



White Paper
September 2001

ENABLING A UNIFIED VIEW OF INFORMATION ACROSS THE ENTERPRISE: *Moving beyond the data integration paradigm*

PRAJA's technology is built on a radical new approach for data access. This white paper will describe this approach and distinguish it from other data access technologies

The challenge facing organizations today is to rapidly adapt to a changing business environment amidst an explosion of data—not only by size, but also by type and by source. To successfully respond to this challenge, organizations must turn vast amounts of data from multiple applications, databases and processes across different systems into well-informed insights

The response to the growing volume of alphanumeric, text, images, graphics, audio, video, and sensory data across multiple locations has been to enhance the efficiency and speed of data extraction and classification into category taxonomies, or to facilitate better data access using data integration and data warehousing.

These approaches produce only incremental gains. They rely on time consuming integration processes, and provide for sequential query and reporting, the results of which can be examined side-by-side, but are not integrated to provide an overarching view. PRAJA believes that despite extensive tweaking, constant jury-rigging and minor enhancements, these solutions remain fundamentally anchored in a paradigm that cannot address the challenge of information access today. The information needs of today's business world demands a radically new approach based on a fundamentally different way of thinking about the growing problem.

While *data integration* approaches are central to current IT strategies, PRAJA believes that such efforts will only serve to extend problems associated with 'information overload,' also termed 'infoglut,' and 'informationitis.' Information value is derived from finding key trends and context between enterprise data silos. Unlike other solutions that try to adapt old approaches for current needs, PRAJA's technology is designed to facilitate unified organization of data and hence, access to highly relevant information in an environment wherein both the data and business rules are changing. PRAJA's technology has been designed and developed with enterprises' basic needs in mind.

Assimilation is not Classification

Data classification applications are designed to automate the organization of text to enable more efficient query and search. These tools determine the category to which a piece of data belongs and expedite the process of populating category taxonomies. Although classification applications help manage text data they do not address key problems:

- Classification is focused on text and does not solve the issue of explosion of data by type. PRAJA's technology is designed to address the needs of data organization when data is not just text and does not derive from a single source.
- Classification solutions are unable to determine the relation between different pieces of data from multiple sources, and hence cannot organize data in a context that makes it relevant to the user. PRAJA's technology goes beyond determining the category to which data belongs in order to show how different data from different categories (and of different types) are related to one another in a context of interest to the end user.

