Chapter 7
Challenges of Data Management in Always-On Enterprise Information Systems

Mladen Varga
University of Zagreb, Croatia

ABSTRACT

Data management in always-on enterprise information systems is an important function that must be governed, that is, planned, supervised, and controlled. According to Data Management Association, data management is the development, execution, and supervision of plans, policies, programs, and practices that control, protect, deliver, and enhance the value of data and information assets. The challenges of successful data management are numerous and vary from technological to conceptual and managerial. The purpose of this chapter is to consider some of the most challenging aspects of data management, whether they are classified as data continuity aspects (e.g., data availability, data protection, data integrity, data security), data improvement aspects (e.g., coping with data overload and data degradation, data integration, data quality, data ownership/stewardship, data privacy, data visualization) or data management aspect (e.g., data governance), and to consider the means of taking care of them.

INTRODUCTION

In everyday business we may notice two important data characteristics:

- **Every business is an information business:** All business processes and important events are registered by data and, eventually, stored in an enterprise information system's data base. In other words: if it is not registered by data, it has not happened.
- **Data is registered in digital form:** The majority of important business data is registered in digital form, e.g. the sales data collected at point of sale, the transaction data at automated teller machine, etc. They are all memorized in digital form in various data bases.

The underlying task of an enterprise information system (EIS) is to link processes on the operational,

DOI: 10.4018/978-1-60566-723-2.ch007

management and decision-making level so as to improve performance efficiency, support good quality management and increase decision-making reliability (Brumec, 1997). The EIS's database, specifically its organization and functionality, play a critical role for the functional use of data and information assets in an organization.

Data management involves activities linked with the handling of all organization's data as information resource. The Data Management Association (DAMA) Data Management Body of Knowledge (DAMA, 2008) defines data management as "the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets."

Always-on EIS supports business agility which is the ability to make quick decisions and take actions. "Agility is the ability of an organization to sense environmental change and respond efficiently and effectively to that change" (Gartner, 2006, p. 2). The measurable features that enable a business system to increase the agility of its performance can be defined as follows:

- **Awareness is knowing what is going on:** Awareness level can be determined by answering these questions: Do end users see the right information at the right time? Is the information easily accessible to the right people?
- **Flexibility is the ability to respond appropriately to expected changes in business conditions.**
- **Adaptability is the ability to respond appropriately to unexpected change:** Does the structure of business data promote or prevent flexibility and adaptability is the key question regarding adaptability and flexibility.
- **Productivity is the ability to operate effectively and efficiently:** It is important to establish whether or not business data increase the efficiency and effectiveness of business operations and decisions.

Effective data management in an EIS is essential to fulfil these tasks. This chapter considers various aspects, possibly not all, of data management that seem to be important for running an

Table 1. Data management aspects and challenges

	Aspect	Challenge
Data continuity	Data availability	Is data available all the time?
	Data integrity	Is data integrity compromised?
	Data security	Is data secure enough?
Data improvement	Data overload	Can we cope with all that data?
	Data integration	How to integrate data? Could the data integrate the business? Do we know organization's data? Do we know how to use organization's data?
	Data quality	Do we know what quality data is? Are we aware that the quality of data is degradable?
	Data ownership/stewardship	Who is the owner/steward of the data?
	Data privacy	Is data privacy a concern?
	Data visualization	Could we amplify cognition of data?
Data management	Data management	Do we need data governance?

EIS. Table 1 shows the considered aspects and challenges classified as data continuity aspects, data improvement aspects, and data management aspects.

The challenges of successful data management vary from technological to conceptual. Technological aspects of data management help business continuity by making EIS data available, complete, consistent, correct and secure. The aspects addressing the problem of EIS data continuity are described in the section on data continuity.

Once the technological aspects are solved, business may run smoothly. Nevertheless, the business may run even better if it is innovated constantly and persistently. Conceptual aspects deal with important data management issues that can improve business data: coping with data overload, solving data integration problems, improving data quality, assigning data ownership/stewardship, maintaining data privacy, and improving insight into data. In fact, these aspects address the issue of EIS data improvement and are dealt with in the section on data improvement.

DATA CONTINUITY

Data Availability

Challenge: Is Data Available All the Time?

Enterprises use their EIS based on information technology (IT) infrastructure to increase process productivity, empower users to make faster and more informed decisions, and consequently provide a competitive advantage. As a result, the users are highly dependent on EIS. If a critical EIS application and its data become unavailable, the entire business can be threatened by loss of customers and revenue. Building of EIS that ensures high data availability is critical to the success of enterprises in today's economy.

Data availability is the degree of availability of data upon demand through application, service or other functionality. Data availability importance varies among applications. It is more important if an organization depends on data to provide business service to its customers. Data availability is always measured at an applications' end and by end users. When data is unavailable the users experience the downtime of their applications. Downtime leads to lost business productivity, lost revenue and it ruins customer base. The total cost of downtime is not easy to determine. Direct cost, such as lost revenue and penalties when service level agreement (SLA) objectives are not met, can be quantified but the effects of bad publicity and customer dissatisfaction are hard to determine.

Setting up a data availability strategy may help to define the aims and procedures to achieve the required high data availability. The steps of data availability strategy involve determining data availability requirements, planning, designing and implementing required data availability. Data availability requirements analysis begins with a rigorous business impact analysis that identifies the critical business processes in the organization. Processes may be classified into a few classes:

- Processes with the most stringent high data availability requirements, with recovery time objective (RTO) and recovery point objective (RPO) close to zero, with the system supporting them on a continuous basis, such as internet banking. RTO (Oracle, 2008; IBM, 2008) is maximum amount of time that a business process can be down before the organization begins suffering unacceptable consequences, such as financial loss. RPO is the maximum amount of data a business process may lose before severe harm to organization results. RPO is a measure for data-loss tolerance of a business process. It is measured in terms of time, for example one hour data loss.

- Processes with relaxed data availability and RTO and RPO requirements, with the system that does not need to support extremely high data availability, such as supply chain.
- Processes that do not have rigorous data availability requirements, such as internal organization's business processes.

