Bay Delta Science Consortium

Concept for a Monitoring & Modeling Data Aggregation, Storage, Retrieval, Integration and Distribution System

Introduction

Most Bay Delta Science Consortium (BDSC) members, including multiple agencies, academic, private, and stakeholder entities collect large amounts of environmental data. These data currently exist in diverse formats and in different databases with inconsistent, and in some cases, difficult means to access. The goal of BDSC is to promote collaboration and interaction among its members. Therefore, environmental data need to be easily and seamlessly stored, integrated, versioned and distributed to data users for analysis, GIS and modeling applications. Sharing of data will facilitate the development of a comprehensive understanding of the status, trends, and environmental processes and mechanisms in Central California, and help guide adaptive management of natural resources in the estuary and watershed. BDSC will not require various data/information providers to adopt particular sets of vocabularies, software, etc, but can provide a suite of technological solutions for various groups interested in sharing data and information.

This report identifies the users of BDSC, principals guiding the design of the BDSC data management infrastructure, and the candidate technologies, which may be used to meet the user needs.

BDSC Composition

BDSC is comprised of members, which can be broadly divided into data providers, data users, and data aggregators. These members have different needs and requirements with respect to data management. BDSC has to develop a complete understanding of the needs of these members prior to devising an information storage, retrieval, integration and distribution system, which address the requirements of all members. The initial assessment of these needs is expressed in this section.

Data Providers

Data Providers are the BDSC members, which collect and manage data of interest to BDSC. Data providers use local systems for management of collected data. These systems could be as simple as an Excel Spreadsheet or an Access databases or as complex as an enterprise system built using Oracle or Informix databases. They require a reliable connection and a mechanism to transmit their data to the data aggregators.

Data Aggregators

Data Aggregators are typically the recipients of the data, which is collected by the data providers. Data Aggregators often have advanced data management practices, which includes organizational and infrastructure support with hardware, software, and information technology staff. Data Aggregators often provide data management services to Data Providers. Note that a data provider may also be a Data Aggregator if they collect data and receive and manage data from multiple sources. Data Aggregators require reliable mechanism to upload the data provider's data. In addition, they require mechanisms to combine the data from different providers into a seamless interface, which provides access to all Data Users. The required mechanism provides for a common data dictionary and the associated vocabularies to translate between the different naming and referencing conventions used by the various data providers. This requires an intensive and ongoing effort to translate data from data providers.

Data Users

Data Users are the BDSC members and non-members, which use the data stored in the Data Aggregator's database for analysis and research purposes. This group is comprised of the casual, which is the more dominant user group, and the advanced users. The advanced users are characterized by operational/planners, researchers, which include graduate students, consultants, many other non governmental organizations and activist groups. Operational/planners include south Delta water conveyance groups, Delta simulation modelers, and Geographic Information System (GIS) users. Casual users include those with simple queries on the Internet.

Advanced users use the data in conjunction with user supplied analytical tools, such as simulation models, to help make decisions that can greatly affect water resources in the Delta region (i.e. drinking water quality, incidental take, project yield, etc.) This group in most cases needs large volumes of high quality multi-agency data distributed to them. Advanced users may extract data in bulk from the aggregated databases and other distributed data providers. Alternatively, they may need a representation of this data in other formats such as charts, plots, or maps. These users are typically willing to spend a couple of hours to learn how to use sophisticated query tools that may also have utilities to convert data into. And in many cases, they may develop their own custom applications to retrieve and display data.

Casual users require simple web based tools. They expect to obtain data shortly after reaching the web site. These tools may be provided by the data aggregators themselves, but could be third-party portals accessing the same corpus of data (anything from Google, through education sites creating language-appropriate interpretations K-6 graders, to supercomputer center-based simulated flyovers).

Although advanced users can make substantial contributions to planning studies, water exports, adaptive management, etc, and use larger amounts of data, they represent a small percentage of the user community, at this time, which leaves the majority of users in casual user group.

