UC-Berkeley Data Warehouse Roadmap

Data Warehouse Architecture

Table of Contents

UC-Berkeley Data Warehouse Roadmap

Data Warehouse Architecture

Introduction

The Roadmap Project

In the summer and fall, 2005, the Berkeley Campus of the University of California undertook a project to develop a plan for expanding its existing reporting systems into an enterprise data warehouse. This initiative originated as a project of the campus Data Stewardship Council. Many members of the Council, as well as other campus leaders, felt there were many opportunities for the campus to realize large benefits by reusing information collected by operational processes in order to support better decision-making and to improve processes. The roadmap project was chartered by executive sponsor William Webster, Vice Chancellor—Business and Finance. The campus CIO, Shel Waggener, was a co-sponsor. The project consisted of two phases.

The first phase was an interview-based study which focused on clarifying the need for an EDW and identifying specific opportunities for an EDW to contribute to important campus goals. The object of this phase was to help the campus' executive management understand the business case for investing in an EDW and to agree on priorities for guiding an incremental development expected to occur over several years. This business study phase was led by functional managers and staffed with analysts from several campus units as well as IS&T.

The second phase focused on developing an architecture to deliver the EDW specified in the business requirements study. The architecture was developed by a cross-functional team including functional business analysts and technical personnel. It was led by the campus data architect. This document summarizes the architecture developed by that team.

Campus Functional Needs

During the study phase, the analysis team met with a cross-section of the campus leadership, including both academic and administrative functions. The Data Stewardship Council oversaw selection of the interview participants. Those included three different kinds of informants:

1. Campus executive leadership, including two vice-chancellors and a vice provost.
2. Campus operating management, including faculty and other instructors, researchers, budget managers and leaders of administrative support units.
3. Representative data providers—personnel involved in managing and delivering campus information systems.

The interviews covered the principal operating challenges of the campus and helped identify specific opportunities for well-organized information from operations to contribute to managing those challenges. The scope of the study included the primary functions of teaching, research and public service, plus the support functions which directly support those primary functions. Interviews with data providers helped establish whether required information was available and of sufficient quality. Resulting conclusions were reviewed with informants, other functional experts and the Data Stewardship Council.

The study identified eleven high-priority business opportunities, covering areas as diverse as course planning, purchasing and fund-raising. In addition, the study summarized important criteria critical to the success of an EDW on the campus. These results were forwarded to Vice Chancellor Webster for use in executive planning of the enterprise data warehouse. Working with the campus CIO, Shel Waggener, VC Webster formed an executive group to consider priorities, funding and governance for the EDW.

The results of the architecture were published to the campus community in December, 2005. A copy of the final report can be found on the EDW project website.

Scope of the Architecture

The architecture of the enterprise data warehouse is designed to deliver the analysis capabilities defined in the business requirements document just referenced and likewise to provide the critical success elements defined there. In addition, the warehouse is expected to deliver any additional analysis capabilities delivered by existing campus decision-support systems which were not explicitly documented in the business requirements document. These systems include the following:

- BAIRS
- BIS
- Cal Profiles
- The pilot student data warehouse
- FASDI
- The Office of Student Research database/ reporting file.

Architecture Requirements

Based on the business requirements and success factors established in the business requirements study, the architecture team investigated and documented the more specific architectural requirements which would govern development of the architecture. These included identifying potential users, defining security requirements, skills requirements, etc. Those architectural requirements are summarized and reported in a document which can be found on the EDW project website.

Based on documented functional business requirements and derived architecture requirements, the team worked with personnel knowledgeable about the existing data warehouse to identify important gaps which needed to be addressed in the architecture. That gap analysis is summarized in a document included as an appendix to this document.

Components of the Architecture

Though it's easy to think of the data warehouse as just a big collection of data, in fact delivering an effective data warehouse requires a large set of related capabilities. (see Figure 1).

Certainly data is the fundamental component: cleaned, organized data, mostly extracted from the campus' operational systems. Making that data useful to a variety of campus personnel, though, requires some applications to deliver and explain it. These applications range from predefined reports through query tools to complex tools for analysis and modeling. Delivering data and applications and securing the data as specified by campus data stewards requires a set of technology, most of it centralized in secure computer locations. Equally important, transforming operational data into a shared resource useful across the boundaries of functional

business domains requires a broad set of functional skills, organized appropriately and working through proven processes.

The architecture for the data warehouse is described in terms of four inter-related dimensions:

1. Applications (or the business intelligence layer).
2. Data.
3. Technology and security.
4. Support—processes and organization.

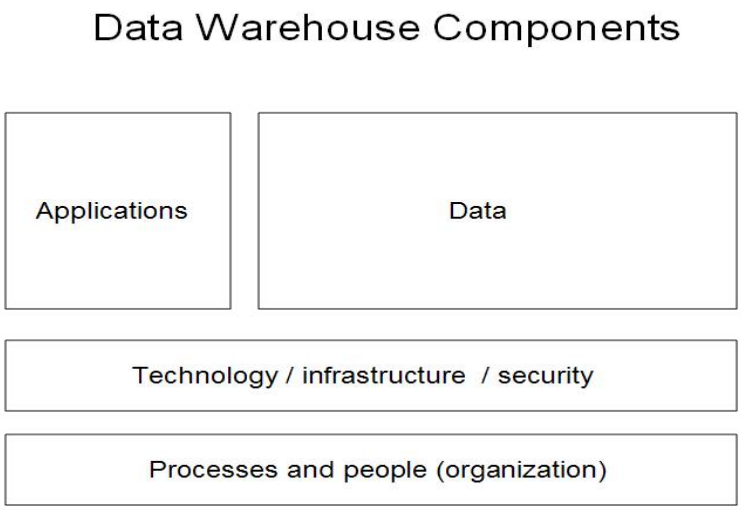


Figure 1: There's more to the data warehouse than just data

Business Intelligence: Applications Architecture

For the information in the data warehouse to be valuable, it needs to be delivered in way that makes it useful to campus personnel in doing their jobs. This is the job of business intelligence applications. For most people, these applications *are* the data warehouse. They are the software systems which help users understand what has happened, identify problems and opportunities, and to make and evaluate plans. The warehouse includes a variety of these tools because there are a variety of users with quite different needs and skills.

Everybody on campus uses information to do their jobs, but they differ greatly in how well they understand available information and its interconnections and also in how comfortable they are in using information to help them make decisions. The data warehouse applications form a toolbox with tools appropriate for the spectrum of campus personnel from deans to departmental administrative assistants, from students to analysts, from researchers to executives.

The following diagram summarizes the EDW applications:

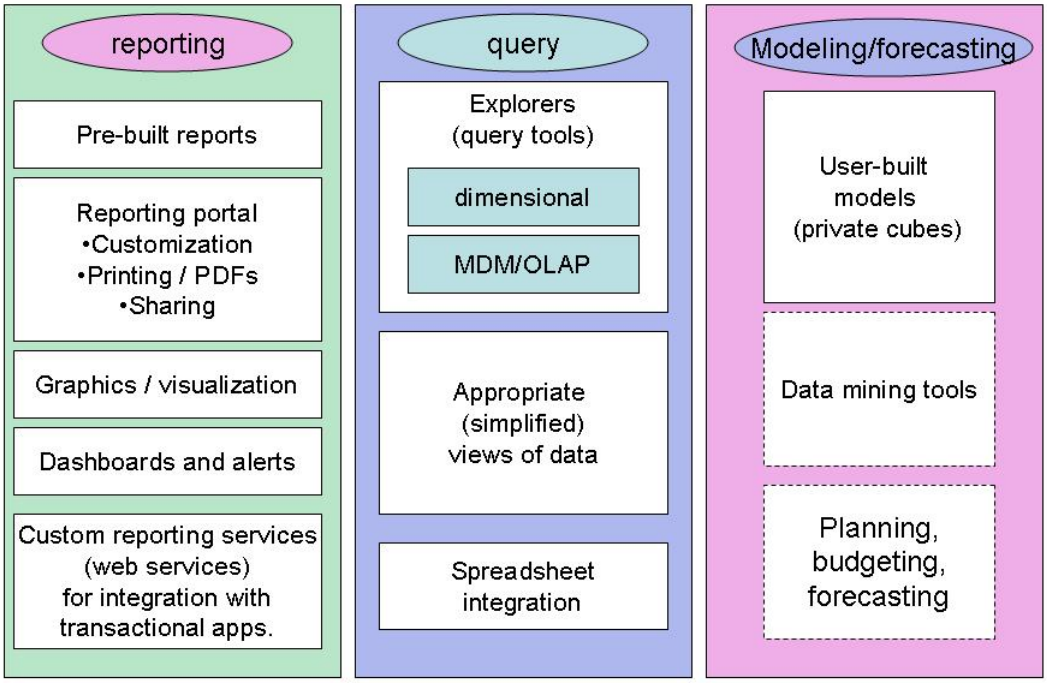


Figure 2--Business Intelligence Applications

The architecture provides three kinds of applications:

1. Reporting, or information delivery, including variants such as dashboards and alerts.
2. Query.
3. Modeling, Planning and Forecasting.

Reporting.

Reporting applications deliver information in a form which is useful to users. In their simplest form—fixed, printed reports—these tools are as old as computers. They imply a partnership between users who need information and specialists who help design reports, displays and graphics to deliver the required information. The best reports provide just the information a user needs for a specific purpose, delivered in a way which makes the information usable and actionable.

Reporting accounts for most of the campus' current data warehouse capability. Systems such as BAIRS and Cal-Profiles focus on reporting.

The UC-B data warehouse will provide a starter set of predefined reports for each distinct group of users. These reports, like the existing BAIRS reports, will be configurable by users to deliver a specific subset of information. They will be delivered through a secure portal which will make it easy for users to further customize the reports and to share both report requests and results with others in their workgroups. In addition, the portal will permit users to schedule routine

reports, to print reports and to share them as electronic documents in other formats, such as Portable Document Format (PDF).

The reports will be designed for easy understandability. In addition, they will be accompanied by high-quality data documentation expressed in language which communicates with the expected users. This documentation will be delivered through the reporting portal.

As appropriate, reports will use graphics, such as charts and data graphs, to make information more comprehensible.

Dashboards and alerts.

In the interest of delivering just the information most useful to support decision-making or action, it will be useful to complement conventional reports with information dashboards and alerts. Dashboards communicate information with quickly-comprehended graphics such as dials and meters. Typically dashboards are used to report on established performance indicators, measured at a predefined intervals. Properly planned, dashboards make it easy for personnel managing a complex process, such as undergraduate admissions, for example, to quickly assess current state and progress against goals. Likewise, sometimes reporting will be confined to exception conditions. Exceptions may delivered periodically as reports. More often, though, they will be delivered as alert messages—emails or similar notifications. For example, departmental administrators might get alerts when course enrollments exceed critical thresholds. Or a grant administrator might get an email when grant expenditures exceed plan for a given period.

Reporting integrated with production applications.

Some reporting will be most useful when it is delivered in the context of some other computer-based process. For example, a student who is registering for courses might benefit from knowing about other students who took similar courses—how many took the same set of courses at once and how well they succeeded; how many completed the courses and their average grades. A researcher applying for a grant might benefit from knowing about previous submissions to the same sponsor in the same research area. To enable such in-context reporting, the data warehouse will provide reporting services which will integrate with other software applications.

Query and analysis.