The challenge in designing a highly available IT infrastructure is to determine all causes of application downtime. Causes of downtime may be planned, such as system or database changes, adding or removing processor or node in cluster, changing hardware configuration, upgrading or patching software, planned changes of logical or physical data structure; and unplanned, such as computer and storage failure, data corruption in database, human errors etc.

Important factors that influence data availability are:

- **Reliability of hardware and software components of the EIS:** hw/sw reliability
- **Recoverability from failure if it occurs:** data protection
- **Timely availability problem detection:** data problem detection

Hw/Sw Reliability

Storage management is very important because data resides on storage media or storage devices. Disk drives or various disk subsystems, such as RAID (Redundant Array of Independent Disks) or JBOD (Just Bunch Of Disks), are dominant storage media. Although disk drives are very reliable (Mean Time Before Failure - MTBF is almost one million hours) their mechanical nature makes them the most vulnerable part of a computer system. If a storage system includes hundreds or thousands of disk drives, data availability problem becomes very severe. Secondary storage media for backup and archive purposes include removable storage, optical storage, non-volatile storage, solid state disks, and tape devices.

Modern IT architecture, referred to as cluster or grid computing, has a potential to achieve data availability. Cluster/grid computing architecture effectively pools large numbers of servers and storage devices into a flexible computing resource. Low-cost blade servers, small and inexpensive multiprocessor servers, modular storage technologies, and open source operating systems such as Linux are raw components of this architecture.

High hw/sw reliability may be achieved by using many features. To give some indication on them and not intending to diminish the features of other vendors, here is a non-exhaustive list of Oracle's features (Oracle, 2008):

- **Efficient storage management, such as Oracle's Automatic Storage Management:** The feature spreads database files across all available storage and simplifies the database files management.
- **Redundant storage with high availability and disaster-recovery solution that provides fast failover, such as Oracle's Data Guard:** The feature maintains standby databases as transactionally consistent copies of the primary or production database. When the primary database becomes unavailable the feature switches any standby database to the primary status, minimizing the downtime. It may be used with traditional backup and restore.
- **Fine replication, such as Oracle's Streams:** The feature includes fine-grained replication, many-to-one replication, data transformation etc.
- **Cluster or grid management, such as Oracle's Real Application Clusters and Clusterware:** The features allow the database to run any application across a set of clustered servers. This provides a high level of availability and scalability. In the case that a server fails the database continues to

run on the surviving servers. Also, a new server can be added online without interrupting database operations.

Data Protection

The main aim of data protection is to recover data from failure if it occurs. High data availability and data vitality can only be ensured using a comprehensive and reliable database backup and recovery strategy, and the ways to locate media loss or corruption. Again, some of Oracle's (Oracle, 2008) features essential to data protection are:

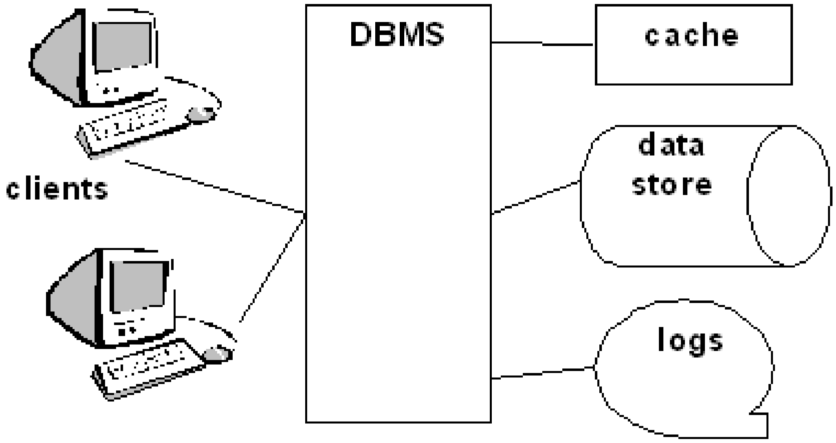
- **Backup and recovery management, such as Oracle's features Secure Backup, and Recovery Manager:** The backup feature must include all modern techniques of local or remote tape devices backup, either on calendar based scheduling or on demand. Data may be encrypted as well. Database recovery is a very demanding job and a good recovery management tool is absolutely necessary. Such a tool may be designed to determine an efficient method of executing backup, restoration or recovery, for example online or offline backup of the whole database or its constituent parts (files, blocks), fast incremental backups (only changed blocks are backuped), and incrementally updated backups (on-disk image copy backups are rolled forward in-place using incremental backups), automatic recovery to control or in time points, etc.
- **Recovery intelligence tool, such as Oracle's Data Recovery Advisor:** The tool which is able to automatically diagnose disk data failures, present repair options and run them.
- **Logical recovery features, such as Oracle's Flashback Technologies:** The feature may analyze and recover data on the row and table level and do fine granular repair, or rewind the entire database to undo extensive logical errors. It includes a fast database point-in-time "rewind" capability, historical viewing and quick recovery.

Continuous data protection (CDP), also called continuous backup or real-time backup, offers some advantages over traditional data protection. CDP data is continuously backuped by saving a copy of every data change, in contrast to traditional backup where the database is backuped at discrete points of time. In CDP, when data is written to disk, it is also asynchronously written to a second (backup) location over the network, usually another computer. Essentially, CDP preserves a record of every transaction and captures every version of the data that the user saves. This eliminates the need for scheduled backups. It allows restoring data to any point in time, while traditional backup can only restore data to the point at which the backup was taken. Furthermore, CDP may use multiple methods for capturing continuous changes of data, providing fine granularity of restorable objects. In the growing CDP products market the leaders are, among others, IBM Tivoli Continuous Data Protection for Files, Microsoft System Center Data Protection Manager, and Symantec Backup Exec.

In fact, CDP does complete journaling, which is essential for high availability (Schmidt, 2006). Journaling is a property of a database system where changes are first written to a non-volatile area before the database I/O operation has finished. Nevertheless, journaling may be restricted by aggressive cashing techniques where write operation places changes in the RAM and cache management eventually writes the data to disk. Thus, DBMS that resides on a database server, shown in Fig. 1 (Schmidt, 2006), employs persistent storage for actual data (data store), cashing for needed performance and persistent storage for auxiliary data such as logs or journaling.

