WHITE PAPER

INTELLIGENT BUSINESS STRATEGIES

Unified Data Integration In A Distributed Data Landscape

By Mike Ferguson
Intelligent Business Strategies
February 2016

Prepared for: DIYOTTA

Table of Contents

The Increasingly Distributed Data Landscape 3
The Exponential Growth of Data Stores 3
Cloud and On-premises Operational Systems 3
Cloud Storage 3
Master Data 3
Analytical Systems 4
Big Data 4
The Internet of Things 4
External Data 4

The Modern Analytical Ecosystem 5

Data Management Issues In A Distributed Environment 6

Data Integration Use Cases In A Distributed Data Lake 8

Data Integration Requirements In A Distributed Data Landscape 10

Managing Distributed Data Integration Using Diyotta 12
Diyotta Data Integration Suite 12
Organising Metadata Specifications For Productivity and Reuse 13
Unifying Data Integration In A Heterogeneous Environment 13

Conclusions 15

THE INCREASINGLY DISTRIBUTED DATA LANDSCAPE

Data complexity is increasing

Over the last several years, data complexity has increased dramatically in many companies and continues to do so as more and more data is captured and more databases and object stores emerge to store it.

Yet the thirst for data continues with new data sources emerging almost on a daily basis. This includes both internal and external data sources.

THE EXPONENTIAL GROWTH OF DATA STORES

Cloud and On-premises Operational Systems

Core transaction processing systems are now on the cloud as well as on-premises

NoSQL databases are being adopted to allow web and mobile commerce applications to capture non-transactional data at scale

Even in traditional environments we have seen complexity increase. Core on-line transaction processing (OLTP) systems have spread outside the firewall as companies adopt cloud-based packaged applications such as Salesforce.com and Workday. In addition, digitalisation has caused explosive growth in the rates at which session data and transactions need to be captured now that web, mobile and social commerce are all occurring. Also customer-facing applications provide a much richer user experience today, storing non-transaction data as well as transaction data. For example, session data, user profiles, shopping cart contents and product reviews. NoSQL databases have emerged to store this kind of data and underpin new scalable operational applications. The Internet of Things (IoT) will undoubtedly increase NoSQL database adoption even further to scale IoT application data capture as the number of devices / things increase. Operational systems therefore now include relational and NoSQL data stores both on-premises and on the cloud and there is a need to move data between them and extract data from these systems for use in on-premises and cloud based analytical systems.

Cloud Storage

The adoption of cloud storage is also increasing – especially for capturing big data

In addition, the adoption of cloud storage is also increasing rapidly. Amazon S3, OpenStack Swift, Microsoft Azure Storage, Google Cloud to name a few. Increasingly many companies are storing corporate data and device data in the cloud. It is also a popular mechanism to enable data to be shared across business units, departments and partners. Also big data such as sensor data is often captured and stored in the cloud and databases like Amazon Redshift have seen significant uptake.

Master Data

Master data is still fractured in many organisations and scattered across systems

Master data and reference data such as Customer, Product, Asset, Employee, Site etc., are among the most widely shared data in any organisation. Many companies have struggled over the years to keep multiple copies of this data consistent, correct and synchronised. Today that problem is still there. Subsets of master data reside in OLTP systems (on-premises and on the cloud) as well as in analytical systems such as data warehouses, operational data stores and big data platforms. Although many companies have implemented one or more Master Data Management (MDM) systems to support different entities, still today, this problem is not fully solved with the need to access this kind of data and provide synchronised copies of it more acute than ever.

Analytical Systems

Multiple data warehouses have been built creating islands of overlapping historical data

In the area of analytical systems, gone are the days of a single enterprise data warehouse. Today, many companies have multiple data warehouses. Furthermore, many have also added data warehouse appliances. The result is that the analytical landscape includes multiple data stores with 'islands' of overlapping historical data.

Big Data

Big data platforms like Hadoop and Graph DBMSs have entered the enterprise extending analytical environments beyond the data warehouse

Big data has entered the enterprise as companies look to analyse it to produce new insights. Examples of this include sensor data (either on the cloud or on-premises), social media data, click stream, machine data (application server logs, database logs, IVR system logs) and much more. This has resulted in new scalable data stores being adopted such as Hadoop and NoSQL databases like graph DBMSs. All of this has increased the complexity of the analytical landscape with different types of data store supporting different types of analytical workloads. In addition, data is flowing between analytical data stores from Hadoop to data warehouses, from data warehouses to Hadoop (e.g. archived data), from MDM to Data Warehouses and Hadoop, from Unix file systems to Hadoop, from NoSQL DBMSs to Hadoop, from Hadoop HDFS to HBase, from HDFS to Hive and even from one part of HDFS to another. It's a flurry of activity 'alive' with data on the move.

