# From raw data in the laboratory to information availability in the enterprise

If pharmaceutical or biotechnology companies could measure their success merely on the volume of raw data accumulated from screening assays and pre-clinical testing, they would be basking in the glow of unprecedented success.

By Dieter Kreusel

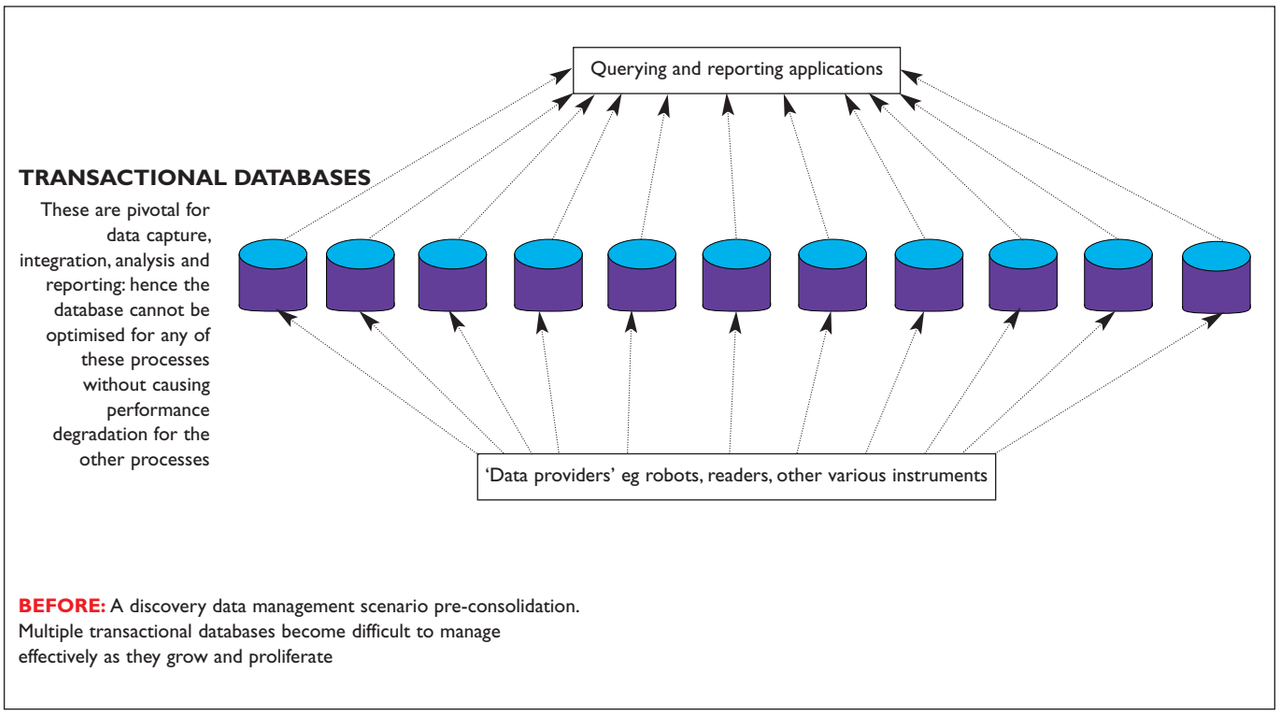
Companies are generating data at an increasingly rapid rate, primarily data generated by ultra high-throughput screening activities. Their databases, which store information about the structure, composition, properties and function of each experimental compound in their vast libraries, are expanding at an extraordinary pace. The ever-increasing size and complexity of these data resources hold great promise for the future of drug discovery and the prosperity of the pharmaceutical industry, an industry in which information, when applied prudently, is a valid measure of wealth.

Electronic data is rapidly becoming large pharma's most valuable asset. It will guide decision-making processes across an enterprise. The raw data itself, though, is of only limited value. Extracting optimal value from large datasets requires consolidation, integration, and organisation of raw data to facilitate data mining, the process of querying a database to identify patterns and trends. This process can transform data into information and information into knowledge, which can then be shared, compared and applied across an organisation. Not unlike a vein of precious mineral buried deep within the Earth's crust, information will only be of value if it is collected, handled and preserved with care and forethought.

Yet the increasing volume of these data stores also threatens to overwhelm the capacity of available data management systems to capture, retrieve and manipulate data efficiently. Continuously expanding databanks could clog these systems, slowing data transport networks, delaying data recall and, in short order, making current data analysis systems increasingly unwieldy and obsolete. As a result, a company's ability to access, interpret, and apply data could be seriously compromised.