A relational database supports transactions as the basic units of works. A transaction consists of a series of low-level operations that are all success-

Figure 1. DBMS and the types of storage



fully executed or not executed at all. Transactions possess ACID properties: *Atomicity* (transaction is either done completely or not at all), *Consistency* (the database does not violate any integrity constraints when a transaction begins and when it ends), *Isolation* (every transaction behaves as if it accesses the database alone) and *Durability* (the successful transaction results are not lost, but remain). In the high-availability environment durability requires special attention. It may be reduced during disaster recovery.

Data Problem Detection

High data availability can only be achieved as a result of preventive action and early data problem detection. It is advisable to use some of the intelligent tools that automatically diagnose data failures, inform users about them, and assess the impact of the failure, such as Oracle's Data Recovery Advisor.

Data Integrity

Challenge: Is Data Integrity Compromised?

Protecting data integrity means ensuring that data is complete, consistent and correct both on physical and logical level. Physical integrity means physical protection from malicious, accidental or natural damage. Logical integrity is preserved if only the permitted combination of data is registered in the data base, in keeping with the data description in the metadata repository.

Data integrity is in an always-on EIS environment, as important as data availability. Data integrity can be compromised in a number of ways. Storage media failure, including cache and controller failures, may cause corruption of data that can be more dangerous than a failure that causes storage to be offline. Corrupted data may be copied during backup operation to backup media without being detected. Consequently, it is extremely important to check data integrity and detect data corruption before it is too late.

Human and logical errors are always possible. For example, a user or application may erroneously modify or delete data. In contrast to physical media corruption, logical errors are difficult to isolate because the database may continue to run without any alerts or errors. Focused repair techniques are thus necessary to combat logical errors.

Repair technology may use the same features that are used in data protection. Again, some of Oracle's features (Oracle, 2008) are:

- Intelligent tool that automatically diagnoses data failures, presents recovery options, and executes recovery, such as Oracle's Data Recovery Advisor.
- Backup and recovery management, such as Oracle's features Secure Backup, and Recovery Manager.
- Logical recovery features, such as Oracle's Flashback Technologies.
- Auditing tool, such as Oracle's Logminer, which can be used to locate changes in the database, to analyse data in the database, and to provide undo operation to rollback logical data corruptions or user errors?

Data Security

Challenge: Is Data Secure Enough?

Data security must ensure that access to data is controlled and that data is kept safe from corruption. The broader and more commonly used term is information security, which is the process of protecting information, and information systems, from unauthorized access, use, disclosure, destruction, modification, disruption or distribution (Allen, J. H., 2001). Protection of confidential data, and information, is in many cases a legal issue (e.g. bank accounts data), ethical issue (e.g. information about personal behaviour of individuals), and business issue (e.g. business enterprise customers' data). For individuals, information security has a significant effect on privacy, which is described in the section on data privacy.

The key features of information security are confidentiality, integrity and availability. *Confidentiality* means that the system must not present data to an unauthorized user or system. Confidentiality is usually enforced by encrypting data during transmission, by limiting the number of places where data might appear, and by restricting access to the places where data is stored. *Integrity* means that data cannot be modified without authorization. *Availability* means that data must be available when it is needed. High availability systems aim to remain available at all times, preventing service disruptions due to power outages, hardware failures, and system upgrades.

Data Protection Act is used to ensure that personal data is accessible only to the individuals concerned. Data should be owned or stewarded so that it is clear whose responsibility it is to protect and control access to data. In the security system where the user is given privileges to access and maintain an organization's data, the principle of least privilege should never be overlooked. This principle requires that a user, i.e. user's programs or processes, is granted only the privileges necessary to perform tasks successfully. Organizations must regularly update their user's privileges database, especially when the user is moved to a new job or leaves the organization.

Due to high probability that security threats will appear during EIS lifecycle it is absolutely necessary to establish a security policy, develop and implement security plans and programs, and manage security risks. Many security techniques are standardized. International Standardization Organization (ISO), the world's largest developer of standards, published the following standards: ISO-15443: Information technology - Security techniques - A framework for IT security assurance; ISO-17799: Information technology - Security techniques - Code of practice for information security management; ISO-20000: Information technology - Service management; and ISO-27001: Information technology - Security techniques - Information security management systems. Professional knowledge may be certified as Certified Information Systems Security Professional (CISSP), Information Systems Security Architecture Professional (ISSAP), Information Systems Security Engineering Professional (ISSEP), Information Systems Security Management Professional (ISSMP), and Certified Information Security Manager (CISM).

DATA IMPROVEMENT

Data Overload: Exploding Data

Challenge: Can We Cope with All that Data?

(IDS, 2008) gives the forecast of worldwide information growth through 2011. It estimates that the digital universe in 2007 numbers 2.25 x 10^{21} bits, and by 2011 it will be 10 times the size it was in 2006. Approximately 70% of the digital universe is created by individuals, but enterprises are responsible for the security, privacy, reliability, and compliance of 87% of the digital universe. The amount of data content is increasing to the point that we individually and collectively suffer from data overload.

Experiments and experience in decision making processes show that data overload negatively impacts decision performance. Decision makers face three problems:

- **problem of timeliness:** high volumes of data constrain them if they are performing both sense making and decision making tasks,
- **problem of throughput:** high volumes of data overwhelm and distract them during sense making tasks, causing "analysis paralysis" and lowering the overall performance, and
- **problem of focus:** decision makers have a finite amount of attention that may be distributed across tasks.

Data fusion is an example of a technique to cope with data overload. By means of fusion, different sources of information are combined to improve the performances of a system (Hall & Mcmullen, 2004). The most obvious illustration of fusion is the use of various sensors, typically to detect a target. The different inputs may originate from a single sensor at different moments (fusion in time) or even from a single sensor at a given moment. In the latter case several experts process the input (fusion on experts).

