WHITE PAPER

INTELLIGENT BUSINESS STRATEGIES

Improving Access to Data for Successful Business Intelligence

Part 3 – Understanding the Key Requirements For Data Access in an Analytical Environment

By Mike Ferguson
Intelligent Business Strategies
March 2014

Prepared for:

Table of Contents

INTRODUCTION

Business is demanding business insight from data acquired from more and more data sources

This series of three white papers on "Improving Access to Data for Successful Business Intelligence" looks at the challenge companies are facing in delivering successful business intelligence (BI) in an era where the data landscape is growing in complexity and there is increasing business demand for more data from more sources to deepen business insight. In addition, business expectation is that BI can and should be delivered more quickly in an agile way, even though data is becoming more distributed.

Business demand to analyse data from new data sources is resulting in a need to support multiple analytical workloads that go beyond the capabilities of a data warehouse

In Part 1 in this series, we looked at how data is becoming increasingly distributed across more systems. For example, in many organisations multiple instances of transaction processing systems, both on-premise and cloud-based, exist. We also looked at the growth in analytical systems, the emergence of big data platforms and NoSQL databases, and how the number of data sources is steadily climbing with much more data coming into the enterprise from external and somewhat unfamiliar data sources (such as social media sites and other external bodies).

In Part 2 in the series, we looked at the impact of growing numbers of data sources on analytical systems and the changing analytical landscape that has emerged from the need to support multiple analytical workloads. We also looked at the growth in adoption of self-service BI tools in use in business areas.

There is also a growth in the deployment of self-service BI tools for visual discovery

In Part 3, the final white paper in the series, we discuss the key requirements for data access in an analytical environment and also how one vendor, Progress Software, is stepping up to these requirements in order to help organisations be successful with their business intelligence initiatives.

KEY REQUIREMENTS FOR DATA ACCESS IN AN ANALYTICAL ENVIRONMENT

Business pressure to remain competitive is causing new requirements to be defined to access data for ETL processing and to produce new insights

When it comes to data access in the context of an analytical environment, there are two areas where data connectivity really matters.

1. Data connectivity to capture data from various sources for data cleansing and integration
2. Data connectivity for business user query, reporting and analysis

Both are equally important. Note that the first one not only needs connectivity to read data but also to write cleansed and integrated data into target systems such as data warehouses, data marts, MDM systems and big data platforms. It is also the case that ETL tools frequently implement ELT processing meaning that they write the data they capture first to a staging area on a target system before data from various sources is cleansed and integrated. This applies to data warehousing and also to big data analytics where Hadoop is increasingly being used as a "data landing zone" and "data refinery" to process raw data before making it available to other systems for further analysis.

Hadoop is becoming a new low-cost landing zone for data captured from inside and outside the enterprise

Data virtualization is also growing in importance to hide complexity of having to access multiple data sources in a new big data analytical ecosystem

Another type of data integration, where good data connectivity is important in analytical environments is data virtualization, shown in Figure 1 below.

