WHITE PAPER

Architecting A Platform For Big Data Analytics

2nd Edition

By Mike Ferguson
Intelligent Business Strategies
March 2016

Prepared for:

Table of Contents

INTRODUCTION

Data and analytics provide the insights for a the key to business disruption

In this digital age, 'disruption' is a term we hear a lot about. It can be defined as

"A disturbance that interrupts an event, activity, or process"

In the context of this paper it means competition appearing from unexpected quarters that rapidly takes business away from traditional suppliers. It is made possible because of data, that when collected, cleaned, integrated and analysed, provides sufficient insights to identify new market opportunities and prospective customers. Those that have the data can see the opportunities and cause disruption. Those that don't have data, or only a subset of it, cannot.

Disruption is accelerating fuelled by people who are increasingly informed before they buy

The speed at which disruption can occur is accelerating. People with mobile devices can easily search for products and services, find available suppliers, compare products and services, read reviews, rate products and collaborate with others to tell them about what they like and dislike all while on the move. This kind of power puts prospective customers in a very powerful position when making buying decisions because they are increasingly *informed before they buy* and, armed with this information, they can switch suppliers at the click of a mouse if a more personalised, better quality product or service is available. The result is that, with this information at hand, loyalty is becoming cheap. People spread the word on social networks and disruption takes hold in the market. In this kind of market, no one is safe. Companies have to fight harder to retain existing customers while also trying to grow their customer base.

Companies are having to fight harder to retain existing customers while also trying to grow

Companies are trying to offer improved quality of service and personalisation to retain customers

Given this backdrop, it is not surprising that many companies are therefore focusing on improving quality and customer service to try to retain their customers. Process bottlenecks are being removed and process errors that impact on the customers' experience are being fixed to make sure that everything runs smoothly. In addition, many companies are trying to improve customer engagement and ensure the same customer experience across all physical and on-line channels. That requires all customer facing employees and systems to have access to deeper customer insight and to know about all customer interactions. The objective (often referred to as an Omni-channel initiative) is to create a *smart* front office with personalised customer insight and personalised customer marketing recommendations available across all channels. This is shown in Figure 1 where personalisation is made possible by analysing enriched customer data using predictive and prescriptive analytics.

Analysing enriched customer data using predictice and prescriptive analytics is the key to personalisation

Insights produced by analysing of transactional activity is no longer enough tocause disruption

Until recently, the way in which we produced customer insight and customer recommendations was to simply analyse transaction activity in a data warehouse. However the limitation is that only transaction activity is analysed. It does not include analysis of other high value data that when pieced together offers a much more complete understanding of a customer's "DNA". Therein, lies the problem. Analysis of traditional transaction data in a data warehouse is insufficient for a company to gain disruptive insight. More data is needed. Therefore new data requirements and new technical requirements need to be defined to identify, process and analyse all the data needed to enable companies to become 'disrupters'. Let's take a look at what those requirements are.

Insights produced by analysing transactional and non-transactional data is now needed across all traditional and digital channels

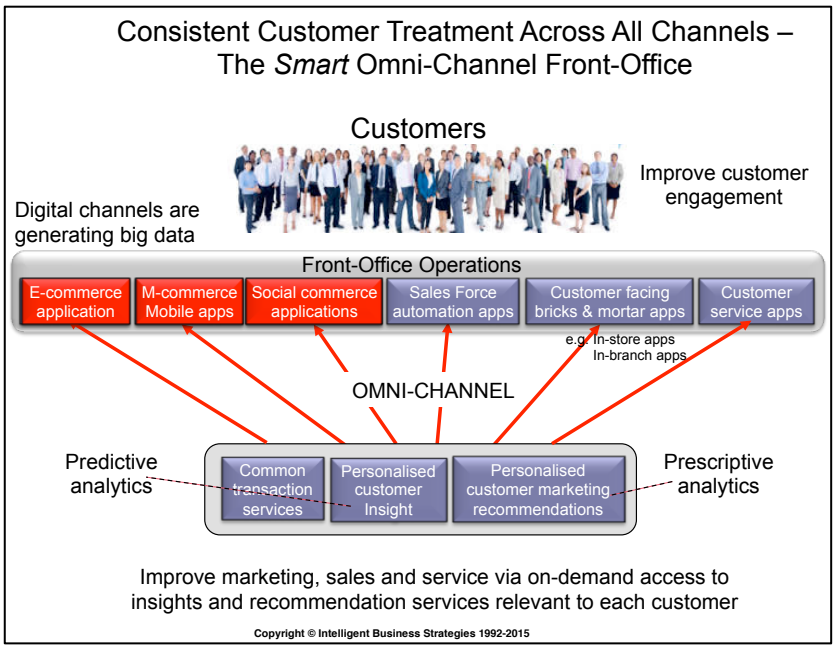

Figure 1

BUSINESS DATA REQUIREMENTS FOR UNDERSTANDING CUSTOMER DNA

Data from digital channels also needs to be analysed to gain a complete understanding of customer behaviour

Given the limitations of traditional transaction data, businesses are now defining new data requirements to attempt to strengthen and enrich what they know about customers in order to see new opportunities and disrupt marketplaces. In addition to the aforementioned transaction data, business departments now want to analyse data generated in the digital channels (highlighted in red in Figure 1), and data from external data sources to gain much more complete understanding of customers. This includes:

- On-line click stream data
- Social network data and in-bound email
- Unstructured data e.g. loan documents, application forms
- Open government data
- Sensor data emitted from smart products being used by customers

ON-LINE CLICK STREAM DATA

Click stream data reveals how customers navigate your website

On-line click stream data is the data generated from visitors clicking on your website. They may land on your home page from a search engine and then start browsing your site moving from web page to web page looking at products and services. Your web servers record every click a visitor makes in a visitor 'session' in one or more weblog files. By accessing these files and analysing clickstream data, it becomes possible to understand:

- Where visitors come from (via geocoding of IP addresses)
- How they reached your site
- Navigation paths taken through your site before buying products
- Navigation paths taken through your site before abandoning it
- What they looked at en route to buying products on your site

The richer the data, the more effective the customer insight will be

This data can be used to improve customer experience and conversion especially if clicks can be associated with customers and prospects. Doing this requires additional data to be integrated with clickstream data. The richer the data, the more detailed questions can be answered.