The Internet of Things

The Internet of Things is spawning a 'Tsunami' of data sources

Also, everything is becoming *smart*. Phones, buildings, cars, watches, household appliances etc., all have 'sensors inside' emitting data 24x365. Billions of things are connecting up to the Internet and emitting data machine-to-machine (M2M). The Internet of Things is causing a 'Tsunami' of new data sources to emerge.

External Data

External data sources are also emerging offering data to enrich what we already know

Finally, external data sources are emerging offering hundreds of thousands of data sets containing Open Government Data, Weather Data, financial data and much more to businesses.

Data now exists in a multitude of data stores creating a distributed data lake with some data now too big to move

Data privacy is now a major issue and is keeping data apart

Looking at all of this, the complexity is now a real challenge and the idea that data is all going in one central data store is far from reality. Like it or not, for many, the so-called 'data lake' is distributed across many data stores. We are now in an era where data is increasingly becoming distributed and the number of data sources is increasing rapidly. Yet despite this, business is demanding more agility. This stark reality is clear. We are facing a totally new challenge in data management with two added complications. Firstly some data is now so big it is too big to move and secondly, the data collected may be stored in different geographies and legal jurisdictions where it is subject to multiple sets of often conflicting data protection laws.

THE MODERN ANALYTICAL ECOSYSTEM

Having painted a picture of increasing complexity, it is worth taking a more detailed look at how analytical systems have evolved. The emergence of big data has resulted in new analytical workloads that are not well suited to traditional data warehouse environments. These workloads, typically being driven by data characteristics (variety, velocity and volume) and the types of analysis required, have caused many companies to extend their analytical set-up beyond the data warehouse to include multiple analytical data stores.

Multiple platforms now exist in the enterprise to support different analytical workloads

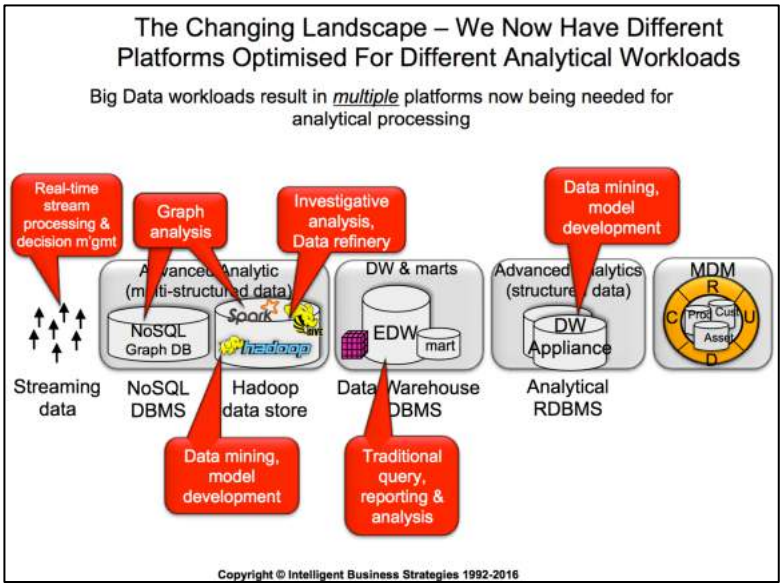

Figure 1

As a result data integration and data movement has increased rapidly across data stores in this new analytical ecosystem

This has resulted in a rapid increase in the amount of data ingestion and movement across the modern analytical ecosystem. Figure 2 shows some of the popular data integration paths that have emerged.

Structured, semi-structured and unstructured data now need to be integrated

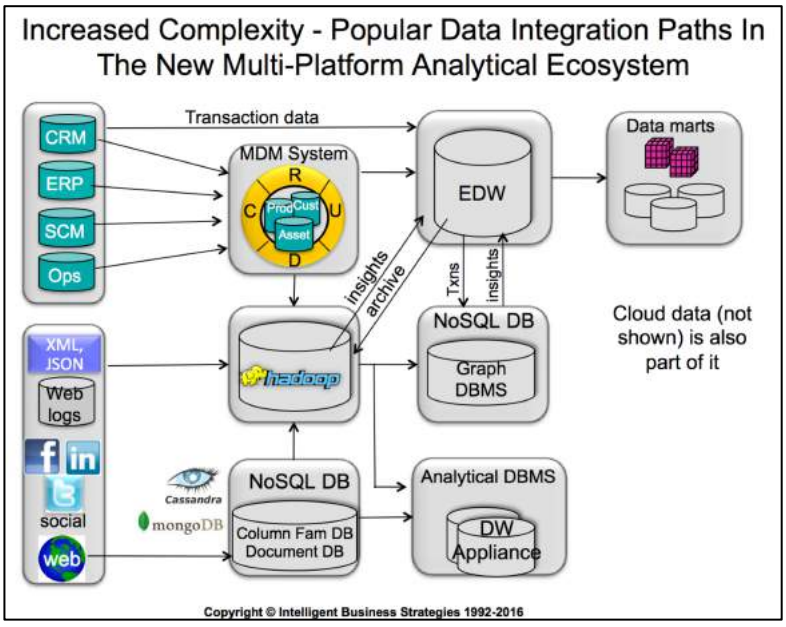

Figure 2

DATA MANAGEMENT ISSUES IN A DISTRIBUTED ENVIRONMENT

The distributed data landscape is causing increased complexity

It is clear from Figure 2 that complexity is increasing and this is without the diagram showing what is happening on the cloud.

