DATA WAREHOUSE CONCEPTS

A fundamental concept of a data warehouse is the distinction between **data** and **information**. **Data** is composed of observable and recordable facts that are often found in operational or transactional systems. At Rutgers, these systems include the registrar's data on students (widely known as the SRDB), human resource and payroll databases, course scheduling data, and data on financial aid. In a data warehouse environment, **data** only comes to have value to end-users when it is organized and presented as **information**. **Information** is an integrated collection of facts and is used as the basis for decision-making. For example, an academic unit needs to have diachronic information about its extent of instructional output of its different faculty members to gauge if it is becoming more or less reliant on part-time faculty.

DATA WAREHOUSE DEFINITIONS

The data warehouse is that portion of an overall Architected Data Environment that serves as the single integrated source of data for processing information. The data warehouse has specific characteristics that include the following:

Subject-Oriented: Information is presented according to specific subjects or areas of interest, not simply as computer files. Data is manipulated to provide information about a particular subject. For example, the SRDB is not simply made accessible to end-users, but is provided structure and organized according to the specific needs.

Integrated: A single source of information for and about understanding multiple areas of interest. The data warehouse provides one-stop shopping and contains information about a variety of subjects. Thus the OIRAP data warehouse has information on students, faculty and staff, instructional workload, and student outcomes.

Non-Volatile: Stable information that doesn't change each time an operational process is executed. Information is consistent regardless of when the warehouse is accessed.

Time-Variant: Containing a history of the subject, as well as current information. Historical information is an important component of a data warehouse.

Accessible: The primary purpose of a data warehouse is to provide readily accessible information to end-users.

Process-Oriented: It is important to view data warehousing as a process for delivery of information. The maintenance of a data warehouse is ongoing and iterative in nature.

Other Definitions

Data Warehouse: A data structure that is optimized for distribution. It collects and stores integrated sets of historical data from multiple operational systems and feeds them to one or more data marts. It may also provide end-user access to support enterprise views of data.

Data Mart: A data structure that is optimized for access. It is designed to facilitate end-user analysis of data. It typically supports a single, analytic application used by a distinct set of workers.

Staging Area: Any data store that is designed primarily to receive data into a warehousing environment.

Operational Data Store: A collection of data that addresses operational needs of various operational units. *It is not a component of a data warehousing architecture, but a solution to operational needs.*

OLAP (On-Line Analytical Processing): A method by which multidimensional analysis occurs.

Multidimensional Analysis: The ability to manipulate information by a variety of relevant categories or "dimensions" to facilitate analysis and understanding of the underlying data. It is also sometimes referred to as "drilling-down", "drilling-across" and "slicing and dicing"

Hypercube: A means of visually representing multidimensional data.

Star Schema: A means of aggregating data based on a set of known dimensions. It stores data multidimensionally in a two dimensional Relational Database Management System (RDBMS), such as Oracle.

Snowflake Schema: An extension of the star schema by means of applying additional dimensions to the dimensions of a star schema in a relational environment.

Multidimensional Database: Also known as MDDB or MDDBS. A class of proprietary, non-relational database management tools that store and manage data in a multidimensional manner, as opposed to the two dimensions associated with traditional relational database management systems.

OLAP Tools: A set of software products that attempt to facilitate multidimensional analysis. Can incorporate data acquisition, data access, data manipulation, or any combination thereof.

COMPARISON OF DATA WAREHOUSE AND OPERATIONAL DATA

HOW IS THE WAREHOUSE DIFFERENT?

The data warehouse is distinctly different from the operational data used and maintained by day-to-day operational systems. Data warehousing is not simply an “access wrapper” for operational data, where data is simply “dumped” into tables for direct access. Among the differences:

OPERATIONAL DATA	DW DATA
application oriented	subject oriented
detailed	summarized, otherwise refined
accurate, as of the moment of access	represents values over time, snapshots
serves the clerical community	serves the managerial community
can be updated	is not updated
run repetitively and nonreflectively	run heuristically
requirements for processing understood before initial development	requirements for processing not completely understood before development
compatible with the Software Development Life Cycle	completely different life cycle
performance sensitive (immediate response required when entering a transaction)	performance relaxed (immediacy not required)
accessed a unit at a time (limited number of data elements for a single record)	accessed a set at a time (many records of many data elements)
transaction driven	analysis driven
control of update a major concern in terms of ownership	control of update no issue
high availability	relaxed availability
managed in its entirety	managed by subsets
nonredundancy	redundancy is a fact of life
static structure; variable contents	flexible structure
small amount of data used in a process	large amount of data used in a process