For example analysing transaction data may provide insights about loyal, low risk customers that are not generating much revenue. Clickstream analysis may identify high revenue products on your website buried under too many clicks. Combining the two may identify new ways to digitally market these products to loyal customers that would result in increased revenue generation.

SOCIAL NETWORK DATA

Social networks can provide new data and customer insights

Popular social networks like Twitter, Facebook, LinkedIn and YouTube can also provide valuable new customer data and insights. This includes:

- Additional customers attributes e.g. about employment, hobbies, interests
- Previously unknown relationships e.g. household members
- Likes and dislikes
- Influencers
- Product and brand sentiment

Social networks provide new data, sentiment and an understanding of who influencers are

Sentiment can also come from in-bound email and CRM system notes. Identifying influencers in social networks is important because it allows marketing departments to run 'target the influencer' marketing campaigns to see if they can recruit new customers, cause revenue uplifts and therefore improve marketing effectiveness.

OPEN GOVERNMENT DATA

Open government data is hugely valuable to understanding customers and risk

Open government data (e.g. Data.gov, Data.gov.uk, Data.gove.be etc.) is any data produced or commissioned by public bodies and published for download. It is free to use and includes information about:

- Businesses e.g. patents, trademarks and public tender databases
- Geographic information (including address information, aerial photos, geodetic networks, geology)
- Planning information
- Legal decisions (of national, foreign and international courts)
- Weather information (e.g. Climate data and models and weather forecasts)
- Social data (including various types of statistics on economics, employment, health, population, public administration)
- Crime data (including various types of statistics)
- Transport information (including traffic congestion, work on roads, public transport and vehicle registration)

This kind of data can be hugely valuable in understanding customers and risk.

SENSOR DATA

Sensor data in smart products is another good data source that can reveal new business opportunities

Sensors can also be used to gain a deeper understanding of customers in that they can be embedded in products those customers own. As a result data can be collected to understand how products are used and to capture data about people's health. The use of sensor data is particularly relevant to customer location, which can be revealed using smart phone GPS sensors. This combined with clickstream data allows telecommunications companies, for example, to monitor what people are browsing on-line and also where they are while browsing. This allows Telecommunications companies to disrupt the advertising industry by offering location-based mobile advertising - a totally new line of business. Sensors can also be used to monitor customer movement and product usage.

Sensor data can help optimise business operations and avoid unplanned operational cost

Beyond customer, sensor data is frequently used to help optimise operations, reduce risk and provide insights that may lead to new products. Sensors allow companies to monitor live operations, prevent problems happening to keep processes optimised and avoid unplanned costs. Typical use cases include:

- Supply/distribution chain optimisation
- Asset management and field service optimisation
- Manufacturing production line optimisation
- Location based advertising (mobile phones)
- Grid health monitoring e,g. Electricity, water, mobile phone cell network
- Oil and gas drilling activity monitoring, well integrity and asset management
- Usage / consumption monitoring via smart metering
- Healthcare
- Traffic optimisation

TECHNICAL REQUIREMENTS TO UNDERSTAND THE CUSTOMER

Structured, semi-structured and unstructured data are all needed to create a more complete view of the customer and optimise operations

Looking at these data requirements, a lot of new internal and external data is needed, over and above traditional transaction data, to create a more complete view of a customer and to optimise business operations. However we need to recognise that the characteristics of the data needed have gone beyond the structured transaction data in traditional data warehouses.

The new data required includes structured, semi-structured (e.g. JSON, XML) and unstructured (e.g. text) data, all of which needs to be captured, processed and analysed to produce new insights. Some of this data (e.g. clickstream and social media data) can be very large in volume and some of it is created in real-time at very high rates (e.g. sensor data). It is not surprising therefore that new technical requirements naturally fall out of the need capture, process and analyse this data. These are shown below and categorised for easy reading.

MULTIPLE ANALYTICAL WORKLOAD REQUIREMENTS

Having already established that additional insight is needed beyond what is produced from analysis of transaction data, it should be possible to:

New kinds of analytical workloads are needed to enable disruption

- Support new kinds of analytical workloads to enable disruption including:
  - Real-time analysis of data in motion (e.g. analyse clickstream while the visitor is on your website, or sensor data to predict asset failure)
  - Exploratory analysis of un-modeled, multi-structured data e.g. Social network text, open government data, sensor data
  - Graph analysis e.g. community analysis, social network influencer analysis
  - Machine learning to:
    - Develop predictive models on large volumes of structured data e.g. clickstream data
    - Score and predict propensity to buy, trends in behavior and customer churn using a more comprehensive set of data
    - Score and predict asset or network failure that would impact negatively on revenue, customer experience, risk or operational cost

Real-time analysis of streaming data, graph analysis, machine learning and exploratory analysis of multi-structured data are all needed

MULTIPLE ANALYTICAL DATA PLATFORMS

In order to capture, prepare and integrate data at scale, and run multiple analytical workloads to enable disruption support is now needed for multiple analytical platforms including:

Multiple platforms are now needed to support different analytical workloads

- NoSQL databases e.g. a Graph DBMS
- Hadoop
- Analytical RDBMSs

These can be on the cloud, on-premises or both

- Streaming analytics platforms

These could be on the cloud, on-premises or both. Also note that Apache Spark massively parallel in-memory processing could be used in analysis of data at rest and so can retrieve data from any of the above listed data stores. It can also be used for streaming analytics on data in-motion.

