A discussion of information integration solutions
November 2005

WebSphere® software

Deploying integration services with WebSphere DataStage SOA Edition.

Contents

Leveraging your IBM WebSphere DataStage investment in a service oriented architecture

A product primer

To compete effectively in today's increasingly crowded markets, companies are shifting their focus to reducing costs and increasing sales through business agility, responsiveness, and the timeliness of information. Businesses need to coordinate activities amongst analytical, operational, and transactional systems, but historically this has been very challenging, hindered by both technological and organizational obstacles.

Much of the recent and well documented innovation and investment in reducing the latency within businesses has been around what Gartner has termed the "Real-time Enterprise.[1]" Despite the recent industry buzz, the concept is not new. It has been the fundamental basis of information technology from the time of its inception. What have changed are the underlying technologies and the renewed emphasis on delivering information in a form to which the business can quickly react – whether via automated or manual response. The timing of the delivery is not always actually real-time, but rather "right-time" – tailored to the appropriate response requirements of the information.

Traditionally, response to business events has been predicated on the ability to recognize these events within the semantics and scope of transactional and operational systems. The fundamental problem with this is that the source systems are limited in their ability to recognize things that might be interesting to the business, particularly when those things transcend the boundaries of a single system. Business process management has enhanced the ability to recognize business events by providing a broader cross-application scope, and some degree of shared semantics. Although new categories of business events can be defined and recognized using this technology, there are still large categories of information, and thus potential business events, which are not considered.

Latency is typically defined as the elapsed time between a business event and an appropriate action or response. So the keys to reducing latency are tied to improving the ability to recognize business events, and the ability to respond to those events. There are actually two related components working together: The first is a process issue manifested in the time lapse from when a business event occurs to when information on the event is recognized. The second component is a process issue manifested in the time lapse from recognition to when an action is taken.

Meanwhile, data-centric technologies have focused on providing context to the data that originates from the transactional and operational systems. This context is provided by separating the data from the source systems, organizing it together in new ways, and looking at its change behavior across time.

The interesting dichotomy here is the fact that in most companies data-centric and process-centric efforts have been separated, with little or no sharing of knowledge, technologies, or approach. The organizational groups who control the transactional and operational systems are often completely separate from those that control the analytical systems, with each side using different tools and technologies. The data-centric approach has focused almost entirely on creating information in a format in which the business can very clearly recognize important and interesting events, while relegating the response to delayed and mostly manual efforts. The process-centric approach has focused on shortening response times to the limited events it can recognize. Clearly, the optimal scenario would involve harnessing the broader and more business-tailored information from the data-centric approach to allow the process-centric approach to respond to it. This paper provides a road map for IBM WebSphere® DataStage® customers to leverage their existing software and skill set investments to achieve this goal and gain greater business agility.

True real-time recognition and response requires overcoming the barriers to analytical, operational, and transactional coordination

Analytical data has been traditionally created in batch because the structure of the data is entirely different from transactional and operational systems, optimized for specific types of analysis. This structure is complicated enough that creating it becomes a very complex and processor-intensive transformation task. In addition, the data is usually derived from multiple source systems, requiring additional matching and merging to arrive at a unified view. It is also often viewed in multiple dimensions (e.g. across time), to provide additional context to the information, resulting in additional processing requirements. All of this processing has dictated that the analytical data is stored separately and that the extraction and transformation processes are completed during batch cycles.

Beyond this data latency issue, is the organizational separation of the groups that control the transactional and operational systems from the groups that control the analytical systems. The access knowledge and technologies for getting to analytical data does not exist on the process side of the organization, and the analytical technologies do not support the same standards. Process-oriented technologies therefore do not have the ability to include analytical data transformation logic as a component of an event process or transaction. In order to overcome these issues, the analytical data and data creation routines need to be published to the process-centric groups in technologies and standards which fit into their enterprise architectures and which do not compromise their real-time performance requirements.

WebSphere DataStage SOA Edition can overcome these technological and organizational barriers

With the 7.0 release of the IBM WebSphere Data Integration Suite, IBM introduced IBM WebSphere RTI™, which harnesses the extensive transformation capabilities of WebSphere DataStage and IBM WebSphere QualityStage™, along with their parallel processing power and publishes them in a standard service oriented architecture (SOA). The 7.5 Release expanded on this baseline, adding a higher level of reliability and extending SOA support to IBM WebSphere DataStage TX. These capabilities have now been packaged together to create the SOA Editions of WebSphere DataStage, WebSphere DataStage TX, and the WebSphere Data Integration Suite.

