

> **PROGRESS[®]**
DATADIRECT CONNECT[®]
DATABASE DRIVERS

ENTERPRISE BIG DATA

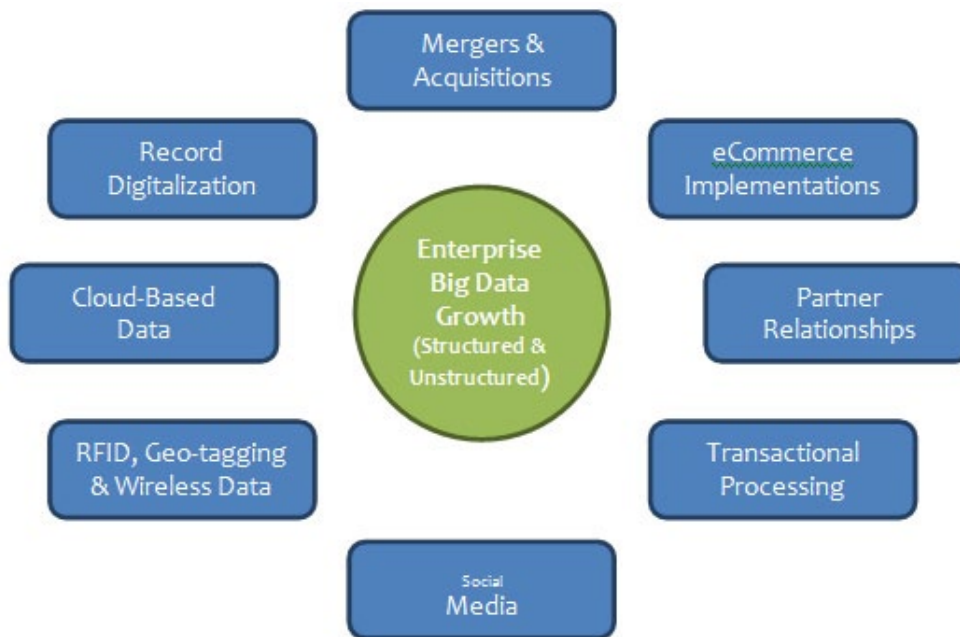
Here to Stay and Requiring
Superior Data Access

AN INTRODUCTION TO BIG DATA

In the years to come, many experts forecast a more than 50% annual growth rate for the 'digital universe'. Gone are the days of the terabyte, as the petabyte now lives at the top of the digital food chain. And if it hasn't already occurred, the zettabyte will soon dominate the data landscape.

The age of Big Data is upon us – and it is here to stay. As data volume, velocity, variability and variety increase, so do the stresses on today's software infrastructures when they can no longer make sense of data deluge.

Big Data continues to grow in popularity and accessibility for a variety of reasons. Big Data is tied to the massive growth in the digital foot print generated as everything knowable becomes digitized and new forms of communication that only exist within the digital realm continue to boom. And as more and more businesses – large and small – move into the online world in the name of ecommerce or transactional processing, the digital universe expands even further.



Enterprises are scrambling to draw out the benefits of working with larger and larger datasets which allow analysts to more closely spot business trends. At the same time, the ubiquitous information-sensing mobile devices, remote sensing technologies, software logs, cameras, microphones, RFID readers, wireless sensor networks and so on of today gather more and more data every day. And then there is social media – an exploding global phenomenon that continues to generate more and more data with every passing second.