The Data Warehousing Process – Part 1

Determine Informational Requirements

- Identify and analyze existing informational capabilities.
- Identify from key users the significant business questions and key metrics that the target user. group regards as their most important requirements for information.
- Decompose these metrics into their component parts with specific definitions.
- Map the component parts to the informational model and systems of record.

The Data Warehousing Process – Part 2

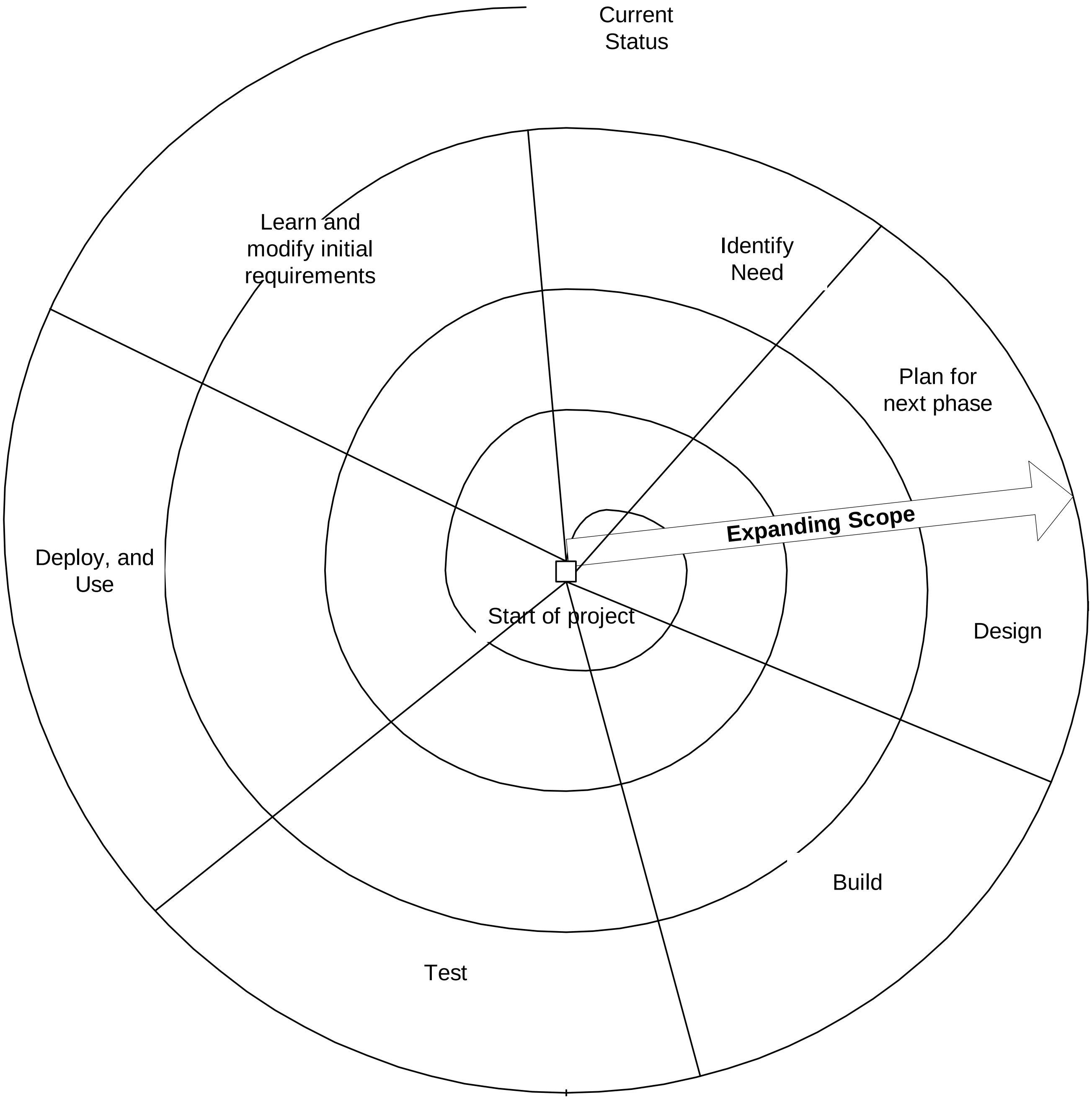
Evolutionary and Iterative Development Process