Different data integration technologies are being used in different parts of the ecosystem

In addition, in many cases, data integration on several of the paths shown across the ecosystem may be happening using different technologies. The technology used for ETL processing for a data warehouse or an MDM system may not be the same as that used to prepare data on the cloud or in Hadoop. One reason for this is because modern analytical environments are bi-modal[1]. 'Traditional' parts to this ecosystem, notably data warehousing and MDM, are classified as production systems today. They are 'nailed down', and heavily governed by IT when it comes to change management and choice of technology used. Yet, other parts of the analytical ecosystem are more agile in nature. Data marts, the big data world and the cloud to some extent are examples of this, where such controls are not necessarily in place and business has a lot more freedom to use whatever tools they want.

Both production and agile self-service data integration technologies are being used and silos have emerged

It is not surprising therefore, that in many cases a project oriented, siloed approach to data integration has emerged with limited collaboration among business and IT. The net result is that a less than optimal set-up has emerged where:

The cost of data integration is too high

- The cost of data integration is too high
- Multiple DI/DQ technologies and techniques are being used that are not well integrated or not integrated at all

Data integration is fractured

- Hand-coding is occurring with scripts that are difficult to maintain
- Development is slow
- There is risk of duplicate inconsistent DI/DQ cleansing and transformations rules occurring for the same data

Multiple tools are in use and metadata is often not centralised

- Maintenance of DI/DQ rules is complex and slow because changes to rules may have to be implemented in multiple places
- Re-invention is occurring rather than re-use
- Metadata specifications are fractured across multiple tools or no metadata at all in some cases

Re-invention rather than reuse is occuring

- Metadata lineage is unavailable in many places especially with hand-coded big data applications that do data cleansing and integration
- Skill sets are fractured across different projects

Also many companies are rapidly reaching a point where a 'data deluge' is occurring in that data is now arriving faster than they can consume it. One

[1] Gartner's Bi-Modal IT (http://www.gartner.com/it-glossary/bimodal) describes Mode 1 and Mode 2 environments. Mode 1 being a production IT environment and Mode 2 being an agile environment.

Business users and IT are both now involved in integrating data

reason often given for this is that IT has become a bottleneck in ETL processing and can't keep place with business demand. There is often merit in this observation given the sheer number of data sources now available to businesses. However, while it is clear that IT need help to cope with the data deluge, just giving out self-service data integration tools to business analysts and data scientists can be fraught with problems if it is not done in a controlled, coordinated manner where both IT and business work together to process and govern data. Self-service data integration may provide more agility and clear the bottleneck in IT but if it is not governed then chaos could very easily set in. Figure 3 shows an example of this where both IT and business are doing data integration that is out of control. People are accessing data in multiple data stores anywhere and everywhere in a distributed data environment.