IBM WebSphere DataStage SOA Edition removes the traditional barriers between data-centric and process-centric projects, and provides a way for analytical processing to be harnessed within real-time business processes. The net result is the ability to automate response across the broadest scope of recognized business events.

WebSphere DataStage SOA Edition enables data-centric processing to occur in real-time. This is possible due to the unique combination of the massively parallel processing capabilities of WebSphere DataStage and the high volume connectivity capabilities of the SOA Edition's J2EE foundation. By leveraging these, WebSphere DataStage SOA Edition is capable of handling the transformation processing demands of analytical data creation, and the request volume and time sensitivity of real-time processing. WebSphere DataStage SOA Edition publishes complex data transformations and access routines as Web services, Enterprise Java Beans (EJBs), or Java Messaging Services (JMS) objects that can be called by Java or .Net application infrastructures, Enterprise Application Integration platforms, Business Process Management tools, or almost any of the products and technologies used within the transactional and operational arenas.

IBM customers are taking advantage of WebSphere DataStage SOA Edition capabilities by blurring the lines between analytical and transactional/ operational data. The result is the availability of the best data at all times, to all people or processes. IBM is consistently seeing four general categories of usage for the SOA Edition.

On Demand data warehousing

The value of analytical data in managerial decision-making has led to increasing pressure to reduce the latency of this data. In a recent TDWI survey, 50% of respondents said they were deploying or planning areal-time data warehouse, and Gartner estimates that by 2006, approximately 30% of BI deployments will require instantaneous data.[2] Most data managers have heard the complaints of executives asking for more up-to-date information. This is often exacerbated when managers see different results in reports originating from transactional and operational systems than what they see in their analytical reports. At the same time, this data is useful within the context of many processes and applications, not just to support human decision-making.

On Demand data warehousing takes advantage of WebSphere DataStage SOA Edition by reducing or eliminating the latency of data within the warehouse and publishing data from the data warehouse in services that are easily consumed by applications and processes. As data is created within source systems, the SOA Edition is able to push this data into WebSphere DataStage for transformation and population to operational stores, warehouses and data marts on an event-driven basis, triggered from applications, Enterprise Application Integration (EAI), or business processes. This allows companies to take advantage of the best-in-class transformation and processing capabilities of WebSphere DataStage without having to wait for a batch window. It also allows data experts to publish services for their data that are easily consumed by applications and processes, without requiring application developers to understand the complex schemas and sources associated with the data warehouse.

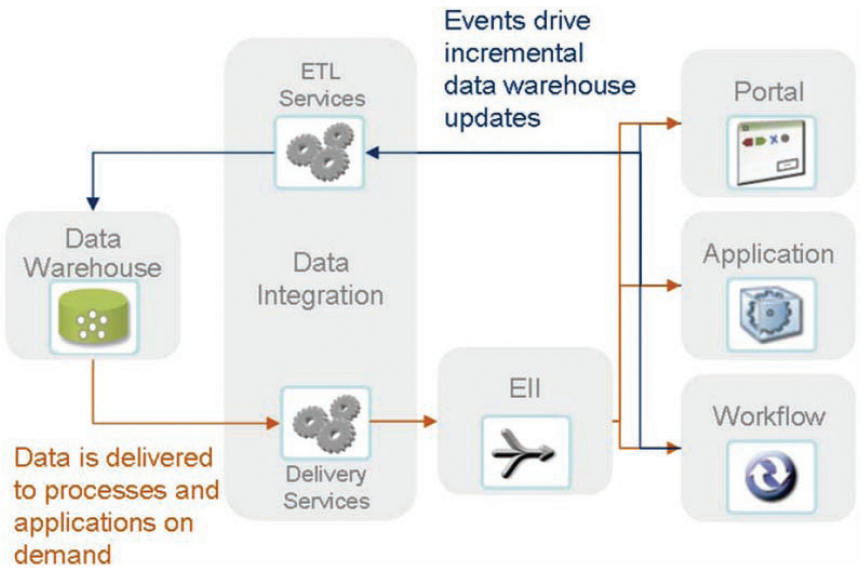

Figure 1. On Demand Data Warehousing with WebSphere DataStage SOA Edition

One of the unique capabilities of WebSphere DataStage SOA Edition is its ability to leverage existing WebSphere DataStage job logic. This means that the logic for data transformation can be reused, rather than having to be re-created. In fact, the SOA Edition allows jobs to be designed without any knowledge of how they will be accessed. This prevents the job developer from having to know anything about JMS queues, SOAP message headers, or EJB structures. It also allows for a quick and smooth transition to on demand data warehousing using WebSphere DataStage, without a great deal of rework or re-design.

