The Bioinformatics System Architecture

by Dr. Richard Casey

Originally published April 26, 2005

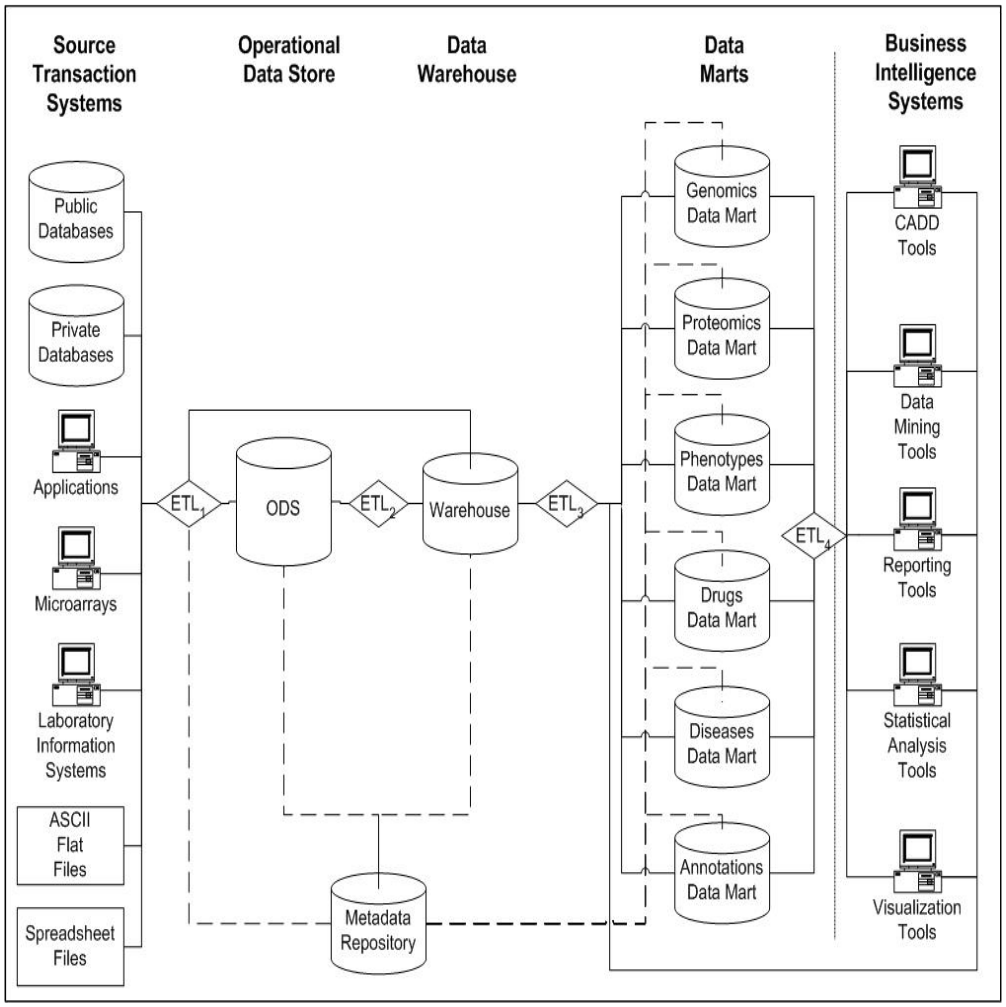
Feeling overwhelmed? If your work involves bioinformatics, then you may have good reason to feel inundated with data and information. A recent report in *Nucleic Acids Research* (2005) identified 700-plus public bioinformatics databases, covering such topics as nucleotide sequences, RNA sequences, protein sequences, small-molecule compounds, genomics and proteomics, metabolism, disease states, microarrays, species and the ever-present “Misc.” We can easily include the many hundreds of proprietary databases and software applications that make up a bioinformatician’s tool chest to this list.

To get a grip on things, organizations that are charged with effectively managing bioinformatics data need a formal structure or framework in which to operate. Without a framework for guidance, bioinformatics data management can easily get out of hand, leading to serious problems with data integrity and information dissemination.

One such framework is the **Bioinformatics System Architecture (BSA**). The BSA is loosely based on system architectures found in other disciplines, such as manufacturing, finance and telecommunications, where it has a proven track record in improving data management. I’ll briefly describe the design of the BSA and some of the benefits that the bioinformatics community can derive from it.

Components of the BSA

The main components of the BSA are shown here in a high-level system architecture diagram:



Each component of the architecture serves a specific function, and they operate together as a cohesive unit to deliver high-quality bioinformatics data and information.

- **Source Transaction Systems.** Data is created by humans, by machines or some combination of the two. Source transaction systems generate, capture and store the data. The source systems can be public or proprietary databases located inside or outside an organization, genomic and proteomic software applications, microarrays, mass spectrometers, LIMS (Laboratory Information Management Systems), flat files, spreadsheets and a host of other systems.
- **Operational Data Store (Data Integration Hub).** Source transaction systems send data to an operational data store (ODS), sometimes called a data integration hub. The ODS or integration hub is used to aggregate data from multiple sources, standardize and format the data if possible and then send it to a data warehouse. This is a good place to enforce company standards on data formats and data content. It is also used to integrate data from multiple

sources, reject bad data and resolve inconsistencies in data before they are propagated downstream.