Un-coordinated use of self-service data integration tools can lead to chaos

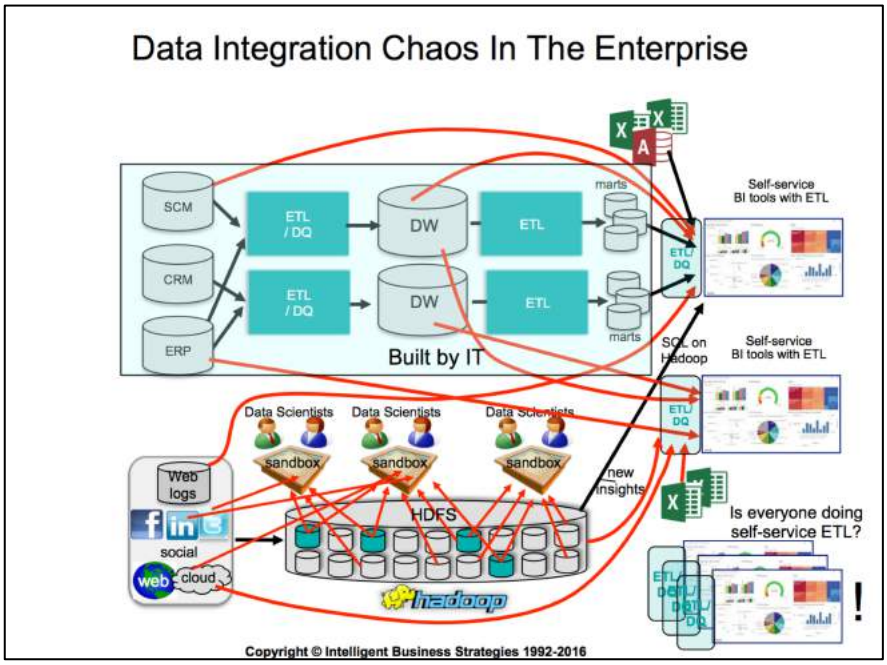

Figure 3

Even on Hadoop data integration chaos has broken out

Furthermore, even on Hadoop, chaos has broken out with data science teams independently accessing any and all files in HDFS including, in some cases, the same data. That means the same data is being prepared differently across projects potentially leading to inconsistency. In addition what happens when data is on the cloud and too big to move? The cloud is not shown on Figure 3, but if IoT data for example is collected in the cloud and is too big to move, do business users start building ETL jobs to move on-premises data outside the firewall to the cloud to integrate with it? If so, who governs that kind of activity? In a distributed environment chaos can breed more chaos.

The conclusion here is obvious. There has to be a better, more governed way to fuel productivity and agility without causing data inconsistency and chaos. Everyone for themselves is not an option.

DATA INTEGRATION USE CASES IN A DISTRIBUTED DATA LAKE

Data is being collected via streaming, batch ingest, replication and archiving with some data too big to move once captured

Figure 4 shows a data reservoir[2] where data is collected in many different ways including streaming, batch ingest, replication and archiving. In this kind of environment, what happens if some data is too big to move once collected? We have already alluded to that just a moment ago with respect to IoT data. If data is too big to move, the data reservoir is distributed by default. Yet we still somehow have to be able to process it.

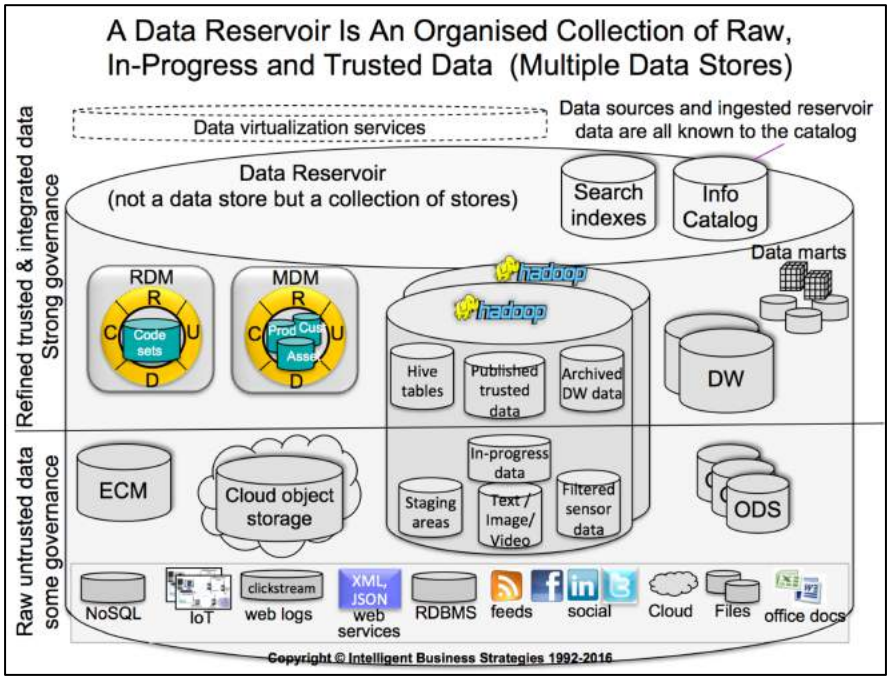

Figure 4

Data lakes / reservoirs are increasingly becoming distributed

Also it is unlikely that all IoT data will be captured and stored centrally before being processed as analysis needs to happen in the network itself

Also, in an IoT environment does it make sense to bring all sensor data to a central data store to integrate it when the analytics need to happen at the edge in order for a solution to scale? Should data integration not also happen at the edge to feed analytics deployed there?

Compliance with different data privacy laws in different jurisdictions around the world is a key reason why some data will be kept apart

Even if data volume is not an issue in your organisation and you don't have an IoT initiative, it is not realistic to say that all data ends up in one huge physical centralised data store. There are many reasons why data may remain distributed. Probably the most dominant reason has nothing to do with data volume. It has to do with compliance caused by different countries around the world introducing different data privacy laws that result in companies having to ensure that they remain compliant in all jurisdictions within which they do business. This will keep data apart.

[2] The term data reservoir is preferred instead of data lake simply because, water flows into a lake and goes nowhere, whereas water flowing into a reservoir is processed for consumption. Therefore data reservoir is seen as a better analogy for this paper.