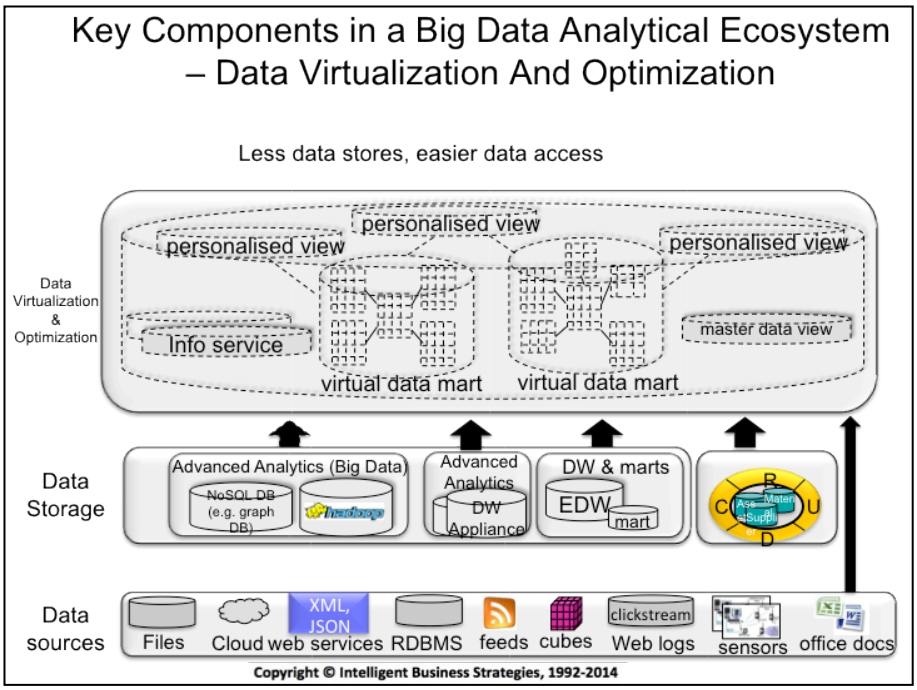

Figure 1.

Data virtualization also helps simplify architecture by replacing physical data marts with virtual ones

Data virtualization integrates data on the fly by federating SQL queries across multiple underlying data sources

Whereas ETL processing consolidates and integrates data to persist in a target data store, data virtualization on the other hand integrates data from multiple underlying data sources at run-time to serve up the results on demand to requesting applications. This could be any application, any BI tool or even an ETL tool wanting access to a "virtual" data source. The virtual data source in this case could be data from multiple physical sources integrated on the fly. Data virtualization also makes it appear as if the data

is persisted in a single database, even though it is not. In other words, a data virtualization server can accept inbound SQL requests from tools and applications that think they are accessing a database. It then breaks each query into sub-queries before accessing underlying systems to get at data to integrate at run time. It is also possible to save popular queries on a data virtualization server as services and invoke them on the fly as web services. Data can also be cached for better performance. There are so many use cases for data virtualization it is easy to see why it is in demand. It can be used:

Data virtualization has many uses and increases agility in analytical environments

- For query and reporting purposes
- To create virtual data marts instead of physical ones
- To simplify access to data in multiple operational systems
- To integrate corporate data with personal data, e.g. Excel data
- To simplify access to multiple BI systems
- To integrate on-premise and cloud data on-the-fly
- To create consistent data services across the enterprise
- To introduce agility into a data warehouse/BI environment
- To integrate corporate performance management (CPM) tools with multiple underlying DW/BI systems
- To provision integrated data into a portal
- To provide on-demand integration of distributed master data
- For heterogeneous replication

As we learned in Part 1 of this series, there is an ever increasing number of data sources that businesses now need to access to build traditional data warehouses, provision data into big data environments for exploratory analysis, or to provision data on-demand for query and analysis.

New requirements need to be defined to access data from data integration and self-service BI tools

Similarly, as the number of data sources grows, business users armed with self-service BI tools require access to multiple data sources to be able to quickly blend together data they need to produce more timely actionable insight in a quick and agile manner.

To fully understand what is needed, the following lists show the requirements for good access to data in a modern analytical ecosystem.

DATA ACCESS REQUIREMENTS FROM DATA INTEGRATION TOOLS

With respect to access to data sources and data types, the following requirements are defined for ETL data integration and data virtualization data integration tools. It should be possible to quickly and easily:

Data integration tools need access to general purpose relational DBMS

- Read data from and write data to all main relational DBMSs products for data warehousing, master data management, data virtualization or big data exploratory analysis. This includes IBM (DB2, Informix), Microsoft SQL Server, MySQL, Oracle, SAP (Sybase ASE, HANA).

Data integration tools need access to specialised analytical relational DBMS

- Read data from and write data to *analytical* relational DBMSs for data warehousing, data virtualization or big data exploratory analysis. This would include Exasol, IBM PureData Systems for Analytics, Microsoft SQL Server Parallel Data Warehouse, Oracle Exadata, Pivotal Greenplum, SAP Sybase IQ, Teradata and Teradata Aster

Access to files is also needed not only on file systems and mainframes but also in Hadoop HDFS

- Access data in flat files (e.g. CSV or tab delimited files, IBM System z VSAM files) for near real-time extraction of data for data warehousing, master data management, or big data exploratory analysis.

Many organisations now have applications deployed on the cloud

- Connect to popular cloud-based transaction processing applications such as Salesforce.com, Workday, NetSuite, Microsoft Dynamics and others, for near real-time extraction of data from these applications for data warehousing, master data management, data virtualization or big data exploratory analysis.