- **Data Warehouse.** The data warehouse is a central repository of data gathered from source transaction systems and the ODS. Data is stored permanently in the warehouse and accumulates over time. Staff members can retrieve many years worth of historical records from the warehouse. And the warehouse can become quite large as a wealth of information is deposited there.
- **Data Marts.** Data marts are used to extract specific subsets of information from the warehouse. The data can be tailored and designed for presentation to individual researchers, specific labs, programs, executive management or groups that need to see only a subset of the warehouse data. Data marts can be designed around specific topics, such as particular disease states, species, genomes, proteomes or almost any area relevant to a single group of end users.
- **Business Intelligence Systems.** Business intelligence systems are at the tail end of the architecture. These systems include data mining tools, statistical analysis tools, visualization applications, molecular modeling tools and a host of applications used by the bioinformatics community. As far as end users are concerned, the business intelligence systems are the most important component of this architecture. Generally, business intelligence systems are the only view end users have into the BSA system, so the quality and integrity of data and information is immediately apparent when they use business intelligence systems.
- **Metadata Repository.** The metadata repository is a key component in this architectural design. It maintains a record of data flow through the system, including such things as the date and location of data creation, the system on which the data was created, who created the data, who owns the data (if anyone!), where the data is stored at any given time, how data was transformed and integrated with other data and so on. The metadata repository is an excellent tool for keeping an audit trail of data, which can be especially useful for adhering to FDA compliance rules.
- **ETL Scripts (Extract-Transform-Load)**. ETL scripts are the plumbing of the architecture. They handle all the data transfer responsibilities between system components. Largely hidden from view and operating in the background, they nevertheless can impact overall system performance and have to be carefully designed and maintained.

Benefits of the BSA

So why would anyone go through the trouble of creating a system like this? From personal experience in designing and implementing this type of architecture in several organizations, I can attest to its benefits and measurable improvements in information delivery.

- **Improvements in Data Integrity and Data Integration.** The BSA applies a rigorous set of filters, standards and monitors to all data flowing through the system. Data integrity rules are enforced at several locations, which makes it difficult for bad data to reach end-users in the business intelligence presentation layer. Also, data integration takes place in the integration hub, well upstream from the business intelligence layer. Data integrity and data

integration combined can result in dramatic improvements in the quality of bioinformatics information delivered to end-users.

- **Improvements in Auditability of Data.** FDA compliance rules and the Sarbanes-Oxley act require that organizations be able to audit, track and report information in a timely and accurate way. The metadata repository is an excellent vehicle for adhering to these rules. With the repository, one can easily monitor the location and progress of bioinformatics data in the BSA. Staff members can maintain complete audit trails over a period of many years, or indefinitely if necessary. This is especially important in the biopharmaceutical industry where drug discovery protocols, lab results, results from clinical trials and a host of other information must be archived, maintained and reproduced at any time.
- **Improvements in Information Dissemination.** The bioinformatics community is experiencing a period of explosive growth in the amount of data it generates. The BSA addresses this issue nicely. The combination of a centralized data warehouse and distributed, specialized data marts gives us the ability to archive vast quantities of data and then disseminate it to specific groups. Data marts can be set up with predefined queries that extract only the most appropriate data for a research group, executive management team or even an individual staff member.

The BSA framework and benefits derived from it make it an appealing architecture for any organization that is grappling with bioinformatics data and information management. If you would like to hear more about my personal experiences with BSA implementations, feel free to contact me anytime.

-  **Dr. Richard Casey**

  Richard is the Founder and Chief Scientific Officer of RMC Biosciences Inc., a firm that offers services in Bioinformatics and Computer Aided Drug Design. Dr. Casey received a Ph.D. in Biological Sciences from Colorado State University. He has 20-plus years experience in Computational Sciences, Information Technology and High-Performance Computing. He has held corporate and academic positions at Hewlett-Packard, Boeing Computer Services, Arizona State University, Colorado State University, the Alabama Supercomputer Center, and the Institute for Computational Studies at CSU and was the founder of a software consulting firm, Alpine Computing Inc. He holds a Project Management Professional Certificate and a Bioinformatics Certificate from Stanford University. Richard can be reached at rcasey@rmcbiosciences.com.

Recent articles by Dr. Richard Casey

- Bioinformatics in Structure-Based Drug Design
- Federated Databases in Bioinformatics and Translational Medical Research
- How Federated Databases Benefit Bioinformatics Research
- Designing Chemical Compound Libraries for Drug Discovery

Related Stories

- No Perfect Answer
- From Business Intelligence to Enterprise IT Architecture, Part 4

Comments

Want to post a comment? Login or become a member today!

Be the first to comment!