PRAJA

Addressing Requirements for Unified Data Access

An IDC White Paper
Analysts: David Sonnen (ISSI) and Henry Morris

Executive Summary

Data-access problems began sometime in the 14th century, when accounting records and inventories became too big to memorize. Responding to the problem, Arab and European scribes wrote records on newly invented paper. Inevitably, paper stacked up and one fateful day some Middle-aged merchant couldn’t find the right record and created the first data-access problem. Those problems have gotten worse ever since.

Today, The New York Times publishes more information daily than a 19th-century businessperson would have seen in a lifetime. Fiber-optic technology lets us transmit all the data ever published in the Times plus about three Libraries of Congress per second down a single cable. Finding the right record is like finding a needle in the Gulf of Mexico.

The information technology industry has evolved toward reasonable data-access solutions, but fundamental problems remain. Finding complete and relevant data is still an issue for business people.

How can information systems be structured to provide more relevant information for the user?

Corporate users and consumers need data that is relevant to their immediate needs. In response, the IT industry has fielded a wide range of technologies that address portions of the data-access problem: universal databases, enterprise information portals, search engines, metadirectories, enterprise data interchanges, and enterprise application integration, among others. These technologies provide a useful, but discontinuous structure for integrating enterprise data and applications. However, the user is still faced with disjointed access that inhibits effective gathering and use of relevant information.

In addressing the data-access problem, PRAJA developed a technology platform that focuses on the events that make up business processes. Its method resolves many application integration and data-access problems inherent in older approaches.

Looking forward, IDC sees the emergence of a new class of technology platforms that provide more robust, unified data access. PRAJA’s concepts and technology provides an important step in that direction.
PRAJA has taken a unique approach to solving data-access problems. First, it manages data about the events that make up business processes. Second, it organizes this data within a contextual framework of time and location. We discuss both these concepts in some detail later in this paper.

PRAJA’s approach solves a number of persistent data-access problems. IDC believes that its technology is an important step toward unified data access. In this paper, we examine current data-access issues and PRAJA’s approach to solving them.

The Data-Access Problem

Data access involves getting a complete, understandable, and relevant answer every time you ask a question. Think of the difference between asking an expert versus asking a librarian for information. An expert gives you a concise, relevant answer. A librarian helps you locate data. Once you’re in the right place in the library, you still have the difficult tasks of finding the right shelves, locating the right books, sorting out the pertinent data, assimilating it, and developing an answer. Current data-access systems are good librarians. Effective data-access systems need to be more like experts.

Why is effective data access challenging? We see four major factors:

- **Data volume.** Data volumes have grown so large that it’s hard to locate pertinent data among all that is available. In 1996, Veronica, an early search technology, could do keyword-in-title searches for all 15 million documents on the Internet. Today, Google, one of many sophisticated search engines, can index and search about 1.5 billion Web pages — impressive but less than half of the Web. The Web is still growing at nearly an exponential rate, and most corporate data is not on the Web but is sequestered in diverse and private data stores.

- **Data diversity.** People understand real-world situations by acquiring data from a variety of sources. Rich digital data — video, photographs, images, unstructured text, audio, data from satellites and many other sensors — can help form a complete picture that approximates the real world. But most information systems are best suited to handle structured text, not rich and diverse data types.

- **Data context.** Every business process consists of events. Every event occurs within a context — time, location, people involved, and so on. Information systems treat events as one-dimensional transactions, stripping contextual information. This approach
allows systems to process transactions with amazing efficiency but leaves the user without valuable contextual information.

- **Ever-changing users and processes.** Increasingly, business processes involve not only employees but also a dynamic set of suppliers and customers. All these users need relevant data that supports all their mutual and constantly changing business processes. So, as processes and users change, information systems must change with them while still providing complete and pertinent data quickly.

In response to these problems, information technology is evolving to the point that more effective data access is possible.

**Evolution of Data Management and the Internet**

To understand potential data-access technologies, it is useful to trace the history of data management. As shown in Figure 1, data management and the Internet are evolving and gradually converging on unified data access. In the following sections we discuss some of the key steps in this convergence.