Access to SaaS application data is needed for data warehousing

- Connect to popular cloud-based data storage (e.g. Amazon S3) to extract data for data warehousing and/or big data exploratory analysis.

A simple approach to accessing cloud data is needed

- Connect to a general-purpose cloud connector rather than have to build or buy connectors for every cloud data source. This general-purpose cloud connector in turn should connect or be capable of being extended to connect to any cloud-based data source required.

- Connect to popular big data NoSQL DBMS sources such as IBM IMS, MongoDB document databases and Cassandra column family databases to capture data for subsequent data transformation, integration and analysis in Hadoop environments.

Big data is in demand

- Connect to web server log data sources to capture clickstream data to capture data for big data processing in a Hadoop data refinery and for big data exploratory analysis.

Clickstream, JSON and social network data are very high priority for many organisations starting big data analytics projects

- Connect to data sources that store semi-structured data such as BSON, JSON, XML and potentially email to capture data for big data processing in a Hadoop data refinery and for big data exploratory analysis.

- Connect to sensor data and to grid networks exchanges to capture data for big data processing in a Hadoop data refinery and for big data exploratory analysis.

Need to access social data

- Connect to social media data sources such as Twitter, Facebook and others to capture data for big data processing in a Hadoop data refinery and for big data exploratory analysis.

Access to Hadoop is also needed

- Read and write to Hadoop HDFS file data. Reading should be possible via Hive or other SQL on Hadoop initiatives (such as Cloudera Impala, Hortonworks Stinger, Pivotal HawQ, IBM BigSQL, etc).

- Connect to popular NoSQL Graph databases (e.g. Neo4J) to write data to these systems for subsequent graph analysis.

Personal data is also in demand in self-service BI environments

- Connect to personal data stores such as Microsoft Office Excel and Access databases to capture data for data warehousing and data virtualization.

- Integrate data from these data sources to create data marts for specific analysis and reporting purposes.

DATA ACCESS REQUIREMENTS FROM SELF-SERVICE BI TOOLS

With respect to data connectivity for BI tool users, the following requirements are defined. It should be possible to easily and quickly:

Access to cloud-based analytical systems is needed

- Connect to cloud applications from self-service BI tools for the purposes of operational reporting and for data blending with corporate data during interactive analysis and reporting.

Operational reporting off NoSQL databases optimised for write processing

- Connect to cloud-based data warehouses (e.g. Amazon Redshift) and data marts from self-service BI tools for the purposes of interactive analysis and reporting.
- Connect to popular NoSQL DBMSs, such as MongoDB and Cassandra from self-service BI tools for simple operational reporting.

Business analysts need to access Hadoop to pick up new insights produced by data scientists

- Connect to cloud big data Hadoop platforms (e.g. Amazon Elastic MapReduce) or any or on-premise Hadoop distribution (e.g. Cloudera CDH, Hortonworks, MapR, PivotalHD) from self-service BI tools (via Hive or other SQL on Hadoop mechanism, such as Cloudera Impala or Pivotal HawQ) to access big data for interactive analysis and reporting.

Some users need access to graph data to identify social network influencers and fraudsters

- Connect to popular NoSQL Graph DBMSs from self-service BI tools for exploratory graph analysis, reporting and visualization.

Traditional reporting and analysis still continues

- Connect to any on-premise analytical relational DBMS-based data warehouses or data mart from self-service BI tools for interactive analysis and reporting.

Access to personal data in conjunction with corporate data is also needed

- Connect to personal data stores such as Microsoft Office Excel and Access databases from self-service BI tools to blend data from other data sources or for interactive analysis and reporting.

SIMPLIFYING ACCESS TO DATA USING PROGRESS DATA DIRECT

Progress is a global company that has been in business for over 30 years

Now that we have clearly defined the requirements for access to data in an analytical environment, we are now in a position to see how one vendor, Progress Software steps up to meeting these requirements to help customers be successful with business intelligence.