People who have analytical skills and jobs requiring analysis will need the ability to explore the information in the warehouse. For example, a member of a committee interested in improving educational outcomes might want to explore the how use of the summer program affects outcome measures and how that varies for students with different majors or different economic backgrounds. Enabling analysis of this kind is one of the great powers of the EDW. Using that power requires understanding the information in the warehouse and knowing how to select data, summarize it or drill down for further detail, and particularly how to combine information across subject areas. The EDW provides two kinds of tools for this function:

1. Explorers—tools for selecting data, drilling down or summarizing data, and combining data across subject area.

2. Special-purpose information views which organize the information in the warehouse into simpler structures which are easily understood and navigated by particular kinds of users.

The information in the EDW is from the outset organized for understanding, consistency and ease in accurately combining information across subject areas. (See Data Architecture). However, many users of the warehouse focus on some aspects of the campus business. For these users, much of the information in the warehouse is irrelevant and unfamiliar. The job-focused views present meaningful subsets relevant to and understandable by particular sets of users. These views are developed over time as new kinds of users begin using the warehouse. In particular, analytical users will have access to views which make it easy to use the dimensional structure of the data warehouse to easily "slice and dice" information. These same views will make it easy for users to summarize information based on relevant characteristics, such as demographics, without having access to protected identifiers such as student-ids or social security numbers.

The EDW will provide a good, general purpose explorer/query tool which is well-suited to the task of selecting and combining data from the dimensional structures of the data warehouse. This tool will have capabilities for presenting selected information in easily-understood report format and additionally as graphs, charts or other visualization tools. Additionally, the warehouse will provide a multi-dimensional analysis tool for delivering data as "cubes" or multi-dimensional structures which lend themselves to easy "slicing and dicing". (This analysis approach is often referred to as On-line Analytical Processing or OLAP). Both tools will allow users to extract information for further analysis with desktop analysis tools such as spreadsheets. Both query tools will deliver clear, easily-understood and user-focused information documentation.

Advanced applications: modeling, forecasting and planning.

Many analytical users want to use the historical data in the warehouse to build models of alternative scenarios—trying to predict, for example, the consequences across a variety of business areas, such as course enrollments, student aid demand, and fees revenue. The warehouse will provide several facilities enabling this kind of modeling:

- Good integration with external modeling tools, ranging from desktop spreadsheets to sophisticated data mining tools.
- Multi-dimensional tools for building large models directly in the warehouse database operating with spreadsheet-like functions but at a scale beyond the capability of spreadsheets.
- Facilities for saving and sharing multidimensional models across a workgroup.

Many campus users need to be able to understand the current status of their finances against plans—not merely budget allocations, but more detailed plans for using the budget. The most frequent example involves research projects. Because staff migrates among research projects, project managers need tools for expressing a spending plan for a research project, especially a multi-year project, showing labor forecasts by individual. Then, with that plan made, the project managers need to track expenditure against plan. The solution for this need involves collaboration between the data warehouse and a transactional application for recording plans. Once plans are recorded, they are migrated into the data warehouse, where they are available

for reporting and query in combination with actual activity. Other units besides research have similar needs: IST, for example, needs to project important financial events such as license renewals and track actual expenditures against such projections. As in the case of salary encumbrances for research projects, the data warehouse is part of the solution for this need. A transactional application records projections, which are carried forward into the data warehouse where they are available for analysis alongside current activity.

There is no provision in the DW architecture for more robust enterprise planning applications, as business users have not identified a need for such applications. However, the warehouse architects should review this area within three to five years, as many organizations of comparable size, including major research universities, use such planning applications to manage the general problem of coordinating plans across business areas under conditions of uncertainty. During campus interviews, informants often talked about the need to manage that problem here at UC-Berkeley.

Summary of differences from current architecture.

- Most of this Business Intelligence architecture consists of new capabilities.
- Today's decision-support, such as BAIRS and Cal-Profiles, are oriented toward reporting. They provide pre-defined reports, with some facilities for configuring general-purpose reports for specific data populations, showing accounting activity reports, for example, for only a specific set of organizational units or funds. Both these systems do a good job of delivering business data documentation alongside the reports. BAIRS provides facilities for exporting report results to spreadsheet tools for analysis.
- The provision of facilities for dashboard presentation and for alerts is new, as are business-intelligence services which can be combined with other applications.
- Most important, all the facilities for query and analysis are new, as are support for modeling and planning. The use of OLAP analysis is new. Business users during the requirements survey talked about their need to query data directly rather than depending on predefined reports.

Data Architecture

There are two important views of the EDW data architecture. The first is the functional view of the information the EDW stores and delivers. This view concerns *content* and *meaning.* The second is the view of warehouse data as a set of software components, mostly interrelated database tables. This view concerns access by query and reporting tools. These two views are described as the *Contents architecture* and the *Implementation architecture,* respectively.

Contents: the warehouse bus and the logical data architecture.

Basically, the enterprise data warehouse is a font of information about *what happened.* It is a history of the operations of the campus drawn from the campus information systems, which record important information in the process of helping to carry out operations. Additionally, the warehouse delivers corresponding information about *plans* such as forecasts and budgets. Being able to investigate and understand in detail what happened enables a cycle of finding problems and opportunities, crafting plans for doing things better, then measuring to learn whether those plans worked out. Tracking plans as well helps campus personnel compare what

actually happened to what was planned in order to identify surprises, allowing for faster responses to the unexpected and also supporting improvements in the planning and forecasting process. So the contents of the warehouse boils down to *history* and *plans* linked to consistent information about the *context* of those events or planned events.

Thus, the contents of the data warehouse have two components:

1. Information about history and plans. These are referred to as *facts,* as they usually consist of discrete facts or measurements.
2. Information about the context in which these events or measurements occur. This context information is organized along consistent *dimensions.* Sample dimensions include time, organization, and student information.

These context *dimensions* provide the mechanism which enables a shared, enterprise data warehouse. Combining or comparing information from different subject areas usually involves lining up different *facts* along the same *dimensions.* For example comparing faculty workload, student teaching contact and resources used involves combining facts about teaching assignments, course enrollments, and expenditures along common dimensions of *organization* and *time.*

Most of the data mismatch or "dueling data" which makes it hard to agree on reports and measurements comes from using different versions of the same dimensions to select and summarize data. Having a common set of dimension definitions, on the other hand, makes comparing data across areas feasible and understandable.

Because context dimensions usually are shared across subject areas, one of the most important and challenging aspects of data design for the warehouse consists of identifying shared dimensions and making sure that everyone is using a set of shared definitions. The set of shared dimensions represents a powerful interpretive grid for making sense of diverse facts. They also enable iterative development of the warehouse. New fact tables can be sourced from operating systems and linked to earlier components along existing shared dimensions. The data warehouse grows by adding *data marts* one at a time to the warehouse. A data mart is a fact table and its associated dimensions. A set of related fact tables makes up a data *subject area.*

Considerable investigation and analysis went into identifying the basic facts which make up the campus data warehouse and their relationship to context dimensions. Together, the basic facts and dimensions represent the base content architecture of Berkeley's data warehouse. That content architecture is summarized in a table of facts and dimensions attached as an appendix to this document.

Following is an extract from that fact dimension matrix:

Key data marts by subject area and their shared dimensions (EXTRACT)

Dimensions

Data Marts	Date (semester, census, add/drop, fiscal, effective, reported,	Facility	Major (all degree-granting programs and interrelatedness)	Course (all courses incl. all sections, offering dept, etc.)	Student (reg status, residency, major,	Applicant for admission	Financial Aid Program	External Institution (high schools, other higher ed, SAT	Research Sponsor (govt agency, foundation, corporation,	Employee (descriptive info fac, staff, affiliates, etc.)	Job Applicant	Budgeted Staff Position (budget dimension for perm	Appointment (title, job code, leave status, etc.)	Job Classification
Teaching														
Teaching assignment	x			x						x				
Classroom space assignment	x	x		x										
Student class enrollment	x			x	x									
Wait list enrollment	x			x	x									
Degree	x		x		x									
Student course evaluation	x			x	x									
Student survey	x				x									
Student														
Undergraduate admissions	x				x	x								
Graduate admissions	x				x	x								
Semester snapshot	x		x		x									
Financial aid awards	x				x		x							
External test results	x					x		x						
Activity membership	x				x									

Implementation: data bases and the physical data architecture.

The implementation data architecture is summarized in Figure 2.

It contains these principal components:

1. The warehouse itself. This is a collection of tables and views consumed by end users, directly or through the applications tools described above. The data in the warehouse has been processed for consistency and alignment with standard data descriptions and value sets. Facts have been aligned with the standard dimensions.
2. A staging environment. This is a set of data bases and files used by the data maintenance process to prepare data for publication in the warehouse as it flows from the operational systems which collect it originally. (This maintenance process is often referred to as Extract-Transform-Load, or ETL). This environment is not directly accessed by users of the warehouse.
3. A collection of *metadata*—descriptive information about the contents of the warehouse, about the source systems, about the business meaning of information, about access and privacy rules, and about the processes which create and transform data before it appears in the warehouse. This metadata is described in greater detail below. Though managed as a consistent set, the metadata itself is stored in repositories distributed among the warehouse, the staging environment and the source systems.

(Though the architecture diagram suggests that the implementation is a single, centralized database environment, in fact any of the components may be distributed across a variety of database servers and even a variety of DBMSs, as discussed under “DBMS platform”, below.)

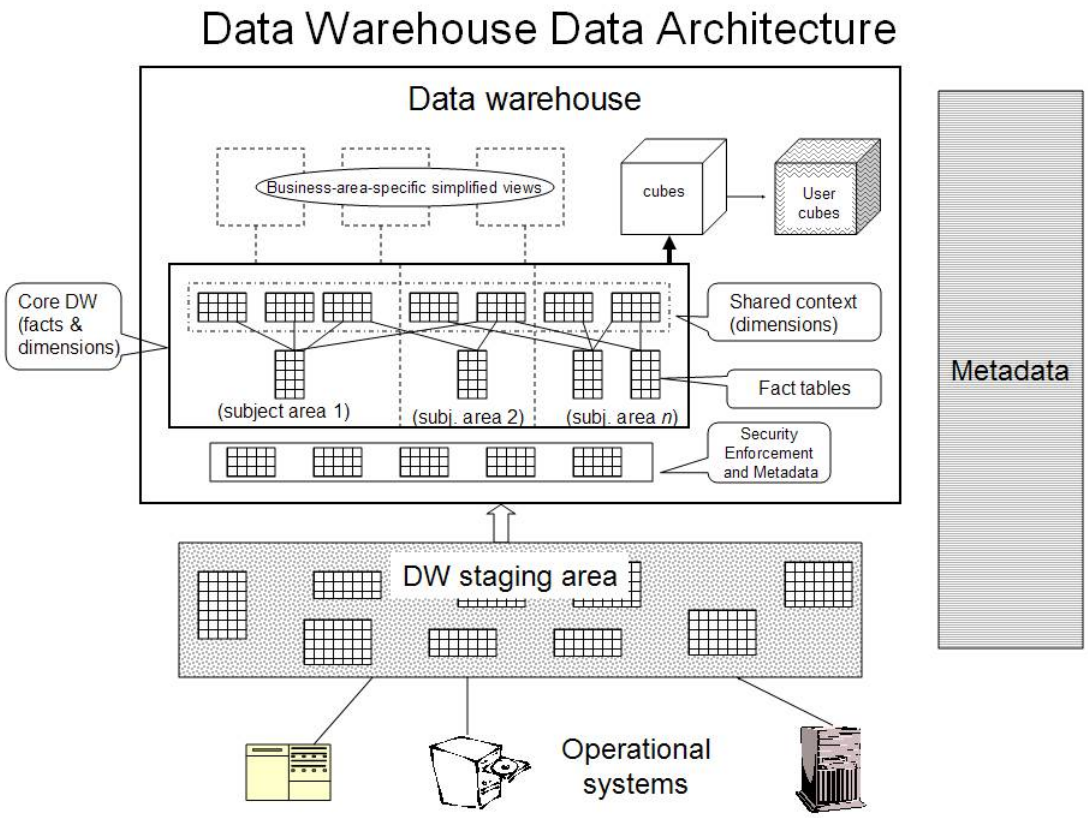


Figure 3: Implementation Data Architecure

The data warehouse component is the largest and most complicated of these components. It is worth describing that component in detail.