If data in a reservoir is distributed then processing it has to be managed as if it were centralised

Not all data will be processed where it is located

If it is accepted that the data reservoir is distributed then it follows that any data management software needs to work across multiple data stores and manage processing *as if it were centralised*. The software needs to sit above all data stores and push processing down to the appropriate places to make the platforms do the work (see Figure 5). We need to take the processing to the data and not the data to the processing. We also need to recognise that some data providers (e.g. devices) are just providers, some repositories are places where we can process data and some others are just consumers. Processing does not have to happen in all locations. Also, if a repository has the ability to scale, then the software should take advantage of the underlying capability and exploit it. If all this occurs then multiple use cases naturally emerge without the need to necessarily pin all hopes on getting all data into a single Hadoop system to process.

Data integration software should exploit the power of underlying platforms to scale ETL

Hadoop, Spark and massively parallel relational DBMSs are good examples of where this could happen

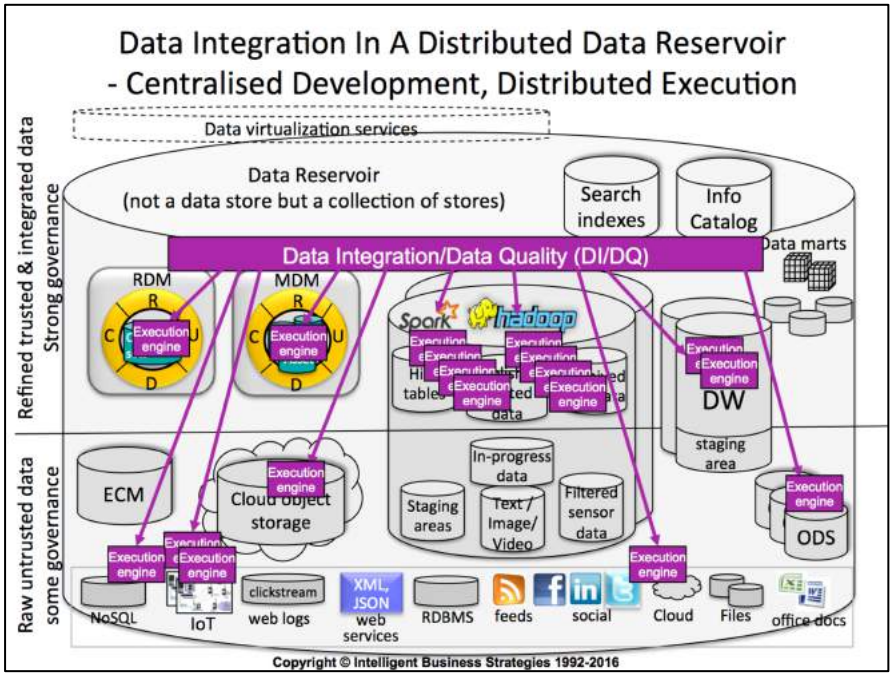

Figure 5

Centralised development, centralised metadata and distributed execution is a flexible, powerful combination

In Figure 5 the software allows for centralised development and distributed execution. That means all data cleansing and transformation rules (metadata specifications) are stored centrally and can be reused. It also means that underlying heterogeneous technologies can be exploited to improve performance. You can build data warehouses, process data in the cloud, bring relational data into Hadoop (using Sqoop for example) and integrate it at scale with multi-structured data in HDFS. You can do all of this while ensuring that as much cleansing, filtering and transformation happens locally before retrieving to minimise network traffic. You can cope with data being too big to move and even invoke local data integration jobs that already exist from a global workflow to unite processing across the entire ecosystem and distributed analytical environment.

DATA INTEGRATION REQUIREMENTS IN A DISTRIBUTED DATA LANDSCAPE

New requirements have emerged to be able to integrate data in a distributed environment

Given this potential possibility, what then are the requirements for data integration in a distributed data landscape? Some key ones follow.

It should be possible to:

Integrate multiple data types in-motion and at rest

- Process structured, semi-structured and unstructured data
- Process streaming data and data at rest

Define once, execute anywhere

- Define data integration rules 'globally' and execute tasks locally
- Centralise metadata specifications so that data lineage for distributed ETL processing is easily accessible from viewer tools and applications

Pushdown processing to exploit scalable platforms

- Push down transformation tasks to exploit the scalability of underlying platforms. Examples here would include, Apache Spark, Hadoop and massively parallel analytical relational DBMSs

Execute in a hybrid environment

- Execute data integration jobs across a hybrid computing environment of cloud and on-premises systems
- Automate tasks such as data profiling, address data cleansing, text tokenization etc., and also to recommend transformations

Nest workflows and invoke 3rd party data integration jobs

- Nest workflows so that one ETL workflow can call another (e.g. via REST or SOAP web service APIs) as a transformation task. This is to enable data integration 'pipelines' to be broken up into re-useable executable components which helps improve productivity, stop re-invention and reduce time to value
- Invoke third party data integration jobs (e.g. via web service APIs) to unite siloed data integration activity across a distributed environment
- Move ETL processing from one platform to another without the need to re-define transformation rules if the data being processed is moved

