Increase Research IT Productivity While Dramatically Reducing R&D Support Costs Through Improved Scientific Data Management

Scientific BI and Informatics and the Benefits to IT

Executive Summary

In today’s science-driven industries, R&D departments are looking to IT departments more and more to help them meet demanding innovation schedules. Companies selling consumer packaged goods, chemicals, materials, and pharmaceuticals count on their R&D teams to develop the breakthrough drugs or new paints that will drive business growth. And, like every other organizational department today, R&D counts on IT to implement and support the systems and solutions that enable improved process efficiencies and slash operational expenses.

Designing the right product requires accessing, integrating and analyzing an enormous amount of data—everything from complex scientific data for the drugs, chemicals, and materials being designed to all the associated information needed to bring the product successfully to market. Without an efficient way to manage it all, the innovation process can easily slow to a crawl, delaying time to market and the profits associated with it.

Typical enterprise-class data management systems have essentially “passed over” R&D, and for good reason. With standard approaches, it is almost impossible to implement an enterprise-wide Business Intelligence (BI) system with the scientific depth and awareness needed to address R&D-specific needs.

The Accelrys Scientific Informatics Platform, Pipeline Pilot™, gives IT departments in R&D-driven enterprises a real advantage in their ability to help scientists and engineers meet the competitive demands that drive business growth. It provides the ability to tie together all the different data sources and types of data needed to answer the questions that drive excellent product and process innovation. Pipeline Pilot is architected so that it is easily deployed globally by IT to enable R&D to quickly and easily access, analyze and report on any data, regardless of where it resides or what format it is in.

Today’s challenges for IT

IT departments in R&D-driven enterprises face unique challenges. In recent years, R&D has been generating rapidly increasing volumes of data, largely driven by new approaches to research, such as high-throughput experimentation techniques. Compounding the mounting data deluge problem is the fact that this data is both structured and unstructured and comes from multiple, disparate scientific applications and equipment. This data is stored in multiple formats and locked in different silos, often with much duplication, making it difficult to access, analyze, report, and share across the R&D enterprise. Managing this data and providing efficient access to the knowledge that can be leveraged from it is fast becoming a mandate for an already overburdened IT department. R&D departments are being asked to do more with less, innovate faster and drive top-line growth while lowering costs and they are looking to IT to help them succeed.

Here are some of the unique challenges that R&D IT departments regularly face:

1. **Data, data everywhere**

   Scientific data is complex and diverse in format. Some of it is structured in a way that makes it accessible to standard databases and BI software. But most of the data is unstructured (or structured in a proprietary format) or industry-specific data, such as graphical representations of chemical or biological compounds, images, etc. Formulation recipes might be saved in a Word document on a researcher’s hard drive or in a home-grown database, while spectroscopic data might need
to be extracted from an analytical device or retrieved from a printout. Since traditional BI software can’t handle such disparate sources and formats, this data is time-consuming and difficult to locate and re-use, resulting in lost knowledge, decreased R&D innovation productivity, and higher IT support costs.

2. **Data volumes keep growing, but knowledge is what is needed**

The total amount of research and scientific data is growing exponentially. Scientific instruments generate streams of data that flow into data silos. New techniques such as high-throughput experimentation and new equipment such as second-generation gene sequencers are only worsening the data deluge.

It is not just about the data you have, though; it is what you want to do with your data. Being able to retrieve research data readily (no small task in itself) is only part of the challenge. It is also important to consider how R&D information is or can be productively used among scientists, engineers, and other key stakeholders across the enterprise. The needs of senior scientists and bench chemists are different from those of process engineers. And eventually even managers across QA, manufacturing, supplier selection, and product marketing will also require their own views of and interactions with the information, once they realize the productivity gains such access can provide. In order to leverage data effectively and efficiently, the ability to extract timely knowledge from all types of research intelligence is key.