Assimilation is Not Integration

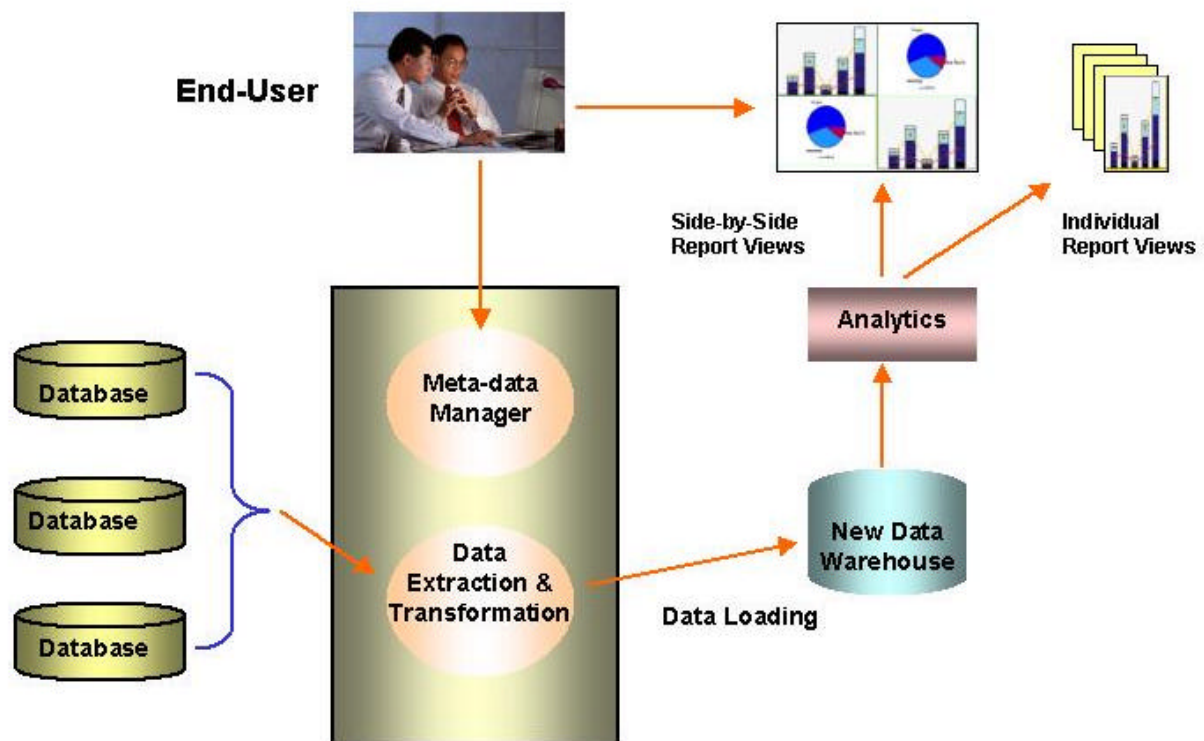
Data integration is often presented as a solution to the problem of accessing information across multiple data sources inside as well as outside of an enterprise. This approach employs Extract, Transform, Load (ETL) technologies to integrate large volumes of data from multiple sources and places them in one location (a data mart), to enable centralized query and analysis, which is performed by *Business Intelligence* applications. The focus of this approach is to ensure rapid access and consistency of data. Data integration does not add semantics to the data, and hence, cannot facilitate understanding the context for data. Information access and business insight solutions in this approach are reduced to integration and report generation activities. There are a number of notable limitations to this approach:

- Building *Data Marts* is a major undertaking, which requires substantial resources and time, especially as the number of data sources increases. It is not a flexible approach, and hence does not provide enterprises with rapid information access solutions in a changing business environment.
- The data integration-analytics approach can by definition capture and hence reflect a snapshot of the business rules and processes of the enterprise at the time of integration process design. Data marts cannot reflect changes in business rules and processes without significant development effort. Existing data integration and their associated analytic tools are focused on the creation of data marts and loading of fresh data into them faster. This, however, does not address the fundamental problem of adaptability of the data marts, which by nature are not flexible to the needs of an agile enterprise.

- Users access data in the form of reports that can at best be displayed side-by-side in a portal environment. To understand the relationship between data, correlations between reports must be established manually, which is both cumbersome and difficult to scale as the number of reports increases. Multiple reports, even when presented in a single view do not constitute an integrated unified view of data.
- Access to and correlation of multiple query results becomes an exponentially more difficult problem as the number of data queries and data sources increase. This is not a scalable solution. Only an approach that moves beyond sequential queries to contextually organize and access data, and presents information of many types and from many sources in a truly integrated and unified view can provide a satisfactory way in which to understand the relationship between results of queries.

Whereas data integration is concerned with data access and consistency, data assimilation is concerned with placing data in the context that is of interest to the end user. Data assimilation is not about moving bits and bytes, but adding contextual semantics—which include time and space. This approach enables data access in a context that comprises not only of the “why” and “what,” but also the “where” and the “when.” PRAJA’s approach differs markedly from data integration, which is *syntactic* in nature.