As a response to the question "Can we ever escape from data overload?" Wood, Paterson & Roth (2002) suggest that the problem can be resolved by cognitive activities involved in extracting meaning from data. They argue "that our situation seems paradoxical: more and more data is available, but our ability to interpret what is available has not increased" (p 23). We are aware that having greater access to data is a great benefit, but the flood of data challenges our ability to find what is informative and meaningful.

Reduction or filtration of data does not seem to solve the problem of data overload. Data which is evaluated as unimportant and non-informative and thus eliminated may turn out to be critically important and informative in another particular situation. Ironically, if the benefit of technology is increased access to data, reduction of data discards some of accessed data.

Data are informative by "relationship to other data, relationships to larger frames of reference, and relationships to the interests and expectation of the observer. "Making data meaningful always requires cognitive work to put the datum of interest into the context of related data and issues" (Wood, Paterson & Roth, 2002, p. 32). Therefore, all approaches to overcome data overload must involve some kind of selectivity. Between positive or negative selectivity we must choose positive. "Positive selectivity facilitates or enhances processing of a portion of the whole" (Wood, Paterson & Roth, 2002, p. 32). Negative selectivity or filtering, which is commonly used, inhibits processing of non-selected areas. "The critical criterion for processes of selection, parallel to human competence, is that observer need to remain sensitive to non-selected parts in order to shift focus fluently as circumstances change or to recover from missteps" (Wood, Paterson & Roth, 2002, p. 32). In conclusion, data overload can only be effectively resolved using a positive form of selectivity and techniques that support focus shifting over the field of data. Consequently, in order to overcome data overload, data processing should create context sensitivity rather than insist of data finesse.

Data Integration

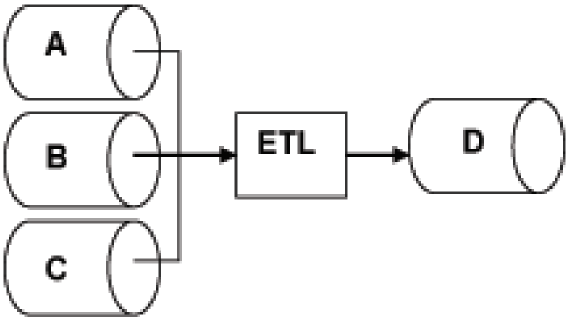
In the eighties of the 20th century we all hoped for a single integrated enterprise's database that would be based on a stable schema. These hopes have never been realised in practice and probably never will be. Data integration techniques are gaining increased importance as a result. Data integration, in a broad sense, is a process of integrating mutually related data which resides at autonomous and heterogeneous sources, and providing a unified view of integrated data through a unified global schema (Halevy, 2001). The need of data integration increases with the overall need to share existing data. The exploding abundance of data and new business needs requires the data to be integrated in order to extract new information and gain new knowledge. Data integration aims to answer complex questions, such as how some market factors influence the marketing strategy of certain products. From the managerial view, data integration is known as Enterprise Information Integration.

Challenge: How to Integrate Data?

Although the data integration problem has been theoretically extensively considered both in business and science, many problems still remain to be solved in practice. Nonetheless, we may recognize three basic types of data integration approach: federation approach, warehouse approach and mediator approach.

In *federation approach* each of data sources talks independently through wrapper to other data sources in order to be mutually "integrated".

Figure 2. Data warehouse approach



Data warehouse approach, which is frequently and successfully used in many commercial systems oriented on data analysis and decision support, is shown in Fig 2. Information is extracted from source databases A, B and C to be transformed and loaded into the data warehouse D by the process known as ETL. Each of data sources A, B or C has its own and unique schema. After extracting data from A, B and C, data may be transformed according to business needs, loaded into D and queried with a single schema. The data from A, B and C are tightly coupled in D, which is ideal for querying purposes.

Nevertheless, the freshness of data in D constitutes a problem as the propagation of updated data from the original data source to D takes some time, called latency time. As business becomes more real-time, the system that supports it needs to be more real-time. The challenge is how to make a real-time data warehouse. Here are some techniques for making data warehouse more or less real-time (Langseth, 2004):

- **"Near real-time" ETL:** The oldest and easiest approach is to execute the ETL process again. For some applications, increasing the frequency of the existing data load may be sufficient.
- **Direct trickle feed:** This is a true real-time data warehouse where the data warehouse is continuously fed with new data from data sources. This is done either by directly

inserting or updating fact tables in the data warehouse, or by inserting data into separate fact tables in a real-time partition.

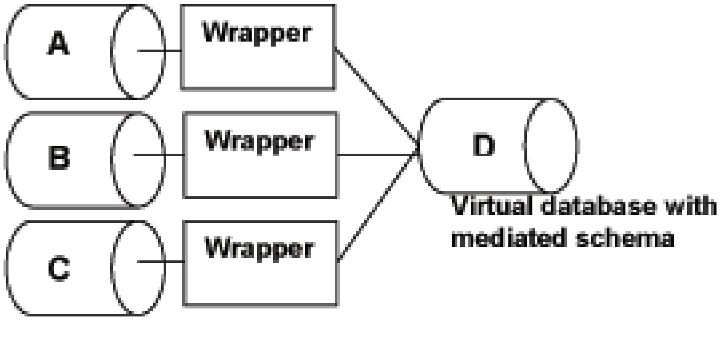
- **Trickle and feed:** The data is continuously fed into staging tables and not directly into the actual data warehouse tables. Staging tables are in the same format as the data warehouse actual tables. They contain either a copy of all the data or a copy of the data of current period of time, such as day. At a given period the staging table is swapped with the actual table so the data warehouse becomes instantly up-to-date. This may be done by changing the view definition where the updated table is used instead the old one.
- **External real-time cache:** this is a variant of trickle and feed where the real-time data are stored outside the data warehouse in an external real-time cache, avoiding potential performance problems and leaving the data warehouse largely intact.