The solution to overflowing data streams and clogged information pipelines lies not in limiting data collection or the size or scope of a database, for every nugget of information is of potential value. Rather, the answer is to establish an enterprise-wide system for consolidating and integrating research data and for storing the data 'offline', outside the laboratory setting. Inherent in such a system would be a mechanism for sorting, cleaning, and assessing the data, making sure it is valid, relevant and presented in appropriate and compatible formats. The cleaning and validation process would eliminate repetitive data stores, link data sets and classify and organise the data to enhance its utility. Finally, the data should be stored in a fashion that makes it readily accessible to users on a local and global level.

There are two major approaches to solving the challenge of corporate data access. Under the 'federated' model, operational databases and other repositories remain intact and independent. Data retrieval and other multiple-database transactions take place at query time, through an integration




**BEFORE:** A discovery data management scenario pre-consolidation. Multiple transactional databases become difficult to manage effectively as they grow and proliferate

**Figure 1**

layer of technology that sits above the operational databases and is often referred to as 'middle ware' or a 'meta layer'. Proponents of the federated approach include IBM, Tripos and Synomics (now part of the Accelrys group), with IBM's DiscoveryLink being the most widely recognised product of the federated type. Database federation has attractive benefits, an important one being that the individual datasources do not require modification and can continue to function independently. In addition, the architecture of the federated model allows for easy expansion when new datasources become available.

The federated approach is less attractive, however, in certain practical areas. For example, it does not address the problem of variable data quality in the different repositories, many of which may contain data collected decades ago and may lack a full complement of contextual scientific data to qualify the various data points. In addition, the federated approach suffers from scalability problems when used with many different datasources. Another issue not resolved by the federated approach is the size and performance of the transactional or capture databases. With the current rate of data production, the transactional systems reach enormous sizes. Data collection rates require the separation of data systems into dedicated capture databases and analysis/reporting databases in order to satisfy the respective requirements of each.

The 'warehousing' approach involves the consolidation of all required data into one or multiple, separate repositories. This approach is common in other industry sectors, where it has proved successful. ID Business Solutions (IDBS) has applied its domain expertise to develop its DiscoveryWarehouse approach, which is based on a datamodel optimised for biopharmaceutical data. Warehousing effectively addresses the separation of transactional and analysis/reporting databases and thereby provides a data management architecture that will be able to cope with increased data demands over time. It also provides a mechanism to clean and quality check the data from the capture databases and thereby guarantees data quality. However, warehousing too has its limitations. Data types generated within R&D (from genomics to clinical trials) are very different, and combining them all effectively in a single data warehouse may not be practical.

A key feature of the data warehouse strategy is the separation of data capture from data storage and analysis systems. While data collection proceeds at the local level using multiple transactional databases, the data in these transactional databases are then consolidated into one or more large data repositories. These data warehouses can

accommodate input from a variety of data sources. The consolidation, integration and organisation of data from multiple sources is, therefore, a natural outcome of the data warehousing strategy.

Another advantage of data warehouses is that they become corporate-wide sites for data querying, reporting and analysis. Queries can proceed at a rapid rate because the warehouse contains only clean and validated data that has been organised for maximum utility. In addition, data warehousing can enhance a company's data analysis capability by applying organisational strategies aimed at optimising data sources for specific querying and reporting purposes.

Aventis developed and began the implementation of a strategy that combines the federated and warehousing approaches. Essentially, Aventis uses warehousing to consolidate and integrate data from multiple operational databases into a manageable handful of large data warehouses, and applies the federated approach to integrate multiple data sources, including the different warehouses. This article discusses the reasoning behind this choice and discusses in detail the implementation of the warehousing portion of the project.

Ultimately, the goal of any data management architecture is easy access to high quality, useful data. Pharmaceutical companies rely on this data to extract information about targets, compounds and biological systems. They then use that information to make predictions that will guide future decision-making. Critical to this process is data mining and the ability to transform raw laboratory data into a globally accessible knowledge base.

## Developing a strategy

A multinational corporation such as Aventis faces enormous challenges in co-ordinating data collection and storage and ensuring optimal accessibility to its data resources and data mining tools on an enterprise-wide scale. Recognising the impending bottleneck caused by the rapidly increasing amounts of data pouring into its localised databases, Aventis developed its strategy of creating a series of centralised data warehouses that would facilitate data retrieval and analysis, freeing the local databases to be enhanced for data capture. The project is now in its final phase and will be fully operational by the end of January 2002.