BIG DATA – VOLUME, VARIETY, VARIABILITY AND VELOCITY

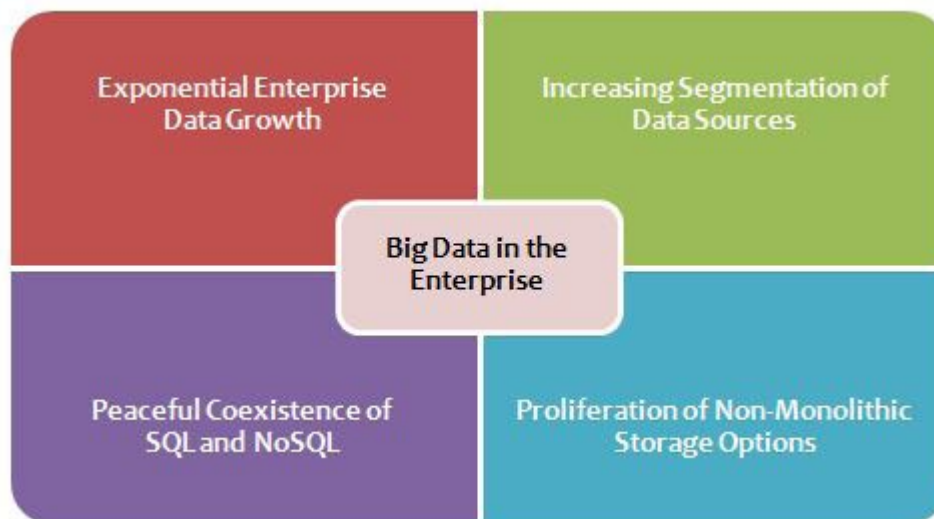
The size of Big Data can be relative to the size of the enterprise. For some, encountering hundreds of gigabytes of data for the first time may mean it is time to reconsider data management options. For others, it may take tens or hundreds of terabytes to cause significant consideration.

Regardless, for enterprise organizations the volume of data they generate, consume, store and access will increase exponentially year over year. Moreover, not only is volume increasing, but data complexity is ramping up in parallel, as is the speed at which data flows into the enterprise. These three Big Data factors – Volume, Variety, Variability and Velocity – hold deep implications.

Data sets are growing so fast that they become awkward to work with using existing database management tools. Difficulties include capture, storage, search, sharing, analytics and visualizing. Additionally, enterprise data exists in a variety of diverse formats, pulled from hundreds of disparate data sources – residing locally or on the other side of the world. It flows in via large batches or real-time streaming. Add in the considerations of cloud-based data, and the plot substantially thickens.

BIG DATA BUSINESS BENEFITS AND IT CHALLENGES

So now, the question becomes whether Big Data serves enterprise organizations as a critical business asset, or cripples operational workflow on a regular basis. As volume, variety and velocity continue to grow; Big Data presents an array of challenges that will resonate through the industry for years to come. One thing is for sure, Big Data is here to stay, and the impact of on everyday IT operations is substantial.



In many organizations, most stakeholders maintain the perspective that Big Data offers tremendous benefits to the enterprise, especially when it comes to more agile business intelligence and analytics. Unfortunately, the days of complete visibility into Big Data are numbered – there is simply too much of it. While we may see companies promoting fancy strategies for managing ‘fire hose data’, only the ones focused on analytics will get close to making meaning from the massive deluge. As a result, more companies will plug into new advancements in relational and non-relational programming frameworks that support the processing of large data sets.

However while Big Data offers potential real-value benefit in the form of enhanced business intelligence, Big Data also presents significant challenges for IT organizations – particularly when it comes to the data connectivity and integration infrastructure – coupling together the emergent set of technologies to allow for rapid data storage; NoSQL and highly-scaled data analysis platforms that use parallel processing such as Hadoop and Map-R.

In other words, these new technologies struggle to maintain the integration point to access, manage and process petabytes of data. And at the same time, they add the significant risk of making application and current skill sets irrelevant.

In the coming years, enterprise organizations will potentially realize significant business benefits from Big Data, but unfortunately few organizations today are fully capable of accessing the full scope of Big Data.

Business expectations are quickly escalating as the data velocity is accelerating, but in many circumstances, IT does not yet have the advanced data connectivity architecture in place to effectively import and export the growing volume of Big Data, nor do they possess the functionality to integrate and transform a wide variety of Big Data formats.

In short, they lack the flexible, scalable data infrastructure needed to exploit Big Data for critical business insights that translate into competitive advantage. Moreover, the inability to seamlessly assimilate the volume, variety and velocity associated with Big Data introduces significant risk. In a worst case scenario, operational visibility muddies, compliance becomes haphazard, customer service levels diminish and revenues tumble.

BIG DATA BULK LOAD USE CASES	
Data Warehousing	Loading bulk data files into a data warehouse
Data Migration	Moving or copying data in tables from one database to another
Data Replication	Taking bulk data files from a server or location and loading them into a database
Disaster Recovery	Moving data into a backup, disaster recovery, or failover database
Cloud Data Publication	Loading bulk data files or tables into a cloud-based database

BIG DATA IN PRACTICE – THE IMPORTANCE OF MIDDLEWARE

But before organizations contend with negative impacts such as reduced visibility or lost revenue opportunities, they must first consider some of the more finite use cases associated with Big Data assimilation today. Primarily, these use cases center on the often overlooked, but ever-so-critical arena of database connectivity and the growing requirements regarding high-performance import and export of data bulk loads.