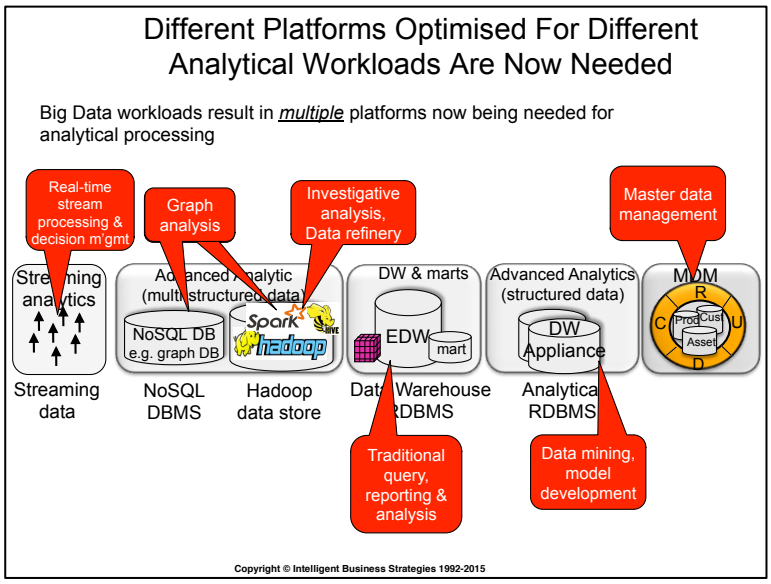

Figure 2

SCALABLE DATA CAPTURE, PREPARATION AND INTEGRATION

Data ingest, data cleansing, data transformation, data integration and analysis all need to scale

To process data with these kinds of characteristics requires scalability. That scalability needs to apply to data ingest (capture), cleansing, transformation, integration and, of course, analysis. There is no point having analytics that scale if the software needed to capture, prepare and integrate data before the analysis begins, does not also scale or vice versa. Therefore the following technical requirements are related to data connectivity, scalable data capture, data processing and analytics.

It should be possible to:

New internal and external data sources need to be supported

- Connect to multiple different data sources where new big data originates. This includes connectors to NoSQL DBMSs (e.g. MongoDB, Cassandra, HBase), social networks (e.g. Facebook, Twitter, LinkedIn), machine data log files, sensor networks (with different types of sensors), web sites, email servers, streaming data sources, cloud data sources, and Hadoop Distributed File System (HDFS). This is in addition to 'traditional' data sources such as RDBMSs, files, XML, JSON, packaged applications, web services etc., that are typically used in data warehousing

Data preparation and integration jobs should be defined once and be able to run on all necessary platforms

- Capture / ingest big data in parallel into a scalable analytical platform either on-premises or on the cloud
- Develop data preparation and integration jobs independently of the platform they execute on and then to be able to run them on the platform required with the necessary parallelism to exploit the full power of the platform selected. This means that business and IT professional productivity should not be slowed down by having to learn new user interfaces and new development languages needed to exploit underlying big data technology platforms. The software should hide that

New data should not impose additional work on IT and business analysts to process it

- Irrespective of whether development of data capture, data preparation, data integration and analysis jobs are done by IT professionals or business analysts, it should be possible to develop once, and run on the platform best suited to do the work irrespective if this is on-premises or on the cloud
- Support the ability to flatten semi-structured data e.g. JSON, XML

Scalable data cleansing, transformation and integration is needed to handle large data volumes

- Support shared nothing parallel execution of each and every data cleansing, data transformation and data integration task defined in a data preparation and integration job. This is needed to be able to exploit the full power of underlying hardware
- Support multiple types of parallelism across nodes including parallel execution of each transformation (as described above) across partitioned data and also pipeline parallelism whereby the output of one task (executing in parallel) can flow to the next task while the first task is still executing

Data cleansing, transformation and integration needs to happen as close to the data as possible

- Push data cleansing, transformation and integration tasks to execute where the data is located and not have to take the data to a central point where cleansing, transformation and integration tasks are located
- Process more data simply by adding more hardware
- Undertake probabilistic matching (fuzzy matching) of sentiment with customer master data at scale to understand customer sentiment

SCALABLE ANALYTICS REQUIREMENTS

It should be possible to:

Analytical models should be able to run in-stream, in-memory in-batch or in-database

- Develop predictive analytical models offline and have flexibility in terms of deploy options to run these models in:
    - Real-time in a scalable streaming analytical environment to analyse data in motion
    - Batch using scalable batch analytical processing e.g. on Hadoop
    - Memory using scalable analytical processing e.g. on Spark
    - An analytical RDBMS e.g. to analyse data in a data warehouse

  This should be possible both on-premises and on the cloud

Invocation of scalable analytics is needed

- Invoke any pre-built analytic algorithms available in-Hadoop, in-Database in-Memory and in-Stream during analytical model execution. For example, exploiting pre-built machine-learning algorithms, text analytics, graph analytics or custom- analytics already available on Hadoop or in-Spark from within a batch data integration job, a batch analytic application or a front-end self-service interactive analytical tool. In the case of text analytics, this would allow high value structured data to be extracted in parallel from large document collections, and then integrated with other data during execution
- Support model management including support for versioning, Champion Challenger and A/B Testing of models
- Automate model evaluation, model refresh and model re-deployment

Analytical models should be able to run in the cloud or on-premises

- Develop predictive analytical models on-premises or on the cloud
- Rapidly explore and analyse new structured, semi-structured and unstructured big data using search

Need to develop analytical applications in popular programming languages

- Create analytical pipelines by stringing together multiple analytic algorithms in a pipeline (workflow) to analyse data at scale
- Develop batch analytic applications in popular languages such as Java, Python, Scala and R that can analyse data at scale using pre-built analytic algorithms that run close to the data in parallel
- Easily develop custom scalable analytical algorithms to add to those available 'out-of-the-box' in Hadoop, Spark and analytical RDBMSs to analyse complex data

Use of tools to generate scalable analytical applications should also be possible to reduce time to value

- Make use of tools that generate code to analyse data at scale using pre-built analytic and custom algorithms that run in parallel
- Analyse streaming data in real-time by invoking in-memory analytics in a scalable cluster computing environment

Automated alerting and recommendations are needed to guide and optimise operations

- Define rules to drive automated decisions and actions when patterns are detected and conditions predicted during streaming analytics
- Automate alerting, recommendations, transaction invocation and process invocation as actions during real-time stream analytics processing

DATA GOVERNANCE REQUIREMENTS

It should be possible to:

An information catalog is needed to track and govern data

- Register new datasets / data streams coming into the enterprise in an information catalog either via API or user interface so their existence is known and the data can be governed
- Automate the profiling, schema detection and semantic analysis of data so that the quality, structure, and meaning of data can be quickly assessed and determined

Automated profiling and classification of data will speed up data processing and help determine how data is governed