BDSC Principles

Any successful approach to integrating data developed by a diverse group of data collectors and programs necessarily will have to be highly distributed, flexible about platforms and formats, readily scalable, and accessible to meet the needs of the different user groups identified above. The data needs to be version controlled, documented, use common (or translatable vocabularies), and have known validation levels.

Highly Distributed

A highly distributed solution is needed to meet the needs of the Data Providers and the Data Aggregators in BDSC. The distributed solution should allow for a reliable replication of data and the updates.

Platform Flexibility

Platform flexibility is required since the BDSC members have different computing infrastructures without a single standard. Different platforms must be able to communicate with each other to achieve the distributed solution.

Readily Scalable

The solutions used for BDSC should be scalable to accommodate more Data Providers as well as the Data Aggregators. The spectrum of solutions should cover small systems using MS Access as well the larger providers with specialized database, GIS, web servers, and other web services, such as on-line models and visualization tools.

Accessibility

Accessibility of the data should be considered as an important factor in implementing this system. Data Aggregators should make their data available on-line using the web technology. Web accessibility standards should be considered in designing the interfaces for retrieval of the data.

Version Control

The authoritative versions of data should be held as closely as possible to those who generate and maintain them, though it is often desirable to distribute copies of data to a more centralized facility. If users can access multiple copies of data sets, it is essential that they be able to track changes in the data and be protected from differing updates to different versions.

Metadata

Data need to be documented by extensive metadata. Accepted metadata standards should be used in documenting the data. At present, Federal partners are mandated to provide documentation compliant with the Federal Geographic Data Committee (FGDC) metadata standards, at least for geospatial and biological datasets, and most important GIS datasets held by BDSC members have FGDC metadata. However, producing full FGDC metadata is burdensome, and much more simplified formulations by ISO and the World-Wide Web Consortium (W3C) are beginning to be widely used. It is an open question whether ISO or W3C (e.g., the "Dublin Core") provide sufficient documentation for Consortium purposes, and it is likely that a more expanded metadata profile will need to be adopted for non-FGDC metadata.

Where feasible, peer-review of both data and metadata should be encouraged.

Use Common Vocabularies

Data providers should be encouraged to share vocabularies, human-readable expressions of data, and crosswalking tools to promote the uses of information. There are several long-term initiatives to develop and adopt shared vocabularies for subject keywords, place names, chemical names, etc., among California natural resource management organizations, but no consensus has been reached. Without further standardization, complete interoperability among different data providers is all but impossible to achieve.

QA/QC Levels

Metadata and data dictionaries need to record the nature of QA/QC for data ("trust" characteristics). The user can then use this information to help identify the suitability of the data for different analysis.

Technological Solutions

There are multiple possible technological paths and combination of technologies consistent with the user needs and the principles outlined. What is presented in this report is a hybrid of technologies that will provide a proven approach to answering the enterprise need for data and information in the Central Valley region, based on the author's experiences managing large data sets and knowledge of the various institutional practices of data collectors in the region.

The hybrid solution will be comprised of data storage component, data exchange component, and the data retrieval of query component. Each component will be further discussed in this report. Candidate technologies will be identified for each component and will be evaluated in light of the user needs and the BDSC principals.

Data Storage Component

Monitoring data needs to be stored using best practices. As mentioned before, each Data Provider and Data Aggregator need to store and manage data reliably in order to contribute it to a distributed data and information system. Coordinated relational database management systems (RDBMS) will greatly assist with Bay/Delta and tributary monitoring data management because of their ability to store and relate the diverse types of physical/chemical (e.g. water quality, hydrodynamics, meteorological), biological, terrestrial, wetland, fisheries, GIS and modeling information collected in the region. Data submitted to a RDBMS would be stored along with those from other providers in tables related to each other according to key fields (location, date/time, data type, etc.) and made accessible online via any computer with Internet access. Data users could then perform simple and refined queries obtaining the data they need from numerous sources quickly and efficiently from a central comprehensive database or database node compared to the hours or days it might take otherwise. In the simple form, RDBMS may be implemented using relatively simple table structure (e.g., with MS Access software or attribute tables of a GIS) for the Data Providers. In the comprehensive implementation, RDBMS will be implemented as a full object relational model using a specific database vendor. The comprehensive system can also be distributed to local nodes (locally situated servers - see below) where large amounts of data are being used, and/or where obtaining data using a browser is slow. Development of a new RDBMS for Bay/Delta and tributaries monitoring data can be avoided by using an existing relational database, the Bay/Delta and Tributaries Database (BDAT). For these reasons, employing BDAT to manage and distribute Bay/Delta and tributaries monitoring data not only may represent a considerable cost and timesaving compared to creating an additional system but also avoids duplication of effort, or "reinventing of the wheel". BDAT is the most obvious candidate technology for the main centralized aggregator and data storage facility for BDSC.