In a well-written, well-tuned application, over 90% of data access time is spent in middleware. And data connectivity middleware plays a critical role in how the application client, network and database resources are utilized.

In any bulk load use case scenario, database connectivity is the cornerstone of performance. Over the years, technology vendors made great strides in database optimization as well as the performance of processors and other hardware-based server components. As a result, the performance bottleneck moved to the database middleware – the software drivers that provide connectivity between applications and databases.

The most popular commercial databases all include data connectivity components – ODBC and JDBC drivers or ADO.NET data providers – at no additional charge. The open source community, too, offers data connectivity software. Attracted to the price tag, architects often use these free or open source components by default when connecting a particular database to various applications.

By choosing the ‘free’ options, architects are choosing drivers that have not been retooled for today’s business data volumes. And when working with Big Data and bulk loading, the use of such ‘free’, but performance-limiting, data connectivity components can actually cost organizations more than they anticipate.

In fact, within the context of bulk loading Big Data, if data connectivity middleware is not designed for maximum streamlined and efficient functionality, database driver performance is a critical risk factor within Big Data use case scenarios.

PROGRESS DATADIRECT CONNECT – MOVING BIG DATA

As enterprise organizations tackle the challenges of assimilating Big Data within existing data infrastructures, high-performance, scalable and reliable data connectivity is an imperative. Moreover, only when IT organizations employ next-generation data connectivity technologies – such as Progress DataDirect Connect ODBC and JDBC drivers or ADO.NET data providers – do they:

- Guarantee the availability of any size data from any source
- Manage ‘single-driver’ connectivity to a wide array of enterprise databases and platforms
- Deliver the best possible bulk load performance, scalability and reliability
- Deploy with no application code changes or database vendor tools
- Reduce the time, cost and risk of making new data sets available to enterprise users

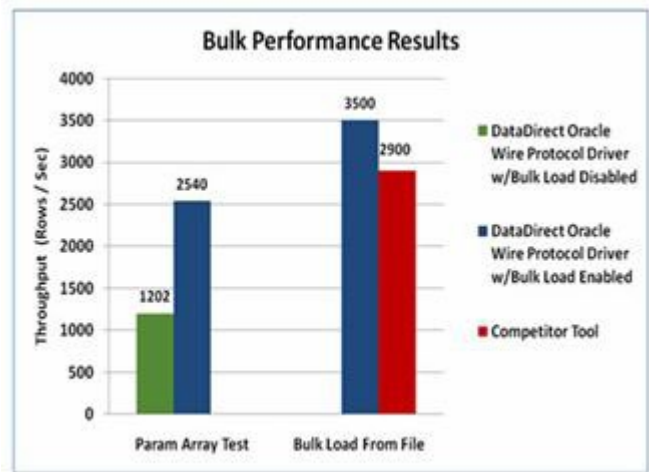
Conventional wisdom often downplays the Big Data potential for drivers compared to individual bulk load utilities. However, DataDirect Connect drivers offer far greater flexibility, and more importantly, functional consistency than individual bulk load utilities and other Big Data relational loaders, which offer highly-variable functionality and unpredictable performance throughput.

With DataDirect Connect driver-based bulk loads, application developers can leverage existing interfaces and bulk load-specific programmatic interfaces to tightly couple their bulk load semantics into applications or platforms.

DATADIRECT CONNECT – HOW IT WORKS

DataDirect Connect Bulk Load consists of two distinct, yet related types of bulk operations – importing and exporting bulk data into or from a database. Depending on API (ODBC, JDBC, or ADO.NET), each type of bulk load operation can use multiple sources of input; can be invoked in a variety of ways with the driver and supports different output destinations. Moreover, DataDirect Connect bulk load provides an array of advanced functionality that ensures superior performance.