Full integration of the data in the business sense is achieved by the warehouse approach supported by an active data warehouse. The *active data warehouse* represents a single, canonical state of the business, i.e. a single version of the truth. It represents a closed loop process between the transactional (operational) system and data warehousing (analytical) system. The transactional system feeds the data warehouse, and the data warehouse feeds back the transactional system in order to drive and optimize transactional processing. Thus, the data warehouse is active if it automatically delivers information to the transactional system.

Recent trends towards *mediator approach* try to loosen the coupling between various data sources and thus to avoid the problem of replicated data and delayed update in the data warehouse architecture. This approach (Fig. 3) uses a virtual database with mediated schema and wrapper, i.e. adapter, which translates incoming queries and outgoing answers. A wrapper wraps (Ullman, 1997) an information source and models the source using a source schema. The situation emerges when two independent data sources have to be merged or combined into one source, such as integration of two similar databases when two organizations merge or integration of two document bases with similar structure content.

The users of the integrated system, i.e. data source D, are separated from the details of the data sources A, B and C at the schema level, by specifying a mediated or global schema. The *mediated schema* is a reconciled view of data in sources A, B and C, so the user needs to understand the structure and semantics of the mediated schema in order to make a meaningful query (Lu, 2006). The task of the data integration system is to isolate the user from the knowledge of where data are, how they are structured at the sources, and how they are reconciled into the mediated schema.

Figure 3. Mediator approach



The source databases are approachable through a wrapper code which transforms the original query into a specialized query over the original databases A, B or C. This is a view based query because each of the data sources A, B or C can be considered to be a view over the virtual database D. The first step in the approach involves setting up the mediated schema either manually or automatically (Rahm & Bernstein, 2001). The next step involves specifying its relationships to the local schemas in either Global As View (GAV) or Local As View (LAV) fashion.

In the GAV or global-centric approach the mediated or global schema is expressed in terms of the data sources, i.e. to each data element of the mediated schema, a view over the data sources must be associated. In the GAV approach mediator processes queries into steps executed at sources. Changes in data sources require revising of the global schema and mapping between the global schema and source schemas. Thus, GAV is sensitive on scalability of global model.

In the LAV or source-centric approach the mediated or global schema is specified independently from the sources but the sources are defined as views over the mediated schema. LAV is better at scalability because the changes in data sources require adjusting only a description of the source view. Nevertheless, query processing in LAV is more difficult. Some approaches try to combine the best of GAV and LAV approach (Xu & Embley, 2004; Cali, De Giacomo & Lenzerini, 2001).

The query language approach propagates extending query languages with powerful constructs to facilitate integration without the creation of the mediated or global schema. SchemaSQL (Laksmanan, Sadri & Subramanian, 2001) and UDM (Lu, 2006) are examples of this approach. They serve as the uniform query interface (UQI) which allows users to write queries with only partial knowledge of the "implied" global schema.

An issue of growing interest in data integration is the problem of elimination of *information conflicts* among data sources that integrate. *Intensional inconsistencies*, often referred to as semantic inconsistencies, appear when the sources are in different data models, or have different data schemas, or the data is represented in different natural languages or in different measures. For example: "Is 35 degrees measured in Fahrenheit or in Centigrade?" this type of inconsistencies may be resolved by the usage of ontologies which explicitly define schema terms and thus help to resolve semantic conflicts. This is also called ontology based data integration.

Extensional inconsistencies, often referred to as data inconsistencies, are factual discrepancies among the data sources in data values that belong to the same objects. Extensional inconsistencies are visible only after intensional inconsistencies are resolved. (Motro & Anokhin, 2006, p.177) argue that "all data are not equal", and "that data environment is not egalitarian, with each information source having the same qualification". In a diverse environment "the quality" of information provider's data is not equal. Many questions arise, such as: Is data enough recent or is outdated? Is the data source trustworthy? Is data expensive? An interesting approach is suggested by (Motro & Anokhin, 2006). In Internet era when the number of alternative sources of information for most applications increases, users often evaluate information about data sources. The meta data, such as timestamp, cost, accuracy, availability, and clearance, whether provided by the sources themselves or by a third-party dedicated to the ranking of information sources, may help users judge the suitability of data from the individual data source.

Challenge: Could the Data Integrate the Business?

A business process is a structured, measured set of activities designed to produce a specific output for a particular customer or market (Davenport, 1993). Even though most organizations have a functional structure, examining of business processes provides a more authentic picture of the way business is run.