Core of this component, and of the warehouse itself, is a set of linked fact and dimension tables. These implement the facts and dimensions described above. Normally there will be a single fact table for any given kind of fact in the warehouse. A single row in the fact table contains the measures related to that fact at the lowest level of granularity appropriate to the fact. For example, a row in the student course enrollment fact table will contain information about a change in a single student's enrollment in a specific section of a course taught in one semester. Each row in a fact table will also contain a foreign key reference to a dimension table for each relevant dimension.

Facts are usually navigated as sets identified by characteristics specified in related dimension tables. An example would be a count of all the course enrollments by seniors in Physics classes meeting before 9 a.m. in the past two spring semesters.

Both fact and dimension tables are identified with system-assigned surrogate keys. This buffers the tables from anomalies in source systems and allows for historical variation. It also makes it easier to mask off natural identifiers—often necessary to meet security goals.

Cubes are specialized views of a set of facts and dimensions. They take a form very similar to spreadsheets, in the sense that they are composed of cells visualized along a set of axes. Unlike

spreadsheets, however, cubes are provided directly through database technology often called OnLine Analytical Processing, or OLAP. OLAP cubes can be much, much larger than any spreadsheet. The axes in cubes are provided by the values of hierarchical dimensions. So for example, rows might be all individual course offerings. OLAP cubes are very useful for certain kinds of analysis and pattern recognition. Personnel such as budget analysts frequently find cubes useful views of warehouse information. Such personnel often use database cubes in the process of making models or projections. Those models can be stored as **user cubes** alongside actual warehouse contents and shared with other users.

Sourcing strategy

By principal, data in the warehouse is sourced from its authoritative systems of record, as certified by the data's organizational custodians. Generally, identifying the appropriate source of *fact* information is not difficult, as fact tables usually are sourced from a single system. Dimension tables, however, are more complicated. Information in dimension tables, particularly for shared dimensions, often is derived from several systems. Moreover, manual intervention may be required to provide data not in any single source system. For that reason, there is a *dimension owner* assigned to each dimension. The owner is a functional custodian responsible for establishing the content of dimension tables and publishing changes to all users. For example, *Fund* is a widely shared dimension. Though much of the descriptive information about funds is sourced from the financial system, other information about fund sponsors, donors, participating principal investigators, etc., needs to be derived from other sources. The owner of the Fund dimension is responsible for providing a process for tracking relevant changes and ensuring that changes to information about funds is published to all users.

The data in the warehouse is subjected to continuing data quality assurance. Before a subject area is added to the warehouse, the data in that subject area is analyzed for consistency, completeness and accuracy. Results are reviewed with data custodians and overall quality is confirmed before the data is made available for use in the warehouse. Routine processes for validating the quality of the data are incorporated in the routine maintenance (ETL) process.

Metadata architecture

The diagram below shows the basic elements of a long-term vision for EDW metadata management. The key to this architecture is the Master Metadata Repository, the central clearing house for all EDW metadata. Built on top of this repository are the applications needed to publish its contents to the platforms and processes shown at the top of the diagram.

The Master Metadata Repository (MMR) is functionally part of the EDW as a whole, refreshed by its own set of ETL processes and dynamic interfaces that extract metadata from the sources shown at the bottom of the diagram. These processes should be designed to seamlessly integrate metadata access and modification with data collection, analysis, maintenance, and development activities.

The MMR serves users beyond the EDW community. Because the MMR contains metadata from systems outside of the EDW, it help users of those systems as well. It also will support IST processes such as systems development, change management, and technical support.

This metadata management framework will need to be implemented incrementally, the first phase being the creation of a Master Metadata repository. Subsequent improvements should

serve to deliver incremental value to the EDW and its customers until the final future state is reached.

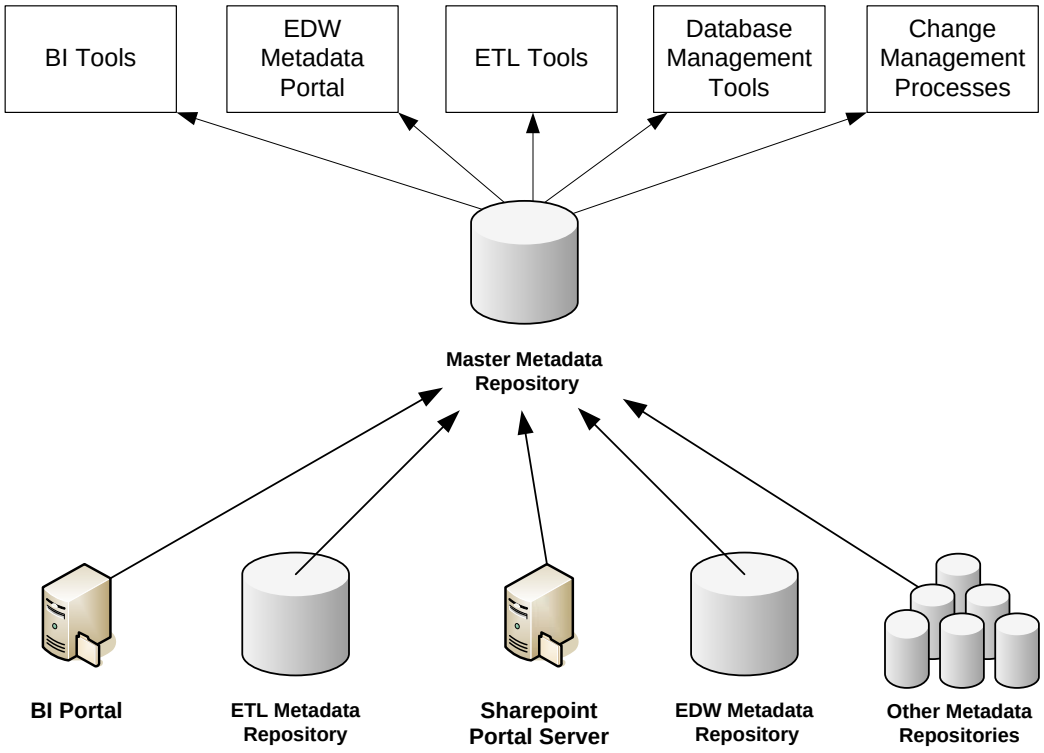


Figure 4--Metadata maintenance

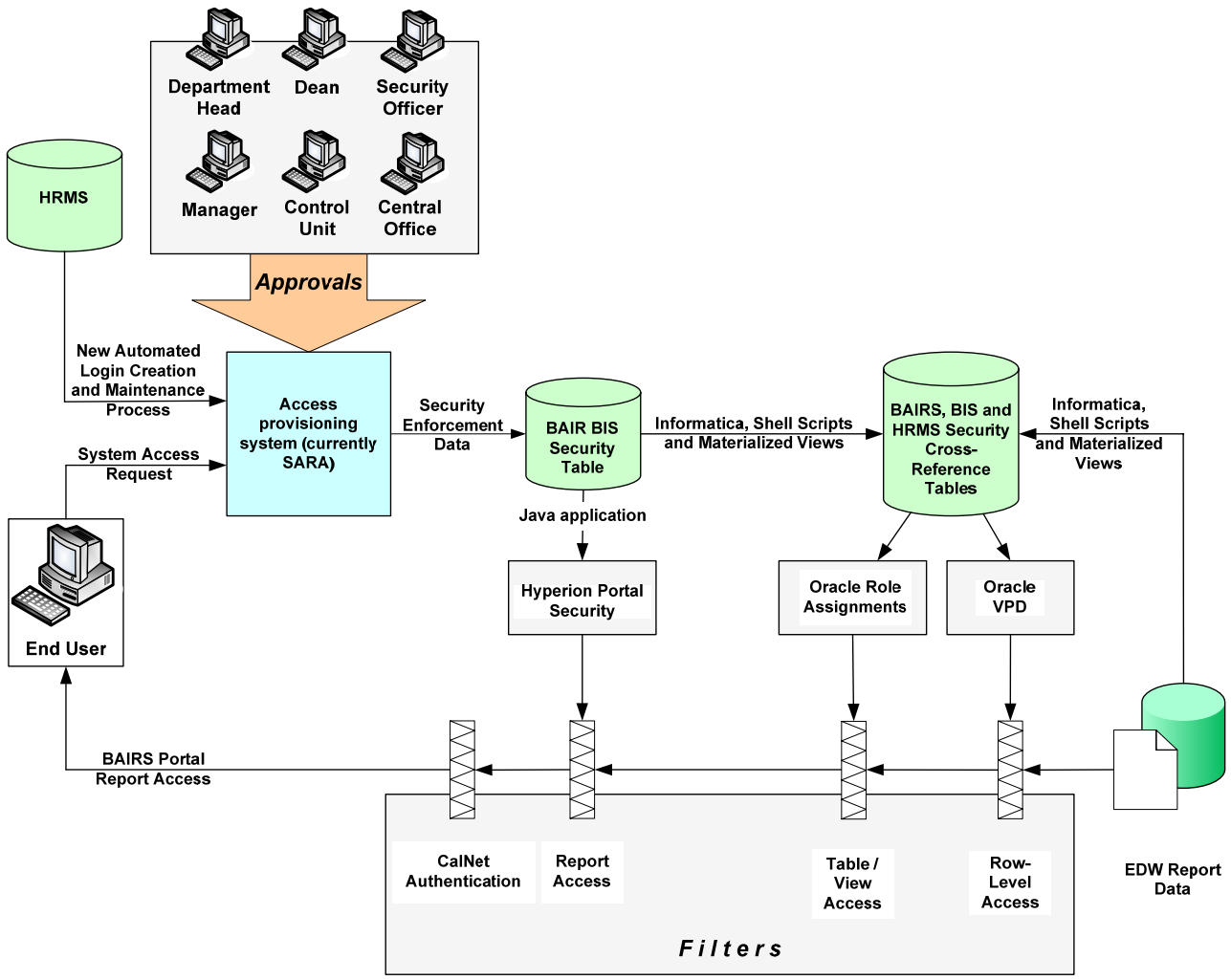
Summary of differences from current data architecture

- The biggest difference from the current architecture is the addition of a large number of new data subject areas, broadening the warehouse to include comprehensive information about activity in the core areas of campus activity.
- The second biggest difference is the provision of an accurate fact-dimension data architecture, implemented with shared dimensions and managed by dimension owners.
- The new data architecture rectifies the overlaps and inconsistencies between the existing BIS and BAIRS data and corrects those systems to align facts with standard dimensions using the standard surrogate key mechanism.
- Additionally, the provision of simplifying database views based on the functional needs of various user groups is an extension of current practices.
- The provision of OLAP cubes is new.
- Consolidated metadata management is new.