When you begin to develop your first data warehouse increment, the architecture is new and fresh. With the second and subsequent increments, the following is true:

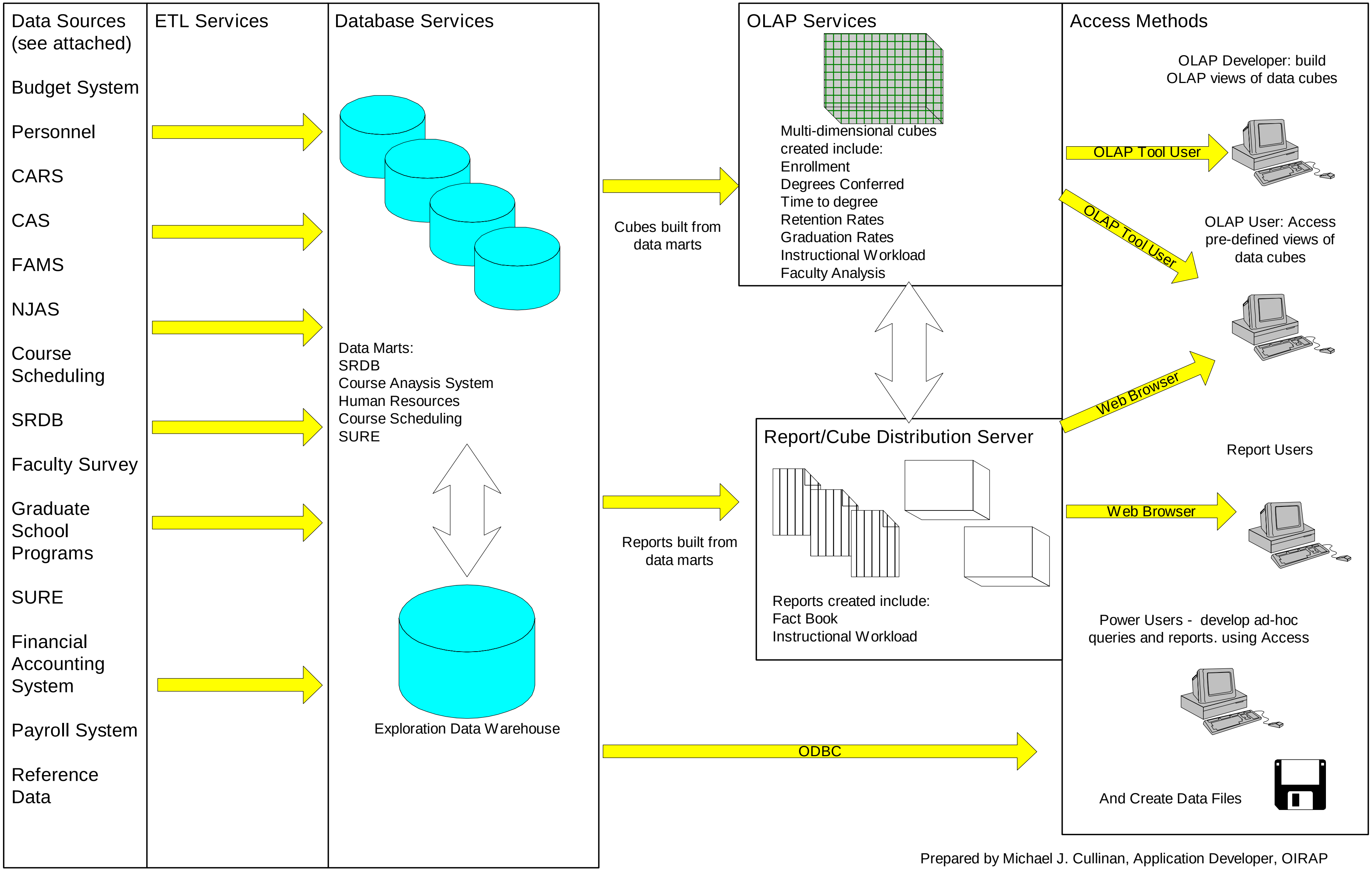
- Start with one subject area (or subset or superset) and one target user group.
- Continue and add subject areas, user groups and informational capabilities to the architecture based on the organization’s requirements for information, not technology.
- Improvements are made from what was learned from previous increments.
- Improvements are made from what was learned about warehouse operation and support.
- The technical environment may have changed.
- Results are seen very quickly after each iteration.
- The end user requirements are refined after each iteration.