- Classify newly registered data in terms of its sensitivity/confidentiality, data quality, trustworthiness, retention and business value
- Define governance policies to govern data based on how it is classified
- Collaborate with others on defining these policies

It should be possible to define rules to govern data irrespective of its location

- Define rules to enforce governance policies
  - Irrespective of data store and location on-premises or in the cloud
  - To support national, regional or other jurisdictional boundary requirements for data

It should be possible to enforce national, regional and other jurisdictional regulations when governing data

- Verify the quality of data and report problems

Auditing of processing tasks is needed to ensure trustworthiness

- Record metadata associated with data cleansing (standardisation, cleansing, enrichment), transformation and integration tasks irrespective of whether the data is at rest or in-motion, on-premises or in the cloud and the tasks are performed by tools or custom applications or both
- View metadata lineage from end-to-end to ensure trustworthiness

Data privacy needs to be enforced irrespective of where data is stored

- Mask and protect data classified as confidential or sensitive and enforce any policies irrespective of whether the data is at rest in Hadoop HDFS, an RDBMS, a NoSQL database, or if it is data-in-motion

- Archive and manage data movement between data stores (e.g. data warehouse and Hadoop) ensuring all governance rules associated with sensitivity, quality and retention are upheld irrespective of data location
- Control access to data by applications, tools and users to protect data

Governance aware execution engines are needed to enforce governance rules anywhere that data to be governed is located

- Have governance aware execution engines to enforce data governance rules anywhere in the analytical ecosystem shown in Figure 2 to enforce defined policies
- Map new insights produced by analysing new data, into a shared business vocabulary in a business glossary so that the meaning of this data can be understood before it is shared

EASY ACCESS TO DATA

It should be possible for IT professionals and data scientists to:

Data should be available as a service within the enterprise and documented in an information catalog

- Publish new datasets, data integration workflows, analytical workflows, and insights, as part of a data curation process, to an information catalog for other users and applications to consume and use

It should be possible for data scientists and business analysts to:

Data lineage is needed to help understand where data came from

- Access an information catalog to see what data exists, where it is located, what state it is in, where it came from, how it has been transformed, whether they can trust it and if it is available for use

Business users should be able to search the information catalog to find and subscribe to data

- Easily search the information catalog to find new datasets and insights so that they can quickly see what exists
- Subscribe to receive new data and insights, published in an information catalog, for delivery to wherever they require it, in whatever format they require it subject to any governance rules being applied to enforce national, regional, or other jurisdictional policies that restrict its use

It should be possible to federate queries across streaming data, Hadoop and traditional data warehouses to produce disruptive insights

- Access authorised data and insights that may exist in multiple analytical data stores and data streaming platforms through a common SQL interface from self-service tools to simplify access to data
- Federate queries data across Hadoop, traditional data warehouses, and live data streams to produce disruptive actionable insights
- Access Hadoop data using SQL from traditional data warehouse data stores as well as via SQL on Hadoop initiatives

It should be possible to hide complexity of data access by creating a logical data warehouse comprised of multiple underlying platforms

- It should be possible to query and analyse data in a logical data warehouse (across multiple analytical data stores and real-time streaming platforms) using traditional and Cognitive Analytical tools irrespective of whether the data in the logical data warehouse is on-premises, on the cloud or both

INTEGRATING BIG DATA INTO YOUR DW/BI ENVIRONMENT – THE LOGICAL DATA WAREHOUSE

In order to bring together new insights with what already exists (e.g. for disruptive customer insight), a new integrated 'logical data warehouse' platform is needed that includes multiple analytical data stores (including Hadoop, the data warehouse, MDM, data warehouse appliances), end-to-end-information management, high volume ingest, data governance, batch and real-time streaming analytics, and simplified access to all data via a SQL-based data virtualisation layer. This logical data warehouse architecture may exist on the cloud, on premises or include data stores spread across both (hybrid). This is shown in Figure 3. Note that access to multi-structured data is simplified because of a 'SQL on Everything' data virtualisation layer that can federate query processing across multiple data stores.

A federated query layer is needed to hide multiple data stores from users in a modern analytical ecosystem

Data in Hadoop can be integrated with data in traditional data warehouses to produce disruptive insights

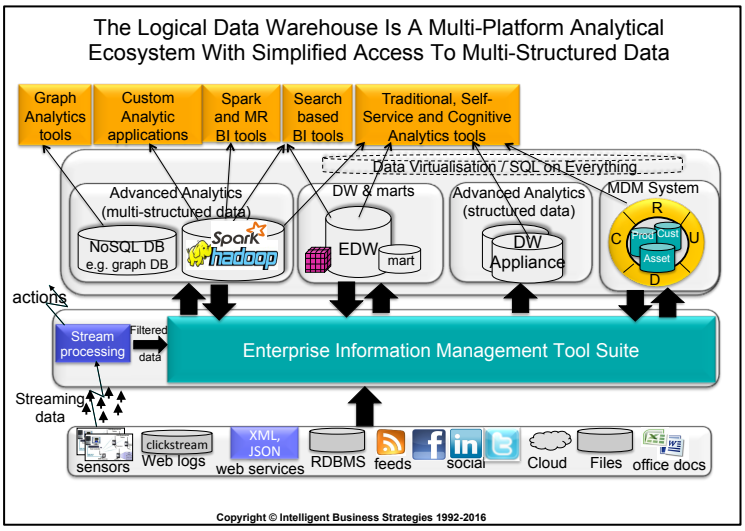

Figure 3

Apache Spark is rapidly emerging as an in-memory analytics layer that can run on top of muliple data stores or even inside DBMSs

Furthermore, Apache Spark can sit on top of all data stores as a general purpose massively parallel in-memory analytics layer primarily aimed at data science. It could also be embedded in database products for in-database, in-memory analytics.

An enterprise information management tool suite enables data curation and governance across a multi-platform analytical ecosystem

Figure 3 also shows an enterprise information management (EIM) tool suite to manage data flows into and between data stores in a hybrid multi-platform analytical ecosystem. The EIM suite includes tooling to support:

The EIM tool suite works across cloud on-premises and hybrid logical data warehouse environments

- An information governance catalog and business glossary
- Data modeling
- Data and metadata relationship discovery
- Data quality profiling and monitoring
- Data cleansing and matching at scale
- Data transformation and integration at scale
- Data privacy and lifecycle management
- Data audit and protection

AGILE BIG DATA DEVELOPMENT

Having understood new data requirements plus the technical requirements that need to be supported to process and analyse that data, the next question is:

"How can you enable development of big data analytics to produce disruptive insight quickly while also dealing with a deluge of new and complex data?"