Security Architecture

The goal of the security architecture is to provide fine-grained control over access to data, administered according to the policies of appropriate data custodians. This includes managing access at the individual data element level, but it also includes considerably more. Current data access requirements mean that sometimes the warehouse has to control access to information within a particular context, such as information about students who have taken certain classes or studied with particular instructors. Restrictions on small sample sizes imply that for some uses, data access is restricted to answer sets large enough that they don't implicitly identify individual persons. Meeting all these requirements is done by a set of facilities, some automated in the data bases and some in the reporting portal.

The security architecture is summarized in the following diagram:



Future EDW Security Architecture

Summary of differences from current architecture

Improve System Access Process for New Hires

Create a new automated process to provision access privileges for new users, especially those from volatile user groups such as students. This can be achieved initially with an application that populates new user roles within SARA based on data obtained from the HR Management System. As the campus Identity and Access Management (IAM) system evolves, the EDW access provisioning will adapt to align with role-based access provisioning mediated by the IAM.

Incorporate Security Process Improvements into EDW Lifecycle Components

Include a security requirement component for all new initiatives, and track any changes through analysis, design, development and implementation. In addition, institute regular review of all existing security requirements by business stakeholders and data stewards, especially when major changes to the EDW environment are being planned.

Improve Maintenance of System Access Privileges as User Roles change

This can be achieved through one of the following methods:

- Provide user lists to supervisors on a regular basis (e.g. once a year) to confirm roles are being properly assigned. Currently, supervisors are not consistently tracked with a high degree of accuracy within HRMS – this will need to be addressed in order to implement this solution.
- Create an automated process that updates user roles within SARA based on changes within HRMS.

Add context-based authorization features to existing security control process

- Modify SARA to process new context-based user roles (e.g. instructor) using new security fields which will be stored and tracked in a revised BAIR BIS Security Table.
- Modify existing applications to load new roles and associated security fields to EDW security tables.
- Modify Hyperion Portal Security, Oracle Role Assignments, and Oracle VPD to implement proper filters to restrict user access to EDW data based on new roles.

Avoid exposing personal information by implementing the following techniques.

- **Use surrogate keys**. In many cases, personal identifiers (e.g. name, student id, employee id, social security number) can be entirely masked without reducing the power of the data warehouse to deliver useful information about student and employee behavior, distribution by demographics, etc. The enabling mechanism is the use of substitute identifiers known as surrogate keys. Using surrogate keys provides many other benefits, including ease in tracking variation over time in information about students, employees, etc. The EDW should use surrogate keys in most cases.
- **Use views to exclude small sample sizes from aggregate queries.** This may require the creation and maintenance of tables which track minimum sample sizes based on subject area or other criteria. This approach should be re-evaluated as detailed requirements for small sample sizes are developed. In particular, views can be used more effectively if they are based on universally-applied rules rather than rules which are contingent on user roles or other contextual information.

Reconcile EDW security architecture with the evolving campus Identity and Access Management (IAM) initiative.

IST is leading a substantial initiative to improve identity management on the campus. The EDW security architecture is designed to use the Calnet and the rest of the campus IAM system. As necessary, the EDW will evolve to align with changes in the IAM. One example of changes to IAM is the move to assign the Calnet userid based on the UID field. This would ensure that campus users who are both students and employees would have a single consistent user-id. As the IAM evolves provide consistent role definitions, the EDW access provisioning would evolve from the existing SARA system to use the IAM roles to help with provisioning access privileges.

Technology Architecture

Technology required for delivering the data warehouse is summarized in the following diagram. This technology architecture does not differ from the technology supporting BAIRS. Indeed, over the past few years, technology to support business intelligence has matured, consolidated, and simplified somewhat. In particular, it is now possible and desirable to eliminate the welter of desktop-based query tools and corresponding database connection protocols which were a feature of older business intelligence implementations.

Some of the architecture components are described in more detail below.

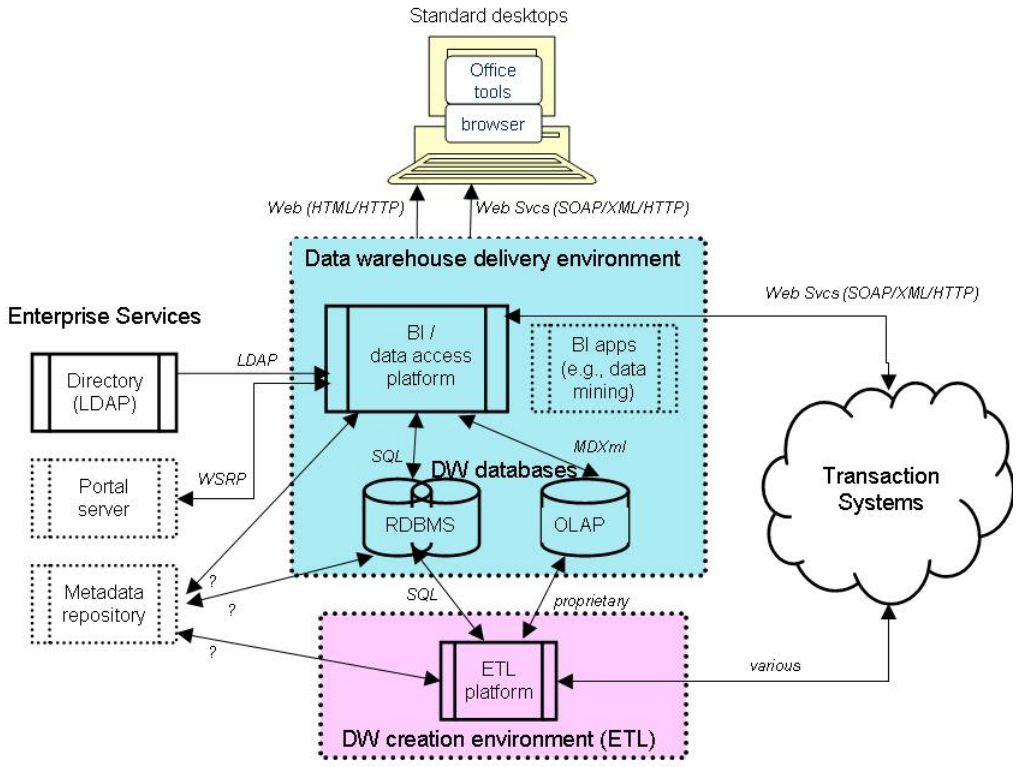


Figure 5: DW Technology Components and Protocols

DW delivery environment

As illustrated above, query and reporting functions are automated by a server-based Business Intelligence platform. That platform interacts with users mainly through browser-based query, reporting and analysis tools. In addition, the platform interacts directly with some office tools, notably Microsoft Excel, again using basic web interfaces.

The BI platform delivers the query and analysis functionality described in the Applications Architecture section above. It interacts with the DW databases, both the dimensional RDBMS platform and the MDDB (OLAP) platform.

Much of the database understanding functionality described above and some of the security features are automated by the BI platform. Desktop-based query tools cannot deliver such functionality alone and therefore are not included in the architecture.

For BAIRS, the functions assigned to the BI platform are delivered by the Hyperion toolset. However, the currently-installed version does not deliver all of the functions assigned to this component. During implementation of the architecture, the campus will need to evaluate Hyperion against competing products and choose an implementation platform. Though there are several capable BI platform products, they are by no means standard. As a result, an effective BI platform for the campus should be a single product set. For that reason, the architecture specifies the BI platform as a single technology component.

In addition to the basic data access/BI platform, the data warehouse will include over time a variety of analysis, modeling and reporting tools. These will interoperate directly with the DW databases and with users through browser-based interfaces. A data-mining tool set is an example of such a tool, as is a comprehensive enterprise planning system.

Database technology

As described in the data implementation architecture above, the data warehouse uses two kinds of data base management systems: relational DBMS, which contains the original copy of all warehouse data, and a Multi-dimensional DBMS, which stores and delivers OLAP cubes.

Today, all the centrally managed warehouse data is stored in Oracle; that is likely to continue to be true. However, there is nothing in the architecture which limits the warehouse to one kind of RDBMS. It may be practical to implement some data subject areas in smaller, less costly RDBMS products such as SQL Server 2005.

On the other had, multi-dimensional databases are only loosely standardized. For that reason, it is undesirable to deploy more than one kind of MDDB. The architecture specifies only a MDDB which will support MDXml; most products do. As the architecture is implemented, the campus will need to evaluate available offerings and select one.

Summary of differences from current architecture

This architecture differs from the installed architecture in three principal ways:

1. It includes support for multidimensional databases (OLAP structures). This is entirely new.
2. The architecture totally eliminates desktop based query tools in favor of server-based data access delivered at the workstation through the browser and through office tools such as Microsoft Excel. This completes a migration which began a while ago.

3. The architecture provides important new interfaces with other campus systems. These are of two kinds:
    a. Web services interfaces which allow business intelligence to be delivered in the course of transactional systems. Some of the applications of this feature are described in the Applications Architecture section, above.
    b. A set of interfaces for delivering business intelligence through a standard portal. These interfaces enable useful dashboards and alerts as well as reporting portlets.

Support Architecture

Most of the critical success factors for the data warehouse identified during business interviews concerned the need for executive engagement and support as well as the need for robust participation by functional experts from all areas of the campus, academic and administrative. These observations are consistent with the experience of organizations who have successfully implemented enterprise data warehouses.

These needs drove the design of the support structure for the data warehouse, made up of skills and organization as well as processes.

The full support architecture is documented in a separate document, Data Warehouse Support Architecture, which is available at the EDW Roadmap website.

Following are the most important features of the support architecture.

Skills and organization

Data Warehouse Competency Center

The successful delivery of Data Warehouse systems requires the development of a support architecture with the capabilities to:

1) Ensure investment and priorities in the data warehouse are aligned with campus goals and objectives
2) Maintain an overall Data Warehouse design and architecture that promotes data consistency, reliability and quality
3) Provide strong data stewardship and governance
4) Manage multiple, concurrent data warehouse projects
5) Maintain a balance between developing new systems (adding new data sources and/or new user groups) and maintaining existing systems
6) Manage multiple sources of funding, with each contributing to building the data warehouse infrastructure

These capabilities can be categorized in to 4 areas:

Governance	General oversight of the Data Warehouse systems including funding, prioritization of initiatives, resource commitments and data stewardship.

Planning and Organizing	Business requirements and analysis, high level design and architecture, data standardization and conformity, end user training and support
Development	Development, or significant enhancements to, data warehouse systems
Support	Implementation and support of data warehouse systems

These capabilities are best developed and managed as a Competency Center. The Competency Center consists of:

a) A virtual organization consisting of both business and IT resources that are responsible for defining, building and maintaining a centralized data warehouse environment
b) A set of processes used to manage the activities of the Competency Center

An underlying assumption of the Data Warehouse Competency Center is that data warehouses are built incrementally. Therefore a program approach is required to manage multiple projects adding new increments to the data warehouse, and to oversee the maintenance of existing data warehouse applications.