Data Warehousing is an evolutionary/iterative process that follows a spiral pattern

- The warehouse architecture is initially developed at the start.
- The first increment is developed based on the architecture.
- Building the first increment causes architectural changes.
- Operation of the warehouse brings architectural changes.
- Each additional increment extends the warehouse.
- Each new increment may cause architectural adjustments.
- Continued operation may cause architectural adjustments.


Current Status
Learn and modify initial requirements
Identify Need
Plan for next phase
Expanding Scope
Deploy, and Use
Start of project
Design
Build
Test

Office of Institutional Research and Academic Planning
Data Warehouse Technical Architecture Design



Data Sources (see attached)

- Budget System
- Personnel
- CARS
- CAS
- FAMS
- NJAS
- Course Scheduling
- SRDB
- Faculty Survey
- Graduate School Programs
- SURE
- Financial Accounting System
- Payroll System
- Reference Data

ETL Services

Database Services

Data Marts:
SRDB
Course Anaysis System
Human Resources
Course Scheduling
SURE

Exploration Data Warehouse

Cubes built from data marts

Reports built from data marts

OLAP Services

Multi-dimensional cubes created include:
Enrollment
Degrees Conferred
Time to degree
Retention Rates
Graduation Rates
Instructional Workload
Faculty Analysis

Report/Cube Distribution Server

Reports created include:
Fact Book
Instructional Workload

Access Methods

OLAP Developer: build OLAP views of data cubes

OLAP User: Access pre-defined views of data cubes

Report Users

Power Users - develop ad-hoc queries and reports. using Access

And Create Data Files

Prepared by Michael J. Cullinan, Application Developer, OIRAP

THE WAREHOUSE POPULATING PROCESS

A data warehouse is populated through a series of steps that

1) Remove data from the source environment (extract).
2) Change the data to have desired warehouse characteristics like subject-orientation and time-variance (transform).
3) Place the data into a target environment (load).

This process is represented by the acronym ETL for Extract, Transform and Load.

Complexity of Transformation and Integration

- The extraction of data from the operational environment to the data warehouse environment requires a change in technology.
- The selection of data from the operational environment may be very complex.
- Data is reformatted.
- Data is cleansed.
- Multiple input sources of data exist.
- Default values need to be supplied.
- Summarization of data often needs to be done.
- The input records that must be read have “exotic” or nonstandard formats.
- Data format conversion must be done.
- Massive volumes of input must be accounted for.
- Perhaps the worst of all: **Data relationships that have been built into old legacy program logic must be understood and unraveled before those files can be used as input.**

Data Warehouse Tools/Software

Component	Product used by Institutional Research	Component Description
Reporting	Crystal Reports	Create presentation style reports with chart and graphs. Can be used to access any type of data source.
Querying	Access 2000	Create complex ad-hoc queries against a variety of data sources using ODBC access to DW databases. Able to export to other types of formats such as text files,
OLAP	Crystal Analysis Professional	Access data cubes for designing views to pivot, filter and aggregate facts on pre-defined dimensions for specific subject areas such as enrollment, degrees conferred, etc.
Data Mining/Statistical Analysis	SAS	Statistical Analysis using ODBC access to IR DW databases.

Recommended System Requirements For Web Access

Internet Explorer for using OLAP web based ActiveX control

Expand video to 1024 X 768 for more viewing space