analytics. InfoSphere BigInsights

[5] FPO = File Placement Optimizer

Enterprise Edition is also available on an appliance in the form of IBM PureData System for Hadoop.

Data Warehousing on IBM System z – DB2 z/OS

Many companies have built data warehouses on IBM System z close to the transaction data sources also stored there

Given the growth in transaction data on IBM System z both in DB2 z/OS and IMS, many organisations have chosen to keep some or all of their analytics components including data warehousing, business intelligence and predictive analytics close to the transaction data on System z in order to make it possible to integrate business intelligence and analytics into operational business processes. In addition many organisations run IBM's zEnterprise Analytics System 9700 and 9710 and IBM DB2 Analytics Accelerator for z/OS to deliver key insights to the business. It is also the case that IBM System z transaction data stores are popular sources feeding into other RDBMS and big data analytical platforms. That may include other data warehouse platforms or IBM BigInsights where transaction data may be archived and analysed.

DB2 z/OS includes a number of features that favour analytic workloads including adaptive table level and page-level compression, multi-temperature data management, temporal queries, a NoSQL graph store and row and column access control. DB2 11 for z/OS can also integrate / interact with Hadoop - invoke an existing HDFS query and fetch the result set which can then be read, and the appropriate table(s) updated.

IBM DB2 Analytic Accelerator

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

IBM DB2 Analytics Accelerator for z/OS is a high performance appliance that integrates System z and Netezza technolocgies. It is specifically designed to offload complex analytical queries from operational transaction processing systems running DB2 mixed workloads on IBM System z. This approach allows customers to combine the best attributes of symmetric multiprocessing (SMP) using DB2 for z/OS, with the best of massively parallel processing (MPP) using Netezza technology to transparently assign the platform best suited to maximise performance for the type of query at hand. It is achieved by re-creating DB2 tables on DB2 Analytics Accelerator using pre- defined administrative DB2 stored procedures and then loading the data from DB2 into DB2 Analytics Accelerator. If necessary, DB2 tables can be locked to prevent update during DB2 Analytics Accelerator loading. Also, queries can be routed and processed by DB2 Analytics Accelerator 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, DB2 Analytics Accelerator 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. In addition, data that is maintained on the DB2 Analytics Accelerator inherits all DB2 for z/OS data attributes, including security and recoverability. All data is loaded, backed up and retrieved through DB2 with no external connections to the DB2 Analytics Accelerator to minimize any opportunity for intrusion.

There are currently three DB2 Analytics Accelerator offerings available supporting 8, 16 and 32 Terabytes of user data. This can be increased with data compression. IBM DB2 Analytics Accelerator Loader for z/OS can load data from multiple data sources including other vendors DBMS products.

The IBM zEnterprise Analytics System 9700/9710

IBM zEnterprise Analytics System is a pre-integrated, sclable, ready-to-use analytical platform that integrates with zEnterprise systems

The IBM zEnterprise Analytics System 9700 is a modular, pre-tested, pre-integrated, scalable and ready to use platform that combines hardware, software and services to deliver business critical analytics that integrate with the zEnterprise system organisations already have in place. It includes DB2 for z/OS, DB2 Analytics Accelerator, and zEnterprise hardware at its core together with three optional add-on packs:

- The Data Analytics Pack - provides a wide range of business intelligence, predictive analytics and performance management capabilities
- The Data Integration Pack - integrates and transforms data and content to deliver accurate, consistent, timely and complete information
- The Fast Start Services Pack - predefined services expedite time to value

IBM offers these optional packs to provide the flexibility required to meet clients' specific analytic needs. With respect to business intelligence and analytics, IBM zEnterprise Analytics System includes support for both structured and unstructured data, reporting, multi-dimensional and statistical analysis, dashboarding, data mining, cubing services and text analytics. Data mining models (clustering, associations, classification and prediction) can be developed using drag-and-drop features.

In addition, for customers who require a smaller footprint solution, IBM provides the entry level IBM zEnterprise Analytics System 9710 offering which includes IBM DB2 Analytics Accelerator and the three add-on packs as options.

IBM IMS

IBM System z also integrates with IBM BigInsights Hadoop distribution

With respect to IMS, IBM is planning on offering solutions that enable the IMS DB data assets, as well as log data, to be integrated into an IBM big data ecosystem as follows:

- Connectors to access IMS data from InfoSphere BigInsights to enable to jobs running on InfoSphere BigInsights to leverage IMS data assets as part of an organization's big data analytics strategy. Technologies like JDBC and Sqoop can be used to extract data from IMS and integrate with the big data platform
- Technology that will enable IMS DB to be a source for InfoSphere Data Explorer – a key element of IBM's comprehensive Big Data platform.
- Enable the Machine Data Accelerator to both ingest as well as run analytics on multiple types of IMS log records. This will assist with the important elements of IT provisioning and log analysis.