Support rule versioning for compliance

- Dynamically enforce different versions of rules depending on the type of data and where the processing takes place. This is to allow organisations to remain compliant across multiple jurisdictions in a geographically distributed data landscape
- Invoke in-database, in-Hadoop, in-memory (e.g. Spark) and in-stream analytics as part of an ETL process to automate analytical processing

Data integration as a service

- Publish data integration and data cleansing workflows and components as services to a catalog to enable business and IT users to understand what information services are available for re-use and what information services can be invoked and scheduled to produce trusted information across a distributed environment

Data integration on-demand

- It should be possible to invoke published information services on-demand (via an API), on an event-driven basis and on a timer-driven basis
- Have the software determine where to best process data integration tasks to get the best performance and produce the information needed

Smart data integration via an optimiser

- Re-ordering data integration tasks if necessary to optimise data integration execution

Business and IT need to work together

- Provide both IT professional and business user self-service user interfaces to the same data integration platform and metadata repository to allow IT and business users to work together to integrate data, produce trusted information and conquer the data deluge facing many organisations today

MANAGING DISTRIBUTED DATA INTEGRATION USING DIYOTTA

Diyotta is a new vendor offering distributed data integration software

Having understood the requirements, this section looks at how one vendor steps up to meet these requirements in order to integrate data in a distributed data environment. That vendor is Diyotta.

DIYOTTA DATA INTEGRATION SUITE

Diyotta is a provider of distributed data integration software that handles the complexity of multiple platforms in a modern analytical ecosystem. Diyotta Data Integration Suite supports a range of on-premises, cloud-based, and external data sources including:

Diyotta Data Integration Suite supports integration of structured, semi-structured and unstructured data

- Structured data from popular relational DBMSs, mainframes, flat files and cloud-based applications like Salesforce
- Semi-structured data such as JSON and XML
- Unstructured data from social networks like Twitter and Facebook

Data integration jobs are executed in a distributed fashion

What's different about Diyotta Data Integration Suite is that it not only exploits the power of scalable platforms to process data at scale but it can do so while also providing the ability to execute filters, data cleansing, and transformations across multiple platforms as part of the same job. In other words it supports distributed execution of data integration jobs effectively allowing it to manage all data movement and data integration across all underlying platforms in a distributed data reservoir.

Data integration jobs are developed centrally and execute locally

Metadata is stored centrallly

Tasks are pushed down to run close to the data

All data is moved point-to-point

Flexible configuration of ELT processing is possible

Diyotta Data Integration Suite operates in a similar way to that shown in Figure 5 whereby data integration jobs are developed centrally via a web-based design studio module and are executed in a distributed manner on one or multiple locations and platforms. All metadata specifications created in the Diyotta Data Integration Suite Studio are held centrally in an RDBMS-based metadata repository. Data integration jobs can then be executed at scale across multiple platforms using agent technology. Tasks are 'pushed down' to different underlying platforms to run close to the data with no need to land data on an intermediate Diyotta server. All data is moved point-to-point.

Diyotta Data Integration Suite effectively allows data integration architects to design data integration jobs and configure them to execute in a centralised or distributed manner to fit the need. This includes pushing down filtering and transformations to execute locally on source systems to minimise data extracted or to process data where it is if it is too big to move. It can also use agent technology deployed on a target system to remotely pull data from a provider by issuing requests that include filters to only retrieve relevant data. The retrieved data is then processed in parallel on a number of platforms to scale data integration in big data environments.

Exploitation of scalable platforms to transform big data

Pushdown execution is achieved by 'agents' that control execution of data integration tasks on whatever platforms are necessary. 'Push down' scalable ELT processing can be executed on the following types of platform.

Diyotta can leverge the power of Hadoop, Spark and MPP RDBMSs

- Massively parallel relational DBMSs - e.g. IBM PureData System for Analytics, Teradata, Oracle Exadata
- Hadoop - e.g. Cloudera, Hortonworks, IBM BigInsights, MapR
- Apache Spark

The Diyotta Data Integration Suite architecture is shown below.

Lightweight software agents execute tasks on underlying systems filtering data locally, procesing at scale and moving data without going through an intermediate server