The BDAT design is a two-level data management system including a normalized and a warehouse database residing on a main server (Figure 1). In addition BDAT includes an interface to upload bulk data and a web-based interface to retrieve and display data by web clients (e.g., desktop web browsers). Data loading into BDAT comes from multiple, local client databases. These local databases are typically developed in MS Access by the BDAT team working together with the data provider, a process that also serves to help providers better organize their data and manage it once the system is in place. Alternatively, if a client is currently using Access or another relational database, data conversion programs could be written. Data in the various client databases are combined and synchronized with existing data on the server using database replication. A record of all transactions between the main server and local databases is kept such that this, combined with the method of synchronization, ensures version control between incoming and stored data.

Because BDAT is already a Geographic Information System (GIS)-aware system, it includes the capability to store and retrieve/display GIS data. Furthermore, its web query interface is capable of formulating queries using both spatial and/or tabular criteria. In addition to the local client databases, other servers containing GIS data can be integrated into the system and GIS information delivered to the Internet for general access or to different locations for collaborative analysis. Future non-propriety expressions of geospatial data (e.g. geographic markup languages) should be obtainable from BDAT, so long as the content (place names, attribute types, etc.) is compatible.

Many of the BDSC members do not currently have a sufficiently developed RDBMS to participate in such a data enterprise with other monitoring groups, so those local databases need to be developed. Common usage of these databases by BDSC members addresses several items that should be considered in a distributed environment. The modern technology and concept of a relational database should improve Data Provider's data management capabilities when compared to the common use of Excel or ASCII files by agencies not currently employing this method. A local MS Access client database is one such system that meets this need while also providing the infrastructure for enterprise wide data management. Agencies adopting this technology are better able to manage their own data. It is not necessary for participants to use the same software, platform, or data model in order to exchange information with BDAT. This solution reflects the preference by data providers to manage their data locally so that they can QA/QC (and control) them prior to uploading them to a central location such as BDAT. Local MS Access client databases readily meet this need for smaller datasets that do not include binary data types. This necessitates a mechanism for maintaining version control of the data and for tracking and ultimately re-propagating these edits. MS Access client database applications can easily be programmed to meet this need and those in production already do. In cases where data providers already have reliable existing databases, other options, including those discussed later in this paper, will be implemented to provide the optimal solution to convey and integrate their data.

The data structure used in any local client database must be flexible enough to allow for user specific extensions or expansions yet also one that provides a standard and consistent format so data can be consolidated after it leaves the local environment with other data sets. Local MS Access client databases already developed for integration with BDAT include all of these features.

Data Exchange Component

Data Exchange component of the BDSC architecture is a key component to enable the transmission of data between Data Providers and Data Aggregators as well as between the Data Aggregators themselves. Many data collectors have their own nomenclature to define species, analytes, etc and these criteria often vary within the same agency. Experience working with these groups has revealed they would like to continue to use their current list of organisms, analytes, etc. Using basic relational database theory, these nomenclature can typically be normalized into tables and unambiguously identified if there is agreement on the underlying content. The common vocabulary in the form of XML protocol will provide a mechanism for different parties to exchange the information between the databases. A critical first step is to promote greater standardization of terms and measurements, in the form of "thesauri", or controlled vocabularies. For example, most of the biological monitoring programs within the BDSC region to not share standard codes for species identity, making it difficult to reconcile species names. Similarly, chemical constituents may be addressed in multiple ways. In some cases, reconciliation is a straightforward look-up. In others, it can be problematic because of heterogeneity in the sources of the underlying data. (Consider measurements of methyl-mercury, organic mercury, total mercury, heavy metals, derived from a variety of analytic methods…) An extremely urgent need is for standardized names for sampling locations. It is recognized that various data collectors can not be forced to adopt a particular set of vocabularies but the node concept provides the opportunity to create crosswalks so queries can be conducted using a known set of vocabularies.