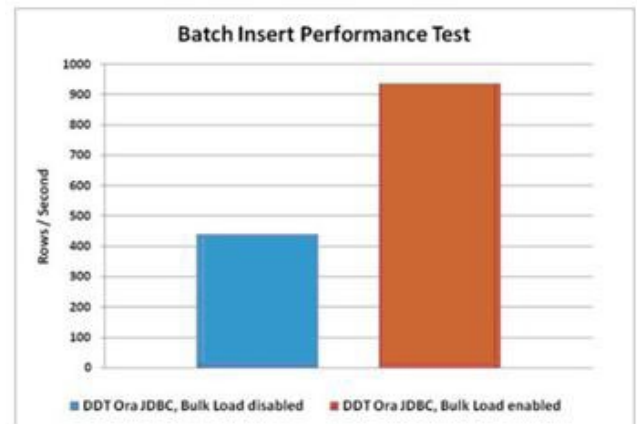
- DataDirect Connect drivers and data providers effortlessly stream data from one database to another. With DataDirect Connect, applications author queries to fetch the data they want. And then using the appropriate API calls in ODBC, JDBC or ADO.NET, the applications redirect the result of that query directly into the target bulk database – essentially streaming the data from one database to another, all without realizing the data on the client. In effect, the application can act as a pipe for data movement with the plumbing defined by the data source query and the target data load.
- DataDirect Connect accomplishes this without an intermediary CSV file by using the EnableBulkLoad connection option in conjunction with a parameter array or batch insert operation. The drivers fetch the desired data from the source database as result set data into an array, and then bind the elements of the array as input parameters for a parameter array insert, which is executed using database bulk load protocols.
- DataDirect Connect drivers and data providers boost the performance of existing parameter array insert operations. The connection option, EnableBulkLoad, allows the driver to use database bulk load protocols for these operations without requiring application code changes. And without requiring client



Enabling Bulk Load in the DataDirect Connect Oracle ODBC wire protocol driver results in the driver inserting over 105% more rows – twice as many – over a given time period. And despite tuning the competitor's tool for maximum performance, DataDirect Connect Bulk Load enables the DataDirect ODBC Oracle Wire Protocol driver to insert over 20% more rows than the competitor's tool.

libraries or bulk load tools from the database vendor, DataDirect Connect drivers and data providers perform a variety of bulk operations.

- DataDirect Connect allows IT developers to write an application that runs as needed or as a batch process to store log files in a database, even as a CLOB or BLOB. Outside of bulk data loads unsuitable for broad-based applications, bulk data loads typically fail when used with data types such as CLOBs and Blobs or data types used to store significant amount of data such as images or large text files. With database-distributed bulk load tools, the bulk load process will fail if these types are encountered. DataDirect Connect compensates for types such as CLOBs and BLOBs and allows the load to continue utilizing non-bulk protocols.
- Developers also use DataDirect Connect to backup critical information to an alternate or failover database by writing an application that runs as a batch process without the need for expensive, difficult-to-manage data replication technology.
- DataDirect Connect extracts table information from relational databases into a database-independent CSV file format which it then uses as input in any bulk load scenario regardless of API, platform or database. While each supported API offers specific means to the bulk export mechanism, the format and governing configuration file it produces is singularly consumable by any of the bulk load (import) implementations. With this consistency, applications can effectively move data between disparate applications and platforms with the underlying guarantee of round trip integrity checking.
- During a bulk export or bulk load, a configuration file is generated to support the resultant bulk data file or bulk import. The file describes the actual data in the bulk data file so that it is fully transportable across the full breath of platforms and software tiers support by DataDirect Connect Bulk Load. Some key features include the ability to define character set conversion to ensure data integrity when moving data across platforms, and a common set of data types so all tiers can correctly compensate and understand the data in the bulk file.



With Bulk Load enabled, DataDirect's Type 5 JDBC driver delivers much more throughput, resulting in over 105% more rows – twice as many – over a given time period. And the time required to execute a batch cycle inserting 10 million rows can be cut by more than half – going from 6.3 hours to less than 3 hours.