Founded in 1981, Progress Software has offices in over twenty countries. It has over 350 ISV organisations and more than 100,000 end users accessing using their technology Progress Software offers three products to help organisations get access and connectivity to data:

Progress has three products that improve access to data

- Progress DataDirect
- Progress DataDirect Cloud
- Progress Easyl

PROGRESS DATADIRECT

Progress offers a family of connectors to data that can be deployed on Windows, Unix and Linux environments

Progress DataDirect is a family of high-performing ODBC drivers, JDBC drivers and ADO.Net providers that facilitate fast access to a broad range of on-premise, cloud and big data sources. Progress DataDirect Connect and DataDirect ConnectXE for ODBC and JDBC are a set of 32-bit and 64-bit drivers that run on Windows, Unix and Linux providing connectivity to:

- Relational DBMSs
  - IBM DB2, IBM Informix, Microsoft SQL Server, SAP Sybase, MySQL, Oracle, PostgreSQL, Teradata
  - Pivotal Greenplum, SAP Sybase IQ

Connectors exist to access relational databases, cloud applications, Hadoop and other data sources

- Cloud Sources
  - Salesforce.com, Microsoft SQL Azure
- Big Data Hadoop
  - Hive interfaces to Apache Hadoop, Cloudera CDH, and MapR
  - Hive interfaces Amazon Elastic MapReduce
  - SQL interface to Cloudera Impala (Parquet, text files, HBase)
- Other
  - Btrieve, dBASE, text, XML

Progress uses low-level wire protocols to improve performance when accessing data

The Progress DataDirect Connect for ODBC drivers use wire protocol to connect to the data sources, which means they communicate directly to the database through the database's own wire level protocol. As a result, they do not require database client libraries such as Oracle SQL*Net or DB2 Connect to be installed on each client machine to access the server. The drivers contain all the software needed to connect to the database.

ODBC, JDBC and ADO.Net connectors are available

Progress DataDirect Connect for JDBC drivers are compliant with Type 4 architecture, but provide features that define them as Type 5 drivers, including application failover distributed transactions and bulk load. Debugging, tracking and logging of driver calls is also supported.

Security is built-in

In terms of security, DataDirect Connect integrates with Kerberos-based or NTLM authentication mechanisms, facilitating secure database access in a Single Sign-On (SSO) environment.

There is also an SDK to build your own drivers

For data sources where connectivity is not available out-of-the-box, Progress provides an OpenAccess software development kit (SDK) to build additional SQL access drivers to proprietary file formats and other data stores as needed.

PROGRESS DATADIRECT CLOUD

Progress DataDirect Cloud allows organizations to connect to cloud data regardless of source using a single ODBC or JDBC driver. This includes connectivity to ODBC or JDBC compatible Software-as-a-Service (SaaS) CRM, ERP and marketing automation applications including:

Progress provides connectivity to many SaaS applications from a single driver

- SaaS CRM applications
  - Microsoft Dynamics CRM
  - Oracle RightNow (beta)
  - Salesforce.com
  - ServiceMax (beta)
  - SugarCRM (in development)
  - Veeva CRM (beta)
- SaaS ERP applications
  - Financial force
  - Workday (in development)
- SaaS Marketing Automation applications
  - HubSpot
  - Marketo
  - Eloqua
- SaaS applications developed on Progress Rollbase

Connectivity to SaaS applications via a single SQL interface gives business users using self-service BI tools the ability to access SaaS applications to produce real-time operational reports and dashboards.

This simplifies access to many cloud-based data sources

Connectors to cloud-based deployments of relational DBMSs are also in development[1], which will include support for Amazon Redshift data warehouse, IBM DB2, Microsoft SQL Server, MySQL, Oracle, PostgreSQL and SAP Hana.

[1] Note that RDBMS access is already available in Progress DataDirect and that cloud deployments of these RDBMSs is what is being added.