3. **Inadequate reporting and collaboration techniques impede results**

Beyond simply gaining access to the proper information, IT also faces the challenge of developing and presenting meaningful reports. Different clients have different demands for the data: a C-level executive, for example, might need a high-level project status view while a scientist might need a much greater level of detail, including the ability to drill down or see side-by-side comparisons of data from different sources, including previously run experiments. In addition, the report consumers might need different delivery mechanisms, ranging from Microsoft SharePoint or browser-based access to files in Microsoft Office formats.

4. **Multiple applications that can't adapt to new innovation processes**

An R&D-driven enterprise might rely on hundreds of different applications, some of them “home grown,” for different scientific processes. These applications often force scientists to manually manipulate files to get data from one program into another, an error-prone process that leads to inefficiencies and productivity losses.

Maintaining and supporting all of these different applications can become burdensome to IT, which is responsible not only for upgrades and changes but also for ensuring that personnel are trained or retrained. Minor change requests may take weeks to be implemented, and any modifications have to be thoroughly tested for accuracy and integrity and, in the end, re-use can be limited.

Because the R&D process is focused on innovation, there is no single way that scientists, chemists, engineers look at, manipulate, and analyze their data. Thus, a flexible approach that empowers all levels of scientists to apply their research in a way that works best for them is required. This approach must accommodate the success of R&D initiatives, and at the same time provide a framework for improved decision making, enhanced results, shorter time to market and ideally, lower costs.

**Look beyond business intelligence**

Traditional business intelligence, Master Data Management (MDM), or data warehousing tools can’t solve the unique challenges of R&D departments. These solutions were built for transactional data, which is generally structured and numerical in nature. While they excel for finance, HR, marketing, and customer service, they are inadequate when it comes to the complexities of the science and engineering world. In order to fully harness innovation intelligence and collaboration, organizations require a solution capable of handling highly complex scientific data in a variety of formats and across many disciplines.
The ideal solution would be a scalable standards-based, “plug and play” platform architecture that can stand on its own or dovetail into a larger, corporate SOA environment. Such an architecture would allow users to decide on their own “best of breed” solutions, be they products developed in-house, legacy applications, third party systems, or proprietary databases. The platform would provide BI and informatics capabilities for integrating, analyzing and reporting on scientific data in near real time, with powerful text and scientific analytics for decision support. The automation, high recall, and reuse of information that such a platform enabled would lead to scientific process improvement and knowledge retention.

**Today’s Solution: Scientific Data Management and Workflow Automation**

“Acclerys has played a pioneering role in developing scientific software for boosting R&D productivity and a leading role in enabling scientific business intelligence systems.”

-Sudeep Basu, Senior Analyst, Frost & Sullivan

The Accelrys Scientific Informatics Platform, Pipeline Pilot, is an enterprise-ready, SOA-based platform for research informatics and scientific BI that supports more efficient, rapid application prototyping, application development, deployment, data management and integration, analytics and collaboration for biological, chemical, and materials sciences. Its data pipelining architecture is specifically designed to meet these industries’ information management challenges.

By bridging multiple data sources and applications, including business intelligence and operations monitoring software, Pipeline Pilot provides unprecedented flexibility in scientific BI. Pipeline Pilot can mine, integrate, and analyze disparate scientific and complex data types and applications (as well as simple numbers and text) and present the results in a single dashboard view with true 3D graphics. All the stakeholders in the process of driving new discoveries to market—including bench chemists and process engineers, managers in charge of manufacturing and supplier selection, and business and marketing executives—can greatly benefit from the ability to view pertinent R&D project information on demand.

New development targets often call for re-analyzing existing data, and Pipeline Pilot’s flexibility and real-time operation make it much easier to revisit data without incurring another lengthy retrieval process.

**Global Architecture, Local Delivery**

Accelrys Pipeline Pilot provides IT with a single, global architecture that allows application delivery in a supportable environment that lowers R&D IT costs. IT can configure hundreds of small applications as Web services that deliver what each scientist needs when he or she needs it. If the scientist has a new requirement a month or a year later, IT can easily reconfigure an existing application without enormous expense, time, or overhead. No longer do you need a rip-and replace strategy to evolve the informatics platform.