**The Legacy Data Management Environment**

In legacy data management environments, the worlds of structured, unstructured, and complex data are like parallel universes. Separate databases were deployed to house each type of data. Separate product
ecosystems grew up around the separate databases to access and manipulate the data. As a result, organizations need to implement different sets of technologies to manage their data assets. Separate applications were built to leverage each database or information repository. In the legacy environment, it is rare for an application to access multiple types of data, largely because of the complex access logic that the application has to manage.

**The Universal Database: A Partial Solution**

Bringing data of different types together in the same database was the principal goal of the unified database, pioneered by Michael Stonebraker with Illustra. Illustra featured access methods for each data type, reducing the complexity for database administrators and application developers.

In Illustra, applications still had to access each datatype separately by making separate calls to each datatype-specific method. Also, the unified database approach assumed that data from specialized databases would be moved to a single database. However, this isn’t always practical given the size of specialized databases and the complexity of information systems. This approach represents unified data management, not unified data access.

**The Unmet Promise of Portals**

Portals have been positioned to become the means for supporting the information access and delivery required for enterprise operations. However, they have not, to date, met the need for unified access to all relevant data.

The demand in the marketplace is there. For example, in IDC’s 1999 Technology Integration Panel Study survey, 65% of data warehouse adopters identified “extracting information from unstructured data” as one of their top 2 data warehousing challenges.

Corporate portals have taken the idea of consumer portals, like Yahoo! and Excite, and adapted them for corporate intranets. These portals partition the “real estate” of the user’s screen, running multiple applications side by side. The burden is placed upon the user to sort out any semantic or data-quality inconsistencies between the meaning of information displayed in one part of the screen (via one application) and that on another part of the screen (via another application).

Advances in text mining, concept extraction, and content management have the potential to allow portals to become more effective. Yet, we expect portals to remain bound by their current limitations, such as keyword search and document orientation, until these advanced capabilities are incorporated.

**Evolution of the Internet as a Tool for Data Access**

Back when the Internet was called ARPANET, people primarily moved information back and forth to each other using email and
newsgroups. Each newsgroup focused on a particular topic. Participants limited extraneous information through culturally enforced “Internet etiquette.” Context was carried in each newsgroup participant’s head.

As the Internet expanded, newsgroup traffic grew to the point that newsgroups became clogged with superfluous junk, and Internet etiquette broke down. To make newsgroups less noisy and more relevant, people resorted to moderated newsgroups and list servers. One technology, gopher, attempted to organize data on the Internet through a unified index. Gopher also introduced a search engine, Veronica. In 1994, Veronica could do keyword-in-title searches for all 15 million documents that existed on all Web and FTP sites. When the Internet exploded in about 1995, Veronica and gopher quickly became obsolete.

The World Wide Web arose in an attempt to make richer data like graphics available to users. In the mid-1990s, the volume of Web-based data exploded. Search engines tried to keep up with the volume, but they rapidly lost ground.

Search engines, like AltaVista, Infoseek, and Excite, sent robots around to each site. These robots do the equivalent of reading every book in a library and allowing users to search on words in the text — better than Veronica’s keyword-in-title method but still cumbersome and unable to deal with semantics.

Gradually, bits and pieces of unified data access have begun to surface. The World Wide Web Consortium (W3C) understood the importance of semantics. It responded with XML and the Resource Description Framework (RDF), significant steps forward. Other groups developed Web-based collaboration tools. Still others developed metadata managers and other more semantic approaches to data access.

The next step is to pull all these pieces together into an infrastructure that enables unified data access.

**PRAJA’s Conceptual Framework**

All business processes consist of events that occur at a time and place. For example, an automobile accident is an event that triggers a stream of data in many processes: police response, medical response, insurance claim, legal proceeding, and, perhaps, a news feed. Another example: a robot makes a weld on a car on an assembly line. That event triggers a number of messages to different information systems — process control, inventory, and accounting.

PRAJA’s foundational concept is that data is both more valuable and easier to manage if an information system maintains the integrity of data about events that make up business processes. PRAJA’s technology platform organizes each event’s data within an appropriate time and location context. This approach simplifies data integration and gives users new tools for organizing and accessing data.
PRAJA’s event concept is both tangible and amorphous. It is tangible when people identify a direct use and realize a direct benefit. It is amorphous because notions like event and context are ubiquitous — part of the invisible information infrastructure that we use intuitively and unconsciously.