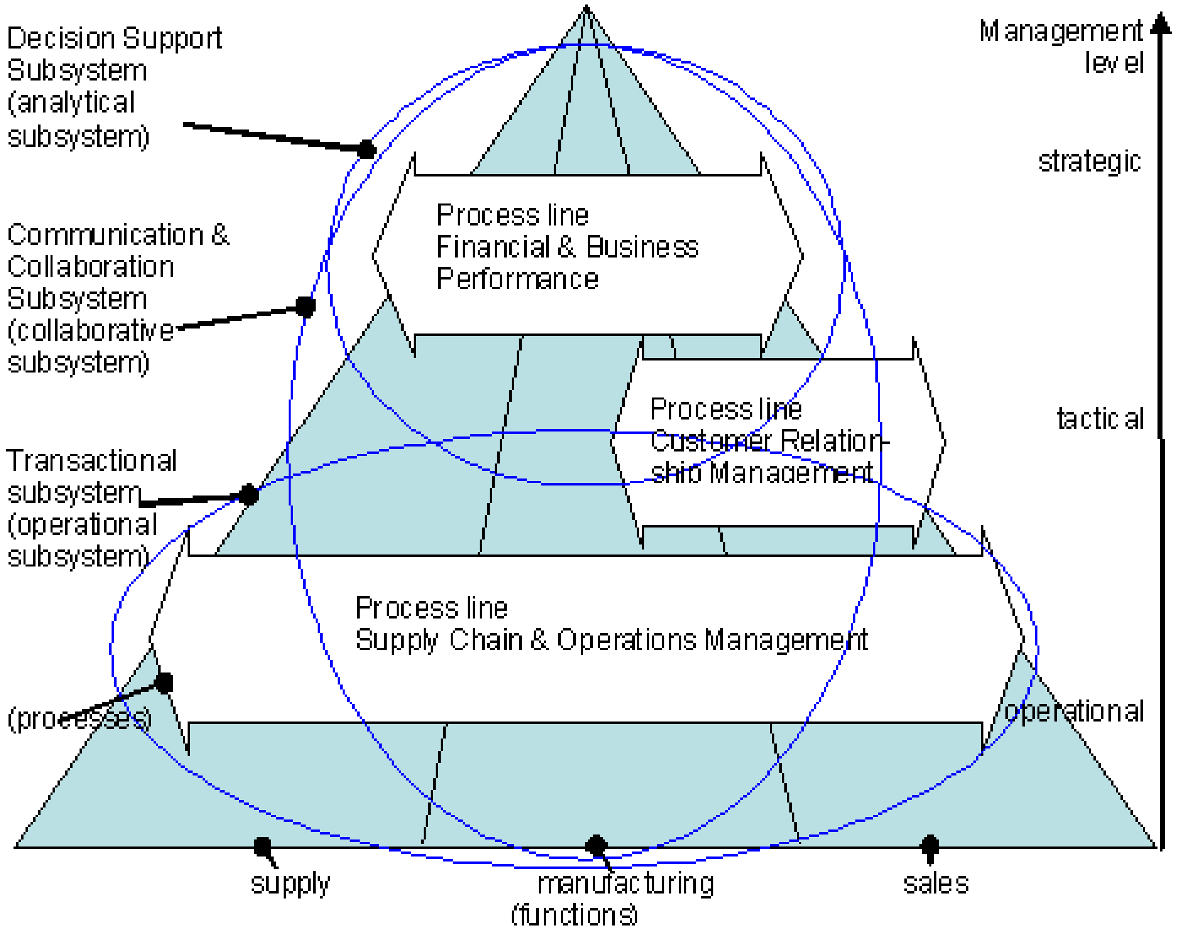
A good information system relies on Enterprise Resource Planning system (ERP) which uses a multitude of interrelated program modules that process data in individual functional areas. In Figure 4, ERP is exemplified by three basic functions: supply, manufacturing and sales. They stretch from the top to the bottom of the pyramid. If a program module covers the whole function, it runs through all management levels: operational, tactical and strategic.

ERP is often supplemented by program modules for analytical data processing which is characteristic for data warehousing and decision support systems. As a result it is perceived as an Enterprise Information system (EIS) or simply an Enterprise System (ES) that represents complete operational and analytical program solutions.

According to business process approach, some typical process lines may be used:

- **Financial and Business Performance Management:** a set of processes that help

Figure 4. Enterprise Resource Planning (ERP) system


Decision Support Subsystem (analytical subsystem)
Communication & Collaboration Subsystem (collaborative subsystem)
Transactional subsystem (operational subsystem)
(processes)
Process line Financial & Business Performance
Process line Customer Relationship Management
Process line Supply Chain & Operations Management
Management level
strategic
tactical
operational
supply
manufacturing
(functions)
sales

organizations optimize their business or financial performance

- **Customer Relationship Management:** a set of processes utilized to provide a high level of customer care
- **Supply Chain and Operations:** a set of processes involved in moving a product or service from supplier to customer

A vertically integrated information system connects activities on the lowest level of a particular function (e.g. a process of transactional retailing) with data analysis and data representation on the decision-making level (e.g. reports on sales analysis for the general manager).

A horizontally integrated information system enables a systematic monitoring of specific business process from end to end. For instance, as soon as a purchase order arrives, an integrated information system can receive it, forward it "automatically" to sales and delivery process which will deliver the goods to the customer and send the invoice to his information system, open a credit account; record the quantity of goods delivered in warehouse evidence; in case that goods first need to be produced in a manufacturing plant the system will issue a work order to manufacture the required quantity of goods; the production process can be supplied by a production plan; sales results will be visible to the sales manager and analysed using analytical tools of the information system; etc. An integrated information system enables recording of all business events so that the data can be used an analysed effectively throughout the organization.

Let us mention some problems related to integrated data:

- If the vertical coverage of data from a particular functional area is insufficient, data does not comprise all levels of this functional area. This is the case when, for example, only operational sales data exist, whereas sales data for the tactical and strategic level of decision-making are missing due to the lack of appropriate sales reports or if there is no possibility to interactively analyse sales data.
- Insufficient horizontal integration of different functional areas results when functional areas are not integrated. In case that production data are not visible in sales application, or they need to be re-entered (although they already exist in production application), operational sales data will not be well integrated with operational production data.
- Finally, if data in an information system are not sufficiently integrated, the information system will be comprised of isolated information (data) islands. An information system is not integrated unless individual functional areas are mutually integrated with other functional areas by data. For instance, sales data may not be integrated i.e. connected with production data; or they may be integrated in indirect and complicated ways. Problem occurs because some older applications were designed for individual functional areas without concern for other functional areas.

Challenge: Do We Know Organization's Data?

If we do not know organization's data well enough, we will use them rarely and poorly. Users often do not know what information is available and, as a consequence, they do not use the information system frequently enough. Users should be briefed about the information system and the ways to use it. A (central) *data catalogue* should be provided to present the complete data repertoire.

Challenge: Do We Know How to Use Organization's Data?

If we do not know how to use the information system well enough, we will not use it sufficiently and we will not exploit all of its capabilities. Even if users know that the organization has data, they may not know how to use it. Therefore, the information system should be documented and ways of using it described in a *catalogue of functions*.

Data Quality

Challenge: Do We Know What the Quality Data is?

(Juran & Godfrey, 1999) defines data to be of high quality if the data is fit for their intended uses in operations, decision making and planning, i.e. if it correctly represents the real world to which it refers. Among various categories of attributes describing data quality the most commonly used are *accuracy* (degree of conformity of the data to the actual or true value to which it refers), *completeness* (if nothing needs to be added to it), *relevance* (the data is pertinent and adequate to the context and the person who use it) and *timeliness* (the data is given in timely manner). Naturally, it is desirable that the data supported by EIS fulfil the mentioned quality characteristics.