Progress DataDirect Connect Benefits to Big Data

- Structured Big Data Movement & Streaming
- Resource Utilization & Throughput Superiority
- Data Source Selection & Ubiquitous Platform Availability
- Flexibility Across Data Sources & Types

- DataDirect Connect drivers and data providers consume or shred data within a CSV file as input to a relational database. DataDirect Connect defines the metadata of the CSV file format just once and repeat the consumption/shredding process over and over.
- Developers employ DataDirect Connect to perform a CSV file data bulk load into a database without writing an application. Using the driver setup dialogs available in the Data Source Administrator on Windows and Linux, developers bypass the need to write a script or program to perform the bulk load.
- In the case of failure, DataDirect Connect records the precise location where the bulk load failed. By setting a simple configuration option, in the bulk load governance file, a log file is generated during the bulk load. The timestamp recorded at bulk load failure contains associated row number, which can be used as the future starting point for resuming the bulk load process.

CAPABILITY	DATADIRECT CONNECT DRIVERS	DATABASE VENDOR TOOLS
Performance	<p>Best Possible Performance</p> <p>DataDirect Connect offers the fastest performance of bulk load into a relational database from a file. In addition, you get a dramatic performance boost for applications doing standard parameter array binding and batch execution.</p>	<p>Marginal Performance</p> <p>Database vendor bulk load tools provide dramatically lower bulk load performance and no performance benefit for standard applications performing parameter array or batch inserts.</p>
Implementation	<p>Simplified Implementation</p> <p>DataDirect Connect Bulk Load is easy to use, providing consistent, cross-database and cross-platform behavior and semantics – without making application code changes.</p>	<p>Overly-Complex Implementation</p> <p>In most cases, database vendor bulk load tools are available on only one or two platforms. This limitation severely restricts what platforms the data transfer application can be deployed to.</p>
Standards-Based	<p>Standards-Based Solution</p> <p>DataDirect Connect Bulk Load offers consistent bulk load semantics and robust functionality across all databases and platforms.</p>	<p>Proprietary Solution</p> <p>Database vendor tools require proprietary code changes to the application specific to each connecting database. Thus, these drivers cannot deliver bulk load for most third party applications and pose significant development and certification challenges for any custom-written applications.</p>

CAPABILITY	DATADIRECT CONNECT DRIVERS	DATABASE VENDOR TOOLS
Reliability	<p>Proprietary Solution</p> <p>Database vendor tools require proprietary code changes to the application specific to each connecting database. Thus, these drivers cannot deliver bulk load for most third party applications and pose significant development and certification challenges for any custom-written applications.</p>	<p>Data Corruption</p> <p>Database vendor tools make corruption possible via lack of data character set conversion. Inconsistent error handling and the lack of operational success contribute to significantly decreased reliability of bulk operations.</p>
Flexibility	<p>Flexible Implementation</p> <p>DataDirect Connect makes it easy to build simple, yet high-performance applications that stream bulk data from one database into another. Bulk load features are available through the driver setup dialogs. All implementations are consistent across platform and database, making cross-database extract and load operations quick and simple to create.</p>	<p>Rigid Implementation</p> <p>Database bulk load tools require the knowledge and use of proprietary command-line execution and syntax, as well as a proprietary bulk data file for input. Setting up and managing cross-database bulk data transfers introduces limitations and headaches that cannot be easily addressed.</p>
Functionality	<p>Supports Key Functionality</p> <p>DataDirect Connect Bulk Load exports a database table or result set into a file ready for import into a wide array of database environments. The created file is not only useable by different databases, but can also be used by any Connect product.</p>	<p>Lack of Support</p> <p>Database vendor tools do not support exporting a database table or result set into database-interoperable formats. They do not bulk load for data types such as BLOBs and CLOBs.</p>

DATADIRECT CONNECT – USE CASE SUCCESSSES

With DataDirect Connect Bulk Load, enterprise organizations effectively satisfy the bulk data access requirements for a broad array of data access use cases. In doing so, they simplify the data access architecture; save important resources for other tasks; and improve operational performance.

- **Data Warehousing** – Results prove that DataDirect Connect ODBC, JDBC, and ADO.NET Bulk Load delivers the fastest, best performance for loading bulk data into an Oracle, DB2, Sybase, or SQL Server-based data warehouse while avoiding data latency issues.
- **Data Migration** – DataDirect Connect Bulk Load is ideal for simple extract and load data migration operations, moving bulk data from one database directly into the other by streaming, thus avoiding the need to load the data into memory.