Three data exchange mechanism may be considered to meet the Data Exchange Component requirements; Direct Database Connection to communicate between the Data Provider's MS Access client and the Data Aggregator's database, Replication to communicate between the Data Aggregator's which have common systems, and Web Services to communicate between Data Aggregator's which do not have similar hardware/software. These are further described in this section.

Direct Database Connection

Direct Database Connection is the simplest mechanism, which allows a Data Provider's database (e.g., MS Access) to communicate with a Data Aggregator's system. MS Access clients in the BDAT setup are currently using Direct Database Connection. This protocol is adequate when the two parties can fully coordinate their database administrative actives. Altering the design of the database on each end needs to be coordinated by both parties. This technology is readily available and typically does not require extensive software or system administration support. It relies on the connection mechanism between the two systems to be made available through the firewall for both parties. Data is moved from the local client to the Data Aggregator's system using programmatic replication.

Database Replication

Database Replication will use Data Nodes (Nodes) by BDSC members and other groups to exchange the data among themselves. This approach uses regional data nodes that serve as localized points of consolidated data and information. These nodes can be linked as a physical database using replication. A fully developed data node will include a server with data and information provided by several groups who have data to share. Nodes will have the ability to crosswalk to various controlled vocabularies including those suggested by the BDSC. Each collector of data will have local data management systems. Data from local collectors will be provided to the node and the various nomenclatures crosswalked to the requested controlled vocabularies, including those requested by BDSC or other parties who would like to use these data. Users of data could potentially select from a list of controlled vocabularies eliminating the need for an authoritarian list and alienating data providers and their affiliated agencies. Custodians of data nodes should have a viable rapport with their data providers and serve as regional centers to provide necessary services to data collectors so they can manage and share their data. Nodes can also be set up strictly for data users. The incredible volume of data collected in this region can make access to it from the WWW too slow for groups who use large volumes of data. In these instances, local servers or nodes can be set up to provide direct access to data.

The symmetrical replication model employed by BDAT can permit linking between the various nodes distributed among locations in the Bay/Delta region. These nodes can then serve as the first point of data consolidation from the local desktop client databases according to their physical proximity or program. For example, a node may be used to consolidate data from various fisheries programs within a regional office. This model essentially provides more localized access to large amounts of consolidated data, an asset that may be desirable for users of large data sets. (In current practice, local replication of highly distributed systems is generally required for adequate performance.

Development of data nodes is most beneficial when there is a single entity who manages data for several distributed groups, and who also, in turn, needs very large volumes of high quality multi-agency data distributed to them. Typical groups currently using nodes require physical representation of these comprehensive data on-site to conduct continual analysis with large data sets. Nodes currently being developed will have the utilities to store and replicate spatial and regular time-series objects for those groups trying to consolidate large volumes of spatial and time series data. Nodes can be standalone data centers or distribute those data to BDAT.

Web Services Based Data Exchange

For cases where Data Aggregators have existing databases and are not interested in being a symmetrically linked node and do not need physical copies of comprehensive data residing on their server, the plan is to supply another mechanism to exchange data and information. Web Services is an infrastructure strategy that promotes development of individually accessible distributed components using a loosely coupled and reusable software architecture**.** For example, Web Services can be used in an Internet/Intranet configuration to develop web and standalone applications. BDSC users can take advantage of the web services in the following areas:

- Provide a mechanism to develop data exchange among heterogeneous data storage systems, such as Oracle, Informix, SQL Server, DB2, , MS Access, Word, Excel etc. Over time, new standards will allow these services to specify rules and formats for access in machine-interpretable forms, permitting relatively automatic access, even as fields and services are expanded or modified.