(Bocij, Chaffey, Greasley & Hickie, 2006) gives a few additional lists of categories of data quality concerning time dimension (timeliness, currency, frequency, time period), content dimension (accuracy, relevance, completeness, conciseness, scope), form dimension (clarity, detail, order, presentation, media) etc. However, the end user of data requires quality data in quantities that support the decision-making process. Too much data can be a burden for the person who needs to make a decision.

Most companies view data quality control as a cost problem, but in reality it can drive revenue. The problem of data quality drives the companies to set up a data governance function whose role it is to be responsible for data quality.

Challenge: Are We Aware that the Quality of Data is Degradable?

Many databases behave as growing organism. Data is added, updated or deleted. Due to new applications the structure of a database may be altered to meet new business requirements. During the database lifetime many alternations keep the database functional. But over time, with continuing alterations to its structure and content, the logical integrity of the database can be so degraded that it becomes unreliable. It often happens that data integration is negatively affected by the lack of proper documentation regarding changes, the fact that the people responsible for the application have moved on, or that the vendor of the application no longer exists. In many cases the organization does not realize that the database has been degraded until a major event, such as data integration, occurs. It is time to determine the true quality of data in the database.

(Healy, 2005) suggests that the only reliable way to determine the true nature of data is through a *data audit*, involving thorough data analysis or profiling. There are manual and automated methods of data profiling. In manual data profiling analysts assess data quality using the existing documentation and assumptions about the data to anticipate what data problems are likely to be encountered. Instead of evaluating all data they take data samples and analyze them. If such an analysis discovers wrong data, the appropriate programs have to be written to correct the problems. The manual procedure can be a very costly and time-consuming process requiring a number of iterations. The more efficient and accurate way of examining data is to use an automated data profiling tool that is able to examine all the data. Data profiling tools can identify many problems resident in the data and provide a good picture of data structure including metadata descriptions, data values, formats, frequencies, and data patterns. Automated data profiling solutions offer to correct data errors and anomalies more easily and efficiently.

Data Ownership/Stewardship

Challenge: Who is the Owner/ Steward of the Data?

Data ownership refers to both the possession of and responsibility for data. Ownership implies power over and control of data. Data control includes the ability to access, create, modify, package, sell or remove data, as well as the right to assign access privileges to others. This definition probably implies that the database administrator is the "owner" of all enterprise data. Nevertheless, this is not true and must not be true. The right approach to ownership aims to find the user who will be responsible for the quality of the data.

According to (Scofield, 1998), telling a user that he/she "owns" some enterprise data is a dangerous thing. The user may exercise the "ownership" to inhibit the sharing of data around the enterprise. Thus, the term "data stewardship" is better, implying a broader responsibility where the user must consider the consequences of changing "his/her" data. Data stewardship may be broken into several stewardship roles. For example, definition stewardship is responsible to clearly define the meaning of data and may be shared between analyst, database administrator and key user. Access stewardship is responsible to permit or prevent access to enterprise's data. Quality stewardship is responsible to fulfil the broad term "quality data" and may be shared between analyst and key user. Thus, data stewardship may be seen as distributed responsibility, shared between IT people (analyst, database administrator) and key users. Responsibilities assigned to all data stewardship actors, i.e. data stewards, must be registered in a data catalogue.

Data Privacy

Challenge: Is Data Privacy a Concern?

Data privacy is important wherever personally identifiable information (in digital or other form) is collected and stored. Data privacy issues arise in various business domains, such as healthcare records, financial records and transactions, criminal justice records, residence records, ethnicity data, etc. Information that are sensitive and need to be confidential are, for example, information about person's financial assets and financial transactions.

The challenge in data privacy is to share data while protecting personally identifiable information. Since freedom of information is propagated at the same time, promoting both data privacy and data openness might seem to be in conflict. Data privacy issue is especially challenging in the internet environment. Search engines and data mining techniques can collect and combine personal data from various sources, thus revealing a great deal about an individual. The sharing of personally identifiable information about users is another concern.

Data protection acts regulate the collecting, storing, modifying and deleting of information which identifies living individuals, sometimes known as data subjects. This information is referred to as *personal data*. The legal protection of the right to privacy in general, and of data privacy in particular, varies greatly in various countries. The most regulated are data privacy rights in the European Union (EU), more restrictively than in the United States of America. Data protection in EU is governed by Data protection Directive 95/46/EC (EU, 1995), where personal data is defined as "any information relating to an identified or identifiable natural person". Personal data must be, according to the Directive "(a) processed fairly and lawfully; (b) collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes; (c) adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed; (d) accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that data which are inaccurate or incomplete, having regard to the purposes for which they were collected or for which they are further processed, are erased or rectified; and (e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed."

Privacy and data protection are particularly challenging in multiple-party collaboration. If heterogeneous information systems with differing privacy rules are interconnected and information is shared, privacy policy rules must be mutually communicated and enforced. Example of a platform for communicating privacy policy in the web-services environment is Web Services Privacy (WS-Privacy). P3P (W3C, 2008) is a system for making Web site privacy policies machine-readable. P3P enables Web sites to translate their privacy practices into a standardized, machine-readable XML format that can be retrieved automatically and easily interpreted by a user's browser. Translation can be performed manually or with automated tools. Eventually, the Web site is able to automatically inform P3P user agents of site privacy practices in both machine- and human-readable formats. On the user side, P3P user agent may read P3P site's privacy policy and, if appropriate, automate decision-making based on this practice.

Data Visualization

Challenge: Could We Amplify Cognition of Data?

Many data acquisition devices, such as various measure instruments, scanners etc. generate large data sets, which data is collected in large databases in textual, numerical or multimedia form. The results of data analysis of the large data sets are not acceptable and attractive to the end users without good techniques of interpretation and visualization of information.