Data integration requires a highly structured description of the data to map data from different sources to a new integrated data format and to create query systems. There are a variety of solutions that promise to solve enterprise data access problems and to create data marts and data warehouses that



**Data Integration: ETL + Query-Report.
No Unified View of Results**

facilitate querying and business analytics. Despite its prevalence, data integration is far from a solution to data access needs of enterprises:

- Not all business data are structurally defined. In fact, the size of semi-structured, unstructured and varied-type data in enterprises is rapidly increasing. Enterprise data now includes rich media and data of different formats from outside of an enterprise. The growth in B2B transactions, collaboration across organizations and online markets, and generally greater access to data sources external to enterprises has made it more difficult to quickly integrate data. As a result, enterprises have turned to heterogeneous database systems to solve their data access problems.
- Not all business data reside in static repositories. Traditionally, the integration of data repositories has taken place through either custom integration efforts or replication of data into a central database, or by using intermediary systems that convert multiple queries from varied data sources into a joint query that is then dispatched to the data repositories. These approaches provide only partial solutions; notably, they do not work with streaming data. Since it is not feasible to store unlimited streams of data, the central repository or intermediary system must maintain certain structural information based on the *semantics* of the data stream.
- Important associations between different data cannot be derived from their structural properties. To understand those associations, which are critical to realizing business objectives, data must be endowed with *semantics*. All the more so as the proliferation of communication networks has greatly increased cross-domain data integration. While the syntax of the data with respect to the domain is a given, the semantic characteristics are always defined based on observation of facts and experience. The semantics of the data depend on the context, organizations, and cultural practices of a given application. Semantic *data*

assimilation (as opposed to syntactic data integration) requires that the data be analyzed in the context of the domain in which they will be used. In other words, the same stream of data, for example, regarding the weather, will have a different interpretation and a different association with other data when a shipping company tries to identify potential trouble spots as opposed to when organizers of a conference are looking for a suitable location.

PRAJA's assimilation technology is based on the use of a *domain model*, which specifies *the nature of relations existing between data* in the context of a particular application. The domain model ensures that assimilated data are always placed in a semantic field, in which they can be related to other data based on commonalities and differences in meaning.

This approach provides the necessary data structure for semantic-based navigation and query of data, and for the unified presentation of related data— independent of the medium that originally carried them.

Events as the focus of Assimilation

The challenge facing enterprises today is to realize value from their immense information investments by understanding the relationship between vast amounts of data with varied structures and semantics and monitoring complex business processes to create actions and insights that can positively impact business operations. To do so, enterprises must move beyond the integration of data, which simply extends the problems of information overload, to the assimilation of data resulting in a unified *interface across enterprise data*. This means they must access and organize data in the appropriate context

that makes the data relevant. That context comprises of the “what, “where”, and “when” that determine the relevance of the data. Without that context, it will not be possible to understand complex events, historical trends, and the manner in which various pieces of information or processes interact. Analysis, forecasting, insight into trends, processes and operations are complex activities that require understanding how data of different types and from different sources relate to one another in a spatio-temporal framework. That understanding is furthermore, conditioned by the network of relations and corporate culture that support decision-making in an enterprise.

PRAJA’s technology organizes structured and unstructured data from many sources in the context of the event, which intuitively makes the data relevant to the user—in that it provides a multi-dimensional view of an event or set of events and associated data. Events are the semantics for data. This is a foundational concept that enables domain modeling capability and data assimilation.

PRAJA ensures the right level of information granularity by focusing on *events* as the organizing principle for data assimilation, navigation, and presentation. PRAJA’s unified indexing organizes data around events, *entities* that participate in these events, and other *related data* that provide additional information about the event. Using events, the database unfolds in time, encompasses space, and cuts across the divisions between applications and process that carry, and sources that house data. All data can be indexed and accessed based on the events to which they refer.

For instance, take the case of a car accident. An accident sets in motion many business processes that involve data of different types from many sources: police, insurance, motor vehicles, video, images, medical and the like. The common thread between all of these disparate data from different sources is the accident or “event.” An end user should not have to integrate all of these sources into a single data warehouse in order to access the data he or she needs, nor should the end-user be compelled to manually correlate the result of multiple queries from all of the different data sources. Using PRAJA’s technology the end user has access to all the information that is relevant to the accident in a single unified view through the existing databases and without the need for any additional infrastructure development. It is PRAJA’s data model that enables the end user to access all information pertinent to a single accident and/or facilitate doing the same across many accidents to gain insight into trends or patterns of interaction of various factors.