There are a few key ways to make this possible:

There are several ways to introduce agility into big data development processes

- Use the cloud as a low cost way to get started in create new insights to add to what you already know
- Make use of data virtualisation and federated query processing to create a 'logical data warehouse' across multiple analytical data stores
- Organise for success by creating a 'publish and subscribe' approach to development using an information catalog to facilitate re-use

START IN THE CLOUD

Creating big data analytics in the cloud is a fast and low cost way to get started

Once candidate big data projects have been defined and prioritised, and new data sources identified, cloud computing is a fast, low cost option to initiate disruptive big data analytical projects. Hadoop and Spark are often available as a service in the cloud. Data like social network and open government data can be loaded into cloud storage so exploratory analysis on that data can start quickly. Data preparation and integration jobs can be developed to process data at scale on the cloud before being analysed by scalable in-memory Spark analytic applications that use Spark streaming, machine learning (MLlib) and/or graph analysis (GraphX). This can optionally be switched to on-premises later.

LOGICAL DATA WAREHOUSE

Data virtualisation or federated SQL query engines can simplify data access and introduce flexibility via support for multiple virtual views of underlying data

Figure 3 shows how data virtualisation can be used to create the logical data warehouse. This can either be done by a data virtualisation server or by a 'SQL on Everything' engine that can connect to multiple sources, optimise queries and federate them across non-relational data (e.g. Hadoop, Spark, NoSQL DBMSs) and relational DBMS data sources. By hiding complexity from users, it becomes possible to create the concept of a logical data warehouse with different virtual personalised views across multiple heterogeneous data stores. This of course requires a SQL optimiser to be able to pushdown analytics close to the data and also to work out the best way to join Hadoop to non-Hadoop data.

ORGANISE FOR SUCCESS USING PUBLISH AND SUBSCRIBE

Creating a production line approach to building insight clearly defines roles and encourages reuse of data and analytics to speed up the production of insights

In addition you can organise for success, creating a 'production line' approach so that people in different parts of the production line build on what others before them have created. This can be done via a publish-and-subscribe development approach and an information catalog to keep track of what is available. This is shown in Figure 5 with every task building on the previous by taking services produced by one task as input to the next. The first task is creating trusted data services. These services are then published for others to pick up to develop DI/DQ flows that join data from multiple trusted data flows. Newly developed integrated data workflows are then themselves published as services. Other people can then pick these up as a 'quick start' to creating analytical workflows. Analytical workflows take integrated trusted data and analyse it to produce scores or other insights. These analytical workflows are then published and made available for others to either embed in applications, visualise the results of the analysis, use as input to a decision engine to create decision services (e.g. next best offer recommendation) or turn into analytic applications.

A publish and subscribe production line approach together with an information catalog encourages reuse and can significantly reduce the time to produce disruptive insights

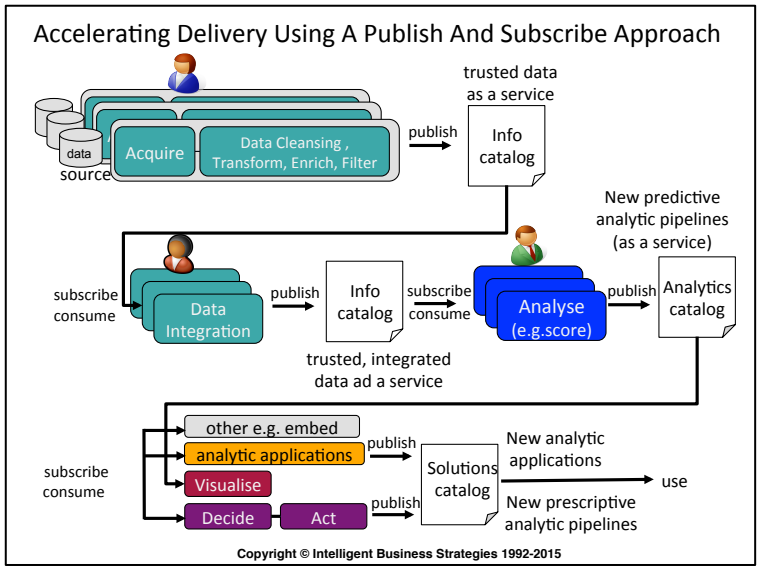

Figure 4

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

IBM has been a provider of big data and analytics technologies for a number of years

Having defined data and technical requirements together with a new logical data warehouse architecture and agile development approaches, this section looks at how one vendor steps up to meet the requirements defined earlier in order to enable disruption. That vendor is IBM.

IBM has a number of technology components that together can be used to build a logical data warehouse architecture combining traditional and big data technologies. These include:

Their portfolio of analytical technology components includes Hadoop, an analytucal RDBMS, streaming analytics, a federated SQL engine, an information management tool suite and a range of analytical tools

- A Hadoop distribution
- Apache Spark in-memory analytics execution engine
- An analytical relational DBMS for enterprise data warehouse
- A streaming analytics platform
- A cloud-based analytical DBMS
- SQL access to Hadoop and traditional data warehouse data
- An EIM tool suite that includes scalable data cleansing and integration

IBM BIGINSIGHTS AND OPEN DATA PLATFORM WITH APACHE HADOOP

IBM's distribution of Apache Hadoop consists of two major components. These are:

Two major components comprise the core of IBM's Hadoop offering

- IBM Open Platform with Apache Hadoop
- IBM BigInsights for Apache Hadoop

IBM Open Platform with Apache Hadoop is a 100% open source common Open Data Platform (ODP) core of Apache Hadoop (inclusive of HDFS, YARN, and MapReduce) and Apache Ambari software. The following components ship as part of this:

IBM Open Platform with Hadoop includes core Hadoop technologies such as HDFS, Hive, Pig, Spark and HBase