- **Data Replication** – Instead of using FTP or similar approaches for pushing files around a network, DataDirect Connect Bulk Load quickly loads the data you need into relational database tables. This approach is faster and provides the added benefit of storing the data as a relational database table easily accessed by reporting or BI applications.
- **Disaster Recovery** – Disaster recovery is all about making sure that when a failure occurs, the backup database you are working with is as close to the original set of data as possible. DataDirect Connect Bulk Load ensures that any bulk data is quickly and easily replicated into disaster recovery databases.
- **Cloud Data Publication** – In cloud-based computing, efficient network usage is critical. As a result, performance is ever-important when moving bulk data files or database tables into a cloud-based database. DataDirect Connect Bulk Load allows developers to quickly and easily build a simple program that publishes bulk data into the cloud.

BIG DATA RESOURCE UTILIZATION – THREATENING THE PROMISE OF VIRTUALIZATION

With Big Data taking on increasing prominence, it is a certainty that this phenomenon will intersect with another developing trend in the enterprise – virtualization. And as demonstrated in everyday bulk load scenarios, data access is typically overlooked when it comes to virtualized environments. So it is here within the data access layer that the potential for performance failure resides – which in short order can jeopardize the otherwise attainable goals of virtualization and consolidation.

Essentially, bulk load application performance in most enterprise architectures is inextricably linked to the performance of database access components (i.e. drivers). When driver performance bottlenecks the downstream effects are quite dramatic. When multiple operating systems vie for the same discrete resources such as processor capacity, memory, storage I/O, and network I/O, the dormant issues of resource contention quickly arise.

To fully exploit hardware resources without harming Big Data application performance, all virtualized components must run as efficiently as possible – data connectivity components are no exception. Virtualization overhead varies slightly from vendor to vendor. Likewise, application requirements are rigid in most circumstances.

Therefore the best way to ensure robust bulk load application performance is to ensure that database connectivity is optimized for a virtualized architecture. If drivers are not efficient in their use of CPU, memory, storage, and network I/O, virtualization efforts will almost certainly fail.

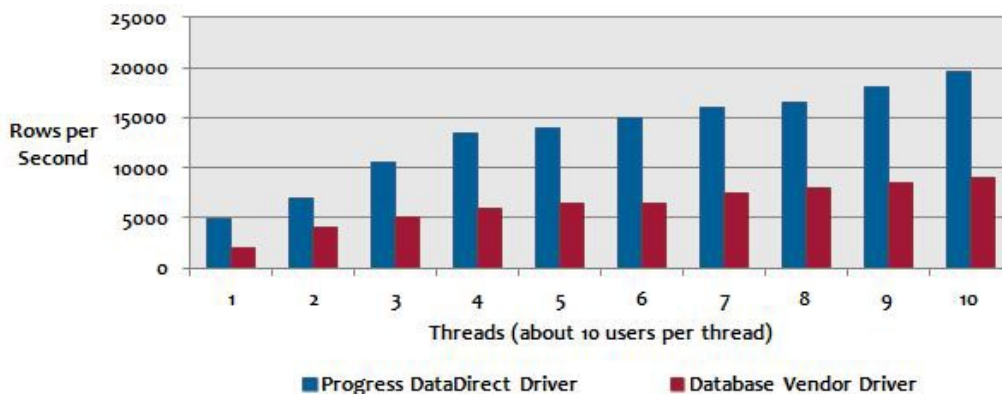
DATADIRECT CONNECT – OPTIMIZING BIG DATA IN VIRTUALIZED ARCHITECTURES

So, with the potential benefits of virtualization at risk, enterprise architects need advanced high volume connectivity components that optimize bulk load throughput performance while minimizing the impact on hardware resources. Without these advanced components, architects sacrifice either virtual machine (VM) consolidation or application performance.

When it comes to virtualized environments, DataDirect Connect ODBC and JDBC drivers and ADO.NET data providers dramatically outperform competitive drivers in network usage, CPU efficiency, and memory footprint. Using DataDirect Connect components means faster response times, more scalable applications and less hardware required to support applications.

- **Network Usage** – When it comes to socket-based operations, DataDirect Connect drivers and data providers minimize how many times applications hit the sockets. Hitting the socket too often increases network activity, so DataDirect Connect buffers up a replies and sends them over as necessary.
- **Memory Usage** – When freeing up memory DataDirect Connect components do not keep variables any longer than necessary. DataDirect Connect components go through application code and free up resources when no longer needed instead of de-allocating memory at a later point before the application is done.
- **CPU Usage** – DataDirect Connect components are tuned to ensure that applications get the most rows per CPU second, or in other words accomplish the most work possible for each allotted CPU time slice.