Aventis outlined several goals at the outset:

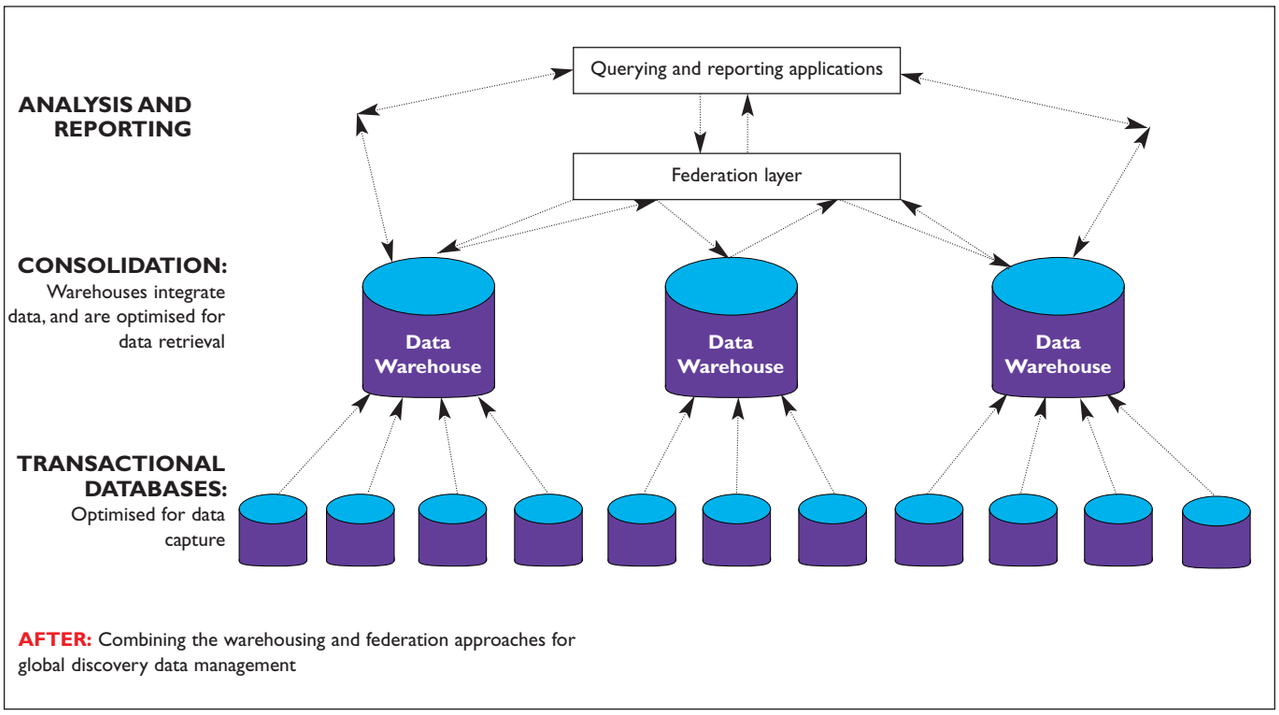
- To harmonise the biological and high throughput screening data emerging from its Lead Generation and Lead Optimisation programmes and spanning the breadth of its therapeutic areas.
- To separate data capture and data retrieval operations, and to be able to apply different optimisation strategies to these two functions.
- To minimise data retrieval costs.
- To introduce improved quality checking procedures.

The company decided that the best and most efficient way to achieve these goals would be to contract with outside vendors for the necessary products and services where possible. By outsourcing, Aventis would have access to the expertise it needed, while ensuring that the ongoing maintenance and development of the system would remain the responsibility of the vendor. IDBS was chosen to develop the data warehousing aspects, while IBM is to implement the federation layer. The model presented here demonstrates how a data warehousing strategy can optimise data organisation and integration to facilitate data mining and maximise the value of a company's data resources. Although this particular implementation suits the specific needs and interests of Aventis, the concept of global data warehousing should have universal appeal across the pharmaceutical and biotechnology industries.

## A practical approach to data warehousing

Data integration and data warehousing can be visualised on three basic organisational levels, which should be accommodated within a dynamic, data centric architecture – the transactional level, the data consolidation level and the analysis level. The transactional level represents data capture from a variety of sources within the research and development and pre-clinical testing areas. The raw data collected may come from genomics or proteomics research within the organisation, combinatorial chemistry, high throughput screening and compound profiling laboratories, pharmacology and toxicology groups, or ADME and other pre-clinical testing programmes.