Also BM recently announced that a single IMS Fastpath database system is capable of processing over 117,000 transactions per second. IBM is investigating low latency streaming analytics for data of this velocity and volume,

Platform Integration in an IBM System z Big Data Environment

Big Data platforms are also a place where analysis takes place on multi-structured data. Data Scientists produce insights from that data and can then move those insights into data warehouses to enrich what an organisation already knows. For those organisations running data warehouses on IBM System z, DB2 can integrate with IBM's InfoSphere BigInsights Hadoop Platform. The first phase of this integration is the availability of connectors between IBM DB2 on z/OS and IBM InfoSphere BigInsights. A JAQL server running on InfoSphere BigInsights can accept JAQL query processing requests from IBM DB2 z/OS via new JAQLSubmit and HDFSRead user defined functions. This means that it is possible to embed a JAQL query (JAQL script) in DB2 for z/OS application SQL statements that execute map reduce applications on Hadoop to analyse data held in Hadoop HDFS.

It is also a direction of IBM to build DB2 Map Reduce functions that can be invoked in DB2 z/OS SQL and offloaded onto IBM DB2 Analytic Accelerator. This would allow a new breed of DB2 z/OS analytic applications to be built to apply advanced analytics to information stored in its native form in IBM InfoSphere BigInsights.

Given the increased level of integration between Big Data platforms and data warehouses, it is also possible to move data from IBM InfoSphere BigInsights Hadoop into data warehouses. With respect to data warehouses developed on DB2 z/OS, map reduce applications running in Hadoop can move data into DB2 z/OS partitioned tables. In order to do this, the map reduce applications running on Hadoop would need access to the DB2 Catalog information to understand DB2 data format and also to match parallel reduce threads to partitions in DB2 for z/OS.

With respect to IMS, it is also possible to access BigInsights from IMS applications and so consume additional insights produced by data scientists using map/reduce processing on Hadoop.

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

Information Management in an IBM System z Big Data Environment

IBM InfoSphere Information Server and Foundation Tools provide end-to-end data management across all data stores

The IBM Information Management platform includes an integrated suite of tools including InfoSphere Information Server, InfoSphere Foundation Tools, InfoSphere Optim and InfoSphere Guardium. Together these tools can be used to govern and manage data in an IBM System z big data environment. They support everything from defining data definitions, data modelling, data profiling, data cleansing, data integration, data virtualisation, data protection, data

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

activity monitoring and moving data across traditional and big data analytical platforms. IBM InfoSphere Information Server supports connectivity to IBM InfoSphere BigInsights, IBM DB2 z/OS data warehouse, IBM DB2 Analytics Accelerator as well as IBM InfoSphere Master Data Management. It also integrates with IBM InfoSphere Streams to pump filtered event data into IBM InfoSphere BigInsights for further analysis.

With respect to information security and protection, policies should be equally enforced across big data environments as well as traditional data warehouse and transaction processing system databases. IBM has extended its security technologies to support multiple analytical platforms in an IBM System z big data environment. In particular, security is supported within IBM InfoSphere Streams, IBM BigInsights, and IBM zEnterprise Analytics System. IBM InfoSphere Guardium has now been extended to audit Hadoop activity via S-TAPs in IBM BigInsights in addition to activity in DB2 z/OS, IMS and VSAM. S-TAPs can be installed on the HBase master, the Hadoop JobTracker, the Hadoop NameNode and Hive Server.

IBM has also combined QRadar with BigInsights to analyse security event data to produce security intelligence.

It is also possible to use InfoSphere 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 IBM System z or analytical Big Data store best suited for an analytical workload. The purpose of this 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 stores in a System z implementation of 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 Cognos BI on IBM System z

IBM has integrated its Cognos BI tool suite with BigInsights, IBM DB2 z/OS and IBM zEnterprise Analytics System

IBM Cognos BI is IBM's flagship business intelligence offering. It is an integrated system for ad hoc reporting and analysis, dashboard creation, scorecarding, production reporting, multi-dimensional analysis, budgeting, planning and forecasting. With respect to IBM System z, IBM Cognos BI runs on Linux on System z and z/OS and can be used to access and analyse structured data housed DB2 z/OS as well as non-IBM data warehouse platforms.

IBM Cognos BI on IBM System z has also been extended to enable analysis and reporting on Big Data on IBM InfoSphere BigInsights via Hive and Big SQL. It also possible to access structured and unstructured data in DB2 z/OS, zEnterprise Analytics Server, IBM DB2 Analytics Accelerator and IBM InfoSphere BigInsights via InfoSphere Federation Server. In addition, InfoSphere Federation Server may be used in combination with InfoSphere Data Explorer to enable access to unstructured and semi-structured data.