**Transform Data into Knowledge**

Improving innovation efficiency involves more than simply accessing and aggregating data. Pipeline Pilot provides the ability to access advanced scientific information related to chemistry, process engineering, materials science, and more. With an extensive array of best practice statistical methods—ranging from simple calculations to advanced modeling methods—manipulated through a drag-and-drop visual programming interface in a data pipelining metaphor, Pipeline Pilot automates the process workflow and drill-down analysis of the research process. Development that used to take hours or days can be done in minutes or hours.
**Immediate ROI**

The modular nature of Pipeline Pilot means that changes in a company’s process requires modifying only the affected component, rather than reworking the entire system. The ROI comes not only from faster, better, and more cost-effective solutions, but also from a more competitive assembly of the applications into workflows that facilitate creativity. In most R&D IT groups, the main goal is to not just enhance bottom-line efficiencies, but also drive top-line growth through IT strategies. Now, it is possible to quickly meet the ever-changing demands of the scientific and engineering community through rapid delivery of applications that can be built at the speed of discovery, without excessive development or legacy support expense.

“The ease of use and modular nature of Pipeline Pilot has helped us reduce the time necessary to complete internal informatics activities by as much as 90%.”

- Head of Informatics for a global consumer packaged goods company

**Reduced Total Cost of Ownership**

By integrating multiple applications, the Accelrys Pipeline Pilot can dramatically reduce the total cost of ownership (TCO) of an enterprise’s application suite. It reduces the resources previously devoted to integrating data and applications, for example by eliminating the need for point-to-point interfaces. Pipeline Pilot comes with an out-of-the-box set of connectivity tools that can tap into various data sources that are common in the scientific workflow, such as on-line laboratory devices, literature databases, image databases, LIMS systems, and structural drawings, just to name a few.

**Real-time Data Retrieval and Integration**

An enterprise often relies on many different repositories of data and scientific information. For example, one data repository might hold image data, while another holds scientific data from instruments in a format that does not accommodate images. Without Pipeline Pilot, for example, process data and chemical compound descriptions may only be combined by including a text summary in a BI report.

Built on a client-server model, Pipeline Pilot bridges the gaps between a company’s information silos. The Pipeline Pilot Server communicates with the various data sources and retrieves information in real time; it also holds the protocols (specific data operations) that embody the rules for data manipulation. Users access the server through either of two clients: The Pipeline Pilot Professional Client or the Pipeline Pilot Web Port (a zero-footprint client requiring only a browser). Pipeline Pilot also supports the creation of custom clients using any of several SDKs. While protocols are created and edited on the client systems, protocol jobs run on the server, so the server is the only machine that needs to connect to a database source. As a result, the client machines do not need to be configured to communicate with the database.

Besides bridging existing data silos, Pipeline Pilot can also be a part of building a new data warehouse as a parallel workflow. While most business intelligence software operates as read-only, Pipeline Pilot can not only read, query, and format data, it can also write the data out to one or more repositories and create a presentation. This aids, for example, the transfer of data from legacy applications to more up-to-date systems.
How Pipeline Pilot Works

Define, Capture, and Automate Scientific Business Processes

In Pipeline Pilot, users build workflows, or sequences of actions to be performed on data, by manipulating protocols in a visual programming environment with “drag and drop” ease. Building a workflow on screen is a matter of connecting the protocols with pipelines in the order of their desired execution. Protocols can be saved for easy access and reuse across the organization to ensure the capture and deployment of best practices.

Figure 1: Integration Architecture with Pipeline Pilot
Integrate Diverse Data Sources and Applications

A wide range of programming interfaces and data interchange formats are supported by Pipeline Pilot so that applications already in place can become components of the entire Informatics platform architecture. It also supports scripting so that non-standard or legacy applications can be easily connected into the system, as well.

Process, Analyze and Report Complex and Scientific Data

It is increasingly difficult to provide tools that don’t just access data but also report it, with the flexibility to create a presentation appropriate to a particular need. Pipeline Pilot can give the C-level executive access to a weekly dashboard view via a Microsoft SharePoint portal, while a scientist gets a detailed PDF document that can be stored for later reference.