Aventis uses IDBS' ActivityBase as its system of transactional databases for collecting data from a variety of different screening assays as well as data generated by therapeutic groups. Maintaining 11 local transactional databases presents a fundamental problem: how to perform an enterprise-wide query in a time and cost efficient manner? Other, less obvious problems arise as well. When transactional databases are housed at corporate sites in various countries, for example, it becomes necessary to apply enterprise-wide data standards. Different sites




**AFTER:** Combining the warehousing and federation approaches for global discovery data management

Figure 2

might have used different nomenclature to express the same terms; for example, terms such as IC50 might be represented as IC-50 in one database, and ic50 or IC 50 in another. These terms need to be harmonised across all the operational databases from which data is to be retrieved and consolidated. Furthermore, one laboratory might enter data into the database using one set of units, whereas another laboratory might use different units to express the same data set (M vs nM, for example). All of these factors may make it almost impossible to apply a single query to a series of independent databases. The inability to conduct a global query could diminish the value of a company's data resources.

The transactional database, therefore, represents the first level of data integration, as it captures and combines data from multiple screening assays and can perform cross-site metrics, historical comparisons, and analyses of screening data for a specific compound or set of compounds. However, as data volumes expand and operational databases increase in size, the rate of data capture slows. Therefore, transactional databases represent not only the first level of data integration, but also the first potential bottleneck in the overall data management process. One solution to these problems is to establish a second level of data systems and to transfer data from the transactional level to the data consolidation level.

## Focus on the warehouse technical architecture

Aventis merged 11 different transactional databases to create a dedicated data storage system comprised of three separate warehouses, one located at each corporate centre to serve the needs of that site. These warehouses share a common structure, data model and data dictionary, but they store only data generated and captured at that particular site, including screening data, profiling data and data emerging from ADME and toxicology screens. An alternative approach would have been to create a single, enterprise-wide data warehouse with multiple copies of the warehouse located at corporate sites worldwide to improve local accessibility. However, this was rejected by Aventis, since nearly three-quarters of all current and anticipated queries only apply to the data generated locally and would not require enterprise-wide data searching. To facilitate the remaining 25% of queries that require global searches, Aventis brings into play the federated approach and links its data sources using IBM's DiscoveryLink.

Whereas relational databases are optimised for data capture and reporting, data warehouses are less adept at data input but are designed for optimum data analysis. In a relational database the key identifiers are compounds and batches of compounds, from which the application compiles

an inventory of their properties and test results. In contrast, the warehouse is designed using the well-established star schema, in which the test result is the key identifier. All other data elements radiate out from the central point of the star, like spokes on a wheel. The warehouse has a more open design and is intrinsically more scalable; it can expand as needs grow, with the addition of new data categories being a simple matter.

The warehouse model can facilitate focused querying and reporting for example by 'spinning off' smaller, temporary databases, or data marts, that accumulate specific types of data, such as screening results, project data, compound properties, or structure/activity information and relationships. The warehouses would periodically update these data marts, which would become highly integrated and valuable resources for targeted query applications.

## Data cleansing and validation

Data transfer from an operational database to the data warehouse occurs only after the data has been cleaned and validated. This involves a series of conversions and quality checks to eliminate discrepancies in nomenclature and presentation, and to remove data that lacks the necessary contextual information. This process would be too cumbersome to include in the data capture phase and would delay data entry.

Aventis scientists enter data into their local transactional databases on a daily basis. The intent of the data warehouse system is to transfer that data to a warehouse at regular intervals and to maintain a relatively constant and easily manageable volume of information in the operational databases. In this way, data consolidation and integration proceed at a steady state, thereby avoiding bottlenecks. All data remain in the warehouses for an indefinite period and are accessible to everyone within the research and development organisation.

## Data consolidation and integration

The use of data warehousing to increase database efficiencies and data consolidation, coupled with a federated layer to integrate all required data sources, creates a firm foundation upon which Aventis can build. The main benefits of this 'best of both worlds' approach are threefold:

- The separation of the transactional databases from the analysis databases allows both to be optimised for their specific functions.
- The consolidation and cleansing of all discovery data into a fully integrated, manageable architecture.
- Global accessibility of discovery data.

With this foundation in place, the next step is to implement significantly more effective data analysis and data mining tools and strategies. The pharmaceutical industry is looking to partner with vendors to focus on this area now that we are together better able to meet the challenges of efficient data collection, organisation and data retrieval. DDW

*Dieter Kreusel is the Global IS Programme Leader for high-throughput screening at Aventis, based in Frankfurt, Germany. Prior to this he was Head of Preclinical IS at the Bridgewater, New Jersey, US site. Before joining Aventis Dieter was a software developer for Hoescht AG where his projects included the development of production planning systems and LIMS. Dieter obtained his MSc in Computer Scienceat the Fachhochschule Darmstadt, Germany in 1990.*