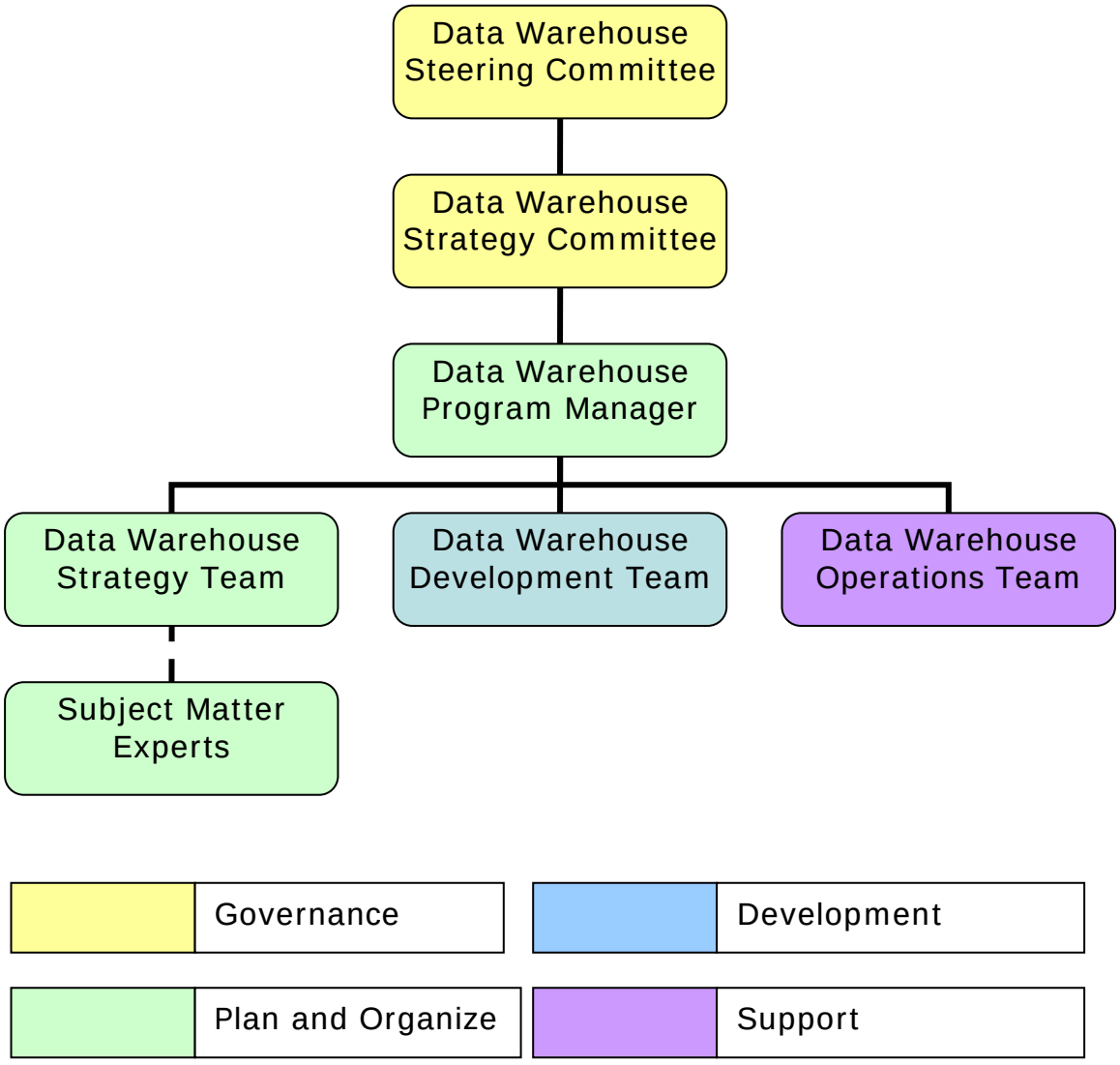
Organizational Components of the Data Warehouse Competency Center

The Data Warehouse Competency Center consists of 7 teams:

Steering Committee	Providing funding and overall objectives for the Data Warehouse Competency Center
Strategy Committee	Responsible for approving the Data Warehouse Roadmap, prioritizing initiatives, championship of the data warehouse program and sponsorship of data warehouse projects
Program Manager	Responsible for overall management and coordination of the Data Warehouse Competency Center
Strategy Team	Responsible for defining data warehouse opportunities, business cases and business analysis, high level design and end-user support
Subject Matter Experts	Representatives of the academic and administrative departments (who are end-users of the data warehouse) responsible for providing requirements and end-user support

Development Team	Responsible for development of new data warehouse increments as defined by the Data Warehouse Roadmap
Operations Team	Responsible for deployment of new increments and support of existing data warehouse systems

Organizationally, the Data Warehouse Competency Center can be depicted as:



Detailed roles and responsibilities for each team are described below:

Data Warehouse Steering Committee

The Data Warehouse Steering Committee is comprised of Cabinet level officers representing all divisions of the UC Berkeley organization. The responsibilities of the Steering Committee are:

- To provide the overall objectives for the data warehouse based on campus strategic goals
- Provide funding for data warehouse initiatives
- Periodically review performance of data warehouse initiatives to determine if campus goals are being met and to review funding and resource needs
- Provide guidance on regulatory requirements that impact the data warehouse

The Steering Committee is not involved in the day to day management of data warehouse projects; however the Data Warehouse Strategy Committee may escalate issues to the Steering Committee if in their judgment the issue warrants attention at the Cabinet level.

Data Warehouse Strategy Committee

The Data Warehouse Strategy Committee consists of senior-officers one level below Cabinet level as well as the Data Warehouse Program Manager. The members of the Strategy Committee represent all divisions of the UC Berkeley organization with an interest and need for reporting and analytics. The responsibilities of the Strategy Committee are:

Managing the Data Warehouse Roadmap

- Review recommendations for data warehouse initiatives from the Strategy Team
- Prioritize data warehouse initiatives (strategic, tactical and operational)
- Ensure data warehouse projects are aligned with the overall University's goals
- Ensure data warehouse systems are compliant with regulatory requirements

Project Initiation

- Approve projects and authorize use of funding
- Sponsor projects

Project Tracking

- Review project status, budget, issues and risks

Monitor Service Levels

- Approve service levels and review actual service levels achieved versus target

Communication

- Provide championship of data warehouse program
- Communication with other senior officers and management (of the data warehouse initiatives and expected benefits)
- Escalate issues to the Steering Committee if necessary

Data Warehouse Performance Assessment

- Review and approval of critical processes use to manage data warehouse initiatives
- Assessment of data warehouse team(s) skills and capabilities
- Assessment of data warehouse program metrics

Data Stewardship

- Review and approve data retention and security policies
- Review and approve data standardization, definitions and naming conventions
- Define data quality standards

Data Warehouse Program Manager

The Data Warehouse Program Manager is a full time position dedicated to coordinating and managing all activities related to data warehouse initiatives. The Data Warehouse Program Manager is a member of the Strategy Committee and directly oversees the work of the Data Warehouse Strategy Team. The responsibilities of the Data Warehouse Program Manager are:

Facilitation of Steering and Strategy Committees

- Prepare and maintain charter for the Steering and Strategy Committees
- Prepare agenda for Steering & Strategy Committee meetings and distribute meeting minutes
- Work with Data Warehouse Strategy Team on presentations for Committee meetings

Manage Data Warehouse Capability Maturity Model

- Conduct and maintain assessment of current skills
- Review and approve DW processes
- Conduct and maintain assessment of data warehouse processes
- Assess current capabilities against established industry best practices
- Define target state and create gap analysis
- Develop training plan and initiatives to reach target state

Alignment with other IS&T Initiatives

- Coordinate and communicate with IS&T Program Office

Program Management

- Prepare and maintain charter for the Data Warehouse Competency Center
- Define data warehouse objectives and metrics
- Define the methodology for prioritization of data warehouse initiatives
- Review and approve the project management methodology
- Provide project status reporting
- Budget management
- Ensure best practices are followed for design and development of Data Warehouse projects

Data Warehouse Strategy Team

The Data Warehouse Strategy Team consists of Business and Data Analysts, End-user Support Analysts, the Data Warehouse Data Architect and the Data Warehouse Technical Architect. The team is responsible for defining requirements, high level design, project management and end-user support and training for all data warehouse initiatives. The Strategy Team works closely with a team of Subject Matter Experts – representatives from the campus or administrative departments who help define and prioritize requirements. Specific responsibilities of the Strategy Team are:

Creating / Maintaining the Data Warehouse Roadmap

- Identification of data warehouse opportunities
- Develop business cases and cost benefit analysis
- Providing recommendations to the Strategy Committee for strategic, tactical and operational projects

Project Initiation

- Requirements gathering
- Business analysis
- Develop project plan

High Level Design

- Solutions architecture
- Service level agreements
- Conceptual and logical data models
- Source data mapping

Project Management

- Status reporting
- Issues management
- Risk management

Technology Evaluation

- Evaluate data warehouse toolsets and technologies
- Recommend technologies that can add business value and enable achievement of the University's objectives

Data Quality Management

- Define data definitions / standards
- Data retention policy
- Security policy
- Data quality management

User Adoption and Training

- Business change management
- User adoption planning
- End-user support and training

Data Warehouse Development Team

The Data Warehouse Development Team is responsible for adding new functionality and subject areas to the data warehouse. Typically the data warehouse is developer as a series of increments defined by the Strategy Team in the Data Warehouse Roadmap. The Development Team does not provide the day to day operations for the Data Warehouse systems. Responsibilities of the Development Team are:

Detailed Design

- Extract, Transform and Load (ETL) functional specifications
- Physical database design
- Metadata package design
- Report / cube specifications
- Security
- Create requirements traceability matrix
- Create test plans

Development

- Create and maintain development standards
- ETL development
- Database setup / scripting
- Metadata development
- Report / cube development
- Portal customization
- Unit testing
- Documentation

Data Warehouse Operations Team

The Data Warehouse Operations Team is responsible for the day to day monitoring, maintenance and support of data warehouse systems. Clear separation of the duties of operations and development limits the impact of production management and issues on the development of new data warehouse increments. Specific responsibilities of the Operations Team are:

Validation

- QA / Integration testing
- User acceptance testing
- Turnover / knowledge transfer to Operations Team
- Turnover / knowledge transfer to Strategy Team

Deployment

- Migrate new increments to the production environment
- Implement security policy
- Source code management
- Create operations documentation
- Create end-user training tools

Operations

- Scheduling
- Data validation
- Backups
- Application logging (front-end and back-end)
- Management of production issues
- Root cause analysis

Performance Management

- Ongoing monitoring
- Reporting on service levels
- Tuning and configuration

Data Quality Management

- Implement data quality retention and security policies
- Measurement of data for accuracy, reliability and consistency

Enhancements

- Front end production enhancements (presentation layer, reports, OLAP)
- Back end production maintenance (ETL, databases, data warehouse, data marts)

Infrastructure Management

- Maintenance of Dev / QA / Prod environments
- Installation of data warehouse tools and software
- Disaster recovery procedures

Capacity Management

- Ongoing monitoring
- Capacity planning

End-user Support

- Technical help desk
- Add / change / delete users
- Technical end-user training

Processes

The Data Warehouse Competency Center requires a standard set of processes to:

- Ensure Data Warehouse systems are initiated, implemented and supported in a consistent manner
- Measure performance of the Data Warehouse Capability Center
- Manage data ownership and stewardship
- Manage the skills and processes needed for continuous improvement

The core processes required by the Data Warehouse Competency Center are:

1) Project Initiation
2) High Level Design
3) Development
4) Validation
5) Deployment

6) User Adoption
7) End User Support
8) Data Warehouse Program Management
9) Data Warehouse Performance Assessment
10) Data Stewardship

Responsibilities for ownership and participation in these processes are defined in the RACI matrix below, where:

R – Indicates the party is responsible for the process
A – Indicates the party is accountable for the process
C – Indicates the party is consulted on the process
I – Indicates the party is informed of the process

Data Warehouse Competency Center Process	Steering Committee	Strategy Committee	Program Manager	Strategy Team	Subject Matter Experts	Development Team	Operations Team
Create and Maintain the Data Warehouse Roadmap	C	A	R	R	C		
Project Initiation	I	C	A	R	C		
High Level Design			A	R		C	C
Development		I	A	C		R	I
Validation		I	A	C	C	R	
Deployment		I	A	C		C	R
User Adoption		C	A	R	C		C
End User Support			A	R		C	R
Data Warehouse Program Management	I	A	R	C			
Data Warehouse Performance Assessment	C	A	R	C			
Data Stewardship	I	A	R	R			

A detailed description of these processes is provided in the Support Architecture document cited above.