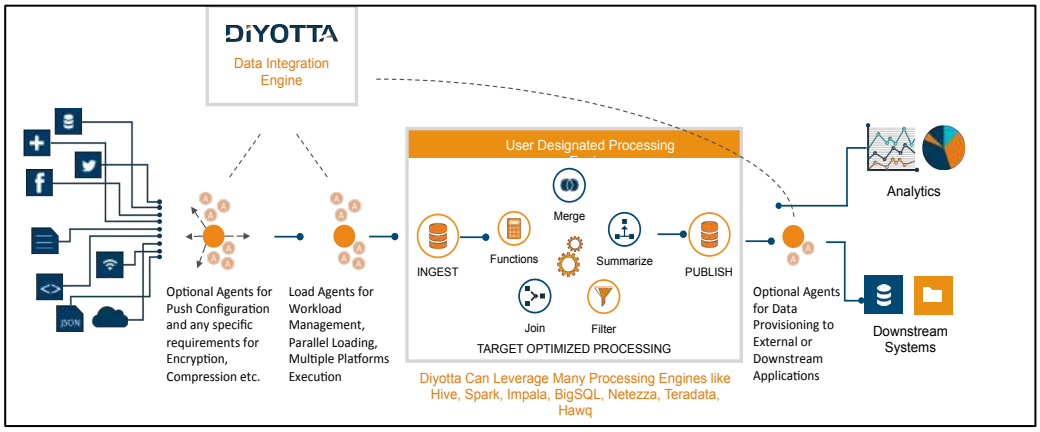

Figure 6

Data integration jobs can be scheduled and monitored

In addition to the aforementioned HTML-based Design Studio, components include other web-based tools to configure data sources and target systems, schedule jobs, monitor their execution, manage user and role-based security, view and analyse metadata.

Organising Metadata Specifications For Productivity And Reuse

Diyotta also enables data integration jobs to be broken up into re-usable components

Data layering enables organisations to standardise design of distributed ELT processing

It also helps to reduce maintenance costs by allowing stages of processing to be maintained separately

Another capability of Diyotta is its ability to break up data integration flows into re-usable components to improve agility and productivity to reduce time to value. It does this by organising metadata into Data Layers. This means that companies can organise themselves to produce trusted information faster which is important in an era where the data deluge is threatening to overwhelm us. By creating a 'production line' of people working on different component pieces (data layers) of data integration flows, it becomes possible to create re-usable components that can be published for others to pick up and run with. Some can be building filtered data source and data landing layers, while others build specific data transformation layers and yet more building data integration layers. If done well, then reuse would dominate and productivity would improve dramatically. Layering work like this also leads to significant reduction in maintenance costs in a distributed environment because data integration job components can be isolated, easily identified and changed without hunting though hundreds of data integration jobs to understand what to change.

Unifying Data Integration In A Heterogeneous Environment

Moving all data to the centre before it can be integrated is not required in a distributed data environment

The way in which Diyotta Data Integration Suite server is architected provides a lot of flexibility in a distributed data environment. Figure 7 shows that it can comfortably manage *point-to-point* data movement and data integration across multiple data stores (as shown in Figure 2) from a single product without forcing everything into a centralised data store. It can therefore potentially cope with multiple Hadoop clusters, deal with data that is too big to move and manage data across multiple legal jurisdictions (a major issue for global organisations). Equally, it can leverage scalable platforms like Hadoop, Spark and MPP relational DBMSs to handle big data if needed. In other words,

Processing can be distributed and re-configured if necessary to provide flexibility

Transformation and integration is taken to the data and not the other way around

Scalable platforms are exploited where it makes sense to do so

Multiple data integration use cases can be supported from a single product in a distributed data environment

Diyotta can also invoke data integration jobs developed in 3rd party tools to help unify silos and coordinate data integration across a distributed data environment

Companies can unify data integration and phase out unnecessay clutter at a pace they can manage to stop data integration chaos from taking hold

Diyotta *centrally* manages data movement and data integration across a distributed data reservoir (lake) meaning that not all data has to be brought into a single Hadoop system. It can also take data integration tasks to the data rather than the other way around via its agent technology *even if* the data is not in Hadoop or a MPP data warehouse DBMS. This bodes well for companies thinking about investment in the Internet of Things as Diyotta could easily build new agents that could be deployed all the way out to the edge of an IoT network to run distributed data integration at the edge as well as in the centre - all from a single, centrally managed product.