The IT industry widely uses concepts like event, context, and data access. Among vendors, we see subtle semantic differences between the concepts and the words used to describe them. To make PRAJA’s concepts clear, we define and discuss three of them: event, context, and unified data access.

**Event**

*Merriam-Webster* first defines event as, “something that happens.” The dictionary then elaborates, “the fundamental entity of observed physical reality represented by a point designated by three coordinates of place and one of time.”

Within business processes, an event is any action occurring within a context of time and location that produces data. This data may be in any form, including rich media, sensory data, structured or unstructured text. As illustrated in Figure 2, an auto accident causes data to be generated in a number of different processes and data repositories. This event is something each data repository has in common, even though each repository is not aware of the others. So the event is useful in organizing and integrating the various repositories to give the user unified data access.
Context

Again, Merriam-Webster is useful. The dictionary defines context as “the interrelated conditions in which something exists or occurs.” The classic reporter’s questions are “what, where, when, who, and why?” “What” is the event. “Where, when, who, and why” provide the context that helps us understand the situation surrounding the event.

Two contextual elements, time and location, are a common reference frame for all events and, consequently, the data associated with events. So, time and location are useful for organizing data about events that is held within disparate repositories.

Further, when people talk, we use context to increase our conversational bandwidth. Here’s an example from the Pacific Northwest. At seven o’clock in the morning, a customer walks into one of the many espresso stands in Seattle and says to the person behind the counter, “Tall, skin-

ny.” The person behind the counter makes a hot, 12oz, nonfat latte and says, “three fifteen.” The customer hands over $3.15, takes the latte, and walks out.

The communication in this simple transaction depends on a lot of implicit situational information. For example:

- Espresso is a drink made by forcing steam at about three atmospheres of pressure through packed, finely ground coffee.
- A latte is a drink made by mixing espresso with steamed milk.
- The transaction took place in Seattle, where espresso drinks are commonplace and well understood.
- The transaction took place in an espresso stand.
- Both the customer and the person behind the counter are adults rather than children.
- The person behind the counter is trained to make espresso drinks and is in the espresso business.
- The person who walked into the business is a customer who wanted an espresso drink.
- At 7:00 a.m., most people want hot rather than iced espresso drinks.
- Lattes are the most common espresso drink, so a drink order without the name of another drink is for a latte.
- “Tall” is jargon for a 12oz espresso drink.
- “Skinny” is jargon for nonfat milk.
- In a point-of-sale transaction in the United States or Canada, 315 expressed as “three fifteen” rather than three hundred fifteen or “three one five” implies $3.15.
Now, imagine walking into an East Coast diner and saying “Tall, skinny” to the person behind the counter. How about trying the same conversation in a health club or with your next-door neighbor? The point is that this brief transaction was successful in Seattle because both people understood the situational context.

Unfortunately, this ability to organize data within context does not transfer well to people interacting with information systems. Most information systems strip contextual information like time and location from data. Consequently, information systems cannot take advantage of the context we find so useful in our person-to-person dialogues.

By improving an information system's ability to provide data in context, we increase the richness of communication between the people who use the system and thus make the information system more effective in communicating with users.

**Unified Data Access**

Within the IT industry, “unified data access” gets thrown around loosely. In older client-server environments, unified data access meant that the user could run different applications against a database. Portal vendors partition the “real estate” of the user’s screen, running multiple applications side by side. The user still has to sort out any semantic or data-quality inconsistencies between the meaning of information displayed in one part of the screen (via one application) and that on another part of the screen (via another application).

Under PRAJA’s definition, unified data access means that data about events is organized by contextual elements, like time and location. So, the user sees data within a consistent reference frame.

**PRAJA’s Data-Access Architecture**

To achieve unified data access, PRAJA pulls together many elements from the leading edge of information technology and standards. PRAJA frames its approach within the following four primary elements:

- Domain model
- Metadata manager
- Navigation/search engine
- Personal interfaces

**Domain Model**

A domain model represents a shared and common understanding of a domain that can be communicated among people and information systems.

In information systems, domain models are frameworks for industries and their business processes. These models separate the fundamental
logic behind an information system from the specifics of the particular software that implements it. A domain model can help applications separate data that is relevant to the domain from data that is not.

PRAJA supports XML and the related RDF as a standard way to represent domain models. They model nonstructured data and rich data types through their Metadata Manager.