With less inherent data latency across the spectrum, WebSphere DataStage can provide data integration services across analytical and transactional/ operational stores. This allows analytical and operational data to be brought closer together, allowing automated operational decision-making to leverage richer analytical data. These data integration services can be easily called from any application, portal, or development project. They can also be called from Enterprise Information Integration (EII) platforms to provide advanced data matching and transformation services to federated queries. Once in a familiar service oriented structure, external applications and development teams are more likely to take advantage of these services to get the best available data and access it in a standard way.

A good example of this can be found in a leading pharmaceutical company who is using the IBM platform to reduce the latency in operational and analytical data stores. This company is using Web services interfaces within WebSphere DataStage processes to access various data sources within and outside the enterprise. They are also triggering WebSphere DataStage processes through Web services calls from their application middleware. By reducing the latency of their operational and analytical systems, they are able to make better decisions. They provide access to this information to their partners through secure Web services, allowing their customers and suppliers to get an “on demand” view of information that is vital to optimizing their supply chain.

Master data management

One of the first things that companies discover when trying to reduce process latency is that their most vital data is often stored across many systems, with little or no consistency. This forces development projects to go to extraordinary lengths to reach the correct sources of data and rationalize it into a single de-duplicated semantic representation. In most cases, it also means that applications and users never have a complete picture of the data. This is commonly seen in customer marketing and customer service initiatives, where obtaining a single view of the customer remains an elusive goal for most organizations.

Master Data Management goes a step beyond on demand data warehousing, by creating authoritative sources of common reference data that can be used throughout organizational operations. The types of data typically targeted for this include things like customer data, product data, and inventory data. Enterprises often choose this common data because it is accessed frequently across many applications, and the consistency of the data is very important to the business. Creating these master stores improves the consistency and reliability of information for everyone, and allows new development efforts to reuse proven standard access mechanisms rather than re-creating them.

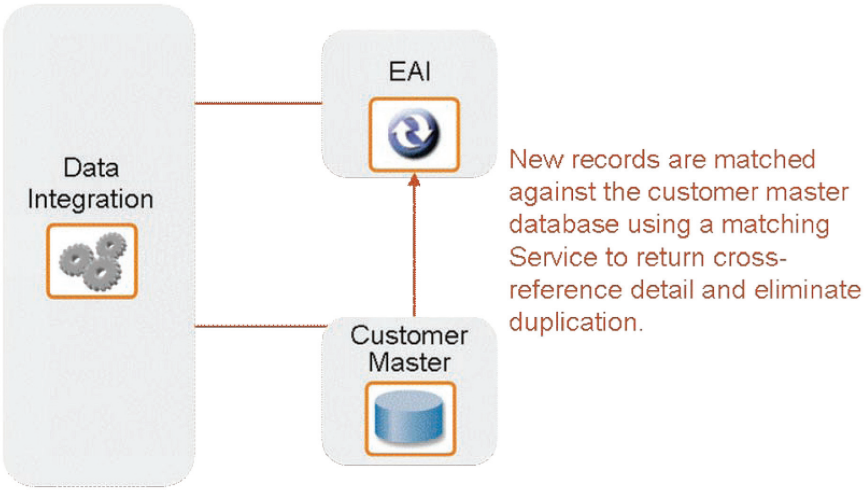

Figure 2. Master Data Management with WebSphere DataStage SOA Edition and WebSphere QualityStage

In this scenario, WebSphere DataStage is used to create a master database, which is kept up to date using the SOA Edition as transactions flow through source systems. The master database can house either complete matched records of reference data, or it can simply contain a cross-reference table of identifiers from the various source systems. In the latter case, when a resource requires the reference data, WebSphere DataStage (or an external technology like EAI or EII) dynamically assembles a complete record from the source systems using these identifiers.