- Provide another mechanism to access service applications (e.g., visualization, simulation, statistical analysis, reporting analysis results etc.) without having to download this data into a local database or reformat the data.

- Provide a mechanism for data providers to share data without enforcing a database structure.

- Provide another mechanism to develop standards based interfaces to read and write data without the need to change the underlying physical data model.

Web services would include a number of modules such as DataProvider and DataRetriever Services. DataProvider Services includes the interfaces for storage and update of data. Data Retriever Services includes the interfaces for obtaining data. Examples include field observation and metadata entry tools, complete with pull-down lists of standardized terms, map browsers for visually specifying locations, on-line bibliographies for references and authorities, and maybe identification aids for biological materials. Figure 2 shows, as an example, the web services based interface to BDAT and other database nodes. The implementation advantage which will be gained from using web services include:

- Allowing for the presentation of the data using different frameworks without having to change the physical structure of the data.

- Use in conjunction with GIS Services to provide a comprehensive interface for building data exchange or standalone applications.

- Additional security through the use of standard web ports (i.e., Port 80) to avoid having to open additional website ports within existing server firewalls.

Data Retrieval Component

Data Retrieval component of the BDSC should address the needs of the Data Users. The candidate technology for the Advanced users using a browser includes the BDAT's GIS enabled technology for extraction of data from BDAT. This technology can be adopted on all data nodes. The BDAT interface allows for extraction of the tabular and the spatial data. The advanced users may also use any of the Web Based Data Exchange interfaces described above as a means of obtaining bulk data from the node of their interest. Many specific web based information sites have been developed that use data from BDSC participants. These include forecast modeling, spring run reporting, simulation modeling, etc. The ability to develop information systems that can convert data into information, based on access to a comprehensive set of Bay/Delta and Tributaries data, greatly enhances many operational, adaptive management and research efforts already underway in the region. Using distribution technologies, therefore, provides the opportunity for many groups to develop customized data retrieval systems that meet their specific needs or to develop processes that convert data into information that they can share with other interested parties via the Internet or other types of media (Figure 3).

Casual users should be provided a new lightweight and easy to use interface to effortlessly obtain data and meta data associated with their queries. This interface should be developed in conjunction with a mapping service (such as those offered by ESRI's, ArcIMS) to allow for development of web-based maps of the data stored in the Aggregator's databases.

Conclusion and Recommendations

The use of existing software technology is proposed to facilitate and access sharing of Bay/Delta and tributaries monitoring and other types of data and information so they can be used interchangeably between multiple monitoring programs and reporting systems. The proposed mechanism is flexible in that it can be set-up to deliver specifically formatted data to decision support groups, comprehensive data for integrated research projects, or customized Internet links to data sub-sets. In addition, data considered preliminary and not ready for broad distribution, can be controlled and made accessible only to specific data user groups.

While the *ad hoc* Data Committee has developed the concepts detailed above, and is preparing proposals to begin implementation, the coordination, implementation, and management of a Data and Information Management System for BDSC will require dedicated leadership. The BDSC should begin to consider the creation of a more formal committee that include representatives from various agency, academic, stakeholder and private sector participants who have experience and expertise in distributed data/information systems. This group should: address tactical planning and implementation of the ideas discussed in this report, help determine how changing technologies should be applied to the effort and determine the staffing needs of a distributed data/information system, including the possible appointment of an Information Technology Manager position to work collaboratively with various participants.

To help move the above ideas forward, a series of common vocabulary workshops, and demonstrations of BDAT installation at another key Bay-Delta node will be proposed as the next steps in implementing the system described above. BDSC data committee is planning to hold a series of workshops in the near future to develop these common vocabularies in collaboration with broader state services (e.g., CERES), standards organizations (FGDC, NBII) and library scientists. Priority applications include taxonomy, geolocation, water quality terminology, and monitoring and analytical methodologies.

There have been several requests to set up regional nodes using many of the technologies employed by BDAT. Implementation of these nodes is another critical step in developing a region wide data and information conveyance and management system.

With interoperable semantics (language), BDAT and other data providers can specify methods for automated access to web services. The sample network resulting from the BDSC setup is shown in Figure 4.