IT Benefits of Pipeline Pilot

Ability to leverage existing tools

Generally, the systems already in use across an R&D enterprise are a disparate collection of software from outside vendors and internally written custom applications. Solutions for exchanging data among these separate applications are often built incrementally, on an ad hoc basis, leaving the IT department with an ever-growing knot of code to maintain. Since that code touches multiple applications, it can be difficult to replace or repair.

By compartmentalizing applications into separate components, Accelrys Pipeline Pilot makes it much easier to diagnose problems and test new solutions in one area without affecting the rest of the enterprise.

Furthermore, the graphical nature of Pipeline Pilot’s programming tools makes it straightforward to analyze an application after the fact and discover how it works. This makes it easier for IT to support and extend existing applications, even after the departure of key people who were originally involved in their development.
Easy definition and automation of business processes
Pipeline Pilot works through the assembly of protocols, or specific data operations, into workflows, or chains of actions. Each protocol performs some operation on the data it receives and passes it on to the next protocol and the next operation. The product comes with predefined protocols for handling data from such diverse fields as statistics, chemistry, and biology, as well as text and images. In addition, members of the IT department as well as informatics professionals can use the Professional Client, which runs on Windows XP or Vista, to define and save new protocols for data manipulation. Other users can access the protocols via the Professional Client, Web Port, or a custom client to build and run workflows.

The ability of end users to both create protocols and build and run workflows enables IT to offload some of those tasks to the user base. Moreover, the ability of IT as well as end users to save, retrieve, and extend workflows permits the capture of best practices for later reuse.

Real-time data integration
Companies generally approach the problem of integrating enterprise-wide data in one of two ways. First is a monolithic approach: choosing a single product to hold and maintain the data. Second is the “roll your own” approach of writing custom applications to pull data together as it’s needed. The problem with both these approaches—besides time and cost—is that the solution built today will likely have to be rebuilt tomorrow.

Pipeline Pilot is installed on a server (Windows Server 2003, Red Hat Enterprise Linux, or SUSE Linux Enterprise Server) and provides interfaces to various types of data sources that include: chemical and biological databases (both text and image databases), analytical data from lab devices, and business intelligence software. Supported data types and interfaces include (but are not limited to) databases via ODCB and JDBC, Web pages via HTTP, Spotfire enterprise analytics, and Microsoft Office documents.

Pipeline Pilot interfaces with all those sources and transforms the output and input to a common data record. It performs an extract, transfer, and load (ETL) operation in real time and processes the data as well, so there is zero latency in moving data from applications to user. And it works equally well with structured data (e.g., from Oracle or SQL databases) and unstructured data (such as Web sites), or even with data contained in documents. It is especially fast at working with data in flat files; for example, it is able to read seven or eight thousand chemical structures a second from a flat-file database.

Figure 4: Pipeline PilotClient server architecture enables users to draw data from multiple sources, manipulate and analyze it, and access the results through one of several client types.
Application integration

Accelrys Pipeline Pilot provides for integrating applications as well as data. For example, an enterprise might have a legacy, command-line application that performs a valuable function but is rarely used because there has been no way to integrate it into the informatics system. Pipeline Pilot works with application objects through a wrapper and can wrap even such a legacy application to make it available as a service to other application objects.

This approach not only enables IT to leverage legacy applications by integrating them into new business processes, it also reduces their support needs. It is no longer necessary to run such applications manually, as a separate operation, so IT does not need to be concerned about losing the staff with that expertise over time.

As SOA middleware, Accelrys Pipeline Pilot offers a flexible system for rapidly tying together existing or preferred components, enabling an organization to select a best-of-breed solution for each separate data management task without worrying about how they will communicate with each other. The result is a single plug-and-play platform for data and application integration. The benefits include not only simplicity but speed: tying together existing components can take as little as a couple of days and requires only a onetime effort.