Contextual Navigation

There is no value in a disjointed assembly of data. Placing lettuce, tomatoes, cucumbers and salad dressing on the same table does not constitute salad. Many data access solutions in the market today merely provide a side-by-side view of different data and reports, claiming to provide “integrated” or “unified” views. Through contextual organization of data, PRAJA is able to create a unified view of disparate data from many sources. This provides for a deep understanding of the underlying relationships between the data, and a genuine experience of the event that serves as the context for the

data. Such distinction between 'experience' and 'information visualization' is summed up in the term "seeing the big picture."

There are three levels of understanding of the term context:

- **Data Context.** Individual pieces of data are inherently related to other data. It is through these relationships that data acquires meaning. For example, the price of a company's stock does not provide insight into the company's performance without examining real-time and historical relationships between other relevant information within the company, economic data, market analysis, and data pertaining to other securities within its industry. This principle lies at the core of all complex information systems that comprise of many processes as well. From a technical point of view, this implies that to provide a total view of an event or to recreate an experience one must pay particular attention to the *structures* that relate individual pieces of data and various processes to one another. As a result, relational databases, in which the only structure is the one that is evident in the set of records, cannot recreate complex events.
- **User Context.** End users seek information in direct relation to other activities—information access is a component of an individual or employee's workflow. As such, information access is bound by personal considerations and preferences as well as organizational needs. Technically, this implies that activity modeling and query rewriting are essential to recreating experiences and that the semantics of context-specific terms like "important," "relevant," "up-to-date," must always be a key consideration.
- **Interaction Context.** A user interacts with an information system through a series of queries and responses that are presented in the form of data. The user can ask the system for completely new results in a fresh query, for a refinement

of the first set of results, or for results to a new query connected with the first set of results. The query and the presentation of results usually occur in two different spaces: a *query space* and a *display space*. This division results in:

- A break in the connection between the user and the data, forcing a context switch between the query and the presentation of results.
- A breakdown between query and result, wherein there is no correspondence between the structure that is suggested in the results and the query.
- Absence of first-person experience as the system that processes the query effectively separates the user from the data. Here the user is barred from discovering the inherent relationships between individual pieces of data by the intermediary system.

PRAJA allows application developers to design context-sensitive information access solutions and move beyond the paradigm of sequential queries to simultaneous discovery.

- PRAJA's data model provides a natural infrastructure for the implementation of data context.
- PRAJA's application development environment provides for extension of the core data model to define the semantic layer that is required for data assimilation.
- PRAJA's application development tools provide numerous ways to interface with data to maintain the context and connection to the user's data management systems enabling implementations of user-contextual applications.

PRAJA's Architecture

PRAJA's architecture aims at providing the development foundation for event-centric and experience-based

applications and, more specifically, of event-centric and experience-based interaction with data rather than information-centric blind queries. The core pieces of PRAJA's architecture are the *semantic structure of the data model* and the *unified index*. While typical data aggregation is based solely on syntactic data composition rules (database schemas, XML tags, etc.), PRAJA assimilates data into a *domain model*, that specifies the semantics of certain pieces of data into the specific context and application domain. Semantic aggregation allows data linking based on commonality of meaning, rather than just by syntactic characteristics.