Access to cloud deployments of RDBMSs and Hadoop are in development

Finally, connectivity to cloud deployments of Hadoop[2] distributions in development includes connectivity to

- Apache Hadoop via Hive
- Cloudera CDH via Impala
- PivotalHD via HawQ
- Hadapt

PROGRESS EASYL

Easyl allows business users to integrate on-premise and cloud data quickly to produce cloud-based data marts

A new product being rolled out by Progress Software is Easyl, which allows business users to create cloud-based data marts from multiple cloud and on-premise data sources. The cloud-based data marts can then be accessed via self-service BI tools. Easyl is aimed at business analysts to help them quickly create data marts by integrating data rather than exporting data to spreadsheets and trying to integrate data manually with crude and error-prone mechanisms such as Excel cut and paste. Easyl also enables the best practice of creating dependent data marts in the cloud that source data from existing on-premise data warehouses. By combining trusted data from existing data warehouses and new cloud data, Easyl allows higher value cloud-based data marts to be created and is intended to help users quickly produce reports on newly integrated data through the use of business templates, further speeding the time to value.

SIMPLIFYING DATA ACCESS IN A SELF-SERVICE BI ENVIRONMENT

Progress is simplifying data access to help users of self-service BI tools get to the data they need

Easyl makes it possible for many users to access integrated data rather than all having to integrate it themselves

The combination of Progress DataDirect, Progress DataDirect Cloud and Progress Easyl are aimed at accelerating time to value in a self-service BI environment. Users of self-service BI tools can easily connect to multiple on-premise and cloud-based data sources and do their own data blending or use Easyl to integrate data and build a cloud-based data mart and then access it without the need for data blending. Note that short-lived data marts could be created this way for departmental or line of business studies and then torn down if there is no long-term value or need for them. Equally, data marts that do prove to be of value can be reused by other business users instead of having to re-invent data integrations unnecessarily.

DATA ACCESS FOR MULTIPLE ANALYTIC WORKLOADS

Progress data connectivity allows companies to access the data needed for new analytical workload

By providing access to on-premise traditional data sources, cloud-based SaaS applications, NoSQL data stores, analytical relational DBMSs and big data platforms like Hadoop, it becomes possible to provide connectivity to data needed to support traditional and new big data analytical workloads. Data integration tools can therefore reach the data sources required by the business (as dictated by new business drivers[3]) to remain competitive. This allows organisations to transition from traditional data warehouse environments to the new big data analytical ecosystem needed to support new emerging analytical workloads as defined in Part 2 in this series.

---

[2] Note that access to some on-premise Hadoop distributions are available in Progress DataDirect and that cloud deployments of these distributions is what is being added.

[3] See the Part 1 in this series for more detail on business drivers

INTEGRATING PROGRESS DATA DIRECT WITH YOUR EXISTING DATA MANAGEMENT AND BI TECHNOLOGIES

It also means that existing investment in data management and new self-service BI tools is preserved as these tools can now reach more data sources

Progress DataDirect and DataDirect Cloud can also be used to expand the number of sources that existing data integration software (ETL and data virtualization) can reach. The technology from Progress DataDirect also provides extended connectivity to existing self-service BI tools, further protecting your investment in data management and BI infrastructure software. It also means that new big data analytical platforms can be added into the analytical ecosystem when needed, with connectivity to these platforms available already out of the box in most cases. Also anywhere where connectivity is not available, a SDK exists to quickly provide that.

CONCLUSION

The need to remain competitive is driving new requirements to access more data

In this paper we have outlined today's business imperatives requiring organisations to obtain deeper insights that help deliver competitive advantage. As a result, organisations must define new technical requirements that enable connectivity and access to new data from new data sources for the purpose of analysis and reporting. These requirements are necessary to capture new data, exploit new big data analytical platforms, and run new analytical workloads to produce new and valuable insights to help the business be more intelligent, more responsive and more competitive.

Progress has recognised the demand for new data and is rolling out new technology to allow organisations to access the data they need to remain competitive

Progress Software has recognised this demand and has rolled out a family of drivers that allows organisations to get to the data they need with existing data management tools and self-service BI tools to prepare data for analysis and to produce new value added insights. In addition, companies can also exploit low-cost cloud computing options to create new data marts from existing on-premise data warehouses and high-value cloud data sources for reuse across the business.

All of this means that Progress Software a useful supplier in putting cloud data, big data and traditional on-premise data to work for competitive advantage.

About Intelligent Business Strategies

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

Author

Mike Ferguson is Managing Director of Intelligent Business Strategies Limited. As an analyst and consultant he specialises in business intelligence and enterprise business integration. With over 32 years of IT experience, Mike has consulted for dozens of companies on 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 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

Improving Access to Data for Successful Business Intelligence – Part 3