Component	Description
Ambari	Hadoop cluster management
Apache Kafka	Scalable message handling for inbound streaming data
Flume	Web log data ingestion into Hadoop HDFS
HBase	Column family NoSQL database for high velocity data ingest and operational reporting
HDFS	Hadoop Distributed File System that partitions & distributes data across a cluster
Hive	SQL access to HDFS and HBase data
Knox	An API Gateway for protection of Rest APIs
Lucene	Java-based indexing and search technology
Oozie	Scheduling
Parquet	Columnar

Pig	Scripting language for data processing
Slidr	A YARN application to deploy distributed applications on YARN
Solr	Search server built using Lucene Core
Spark	A massively parallel in-memory execution environment for Java, Python, Scala and R based analytic applications doing machine learning, graph analysis and streaming analytics. It also supports SQL access to in-memory data
Sqoop	Data mover from relational to HDFS and vice-versa. It also supports other sources and targets
Zookeeper	An open-source server which enables highly reliable distributed coordination

IBM BigInsights includes a set of packaged modules aimed at different types of user

IBM BigInsights for Apache Hadoop is a collection of value-added services that can be installed on top of the IBM® Open Platform with Apache Hadoop or any incumbent Hadoop distribution (e.g. Cloudera, MapR etc.). It offers analytic and enterprise capabilities for Hadoop and includes the following modules:

Data scientists have access to a scalable version of R and machine learning algorithms optimised for Hadoop

- BigInsights Data Scientist - includes native support for the R programming language (Big R) and adds machine learning algorithms that are optimised for Hadoop. It also includes various analyst capabilities along with Scaling R packages to Hadoop, using declarative machine language text analytics and distilling unstructured data into meaningful data. It provides web-based tooling for annotation. Native support for open source R statistical computing helps clients leverage their existing R code or gain from more than 4,500 freely available statistics packages from the open R community.

Business analysts can use BigSheets and BigSQL to join Hadoop to non-Hadoop data

- BigInsights Analyst - helps discover data in Hadoop for analysis. The module includes IBMs SQL-on-Hadoop engine Big SQL and also BigSheets, which is a spreadsheet-like visualisation tool.

Administrators have access to tools to manage monitor and secure a Hadoop cluster

- BigInsights Enterprise Management - includes tools to help administrators manage, monitor, and secure their Hadoop distribution. This includes tools to allocate resources, monitor multiple clusters, and optimise workflows to increase performance.

APACHE SPARK

IBM Technology Integration With Apache Spark

IBM has made a strategic commitment to using Apache Spark

A number of IBM software products now integrate with Apache Spark. Some of these are shown in the table below along with a description of how they integrate.

IBM Product	Description of Integration with Spark
IBM SPSS Analytic Server and Modeler	Can invoke Spark MLlib algorithms from IBM SPSS model workflows
IBM BigSQL	Spark analytic applications can access data in HDFS, S3, HBase and other NoSQL data stores using IBMBigSQL which will return an RDD for processing. Also IBM BigSQL can chose to leverage Spark if required in answering SQL queries

A number of IBM software products are now using Apache Spark

IBM Streams	Add Spark transformation functions, action functions and Spark MLlib algorithms to existing Streams applications
IBM Cloudant on Bluemix	Data in IBM Cloudant can be accessed and analysed in Spark analytic applications on the IBM Bluemix cloud
IBM BigInsights on Bluemix	Data in IBM Open Platform with Apache Hadoop can be accessed and analysed in BigInsights Data Scientist analytic applications using Spark on the IBM Bluemix cloud
IBM Swift Object Storage	Data in IBM Swift Object Storage can be accessed and analysed in Spark analytic applications
IBM IoT on Bluemix	A fully managed, cloud-hosted service that makes it simple to derive value from Internet of Things (IoT) device
IBM Insights for Twitter Service	Allow developers and entrepreneurs to search, quickly explore and then mine enriched Twitter content

In fact, IBM's entire analytics platform is built on Spark (see Figure 5)

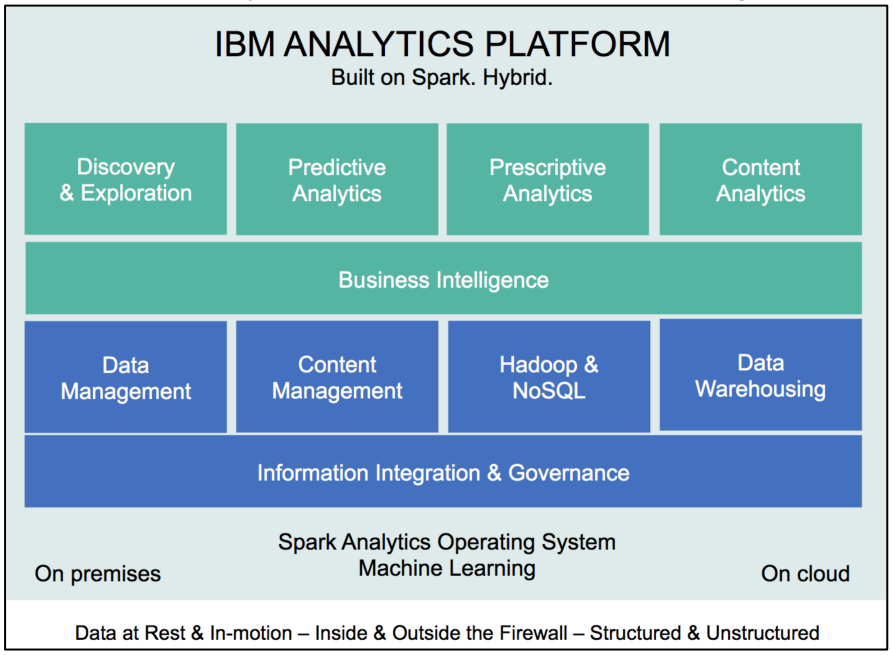

Figure 5

Spark-as-a-Service on IBM Bluemix

IBM has also made Spark available as a service on IBM Bluemix. Analytics for Apache Spark works with commonly used tools available in IBM Bluemix so that you can quickly start tapping into the full power of Apache Spark. The tools include the following:

- Jupyter Notebooks for interactive and reproducible data analysis and visualization
- SWIFT Object Storage for storage and management of data files
- Apache Spark for data processing at scale

IBM PUREDATA SYSTEM FOR ANALYTICS

IBM PureData System for Analytics powered by Netezza technology is a massively parallel data warehouse and analytic hardware appliance running IBM's Netezza analytical relational DBMS optimised for advanced analytical workloads on structured data. In includes a range of in-database analytics so that visual discovery tools can invoke these on the fly to get good performance on large volumes of structured data.