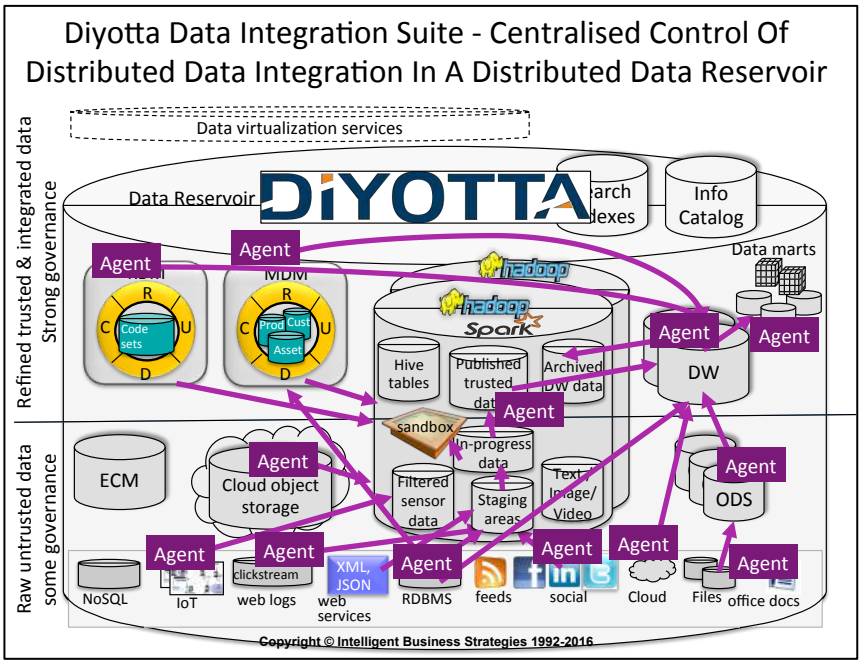

Figure 7

Furthermore, because Diyotta can invoke remote web services, from within a distributed execution, it also means that it becomes possible to unify and integrate multiple data integration technologies being used in different silos across the distributed data landscape. This means that some of the tasks in a distributed execution can reuse jobs already built. This approach applies to both IT developed and business user / data scientist developed self-service data integration jobs. This capability opens up a way to unite data movement and data integration jobs across the enterprise by turning at least some of them into 'proxy' data layers (as discussed earlier) within a distributed data workflow. While the lineage within these external jobs is not available, it potentially opens up a way to isolate these layers and set about reducing complexity in environments where chaos, like that shown in Figure 3, has already taken hold. This means that companies can systematically straighten out the 'spaghetti' shown in Figure 3, get control of the environment and fade out costly complex layers by replacing them with Diyotta at their own pace.

Even if replacement is not required, it means organisations can preserve existing investment and up the silos to improve productivity.

CONCLUSIONS

Multiple workload optimised data stores now exist in a modern analytical environment

Data is being captured in the cloud, on-premises across different geographies and jurisdictions

The need to comply with different data privacy laws is keeping data apart

Data volumes and data sources are increasing

The data landscape is becoming more distributed yet more agility is demanded

New tools are needed to manage and govern data in this complex environment

Business and IT need to work together to deal with the deluge

Yet business must be able to trust the data

There is a seismic shift occuring in data management needs

A centralised data lake is unlikely in many cases and IoT requires data integration at the edge

Diyotta is well placed to help with this challenge

The explosion in the number of data sources, together with the need to analyse new types of data has led many companies to extend their analytical environments beyond the data warehouse to include new data stores and platforms optimised for new analytical workloads. In addition, new data is being captured and stored both in the cloud and on-premises across different geographies and jurisdictions. The result is that data is now housed in multiple data stores inside and outside the corporate firewall in an increasingly distributed data landscape. Also, in some cases, data is now becoming so big that it is too big to move and is subject to potentially conflicting data privacy laws.

As companies continue to instrument business operations with telemetry and increasingly move to lower latency streaming data flowing in from digital channels, sensor networks and the Internet of Things, the challenge of distributed data and the need to integrate it for analysis is set to get worse. Volumes of data are growing rapidly, velocity is increasing and new data sources are appearing everywhere. The result is that data is becoming harder to access because it is in multiple data stores and multiple formats and yet, paradoxically, business is demanding more and more agility, together with the ability to respond much more rapidly than ever before.

In this kind of environment, companies need new tools to manage and govern data ingestion, data integration and data movement across workload optimised analytical systems. They also need the ability to scale to handle volume and velocity as required. In addition, there is a need to integrate structured, semi-structured and unstructured data and make it available in a logical data warehouse to enable rapid production of disruptive insight. Furthermore, companies now need to manage what is clearly a distributed data lake as if it was a centralised system while simultaneously fuelling productivity and reducing the time it takes to produce trusted information and offer it up to consumers as a service. If that is not enough, both business and IT now need to work together on data integration to deal with the deluge but data integration chaos needs to be prevented.

This is a really tough ask. There is a need to comply with legislation in different jurisdictions and to prove this by accessing centralised metadata to be able to know how data was transformed and where it came from.

Without doubt, a seismic shift has emerged in data management requirements and many companies could easily be caught off guard and not realise the magnitude of this problem. It is a major, major issue that has to be dealt with if data-driven strategies are to succeed in business. As data grows the only way to get performance in a distributed data environment is to take integration to the data and not the data to where it needs to be integrated. That applies irrespective of whether or not data is stored in a scalable platform. As we have already discussed, IoT will drive the need to do this all the way to the edge.

Diyotta Data Integration Suite is a clear candidate technology to deal with this new set of requirements, address the data deluge, unify data integration siloes and allow companies to remain agile in a distributed data landscape.

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 independent IT industry analyst and consultant he specialises in Big Data, BI/Analytics, Data Management and enterprise business integration. With over 34 years of IT experience, Mike has consulted for dozens of companies on BI/Analytics, big data, data governance, master data management and enterprise architecture. He has spoken at events all over the world and written numerous 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 IT industry analyst organisation. He teaches popular master classes in Big Data Analytics, New Technologies for Business Intelligence and Data Warehousing, Data Virtualisation 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