Summary of differences from current architecture

- An enterprise data warehouse needs processes and supporting organizations to coordinate evolution of the warehouse with overall campus priorities and needs. The addition of those processes and organizations represent the biggest difference between the support architecture just described and processes currently in place.
- In particular, the two high-level committees which provide funding and objectives and prioritize development work are important components not yet in place on the campus.
- Having a single program manager responsible for the overall management and coordination of the DW competency center on behalf of the whole campus is another important change.
- Also new is having subject-matter experts from the various academic and administrative functions formally engaged in clarifying requirements and helping to ensure data quality.
- Finally, the data warehouse competency center described above is a single virtual team comprised of personnel with functional and IT specialties. This represents an evolution from current practice, as does distinguishing development and operations teams.

Appendices

UC-Berkeley Fact/Dimension Data Matrix

Architecture Requirements Gap Analysis

EDW: Key data marts by subject area and their shared dimensions

Dimensions

Data Marts	Date (semester, census, add/drop, fiscal, effective, reported, etc.)	Facility	Major (all degree-granting programs and interrelatedness)	Course (all courses incl. all sections, offering dept, etc.)	Student (reg status, residency, major, demographics)	Applicant for admission	Financial Aid Program	External Institution (high schools, other higher ed, SAT test,	Research Sponsor (govt agency, foundation, corporation, etc.)	Employee (descriptive info fac, staff, affiliates, etc.)	Job Applicant	Budgeted Staff Position (budget dimension for perm positions)	Appointment (title, job code, leave status, etc.)	Job Classification	Job Requisition	Employee Training (all campus provided training)	Supporter (person, organization, actual or potential, alumni)	Campaign (fund raising programs)	Fund	Organization (department, unit, student groups, etc.)	COA-Business unit	COA-Account	COA-Program	COA-Project	COA-Flexfield	Vendor (descriptive info of individuals & orgs)	Product/Commodity	Service Element (services by recharge units, e.g. tel, data, etc.)	
Teaching																													
Teaching assignment	x			x						x																			
Classroom space assignment	x	x		x																									
Student class enrollment	x			x	x																								
Wait list enrollment	x			x	x																								
Degree	x		x		x																								
Student course evaluation	x			x	x																								
Student survey	x				x																								
Student																													
Undergraduate admissions	x				x	x																							
Graduate admissions	x				x	x																							
Semester snapshot	x		x		x																								
Financial aid awards	x				x		x												x										
External test results	x					x		x																					

Data Marts	Date (semester, census, add/drop, fiscal, effective, reported, etc.)	Facility	Major (all degree-granting programs and interrelatedness)	Course (all courses incl. all sections, offering dept, etc.)	Student (reg status, residency, major, demographics)	Applicant for admission	Financial Aid Program	External Institution (high schools, other higher ed, SAT test,	Research Sponsor (govt agency, foundation, corporation, etc.)	Employee (descriptive info fac, staff, affiliates, etc.)	Job Applicant	Budgeted Staff Position (budget dimension for perm positions)	Appointment (title, job code, leave status, etc.)	Job Classification	Job Requisition	Employee Training (all campus provided training)	Supporter (person, organization, actual or potential, alumni)	Campaign (fund raising programs)	Fund	Organization (department, unit, student groups, etc.)	COA-Business unit	COA-Account	COA-Program	COA-Project	COA-Flexfield	Vendor (descriptive info of individuals & orgs)	Product/Commodity	Service Element (services by recharge units, e.g. tel, data, etc.)	
Activity membership	x				x															x									
Sponsored Research																													
Grant proposal	x								x	x										x									
Grant award	x								x	x									x	x									
Gifts and Fundraising																													
Contacts	x									x							x	x		x									
External events	x																x	x		x									
Memberships	x																x			x									
Pledges	x																x	x	x	x									
Gift receipt	x																x	x	x	x									
Human Resources																													
Recruitment	x									x	x				x					x									
Personnel actions	x									x			x							x									
Training Activity	x									x						x													
Spending																													

Data Marts	Date (semester, census, add/drop, fiscal, effective, reported, etc.)	Facility	Major (all degree-granting programs and interrelatedness)	Course (all courses incl. all sections, offering dept, etc.)	Student (reg status, residency, major, demographics)	Applicant for admission	Financial Aid Program	External Institution (high schools, other higher ed, SAT test,	Research Sponsor (govt agency, foundation, corporation, etc.)	Employee (descriptive info fac, staff, affiliates, etc.)	Job Applicant	Budgeted Staff Position (budget dimension for perm positions)	Appointment (title, job code, leave status, etc.)	Job Classification	Job Requisition	Employee Training (all campus provided training)	Supporter (person, organization, actual or potential, alumni)	Campaign (fund raising programs)	Fund	Organization (department, unit, student groups, etc.)	COA-Business unit	COA-Account	COA-Program	COA-Project	COA-Flexfield	Vendor (descriptive info of individuals & orgs)	Product/Commodity	Service Element (services by recharge units, e.g. tel, data, etc.)	
Payroll	x				x					x			x						x	x	x	x	x	x	x				
Purchasing	x									x									x	x	x	x	x	x	x	x	x		
Budget and Finance																													
Permanent budget	x											x							x	x	x	x	x						
(Temporary) budget	x																		x	x	x	x	x						
Balances	x																		x	x	x	x	x	x	x				
Pre-encumbrances	x																		x	x	x	x	x	x	x				
Encumbrances	x																		x	x	x	x	x	x	x				
Revenue/expense journals	x																		x	x	x	x	x	x	x				
Recharge Billing																													
Communications services	x																		x	x	x	x	x	x	x			x	
IST services	x																		x	x	x	x	x	x	x			x	
Physical plant	x																		x	x	x	x	x	x	x			x	
Other recharge areas																												x	
Facilities																													
Space assignment	x	x																		x									

Data Marts	Date (semester, census, add/drop, fiscal, effective, reported, etc.)	Facility	Major (all degree-granting programs and interrelatedness)	Course (all courses incl. all sections, offering dept, etc.)	Student (reg status, residency, major, demographics)	Applicant for admission	Financial Aid Program	External Institution (high schools, other higher ed, SAT test,	Research Sponsor (govt agency, foundation, corporation, etc.)	Employee (descriptive info fac, staff, affiliates, etc.)	Job Applicant	Budgeted Staff Position (budget dimension for perm positions)	Appointment (title, job code, leave status, etc.)	Job Classification	Job Requisition	Employee Training (all campus provided training)	Supporter (person, organization, actual or potential, alumni)	Campaign (fund raising programs)	Fund	Organization (department, unit, student groups, etc.)	COA-Business unit	COA-Account	COA-Program	COA-Project	COA-Flexfield	Vendor (descriptive info of individuals & orgs)	Product/Commodity	Service Element (services by recharge units, e.g. tel, data, etc.)	
Payments																												
Vendor payments	x																		x	x	x	x	x	x	x	x		
Financial aid payments	x				x														x	x	x	x	x	x	x			
Receipts																												
Student fee payments	x				x																							

Description of dimensions:

Following are definitions of the dimensions listed in the matrix. (in the following a **dimension** name is in boldface; related ***sub-dimensions*** in italic).

- **Date.**

  The date dimension provides a coordinated calendar for the university, relating individual dates to:
  - the academic calendar, including semester and summer session start and end, census dates, etc.
  - the financial calendar, including fiscal periods, processing cut-off dates, etc.

  A related sub-dimension is the set of possible class meeting schedules, ***course meeting schedule***.

- **Facility.**

  Descriptive information about rooms and buildings.

- **Major (degree program).**

  Describes the degree-granting programs of the campus, along with their inter-relationships. Degree programs are linked to the organizations which offer them.

- **Course.**

  This dimension describes all the courses offered by the campus as well as their individual sections, along with all of the relationships among them. Course offerings are linked to the organizations which offer them and to the major courses of study to which they contribute.

- **Student**

  Describes demographic and academic characteristics of anyone who has ever been admitted to the university. An important set of descriptive attributes of students are academic status information, such as registration status, residency, degree goals and major, together form a subdimension with information which is refreshed once or more during an academic term for continuing students. This more detailed subdimension is identified as:

  - ***Student academic status*.** (This is largely the information collected under the title "semester event" in the student data warehouse pilot).

- **Applicant for admission.**

  Demographic, credential and status information about people who have applied for admission to the university, either as undergraduates or graduates.

- **Financial aid program.**

  Information about scholarship and fellowship programs, undergraduate and graduate, including external programs such as Fulbright scholarships.

- **External institution.**

  Information about all the secondary schools and other higher educations which interact with the campus, including all high schools and junior colleges which prepare undergraduate applicants as well as other higher-education institutions. Closely related is the subdimension.

  - ***Testing agency*.** Identification and other information about organizations who furnish standardized test results, such as SAT scores, to the campus.

- **Research sponsor.**

  Descriptive information about organizations such as government agencies, foundations and corporations, which sponsor research.

- **Employee**

  Identifying and descriptive information, such as demographic information, about present and past employees of the campus, including faculty, staff and all manner of affiliates, such as volunteer contributors.

- **Appointment.**

  Facts about an employee job, including such information as title, job code, reporting relationship, leave status. Because it is common for employees, particularly faculty, to have several appointments, some key information about employment status, salary grade, etc., is only accurate by appointment. The appointment dimension obviously needs to be coordinated with the employee dimension, as changes to the overall employment status of an employee cascade to that employee's various appointments. Appointments for permanently-budgeted positions are linked to the budget dimension "**Budgeted staff position**".

- **Budgeted Staff Position.**

  Information about permanently budgeted staff positions.

- **Job Applicant.**

  Demographic and other information about persons who have applied for employment on the campus.

- **Job Classification.**

  Information about defined jobs, including relationships among them, such as job families and progression ladders.

- **Job Requisition.**

  Information about individual job requisitions including coded skills requirements, dates, etc.

- **Employee training.**

  Information about campus training programs, including specific training offerings, related programs and training hierarchies.

- **Supporter.**

  Descriptive information about donors and potential donors (prospect) to the campus, including identifying and demographic information, and fund-raising attributes. Closely related to this dimension is the following subdimension:

  - ***Supporter student interest***

    Describes the relationship of a supporter to an individual student; e.g., parent of the student, along with other related descriptive information.

- **Campaign.**

  Descriptive information about organized fund-raising programs.

- **Fund.**

  Information about sources of funding for the campus, including BFS funds and fund hierarchies. Includes linkages when appropriate to Sponsors and Supporters.

- **Organization.**

  Information about campus organizations. Includes all internal organizations, including cross-functional organizations, and their interrelationships and reporting hierarchies.

- **COA-Account.**

  Information about GL accounts and account hierarchies, including information needed to reconcile BFS accounts with UCOP accounts.

- **COA-Program.**

  Identified programs from BFS, along with any hierarchies or inter-program relationships.

- **COA-Project.**

  Information about projects used to identify financial information, including descriptions and relationships among those projects.