Copyright © Intelligent Business Strategies Limited, 2013, All Rights Reserved

The IBM Cognos BI family of products starts with a personal edition scaling up through workgroup capability to its enterprise edition. The IBM Cognos BI family of products include:

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

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

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

IBM Cognos BI can also be extended to mobile environments as a native application. In addition, IBM also offers **IBM Cognos BI 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 BI 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 BI 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 SPSS for IBM System z

IBM SPSS is used to build advanced analytics that can be deployed in InfoSphere Streams and IBM zEnterprise Analystics System 9700 / 9710

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 zEnterprise Analytics System 9700/9710 for in-database advanced analytics and operational BI.

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 DB2 QMF

QMF can be used to analyse data in an IBM System z Big Data environment

IBM DB2 Query Management Facility (QMF) is a query tool that allows users to execute SQL queries on IBM DB2 z/OS, and other non-IBM databases. It has a semantic layer, can display query results in visualisations and dashboards and is available as:

- QMF for Workstation – a java based rich client application
- QMF for WebSphere – a thin client application running in WebSphere Application Server on z/OS and Linux on System z
- QMF Dashboards

QMF supports enhanced data access and data federation out of the box.

It is a direction of IBM DB2 QMF to interface with Hive and BigSQL to access data in IBM InfoSphere BigInsights.

Copyright © Intelligent Business Strategies Limited, 2013, All Rights Reserved

Third Party Analytical Tools

Third party BI and analytics tools can also be used to analyse data in an IBM System z Big Data environment

In addition to IBM's own analytical tools, IBM System z also supports 3rd party analytical products such as SAS and Information Builders webFocus. These third party BI and analytical platforms can also be used to analyse data in IBM DB2 z/OS, IBM zEnterprise Analytics System, IBM DB2 Analytics Accelerator and IBM BigInsights. SAS can be used to create statistical and predictive models. These models can be deployed in IBM InfoSphere Warehouse running on DB2 z/OS for in-database advanced analytics and operational BI. In addition SAS LASR analytics server can be deployed in Hadoop clusters for massively parallel in-memory analysis of big data help in Hadoop.

CONCLUSIONS

Companies that have invested in IBM System z can now extend this environment to make use of big data for competitive advantage

Companies that have built core transaction processing systems on IBM System z and who have data warehouse(s) on IBM DB2 z/OS, can now extend their architecture along the lines of Figure 2 to deepen customer insight to facilitate increased retention and growth. They can also leverage IBM zEnterprise Analytics System 9700/9710 and IBM DB2 Analytics Accelerator to deliver business critical analytics, as well as IMS and DB2 z/OS integration with IBM BigInsights to help deliver improvements in operational effectiveness. Both of these business needs are high on the agenda in many board rooms.

Workload management and high availability will be critical in integrating transaction processing with on-demand analytic workloads

In this extended analytical environment, IBM System z workload management and high availability become even more important in a world where core transaction processing systems are operating on a 24 x 365 basis and requesting actionable insights on-demand from traditional and Big Data platforms.

IBM information management now encompasses new big data platforms

Existing and new BI tools can be used to analyse big data in a IBM System z environment

New insights produced can be added to what companies already know to increase customer insight and improve operational effectiveness

With respect to information management, IBM has extended its information management tool suite to capture data from InfoSphere Streams and unstructured data sources as well as traditional structured data sources. It can also integrate and protect data in IBM BigInsights and move data between IBM BigInsights and data warehouses running on DB2 z/OS and zEnterprise Analytics System. In addition the ability to deploy IBM SPSS models in IBM InfoSphere Streams and IBM BigInsights as well as in-database within DB2 z/OS means that predictive and pro-active analytics can now be deployed to analyse data-in-motion as well as structured and multi-structured data at rest. Deploying models in InfoSphere Streams introduces automated decision management for operational effectiveness while deploying them in IBM BigInsights, IBM DB2 z/OS and IBM zEnterprise Analytics System makes it possible to deepen insights and make solid recommendations. Finally accessing all of this from IBM Cognos, IBM SPSS and IBM DB2 QMF on System z allows integration of insights produced by the entire analytical ecosystem into easy to use BI and predictive analytics tools. This allows data scientists and business analysts to work together to provide actionable insights to information consumers and people in operations for more comprehensive accurate decision-making.

All of this makes IBM's end-to-end big data offering in a System z environment a very competitive offering.

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 32 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.

INTELLIGENT
BUSINESS
STRATEGIES

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

Transforming Big Data into Insights in an IBM System z Environment
Copyright © 2013 by Intelligent Business Strategies
All rights reserved