RAW PERFORMANCE - ROWS PER SECOND IN VIRTUALIZED ENVIRONMENT

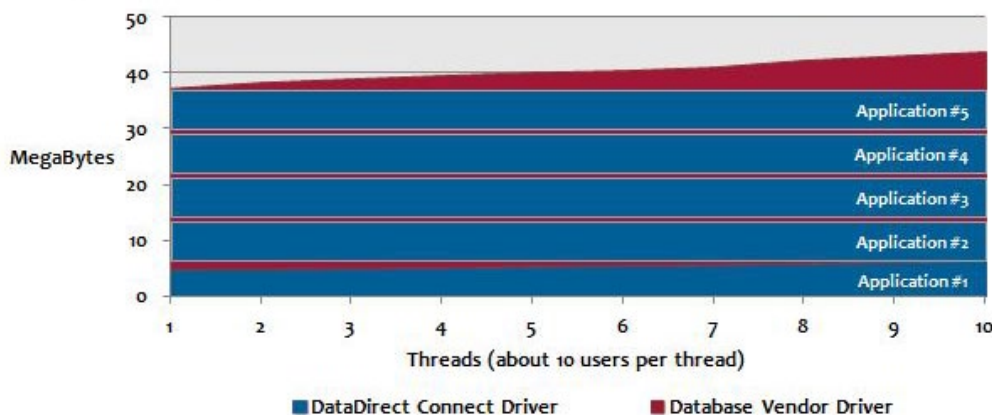


For example in a virtualized scenario, the performance gap when comparing DataDirect Connect to a major database vendor tool becomes highly significant when considering additional VMs on a single physical machine contending for finite hardware resources. In a virtualized architecture, hardware contention issues will either curtail the number of VMs a particular machine can support or dramatically impact application

performance. For example, the figure below details the impact on rows per second when the same two drivers run in a virtualized environment.

And when it comes to CPU and memory usage, the DataDirect Connect driver consumes far less memory than the competing database driver. In fact, as the figure below illustrates, if more efficient virtualization and consolidation is the goal, architects easily place additional applications on the same virtual machine – while still consuming less memory than the free driver alternative.

MEMORY USAGE-MEGABYTES IN A VIRTUALIZED ENVIRONMENT



As the figure shows, with the DataDirect Connect driver, architects can actually host 4 more applications upon the same virtual machine, providing they require the same level of resources as the first application. As the raw performance data shows, in a virtualized environment the DataDirect Connect driver delivers far better throughput than the database vendor driver – while exploiting significantly less resources.

In effect, the DataDirect Connect driver provides architects with substantial Big Data access runtime efficiencies. These efficiencies mean less strain on the virtualized architecture, and as a result mean far fewer Big Data headaches for architects as they attempt to reap returns on their virtualization investment.

BIG DATA – DESERVING OF SUPERIOR DATA CONNECTIVITY

Big Data is here to stay – there is no denying this fact. And as enterprise organizations attempt to reap the benefits of Big Data, they must come to grips with the inherent limitations of most of the existing data connectivity tools on the market today.

They must embrace the fact that Big Data and all of its potential to improve everyday intelligence and operations are deserving of a more robust, yet flexible approach to database access. Only in doing so are they able to develop and deploy an advanced data architecture that enables the seamless and uninterrupted flow of Big Data throughout the enterprise.

DataDirect Connect is the industry leader when it comes to high-performance, scalable and reliable data access. DataDirect Connect ODBC and JDBC drivers and ADO.NET data providers are best-in-class technologies that deliver unrivaled functionality and superior runtime performance – across all major databases and application deployment platforms.

Moreover, by offering a more resource-efficient methodology to data access, DataDirect Connect enables enterprise organizations to realize the bottom-line benefits of complementary technologies such as virtualization. As volume, variety and velocity continue to grow, Big Data presents an array of challenges for enterprise IT organizations. One thing is for sure, Big Data is here to stay, and with DataDirect Connect, Big Data immediately becomes an organizational asset, rather than an operational liability.

DOWNLOAD NOW

Ready to get started? Progress DataDirect offers a free, fully functional, 15-day trial on all products. www.datadirect.com/download