In order to populate the master database, WebSphere DataStage leverages WebSphere QualityStage, with its best-in-class data matching technology. These matching routines are typically used in batch during the initial data population, and then reused in real-time to match new records as they are entered into source systems. WebSphere QualityStage is also used to ensure that data is centrally standardized, resulting in more consistent and accurate data. WebSphere DataStage is able to easily involve systems that are traditionally very difficult to reach, like legacy mainframe systems, providing a complete organizational view of matched reference data.

The advantage of this is not only to provide a single place to go to get reference data, but also to reduce the burden on development projects. Development projects are often hampered by the difficulty in understanding and accessing all of the different sources of data. Even when they know where all of the data is and can actually get to it, the task of matching and merging records of information is often very tedious and difficult. WebSphere DataStage allows data managers, the people who know the data the best, to focus on these tasks and easily create reusable services that development projects can leverage. Even better, because these services are based on a service oriented architecture, they abstract the underlying complexities of data sources and integration from the developer. This means that developers are presented with business-level services that do not require them to be experts in source systems or data. In fact, when they use these services, they don't even need to know anything about the WebSphere Suite. This insulation of application development tasks from data management tasks best leverages the competencies and skills of each group, and leads to more unified and standard systems.

An excellent example of Master Data Management can be found in one of the world's largest technology companies, who is using the IBM platform to consolidate customer information from multiple internal systems. The goal of providing a single point of reference for the authoritative data on each customer was met by first rationalizing and profiling the source systems, and then merging the existing data into a single master store. For ongoing data entered into various systems, WebSphere DataStage SOA Edition with WebSphere QualityStage is used to run the same standardization, matching, and transformation processes to control information at the point of entry. With the new database in place, the company has abetter understanding of who their customers are across all of products, services, and locations. They can also provide an improved level of support, and better tailor marketing and sales activities to their customer base.

In-flight enrichment

As companies develop new applications, the requirements for cross-system data validation, standardization, matching, and transformation continuously arise. Much of the logic required for this has already been built into existing data warehousing applications. The cost of trying to re-develop these in other technologies is very high, requiring high level developers and producing another point of maintenance when data structures eventually change.

In-Flight Enrichment publishes a more granular set of WebSphere DataStage or WebSphere DataStage TX capabilities as services that can be leveraged by development efforts. This expands on the master data management theme, by offering standard services for capabilities like validation, standardization, matching, transformation, and enrichment. Organizations publish these services to foster reuse and encourage consistent data standards. Development projects can then reuse these services, instead of trying to develop them from scratch. For example, when creating a portal application, a development team does not have to re-create an address validation routine. Instead, they can reuse one published as an in-flight enrichment routine through the SOA Edition.

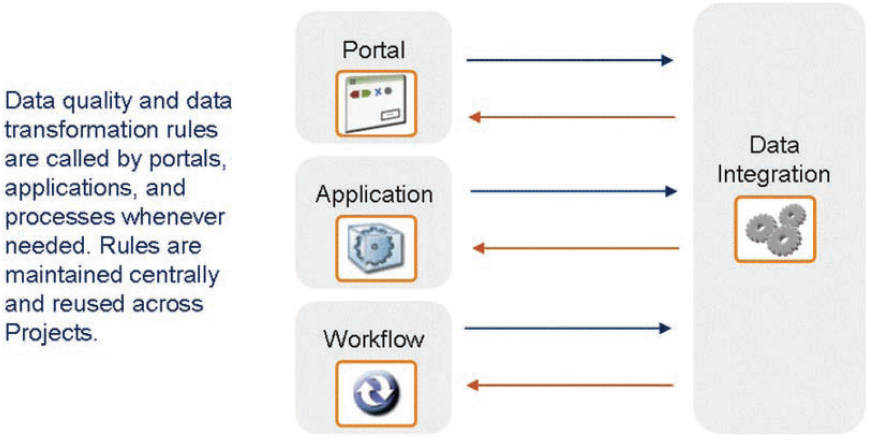

Figure 3. In-Flight enrichment with the WebSphere Data Integration Suite SOA Edition

In-Flight Enrichment services are created by using the SOA Edition to publish discreet WebSphere DataStage, WebSphere DataStage TX and WebSphere QualityStage jobs for specific data. These services can be very granular in nature, like a simple address validation, or more comprehensive, like a service that validates, standardizes, and inserts a customer record. Much of this logic already exists in analytical applications, however it is not in a form that it can be leveraged by external applications. Using the SOA Edition, this logic can be exposed as services, allowing it to be universally utilized.