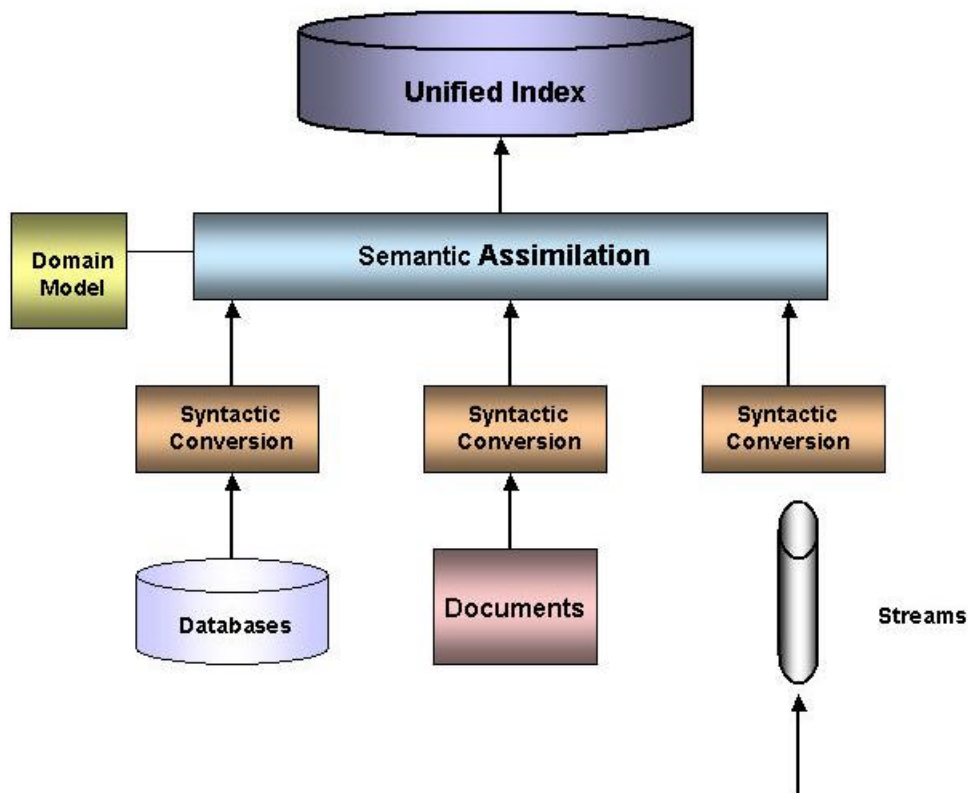
The specific domain models are created to describe a given domain that constitutes the fundamental unifying language of all PRAJA applications. PRAJA uses *space* and *time* along with the taxonomy of the data as the foundation for domain models. The domain model allows for real-time

assimilation, organization and presentation of data from multiple sources in the context of the event and experience that is captured by the domain model.

Central to PRAJA's semantic unification are the notions of:

- *Event*—a situation, incident or development that is defined by its attributes (what is it), time (when it occurred) and space (where it occurred) **both physically and conceptually**.
- *Entity*—a piece of data with no time reference associated with it—a timeless event.

The data that is assimilated into PRAJA's system revolve around central events and their related entities. The data structure of the system is therefore based on the relations between events and entities (or their categories). The focus on a nucleus for assimilation and organization of data is central to the *unified indexing* system, which for



PRAJA consists of an index of events, entities, and their structural relations.

PRAJA's data model and tools allow for the definition; of domain-specific events and entities, of complex assimilation and aggregation of data types containing events and entities (such as lists that represent sequences, or trees that represent hierarchies), and of associations between events and multimedia data. PRAJA's unified indexing is capable of efficiently mapping the data model into relational or Object Oriented databases for persistency.

Data from databases, text documents, data streams, and/or multimedia streams can be analyzed online to detect and store events. PRAJA's event language allows for a definition, within the framework of any domain model; of (1) the events of interests, (2) any diverse source of data that contribute to the existence of an event, (3) the condition on these multiple sources (4) multiple time intervals that must be satisfied in order to detect an event, and (5) the associations and dependencies between groups of events, events and entities, events and media.