Pipeline Pilot provides several options for integrating applications in this way. Among the points of integration available are

- Java SDK and .NET SDKs
- Portal environments (JSR168 portals) and SharePoint
- Java custom code
- Web Services
- SOAP and WSDL
- Perl, Python, JavaScript, VB (Windows), and R custom scripts
- Pipeline Pilot’s own scripting language, similar to Perl

Easy integration into IT infrastructure

Pipeline Pilot’s flexible Service Oriented Architecture (SOA) enables it to stand on its own or dovetail into a larger, corporate SOA environment. Its ability to define processes that consume services means that the various applications already in place in an enterprise can become components of the entire Pipeline Pilot-based architecture. These components can be joined together to enable workflows that are simple (e.g., a one-click check against molecular weight) to extremely complex (e.g., homology modeling or docking workflows). The components can be modified, repaired, or replaced without risk to the rest of the system.

This flexibility works in both directions. Pipeline Pilot can be used to generate services as well as data output, so it can be used to create science-aware complex-data Web services that plug into an existing SOA architecture.

Pipeline Pilot is also scalable, and it supports clustering as well as distributed and grid-based computing. Jobs can draw on the resources of multiple servers at the same time, and the software provisions hardware and software on demand, on an as-needed basis. It works with Sun Grid Engine and other workload management and provisioning platforms, including Platform LSF and PBS Professional.
Single platform
As a consolidated platform, Pipeline Pilot provides a single user space and a single programming environment for the multiple tasks required for a true enterprise informatics system. The single plug-and-play platform for data access, analysis, and reporting eliminates the need to master multiple products, resulting in more flexible staffing options. One customer has reported replacing more than 27 separate applications with Pipeline Pilot.

Robust security
Security is role-based: access to resources is based on user identities and the roles assigned to them. The security settings for users’ roles determine the applications they can run, the databases they can access, and their file permissions. Single Sign-On simplifies security management for both the users and the IT department.

Pipeline Pilot further extends its security capabilities through user authentication, impersonation, and authorization.

- Authentication provides a way to control server access by verifying network user identities based on login names and passwords.
- Impersonation provides running protocol jobs with exclusive user-based access to resources and prevents excess permission being granted.
- Authorization defines the roles of Pipeline Pilot client users.

Speedy and flexible deployment
The demand for integrated information often comes with significant time pressure, as users are impatient for the data they need. By relying on a SOA, Pipeline Pilot dramatically shortens the time to delivery. Tying together existing components can take as little as a couple of days and requires only a onetime effort.

And Pipeline Pilot can be used to generate a wide range of output. It can create reports, provision SharePoint portals, provide services, and populate databases.

Pipeline Pilot can also be used to create Web-based applications that can be deployed to the end user base. For example, a user could use a browser to access a particular set of data without ever knowing that Pipeline Pilot is active behind the scenes.
A familiar scenario in many enterprises is the one in which a project, after months of development, is rejected by its intended users. Pipeline Pilot can shorten the development process to weeks. Many customers use Pipeline Pilot as a rapid prototyping tool before actual development starts, in some cases actually starting construction of the prototype with the end user present.

Furthermore, informatics professionals can use the Pipeline Pilot Professional client to develop or customize some of the application themselves by writing new protocols. And all users can create new workflows without requiring application development at all. All this means that IT can deliver application solutions right away rather than the months or years after installation required with other business intelligence tools.

Summary
Accelrys Pipeline Pilot is an enterprise class platform for integrating, analyzing, and reporting scientific data and processes. For end users, Pipeline Pilot enables the presentation of integrated data from multiple applications in near real time. For IT managers, Pipeline Pilot addresses many of the issues facing today’s science-based enterprises:

- It provides a single platform and programming environment for data access, analysis, and reporting
- It enables quick deployment of new applications
- It can present information in complex reports or simplified dashboards.
- It works with integration architectures already in place, or provides an integration architecture where one is needed.
- It reduces the load on already overburdened IT resources.
- It lowers TCO by simplifying and integrating the data extraction and presentation process.