The primary advantages of in-flight enrichment are the reduction of development burden and increased data consistency. Since they are based on the common standards of a service oriented architecture, these services are extremely easy to incorporate into applications. From a development perspective, these services are available in a searchable directory, effectively creating a library of data integration components that can be reused from project to project. In addition, from an ongoing maintenance perspective, changes to these routines do not necessitate changes to the calling applications. This allows these services to be maintained separately, without involving the applications teams.

Data consistency is improved by all applications sharing the same mechanisms for validation and standardization. This ensures that all data entering the database goes through the same set of rigorous requirements. In this way, the chances of errant or mismatched data are reduced.

An example of in-flight enrichment in practice can be seen in an international insurance data services company who is using IBM's platform to provide Web service based validation and enrichment of property addresses. As insurance companies submit lists of these addresses to them for underwriting, the real-time services standardize the addresses based on their rules, validate that the address is accurate, match the addresses to a list of known addresses, and enrich the addresses with additional information to assist in underwriting decision-making. As a result, they are able to automate 80% of the process of property address research, and eliminate the errors typically associated with it. All of this was made easy using the publishing capabilities of WebSphere DataStage SOA Edition and the standardization and matching capabilities of WebSphere QualityStage.

Enterprise data integration services

Within many organizations, the number and complexity of applications requiring access to data is continuously on the rise. In most cases there is very limited reuse between these applications, meaning that much of the effort, particularly in data access mechanisms, is reproduced in each project. This leads to high development costs, high maintenance costs, and inconsistencies in how data is accessed. Data managers quickly lose control of the data, and are no longer able to enforce data standards and ensure data quality.

In the Enterprise Data Integration Services scenario, businesses create a standard data integration layer within their enterprise architectures, to improve consistency, reuse, and control. This is a simple extension of the logic behind in-flight data enrichment, expanding the scope to all data. The mechanisms and the benefits are essentially the same, only magnified to encompass a larger set of data. The primary difference is the true separation of data management and application development.

This separation allows each group to focus on their core competencies, using the tools and technologies with which they are comfortable, without restricting or inhibiting the abilities of the other group. It also allows each group to independently make changes within their scope, without impacting the other group. When changes do need to cross the boundaries, metadata linkages provide an end-to-end understanding of the impact of change, ensuring that nothing is missed.

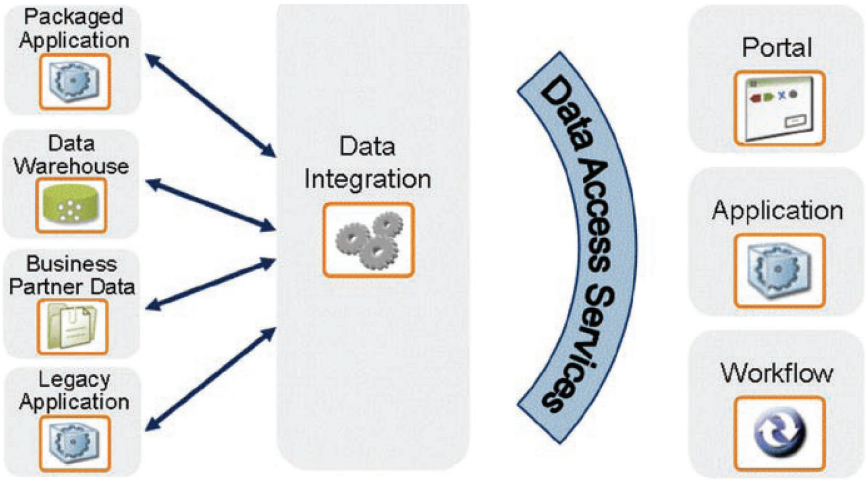

Figure 4. Data integration services with the WebSphere Data Integration Suite SOA Edition

Enterprise Data Integration Services truly delivers on the promise of service oriented architectures, taking full advantage of the inherent capabilities. It also fully leverages the capabilities of the WebSphere Data Integration Suite. The net result is better standardization and consistency, reduced development cycles and project risk, lower cost and risk in maintenance, and overall better and more available information for business users.