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

The IBM PureData System for Analytics 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 requires no indexing or tuning which makes it easier to manage. It is designed to interface with traditional BI tools including IBM Cognos Analytics and also runs IBM SPSS developed advanced analytical models deployed in the database on large volumes of data.

IBM PureData System for Analytics provides free in-database analytics capabilities allowing you to create and apply complex and sophisticated analytics right inside the appliance.

Complementing the IBM PureData System for Analytics, is an advanced analytics framework. In addition to providing a large library of parallelised 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, and Fuzzy Logix.

IBM PureData System for Analytics allows you to create, test and apply models to score data right inside the 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 another computer.

IBM dashDB

IBM dashDB is a new massively parallel cloud-based DBMS offering data warehouse-as-a-service

It can also be deployed in a private cloud

IBM dashDB is a fully managed, cloud-based MPP DBMS enabling IBM to offer data warehouse-as-a-service. It provides in-database analytics, in-memory columnar computing and connectivity to a wide range of analytical toolsets, including Watson Analytics and many third-party BI tools.

A second deployment option of IBM dashDB, currently in early access preview, is also available for fast deployment into private or virtual private clouds via Docker container.

IBM Data Integration For The Big Data Enterprise

In terms of scalable data cleansing and integration, IBM provides the following new and extended tooling capabilities.

IBM BigInsights BigIntegrate and BigInsights BigQuality

IBM has extended its data integration and data cleansing software to support big data sources by releasing IBM BigIntegrate and BigQuality

IBM BigInsights BigIntegrate and BigInsights BigQuality are data integration and data cleansing components that run on a number of Hadoop distributions. These include IBM BigInsights and IBM Open Platform with Apache Hadoop, Hortonworks HDP and Cloudera. BigInsights BigIntegrate and BigInsights BigQuality are capable of supporting large data integration workloads. This is because the execution architecture is a shared nothing massively parallel implementation. They include the following capability:

- Connectivity to multiple data sources. These include:

IBM BigIntegrate and BigQuality run in parallel across a cluster and can process data in HDFS, Hive, NoSQL DBMSs, relational DBMSs, files, and packaged applications

Streaming data can also be accessed

In-Hadoop analytics can also be invoked during data intrgration processing

IBM BigIntegrate and BigQuality can run on a range of multi-processor environments

IBM BigIntegrate and BigQuality also integrate with IBM business glossary and data modelling software

Data integration jobs can also be published in an information catalog to offer trusted information services to information consumers

- The Hadoop Distributed File System (HDFS)
- Hadoop Hive tables
- NoSQL DBMSs e.g. Apache HBase, Mongo DB and Cassandra
- JSON data
- Streaming data from IBM Streams to pump filtered event data into IBM BigInsights for further analysis
- Java Message Service [JMS]
- Relational DBMSs e.g. IBM PureData System for Analytics, IBM Distributed, DB2 IBM DB2 z/OS data warehouse, IBM DB2 Analytics Accelerator, 3rd party RDBMSs
- Flat files
- IBM InfoSphere Master Data Management
- Popular packaged applications

- Change data capture from big data sources
- Ability to invoke in-Hadoop analytics including Text Analytics via Java APIs to invoke AQL
- Scalable data processing that includes:
  - Data partitioning and distribution of data across Hadoop nodes
  - Parallel execution of all data cleansing and data transformation tasks on any hardware configuration. This includes partition *and* pipeline parallelism with repartitioning of data between stages and between nodes without needing to persist intermediate data to disk
  - Ability to run unaltered on single processor MPP nodes, clustered multi-core, multi-processor SMP nodes and full grid computing environments
  - No upper limit on data volumes, processing throughput and numbers of processing nodes
  - Ability to run natively on Hadoop YARN bypassing Hadoop frameworks such as MapReduce, Tez and Spark
- Integration with business glossary and data modeling software in the same EIM platform via shared metadata
- An information catalog and the ability to publish data integration jobs as data services in InfoSphere Information Governance Catalog so that information consumers can see what data services are available for them to shop for and order via InfoSphere Data Click

IBM STREAMS – REAL-TIME OPTIMISATION USING BIG DATA

IBM Streams is IBM's platform for developing real-time analytic applications on streaming data

This includes sensor data from supply chains, production lines, assets, IoT data, financial markets data, and also unstructured

IBM Streams is a platform 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, monitoring sensor data to understand product performance and usage in the Internet of Things (IoT), neonatal ICU monitoring, real-time fraud prevention and real-time multi-modal surveillance in law enforcement. IBM 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 Streams 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 IBM Streams analytic application workflows to predict business impact of event patterns.

Events can be analysed and acted upon in real-time or filtered and stored for further analysis and replay in IBM BigInsights

Scalability is provided by deploying IBM Streams applications on multi-core, multi-processor hardware clusters optimised for real-time analytics and via integration with Apache Spark. Events of interest to the business can also be filtered out and pumped to other IBM analytical data stores for further analysis and/or replay. IBM Streams can therefore be used to continually ingest data of interest into IBM BigInsights to analyse. It is also possible to summarise high volume data streams and route these to IBM Cognos Analytics for visualisation in a dashboard for further human analysis.

ACCESSING THE LOGICAL DATA WAREHOUSE USING IBM BIG SQL AND IBM FLUID QUERY

IBM can also simplify access to multiple data stores via a federated SQL interface

Users and analytic applications need access to data in a variety of data repositories and platforms without concern for the data's location or access method or the need to rewrite a query. To make this possible, IBM provides Big SQL and Fluid Query.