- **COA-Flexfield.**

  Any descriptive information about flexfields used to classify financial information. This is probably a slight dimension, since information about flexfields is local to individual organizations and may not be well documented/

- **Vendor.**

  Descriptive information about parties—organizations and individuals—paid through the campus payments system, including hierarchical relationships among them. Many employees are vendors.

- **Product/commodity.**

  Description of the products and services bought by the university through the purchasing system.

- **Service element.**

  This is a place holder for a family of dimensions describing the service elements for which recharge organizations charge, such as telecommunications services. There is one or more such dimension for each recharge organization and no expectation that they are aggregated into a single linked dimension.

Notes on some highly-shared dimensions:Integration notes and linkages among dimensions:

1. Person.

   An identified member of the campus may be a student, an employee or an affiliate or all of those. Every such person can be identified with a unique University ID (UID), which serves to relate the student and employee dimensions. There is an important need to link the student and employee dimensions in order to understand, for example, all the sources of support for a student. Though the UID may provide an umbrella identifier linking those dimensions, the completeness and accuracy of that linkage mechanism needs to be confirmed.
   Identifying and descriptive information, such as demographic information, about present and past employees of the campus, including faculty, staff and all manner of affiliates, such as volunteer contributors. There are two independent dimensions which are very closely linked to the employee dimension. They are the following:

   - Appointment.

   Facts about an employe job, including such information as title, job code, reporting relationship, leave status. Because it is common for employees, particularly faculty, to have several appointments, some key information about employment status, salary grade, etc., is only accurate by appointment. The appointment dimension obviously needs to be coordinated with the employee dimension, as changes to the overall employment status of an employee cascade to that employee's various appointments.
   Appointments for permanently-budgeted positions are linked to the budget dimension "budgeted staff position".

   - Academic interests and expertise.

   The following dimensions describe persons who do not necessarily have UIDs, student id or employee ids.

   - Job applicant
   - Applicant for admission.
   - Supporter.
     Describes a party (person or organization) identified as an actual or potential supporter of the campus, including alumni. Most supporters are persons but supporters which are organizations represent an important source of donations. Supporters often are related to students; in the most frequent case by having been Cal students.

2. Campus organization.

   Includes all internal organizations, including cross-functional organizations, and their interrelationships and reporting hierarchies. Several degree programs and majors represent joint offerings of several academic organizations and are not themselves organizations.

3. External organization / party.
   A group of dimensions describe external parties, mostly organizations. Though there is potential for some organizations to be included in more than one of these; e.g., some organizations might be both Sponsors and Vendors, no process has been identified to establish that linkage.

   - Sponsor.
   - Secondary school.
   - Higher education institution
   - Standardized testing service (e.g., SAT)
   - Vendor

4. Date.

   The date dimension provides a coordinated calendar for the university, relating individual dates to:
   - the academic calendar, including semester and summer session start and end, census dates, etc.
   - the financial calendar, including fiscal periods, processing cut-off dates, etc.

   Because of university practices of making some important changes retroactively and because of requirements for reporting to UCOP and some external stakeholders according to census dates, many data marts have two links to the date dimension:
   - effective date (the date as of which information is considered to be final)
   - reported date (the date information was available to be reported externally)

   In addition, many time-varying dimensions, such as student and employee, will have different values for different combinations of effective and reported date.

   For example, consider the following example of course enrollments:

   In Fall 2005 there is a course and section AA11. At the beginning of the term, September 7, 2005, it has two student enrollments, for these two students:
   - student 123 who is a resident graduate student
   - student 456, who is a non-resident graduate student.

   On October 1, student 456 is granted resident status, retroactive to the beginning of the semester.
   On November 1, student 456 drops the class.

A count of course enrollments by residency as reported on date n should return three different values for n = 9/10, 10/10 and 11/10, as follows:

Date reported	Resident enrollments	Non-resident enrollments
9/10	1	1
10/10	2	0
11/10	1	0
Today	1	0

5. Fund.

   One of the most-highly shared dimensions is Fund, which serves to associate activity, such as research spending, employment, other spending, and financial aid payments, with funding sources and fund-producing activities. For example, funds associate research spending with sponsors, research proposals and grants. Likewise funds associate supporters and gifts with the financial aid payments and other spending which represent the use of grants.

UC-Berkeley Data Warehouse Roadmap

Summary of Gaps

Existing Data Warehouse Environment vs. Architecture Requirements

As illustrated in the following diagram, there are many components that work together to deliver on the data warehouse's mission of delivering accurate, understandable information to support decision-making. Certainly data is the fundamental component: cleaned, organized data, mostly extracted from the campus' operational systems. Making that data useful to a variety of campus personnel, though, requires some applications to deliver and explain it. These applications range from predefined reports through query tools to complex tools for analysis and modeling. Delivering data and applications and securing the data as specified by campus data stewards requires a set of technology, most of it centralized in secure computer locations. Equally important, transforming operational data into a shared resource useful across the boundaries of functional business domains requires a broad set of functional skills, organized appropriately and working through proven processes.

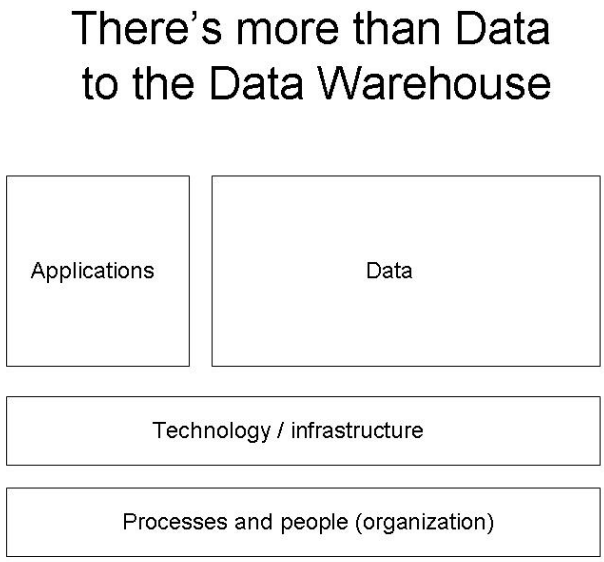


Figure 1: Components of the data warehouse

The architecture effort examined these four important components of the data warehouse:

1. Applications
2. Data
3. Technology and security
4. Support—processes and organization

Scope of the Architecture:

The UC-Berkeley Enterprise Data Warehouse (EDW) is intended to meet the consolidated analysis and reporting needs of the campus. The EDW will evolve over time. However, the capabilities needed by the EDW in the practical planning horizon were defined in a requirements process conducted in late summer and fall, 2005. The requirements established in that process are summarized in the EDW requirements document, which is available at the following link:

https://bearshare.berkeley.edu/sites/RAPO/EDW/DWStrat/Project documents/EDW Phase I FINAL Report1 12-16-05.doc

The architecture of the enterprise data warehouse is designed to deliver the analysis capabilities defined in that document and likewise to provide the critical success elements defined there. In addition, the EDW is expected to incorporate over time the capabilities of the existing campus reporting systems. Those decision-support systems include the following:

- BAIRS
- BIS
- Cal Profiles
- The pilot student data warehouse
- FASDI
- The Office of Student Research database/ reporting file

Architecture Requirements:

Starting with the functional requirements articulated in the campus interview process and recorded in the requirements document referenced above, the architecture team analyzed resulting architectural requirements in the four dimensions of applications, data, security/technology and process/organization. The resulting architecture requirements are summarized in the EDW Architecture Requirements documents, which are available at the following link:

https://bearshare.berkeley.edu/sites/RAPO/EDW/DWStrat/Architecture/Document Library1/Forms/AllItems.aspx

Based on those requirements, the team analyzed existing warehouse capabilities, with particular emphasis on BAIRS, which is the platform which will serve as the basic foundation from which to build the EDW. Requirement gaps were identified and described for each of the components of the EDW; they are documented in individual documents at the architecture website:

https://bearshare.berkeley.edu/sites/RAPO/EDW/DWStrat/Architecture/Document Library1/Forms/AllItems.aspx

Following is a summary of the gaps.

Architecture Gaps:

1. Applications gaps.

   - Better support for query in addition to reporting.

   The current reporting systems, including BAIRS, emphasize delivery of information in the form of predefined reports, many of which provide facilities for filtering the report to include only specified information. However, for many users and many purposes directly exploring the data warehouse is more valuable. This is particularly true for analytical uses of the information, such as identifying patterns or opportunities. Many of the questions managers or executives ask require this kind of analysis. For example, a department chairman evaluating course plans might want to look at background and outcomes for a cohort of students who took a specified set of courses. A lot of the value of a data warehouse derives from its ability to answer such analytical questions.

   The data in the data warehouse and the basic access toolset (currently Hyperion Intelligence) provide the basic capabilities needed for query and analysis. However, many users don't understand the data in the warehouse well enough to explore it confidently and accurately. They need some organized aids for understanding, including:

   o Improved data documentation (Metadata).

     The data in the warehouse should be described in clear business language. This certainly includes definitions of individual data elements, but also includes explanations of the way data is organized and combined. Data is organized as facts measured within a context described by well-organized dimensions. The fact-dimension structures should be clearly described, data mart by data mart so that users can confidently select and combine the information they need. Documentation should include description of the subject areas in the warehouse, as well as information about how warehouse data is derived from source systems and about how recently the data was extracted.

     To be most useful, data documentation should be delivered directly through the query tool.

     Data documentation must be developed by personnel very knowledgeable about the data. Subsequently, the warehouse support team must work with query users to ensure the documentation is adequate and to improve it as necessary and kept it current.

   o Organized view-layers appropriate to groups of users.

     Many users need and understand only a part of the information in the warehouse. Those users should see a simplified representation of the warehouse, in which data have been filtered, organized and pre-joined as required to present the portion of the data those users need. Various groups of users, such as academic advisers, will need views appropriate to them. Part of the process of bringing the warehouse to a new functional area

consists of identifying such user groups and building view-layers appropriate for them.

- Pre-developed navigation and drill-down.

  Joins between facts and dimensions can often be implemented in advance, as described above; in other cases, the join mechanisms can be made crystal-clear so that users can combine data confidently. Also, there are many hierarchies contained in the dimensions organizing the warehouse, such as organizational hierarchies, fund trees, and hierarchical course offerings. These hierarchies can be made explicit through the query tool so that users find it easy to drill-down into details or to aggregate into categories. This kind of navigation can be provided through existing query tools. In addition, some specialized data base technology (described under Technology gaps, below) can simplify presentation of data as multi-dimensional structures, often referred to as cubes. This technology, OLAP, is particularly useful for supporting modeling, planning and forecasting applications.

- Applications support for modeling, planning and forecasting.

  There are specific application tools which help budget makers, contract negotiators and other analysts to use data in the warehouse to build models of future activity based on history. These include planning and budgeting applications as well as data mining and general-purpose modeling tools. These are needed in the near term by Budget Office and HR personnel and over time by many others.

  There is a strong need for tools which allow users, including principal investigators and financial analysts, to enter projections of labor costs and other financial transactions and then to combine these projected costs with actual costs and revenues as they occur in order to track status of budgets. This is often needed, for example, in order to understand status of budgets for multi-year research grants.

- Standard reporting and query support for subject areas other than G/L, budgeting, human resources.

  The data warehouse will grow by adding new data subject areas, such as teaching and course enrollments, admissions and student registration, donors and gifts, purchases and recharges. Along with each new data increment, the warehouse should add new standard reports and new facilities, such as view layers and drill-downs, to make that data useful.

- Proactive business intelligence (dashboards, alerting, etc.).