Figure 1. Bay/Delta and Tributaries Database (BDAT) components

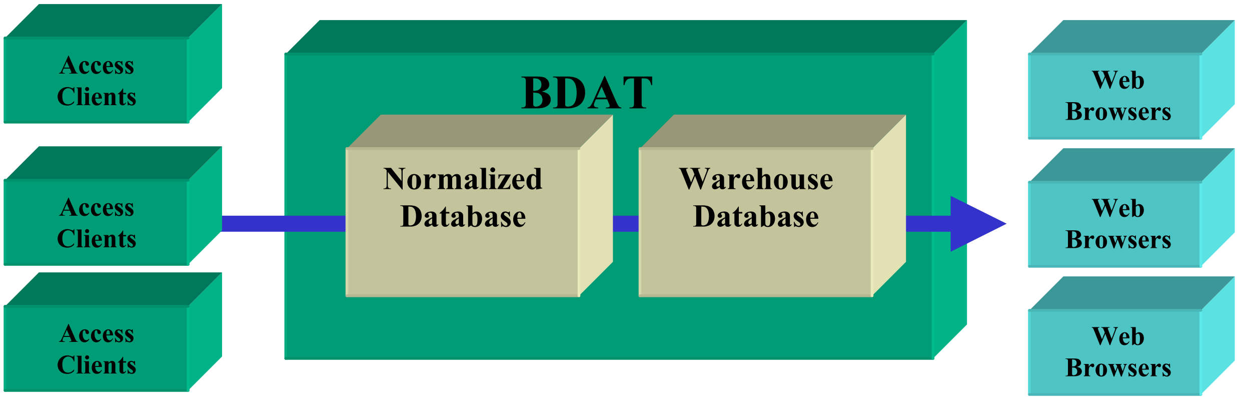


Figure 2. Bay/Delta and Tributaries Database (BDAT), or other more localized servers (nodes), with web services