IBM Big SQL

IBM Big SQL can access data in both Hadoop and relational DBMSs

IBM Big SQL can be used to create a Logical Data Warehouse layer over the top of multiple underlying data stores

IBM Big SQL is IBM's flagship multiple-platform SQL query engine for accessing Hadoop and non-Hadoop data or both. It therefore creates a Logical Data Warehouse layer over the top of multiple underlying analytical data stores and can federate queries to make those platforms work locally on the necessary data. Business analysts can connect directly to Big SQL from self-service BI tools that generate SQL. Data scientists and IT developers who want to access Hadoop and non-Hadoop data using SQL from within their analytic applications can also use it.

IBM Big SQL is a Spark compliant massively parallel SQL engine that can be used by Spark analytic applications and BI tools to query Hadoop and non-Hadoop data

IBM Big SQL provides support for the full 2011 ANSI SQL standard

It supports complex data types, OLAP functions and can use UDFs to analyse unstructed data

When processing in-bound SQL, IBM Big SQL bypass Hadoop MapReduce, Tez and Spark execution environments. Instead it runs natively under YARN on a Hadoop cluster with direct access to HDFS and HBase data. Big SQL is fully integrated with the Hive metastore and can therefore see Hive Tables, Hive SerDes, Hive partitioning and Hive Statistics. It is also fully compliant with Spark but does not require Spark. By that we mean that IBM Big SQL can be used by Spark analytic applications written in Python, Java, Scala and R to access data as an alternative to Spark SQL. This works because Spark applications can access data via Big SQL and get back Spark RDDs to analyse data in memory across a cluster. The difference is that there is more functionality in Big SQL, it is fully ANSI 2011 compliant and has optimisation capability to perform query re-write to improve performance. It supports aggregate, scalar and OLAP functions, virtual tables, JAQL UDFs for analysis of unstructured data, and data types such as STRUCT, ARRAY, MAP and BINARY to handle more complex data. In addition, it supports Hadoop columnar file formats such as ORC, Parquet, and RCFile and has no proprietary storage format of its own. In terms of security, IBM Big SQL offers role-based access plus column and row security. IBM Big SQL can also potentially 'push down' query functionality into Spark to execute if it deems it necessary (e.g. to make use of GraphX functions).

IBM Fluid Query

IBM Fluid Query is included with IBM PureData System for Analytics

IBM Fluid Query is included with IBM PureData System for Analytics. It provides access to data in Hadoop from IBM PureData System for Analytics appliances and enables the fast movement of data between Hadoop and IBM PureData System for Analytics appliances. IBM Fluid Query allows queries against PureData System for Analytics, Hadoop or both by merging results from PureData System for Analytics database tables and Hadoop data sources thus creating powerful analytic combinations. Therefore existing queries, reports, and analytics can run against data on Hadoop, in addition to the data in a PureData System for Analytics appliance.

IBM Fluid Query allows BI tools and analytic applications to query data in PureData System for Analytics, Hadoop or both

Data in PureData System for Analytics, Hadoop can be joined to data in Hadoop

This means that IBM Fluid Query allows IBM PureData System for Analytics to route a query (or even part of a query) to the correct data store. It provides IBM with the ability to combine data in PureData System for Analytics with Hadoop data without having to know how to connect to Hadoop. IBM Fluid Query connectivity to a 'back end' Hadoop can be via Hive, IBM Big SQL, Cloudera Impala or SparkSQL.

Analytical Tools

A wide range of tools can leverage scalable analytics running in the IBM Analytics Platform

A wide range of third party and IBM analytical tools like IBM Cognos Analytics, IBM Watson Analytics, IBM SPSS, BigSheets, IBM BigR and IBM BigSheets can all make use of BigSQL and Fluid Query to access data and invoke in-database, in-memory, in-Hadoop and in-stream scalable analytics running in the IBM Analytics Platform.

CONCLUSIONS

Organisations need to understand what they want to achieve though the production of disruptive insight

With technology advancing rapidly companies need to work out how to piece together and integrate the necessary components to maximise their ability to produce disruptive insight. They need to be able to define what they need to produce to cause disruption, understand their data and analytical requirements and then select the technology components necessary to be able to get started. It should be possible to get started quickly in the cloud and then move on-premises if needs be.

They need to understand data and analytical requirements before selecting technology components

It should be possible to start in the cloud and move on-premises or create a hybrid solution

In addition they also need to be able to produce disruptive insight in a productive manner without the need for major re-training to use new technologies like Hadoop, Spark and streaming analytics. To that end, if companies can make use of existing tools used in traditional data warehousing to also clean, integrate and analyse data in big data environments then the time to value will come down significantly. Also, tools should be able to exploit the scalability of underlying hardware, when data volumes and data velocity is high, without users needing to know how that is done.

Exploiting existing tools and skillsets that can exploit traditional and big data technologies fuels productivity

It should also be possible to simplify access to data in multiple data stores and join data across multiple data stores so that complexity is kept to a minimum. This is true irrespective of whether a business analyst needs to access data from a self-service analytics tool or whether it an IT developer or data scientist needs to access data from a custom built analytical application. As such a common federated SQL layer is needed to create a 'Logical Data Warehouse'.

IBM is building an architecture for Big Data and traditional analytics on the cloud and on-premises

IBM is pursuing all of this both on the cloud and on-premises with the ability deploy BigInsights, Apache Spark Open Platform with Apache Hadoop both on the cloud and on premises. In addition it is creating a common scalable analytical RDBMS code base that works across its dashDB, Big SQL, DB2, PureData System for Analytics Appliances and as a software version on private cloud with dashDB for software defined environments. Also Spark is being integrated everywhere to push down analytics to make them run as close to the data as possible. In addition, Big SQL and Fluid Query simplify access to traditional data warehouses and Hadoop, helping to create a logical data warehouse layer. And there is more to come.

Data virtualisation and federated query support simplifies access to traditional and big data stores by creating a Logical Data Warehouse

Today we are beyond the point where big data is in the prototype stage. We are entering an era where automation, integration and end-to-end solutions need to be built rapidly to facilitate disruption. Companies need to architect a platform for Big Data (and traditional data) analytics. Given this requirement, IBM would have to be a short list contender to helping any organisation become a disrupter whether it be on the cloud or on-premises.

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 34 years of IT experience, Mike has consulted for dozens of companies on business intelligence strategy, big data, data governance, master data management and enterprise architecture. 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