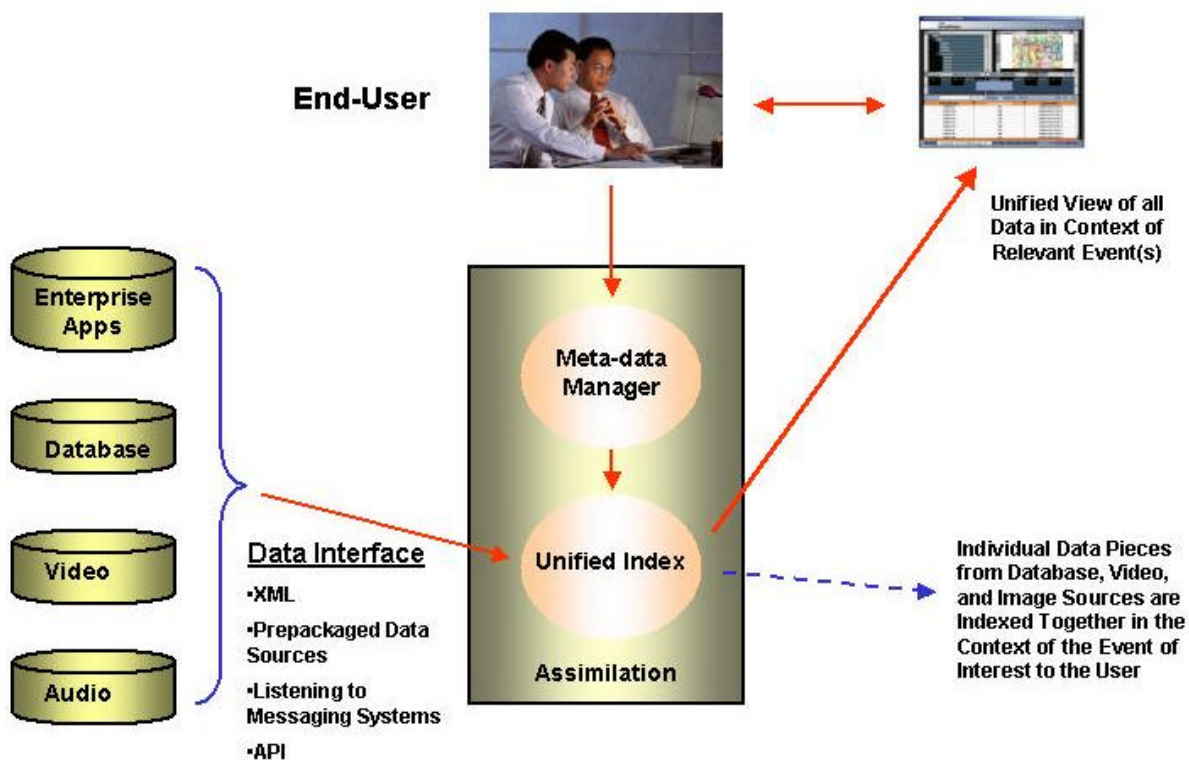
PRAJA's query language allows users to include sophisticated stateless queries into the data model, and to employ operators for navigating the data

structure starting with the results of a query.

The unified indexing, event definition language, and query language provide a powerful data infrastructure for semantic-based navigation of complex time-dependent data sets. They are complemented by a powerful development environment, which allows the definition of user-facing experience-based software or use of PRAJA's own time and space-based browser.

PRAJA's unified information access represents a paradigm shift away from current data access and organization technologies. These technologies do not provide data to users in the context that is relevant to them, nor can they view the data in a genuinely integrated fashion. The limitations of these technologies prevent users from fully exploring complex relationships between data. *PRAJA provides data* assimilation in place of data integration, and *navigation and exploration* in place of query and search. PRAJA focuses on enabling users to discover whereas existing technologies can only provide answers to predefined questions. PRAJA provides a multi-dimensional, unified view that cannot be obtained from sequential one-dimensional queries and searches.

The tremendous promise of PRAJA's technology is evident in the many plausible applications enabling users to develop a clear and complete view through the clutter of data, to discover and understand interdependencies, identify new paths, and to do so quickly and without need for new infrastructure investments. PRAJA's technology has been engineered to operate in an environment of constantly changing processes and increasingly complex interrelationships of data of diverse types. It is a technology that enables a more agile enterprise.



Data Assimilation: Unified Indexing + Navigation-Discovery. Unified View of Data in Context

PRAJA's technology is a radical departure from the ETL-query/report solutions. It represents a fundamentally new approach: **Data Assimilation**.