  Many campus personnel use standard reports to help identify exception conditions requiring action. It's increasingly common for data warehouses to automate this process. Users specify measurements, frequencies and thresholds; the warehouse infrastructure periodically tests data and alerts users when values pass thresholds. Alerts are delivered as notifications on the report portal or as emails.

  In similar fashion, users can define key performance indicators which indicate successful process execution. The data warehouse tests the indicators and displays the status of those KPIs as graphical dashboards either in the reporting

portal or in a user-specific portlet within a general portal such as the campus' business portal, blu.

- Custom BI web services.

  Not everyone will use the data warehouse through standard reporting and query tools. In some cases, custom applications will use information from the warehouse and present it in context to users trying to make decisions. For example, students in the process of enrolling for classes might be furnished information about how other students with similar degree goals fared with alternative courses. This would be something like the way Amazon helps customers see how other customers evaluated similar choices. Providing such services will require providing web services interfaces to the data warehouse.

- Reporting refinements.

  Some users need better facilities for scheduling routine reports. Some need improvements in the way the reporting tool prints reports.

2. Data gaps.

- Missing data subject areas.

  The data required to meet the requirements of the scope identified for the data warehouse has been summarized in the outline data architecture, also known as the data bus or fact/dimension matrix. (See *https://bearshare.berkeley.edu/sites/RAPO/EDW/DWStrat/Architecture/Document Library1/1/consolidated data matrix for review 3.1.doc).*

  Subject areas identified there include Teaching, Student, Sponsored Research, Gifts and Fundraising, Human Resources, Facilities, Budget and Finance, Spending, Payments, Receipts and Recharge. Of those, only Budget and Finance and some Human Resources data are delivered on a platform similar to the target EDW platform. The remaining subject areas need to be added to the data warehouse incrementally in several stages.

- Conformed dimensions (as summarized in data matrix).

  The key mechanism which allows data in the warehouse to be integrated across subject areas, as well as combined and aggregated unambiguously, is the use of shared, "conformed" data dimensions.

  Data in the warehouse is delivered as discrete facts, usually measurements, linked to dimensions which describe the context in which those facts were recorded. To ensure that facts can be summarized consistently and combined with other facts, it is important that the dimension data used to organize the facts be consistently defined and delivered as a set of shared, "conformed dimensions".

  Building and maintaining the conformed dimensions requires reaching agreement among functional proprietors and users of the data. The work required to do this pays off in improved integration and avoidance of "dueling data".

- Improved implementation of dimensional architecture (BAIRS).

  Though BAIRS and BIS systems are structured as facts and dimensions as required, these systems require some rework to deliver the flexibility and

extensibility provided by the fact/dimension architecture. Specifically, existing dimensions need to be conformed for sharing and in several cases redesigned to deliver more straightforwardly the hierarchies which are natural to some dimensions, such as organization or time. Additionally, the linkage between fact and dimension tables needs to be reworked. For reliability and flexibility, links between facts and dimensions should be specific warehouse keys, or surrogate keys.

- Consolidation of BAIRS and BIS.

  There is a great deal of duplication and overlap between BAIRS and BIS. These two data collections need to be rationalized so that they share common, conformed dimensions and so that redundant fact tables are eliminated.

- Migration of purchasing data to an independent data mart.

  In the current BAIRS system, purchase order data is recorded as details accompanying general ledger transactions. To be more generally useful and to meet the requirements documented during the business requirements process, purchasing data should be migrated to an independent purchasing data mart.

- Incorporation of plan data (e.g. salary and budget projections) into the data warehouse.

  Budget tracking is one of the most widely-used applications of the data warehouse. Reporting current status against budget requires summarizing actual expenses along with known commitments and spending plans. The current budget warehouse, BAIRS, contains some but not all of the information about known future expenses. BAIRS budget reports include purchasing commitments (requisitions and purchases) but don't include other projected expenses. Users would like to see foreseeable expenses, such as large software license costs or travel expenses, along with other commitments and actual expenses in order to know where they stand with budget.

  This is a particularly important need in tracking budgets for research grants, which often have multi-year budgets. In particular, it's important to include projected salary costs along with actual expenses and commitments to understand current budget status of research grants.

  Adding projections and salary encumbrances to the data warehouse depends on first creating an application which allows users to record them. This will probably be most easily developed as a custom function within the Berkeley Financial System (BFS).

- Improved business-facing metadata.

  As indicated above (see Applications gaps), more extensive documentation of warehouse data is needed.

- Improved technical metadata.

  The technical personnel who maintain the data warehouse require good, consistent documentation of the processes by which information in the warehouse is sourced from operational systems. Today, documentation of the mappings of individual data elements is good; however, documentation of higher-level mapping is needed, including the following:

- Standard documentation of ETL designs and processes.

  The end-to-end design for populating a data subject area needs to be clearly expressed at a summary level.

- Improved documentation of shared dimension ETL process.

  Data in shared dimensions is created once and then linked to several data subject areas. Although this process creates powerful integration opportunities, it is also rather complex. Dimension changes often cascade across data subject areas, with implications for testing. For this reason, dimension-maintenance software needs clear and careful documentation.

3. Technology / Infrastructure / security gaps.

   In general, the technology platform supporting the existing BAIRS system is current and highly capable. Likewise, there is a well-developed security architecture controlling access to information in accordance with business rules. However, both the security system and the technology platform will require enhancements to meet the needs of the Enterprise Data Warehouse.

   - Security enhancements.

     New data subject areas, such as student and employee data, have somewhat more complex requirements for controlling access. In particular, these subject areas contain much sensitive personally-identifiable data, highly regulated by government statutes and University policy. Meeting the requirements of these subject areas will require an expanded access-control system.

     - Avoid exposing personal information by using surrogate keys.

       Much of the information about students and employees becomes highly sensitive when associated with names, student or employee ids, social security numbers and other identifiers. In many cases, all these identifiers can be entirely masked without reducing the power of the data warehouse to deliver useful information about student and employee behavior, distribution by demographics, etc. The enabling mechanism is the use of substitute identifiers known as surrogate keys. Using surrogate keys provides many other benefits, including ease in tracking variation over time in information about students, employees, etc. The EDW should use surrogate keys in most cases.

     - Support for filtering of data to exclude small sample sizes.

       In some situations, identifying a small cohort of people by shared characteristics such as ethnicity may be tantamount to identifying them directly; e.g., a list of grades by gender within ethnicity may be quite allowable for a college but inappropriate for a class with only one Asian woman. Some mechanism is needed to filter query results and substitute a “N/A” value when resulting numbers fall below appropriate thresholds.

     - Support for context-based authorization.

       In the existing security mechanism, access is controlled principally through organizational affiliation; i.e. users are granted access to a slice

of information which is associated with some level in the campus organization.

As more data about students, employees and grants is added to the warehouse, there will need to be a mechanism for restricting access in terms of a context defined by some principle other than organization; e.g., an adviser may need access to complete student records for all students he advises; a grant administrator may need access to salary information about all the personnel working on the grants the administrator is administering.

- OLAP functionality.

  As mentioned in Application gaps, above, some users who need to analyze large data sets rapidly along various dimensions will benefit from using a special OLAP database management product. Users such as budget analysts and labor negotiators have these needs. In addition, many such users need to be able to build models created by manipulating data from the warehouse. These needs, too, are met by OLAP technology. Some key requirements include the following:

  - Support for presentation of data as cubes.
  - Support for multi-dimensional analysis and MDX.
  - Support for “writeback”—user updating and saving of cubes. This enables model-building.

4. Support/Process/Organization gaps.

   Some of the biggest, most important gaps between the EDW specified by campus interviews and the existing data warehouse involve organization, skills and processes. Most interview subjects commented on the need for changes in this dimension of the warehouse. The EDW serves the whole campus—academic units and administration—and delivers information drawn from across the activities of the campus. As a result, it needs cross-campus leadership, processes which effectively draw out the collective institutional knowledge of the campus, and a skilled, professional warehouse team. Following are some specific gaps:

   - EDW governance.

     The EDW is a multi-year investment which needs to be directed and paced by overall campus priorities. It is important that the executive leadership of the campus, including both academic and administrative leadership, participate in setting priorities and success criteria for the effort. The same leaders must ensure that the warehouse effort has a budget appropriate to those priorities. A high-level executive steering committee is required, as well as a delegated executive owner. This steering group must charter the work of the rest of the EDW team, ensure that the team is competent to meet that mission, and appropriate the needed funds.

     In addition, the executive steering group needs to ensure that the whole campus is participating appropriately in evolving the warehouse, either by participating in the operational management of the warehouse effort or by supplying data, and data knowledge and quality oversight.

   - BI Competency Center.

     The data warehouse is a campus information tool with many components. Delivering it requires a blend of campus business knowledge, management and technical skills. These skills are most effective when collected together and managed as a single organization. This collection of skills is often called a Business Intelligence Competency Center. What's most important is that all the required skills are coordinated as a single set and harnessed to help the whole spectrum of campus users get their information needs met. Having these skills operate as a single organization helps ensure that the process is needs-driven; it also promotes uncovering options for meeting those needs.

     An effective competency center should include the following components:

     - Management.
     - Support for existing DW.

       This includes both user support and technical support. Support personnel maximize effectiveness of the existing warehouse and make required enhancements to existing components.

     - Architecture and release planning.

       Experienced personnel knowledgeable about overall campus needs and available data maintain a migration plan, composed of practical implements, for implementing the overall data warehouse.

     - Development.

This includes the skills and personnel needed to design and implement new releases, including functional requirements definition, technical development and change management.

- Create a panel of functional data experts.

  Delivering shared data useful across functions is only possible if functional personnel knowledgeable about the data collaborate in defining the required data and establishing associated business rules for using and combining it. This requires a delegated group of functional data experts who work together on the business-level specifications of the warehouse. The same personnel are needed to clarify security and access rules.

  Adding new data to the warehouse usually drives out substantial issues with the quality of existing data. These functional experts need to help evaluate these quality issues and what to do about them. Often they will require changes to existing business practices, processes or systems.

  The existing Data Stewardship Council might be a suitable group to identify and delegate this panel of functional experts.

- Implement a release process.

  The EDW is too large to be developed in a single monolithic project. Instead, it needs to be built over several years in a set of projects which add new capabilities as dictated by campus priorities and resources. Adding new capability in increments requires a formal release process which scopes and manages development of new capabilities without interference to the smooth operation of existing capabilities. This process is quite different from the processes which make minor enhancements to existing functionality. In particular, it incorporates two new functions:

  - Architecture.

    Planning a roadmap of projects which will make the required changes to data, applications, security and processes in order to deliver increments of value requires architectural understanding and design of the whole program.

  - Release planning.

    Implementing a new release requires change management which begins with identifying new user populations and their needs and progresses through measurement of the value of new capabilities. In addition, substantial project management is required to coordinate the change management along with the development of technical components, user training and development of associated documentation and metadata. This is a different process from the day-to-day management of the operating warehouse.

- Standard development process.

  Delivery of new increments requires a thought-out and repeatable process for coordinating all the activities and personnel involved in their development. For effectiveness and efficiency, this process should be formalized and well understood by participants.

- Document support/maintenance approach (making changes to existing DW).

  A corresponding process is needed for supporting and enhancing existing elements of the warehouse. As with development, support involves a variety of handoffs and reviews. When this process is formalized and well understood by participants, those personnel can deliver changes with more confidence, greater efficiency and fewer outages. Rudiments of this process have been worked out by the personnel who support the existing warehouse. The process needs to be codified, expanded where necessary and reinforced by operating management.