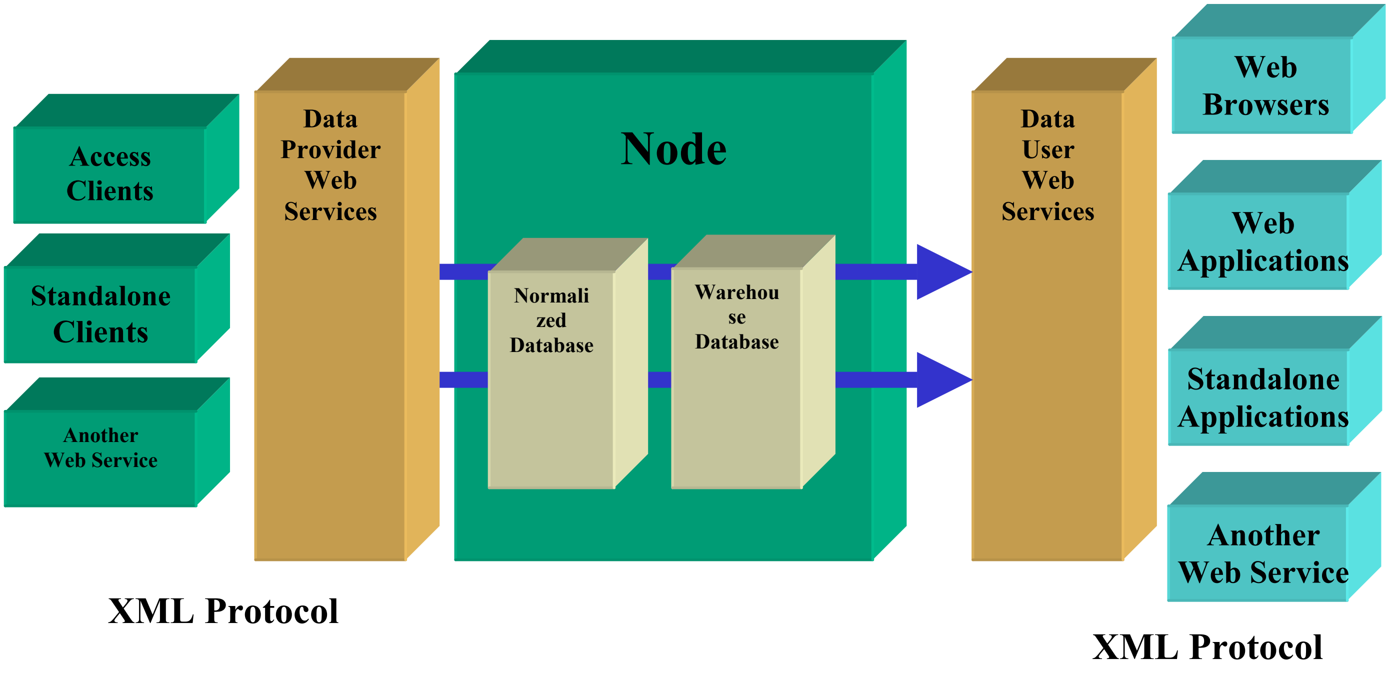


Figure 3. Sample web services application

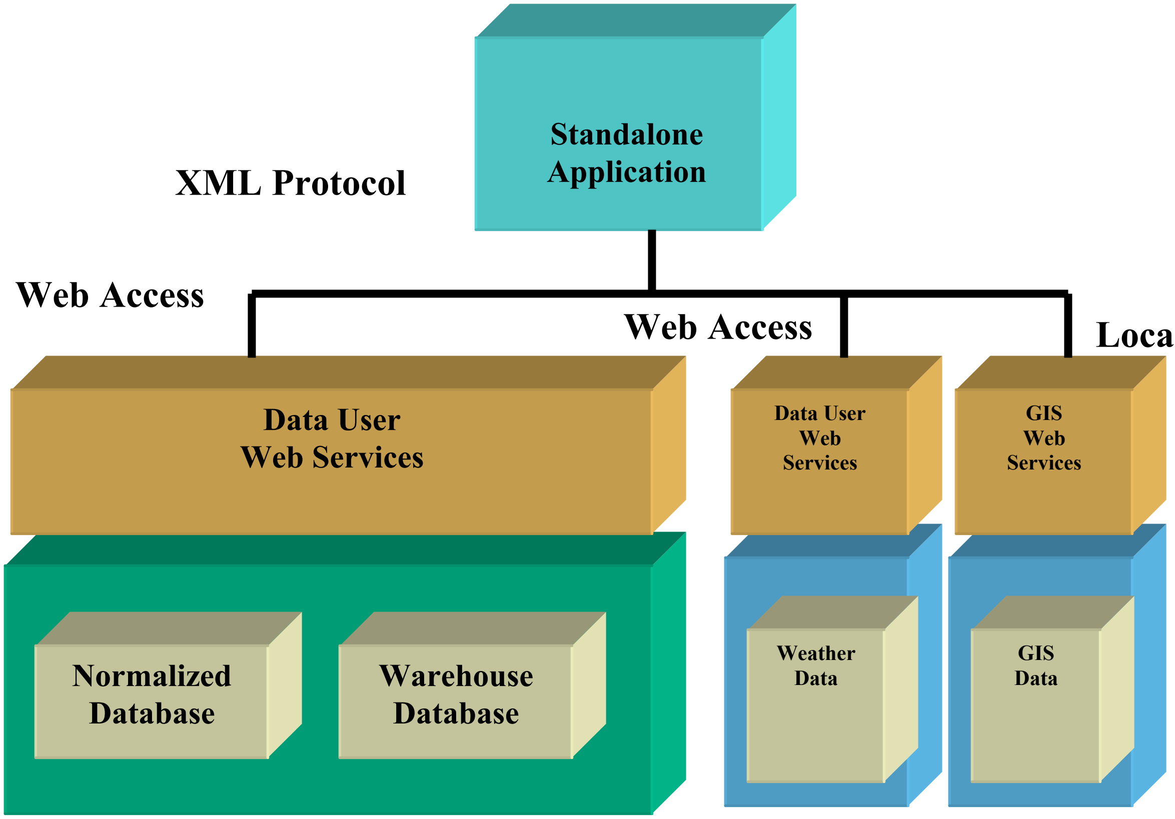
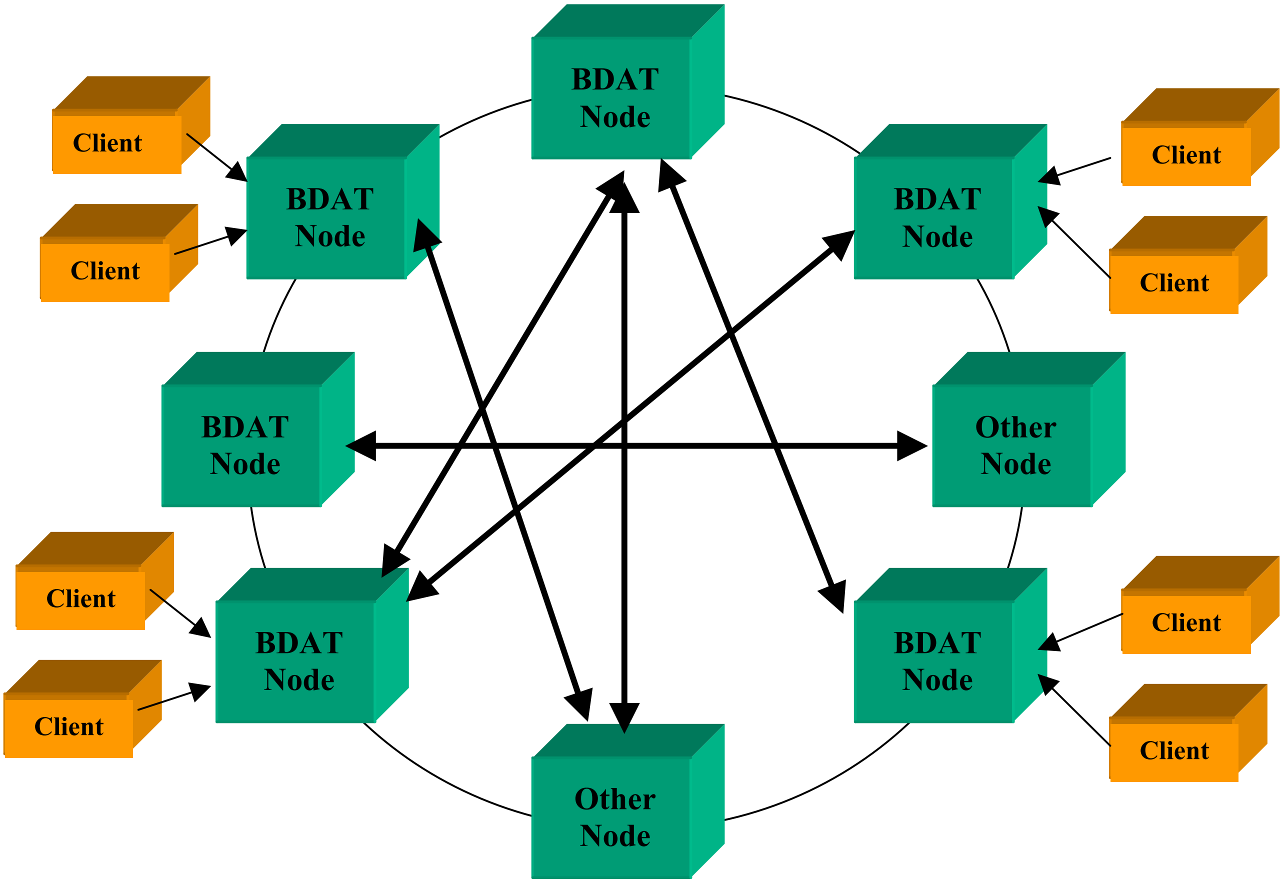


Figure 4. Sample Network of Nodes in BDSC System


Client
Client
BDAT Node
BDAT Node
BDAT Node
Client
Client
BDAT Node
Other Node
Client
Client
BDAT Node
BDAT Node
Client
Client
Other Node