**Metadata Management**

PRAJA defines metadata as an information mechanism that provides the user a common and integrated view of an organization’s information within a consistent contextual framework.

As shown in Figure 3, PRAJA’s Metadata Manager sits over a set of data structure definitions that describe the structure of each data source. The metadata manager maps these data structure definitions to a unified index. Constraint tables and a rules engine define the appropriate uses for each data source. Each user sees the most current data because metadata linkages are dynamic.

Correctly implemented, PRAJA’s metadata facility has the following capabilities:

- Aggregates all data that is potentially relevant to a user’s requirements, regardless of the data’s location or form; this includes structured and unstructured text, video, audio, or any other relevant data type
- Captures appropriate contextual information for each transaction
- Prescribes appropriate information exchange among other enterprise applications
- Provides a dynamic dictionary of data structures and data context represented by data elements
- Provides data history — captures the history of data over time
- Facilitates data exploration and extrapolation through integrated search and navigation

**Integrated Navigation and Search**

Current search approaches are inherently limited to locating keywords within text. Search results are presented in a sequential list, removing contextual information like time and location.

Integrated navigation and search makes the query and presentation environment the same. The user can look at the results of a query and then refine the query, rather than starting a new query from scratch. This approach reduces the latency in query processing and results display. So, the user can more quickly find the right results. PRAJA’s navigation/exploration technology gives the user options to organize and display results by time, location, or in a traditional sequential list.
Personal Interfaces

PRAJA provides a customizable user interface (UI) that includes time and location organizers. It also allows standard HTML/XML interfaces to any standard browser. PRAJA's UI development tools let developers create interfaces that match each user's individual requirements.

Role for PRAJA Within Business Processes

As shown in Figure 4, unified data access can facilitate a number of applications within a business process. For PRAJA, this is both good and bad news. PRAJA can potentially address a wide range of application areas and markets. But each prospective application area requires PRAJA to have at least some specific expertise and experience. To its credit, PRAJA has implemented several vertical applications (see www.PRAJA.com for details).
Market for PRAJA’s Technology

In vertical markets, PRAJA faces market requirements similar to those faced by analytic applications vendors. It follows, then, that the market dynamics that drive PRAJA’s market are similar to those that drive the analytic applications market. If PRAJA can address these significant challenges and execute its business plan, we expect that PRAJA can participate in the analytic applications market. Within the analytic applications segment, we see interesting possibilities for PRAJA within two fast growing sectors: customer relationship management (CRM) analytics and supply chain management (SCM) analytics.

IDC estimates that the worldwide packaged analytic applications market will increase at a compound annual growth rate (CAGR) of 20% and reach revenue of about $6 billion by 2005.

We cannot predict how much revenue PRAJA will capture or how fast it will grow. We can say that its revenue potential is not constrained by its market space and that it is positioned for favorable growth.

Challenges

We see three specific challenges for PRAJA as it introduces its concepts and technology to the market:

- **Maintaining focus.** PRAJA’s technology platform can be used for a wide range of applications. The good news is that the company has a large addressable market. The bad news is that it cannot address a broad market without first focusing on a few areas to develop expertise and revenue. PRAJA will have to maintain a tight focus and execute cleanly to survive.
• **Changing established assumptions.** As we discussed earlier in this paper, PRAJA’s technology challenges a number of established assumptions about data access and data management. To enter the market, PRAJA will have to demonstrate striking success each time they implement a new system or establish a relationship with a partner until the IT industry believes that its approach is viable.

• **Building a strong value-added reseller/original equipment manufacturer (VAR/OEM) channel.** Because PRAJA’s technology complements other applications, it will have to establish a strong VAR/OEM channel with those complementary vendors. PRAJA’s recent partnership with IBM’s Global Services group is a solid step in the right direction.

**Conclusion**

PRAJA’s technology platform provides unified data access within a contextual framework of time and location. Its approach is a significant step forward from current data access approaches, like enterprise portals, federated search, metadirectories, and enterprise data integration. PRAJA has been able to solve data-access problems that have vexed large companies for years.

PRAJA’s primary challenge is staying focused and executing its business plan. To its credit, the company has implemented significant systems at companies like GM and Zurich Financial Services. If PRAJA continues along its current path, it will be a company to watch.