Enterprise Data Integration Services are particularly pertinent to the burgeoning area of regulatory compliance. Many of the regulations that companies are facing involve understanding the content and lineage of data, and controlling how it is used. For example, Sarbanes-Oxley includes a strong focus on understanding how financial data was calculated and where it came from. This is an ideal scenario in which to apply enterprise data integration services. Not only do these services ensure a consistent enforcement of logic to data, but they also provide a metadata audit trail for where the data came from, and what happened to it along the way. In addition, policies can be applied to the services to ensure that they are used appropriately and securely throughout the enterprise.

An example of an IBM customer who is already implementing enterprise data integration services is one of the largest automotive manufacturers in the world. This company is consolidating all of their data integration logic into a common services layer that can be reused across their EAI infrastructure and their ETL processing, and can be called directly from their applications. This reduces the development burden on individual projects by simplifying data integration tasks to a simple service call, and providing a library of reusable components. It also improves the consistency of data, since the same integration logic is applied, regardless of the project. WebSphere DataStage greatly simplifies the task of accessing and integrating these data sources using a visual development paradigm rather than hand-coding. The result is a tangible reduction in project development effort, and a higher level of consistency in data across the enterprise.

Getting started

Although these examples may seem like fairly large-scale undertakings, each of them was implemented on an incremental basis, with iterative project goals producing short-term results. In most cases, these iterations produced tangible results within days or weeks, quickly showing a positive return on investment.

A successful incremental approach is predicated on selecting the appropriate starting point. Many customers begin by adding real-time capabilities to an existing warehouse implementation, leveraging existing ETL logic. Other customers focus on specific opportunities related to master data management or in-flight enrichment. In any case, it is important to choose a manageable unit of work that can quickly demonstrate business value aligned with your ultimate objectives.

A good place to start is to simply ask the question, "Would my operational processes benefit from having better access to pieces of my analytical data?" Often there are multiple pieces of analytical data that would be valuable to operational systems, but in the past there wasn't an easy way to access them. Business users can usually quickly identify these data elements, so they are great resources to identify starting points. Publishing services that provide this data to operational processes whenever they need it can be extremely valuable to the business. Managers of analytical systems, business intelligence systems, and data warehouses can approach application integration groups with these services and offer them as valuable components to help in operational processing. Once the initial service oriented relationship is started, additional areas of synergy will naturally flow out of it.

WebSphere QualityStage often offers another good starting point, particularly for in-flight enrichment scenarios. Often Web applications, portal projects, or application development projects can immediately benefit from reusing existing validation and standardization routines that have been built in WebSphere QualityStage.

The first project should focus on a discreet number of related capabilities that can be delivered in a short time, usually within 30 days. It should also attempt to leverage existing WebSphere DataStage, WebSphere QualityStage, or WebSphere DataStage TX logic as much as possible. Subsequent projects will expand on this baseline, providing a more complete set of functionality.

The advantages of improving your business agility are clear, while the means are sometimes difficult to achieve. Business agility can only be improved by reducing the time it takes to recognize business events, and the time it takes to respond with an appropriate or better defined reaction. This requires coordination amongst analytical, operational, and transactional systems, which has historically been hindered by both technological and organizational obstacles. The IBM platform, with the SOA Edition, can assist in overcoming these obstacles by providing the technological capabilities required to bridge these systems, and publishing the results as services that application developers will want to use.

Next steps

WebSphere DataStage SOA Edition is available starting in version 7.5 for WebSphere DataStage and WebSphere DataStage TX customers. For more information on upgrading to SOA Edition today, or for information on other IBM product or services offerings, contact your local IBM representative, and ask about our IBM Workshop programs for SOA.

IBM Software Group
Route 100
Somers, NY 10589
U.S.A.

Produced in the United States of America
11-05

[1] "The Real-Time Enterprise: Key Issues for 2003" January 2003, David Flint and Mark Raskino

[2] "Real-Time Management Will Demand Changes in Business Culture" April 2004, David Flint.

References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates. Offerings are subject to change, extension or withdrawal without notice.

All statements regarding IBM future direction or intent are subject to change or withdrawal without notice and represent goals and objectives only.

The IBM home page on the Internet can be found at **ibm.com**

GC-18-9725-00