Information visualization is a rapidly advancing field of study both in academic research and practical applications. (Card, Mackinlay & Shneiderman, 1999, p.8) define information visualization, as "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition. Cognition is the acquisition or use of knowledge. The purpose of the information visualization is insight, not picture. The goal of the insight is discovery, decision making and explanation. Information visualization is useful to the extent that it increases our ability to perform this and other cognitive activities."

Information visualization exploits the exceptional ability of human brain to effectively process visual representations. It aims to explore large amounts of abstract, usually numeric, data to derive new insights or simply make the stored data more interpretable. Thus, the purpose of data visualization is to communicate information clearly and effectively through graphical means, so that information can be easily interpreted and then used.

There are many specialized techniques designed to make various kinds of visualization using graphics or animation. Various techniques are used for turning data into information by using the capacity of the human brain to visually recognize data patterns. Based on the scope of visualization, there are different approaches to data visualization. According to (VisualLiteracy, 2008) the visualization types can be structured into seven groups: *sketches* ("Sketches are atmospheric and help quickly visualize a concept. They present key features, support reasoning and arguing, and allow room for own interpretation. In business sketches can be used to draw on flip charts and to explain complex concepts to a client in a client meeting."), *diagrams* ("Diagramming is the precise, abstract and focused representation of numeric or non-numeric relationships at times using predefined graphic formats and/or categories. An example of a diagram is a Cartesian coordinate system, a management matrix, or a network. Diagrams explain causal relationships, reduce the complexity to key issues, structure, and display relationships."), *images* ("Images are representations that can visualize impression, expression or realism. An image can be a photograph, a computer rendering, a painting, or another format. Images catch the attention, inspire, address emotions, improve recall, and initiate discussions."), *maps* ("Maps represent individual elements (e.g., roads) in a global context (e.g., a city). Maps illustrate both overview and details, relationships among items, they structure information through spatial alignment and allow zoom-ins and easy access to information."), *objects* ("Objects exploit the third dimension and are haptic. They help attract recipients (e.g., a physical dinosaur in a science museum), support learning through constant presence, and allow the integration of digital interfaces."), *interactive visualizations* ("Interactive visualizations are computer-based visualizations that allow users to access, control, combine, and manipulate different types of information or media. Interactive visualizations help catch the attention of people, enable interactive collaboration across time and space and make it possible to represent and explore complex data, or to create new insights."), and *stories* ("Stories and mental images are imaginary and non-physical visualizations. Creating mental images happens trough envisioning."). Each of mentioned visualization types has its specific area of application.

DATA MANAGEMENT

DAMA's functional framework (DAMA, 2008) suggests and guides management initiatives to implement and improve data management through ten functions:

- **Data Governance:** planning, supervision and control over data management and use
- **Data Architecture Management:** as an integral part of the enterprise architecture
- **Data Development:** analysis, design, building, testing, deployment and maintenance
- **Database Operations Management:** support for structured physical data assets
- **Data Security Management:** ensuring privacy, confidentiality and appropriate access
- **Reference & Master Data Management:** managing versions and replicas
- **Data Warehousing & Business Intelligence Management:** enabling access to decision support data for reporting and analysis
- **Document & Content Management:** storing, protecting, indexing and enabling access to data found in unstructured sources (electronic files and physical records)
- **Meta Data Management:** integrating, controlling and delivering meta data
- **Data Quality Management:** defining, monitoring and improving data quality

Using DAMA's functional framework (DAMA, 2008) each function may be described by seven environment elements: goal and principles of the function, activities to be done in the function (i.e. planning, control, development and operational activities), deliverables produced by the function, roles and responsibilities in the function, practices and procedures used to perform the function, technology supporting the function; and organization and culture.

A similar term is enterprise information management. (Gartner, 2005) defines it as an organizational commitment to define, secure and improve the accuracy and integrity of information assets and to resolve semantic inconsistencies across all boundaries to support the technical, operational and business objectives of the company's enterprise architecture strategy.

Challenge: Do We Need Data Governance?

According to DAMA (DAMA, 2008), the beginning and central data management activity is data governance which is exercise of authority, control and decision-making over the management of organization's data. It is a high-level planning and control mechanism over data management. (IBM, 2007, p.3) defines data governance as "a quality control discipline for adding new rigor and discipline to the process of managing, using, improving, monitoring and protecting organizational information". It may be a part of organization's IT governance.

All organizations take care of data governance, whether formally or informally. Tendency is to manage data governance more formally. The objectives of data governance are implemented by various data governance programs or initiatives. Factors driving data governance programs include the need for implementation of EIS as a core system with a consistent enterprise-wide definition of data, the problems of organization's data overload, data integration, data security etc. Some external drivers are regulatory requirements for more effective controls for information transparency, such as Basel II and Sarbanes-Oxley, the need to compete on market using better analytics and having the right information to provide reliable insights into business.

Data governance is not a typical IT project as it is declared at the first Data Governance Conference in 2006. Based on the experience of successful data governance programs it is agreed that data governance is more than 80 percent communication and that data governance must focus on people, process and communication before considering any technology implementation. Organization's data governance body may consist of executive leadership, line-of-business and project managers, and data stewards.

Data governance must follow a strategic approach considering first the current state, devising a future state and designing an implementation plan to achieve the future state. A data governance project may be difficult, often with the tough task to change the culture of an organization. Therefore, a phased approach should be applied. Each step must be small enough to be successfully and relatively quickly implemented, the achieved results analysed and lessons learned implemented in the following steps.

According to the well known Software Engineering Institute, Capability Maturity Model for organization's software development process which describes five-level graduated path from initial (level 1), managed (level 2), defined (level 3), quantitatively managed (level 4) to optimizing (level 5) (IBM, 2007) describes a similar data governance maturity model. The model measures data governance competencies of an enterprise based on 11 domains of data governance maturity: organizational structures and awareness, stewardship, policy, value creation, data risk management and compliance, information security and privacy, data architecture, data quality management, classification and metadata, information lifecycle management; and audit information, logging and reporting.

CONCLUSION

Challenges of successful data management are numerous and diverse, varying from technological to conceptual. The chapter presented the most challenging aspects of data management classified into three classes. The first combines data availability, data integrity and data security, which serve as data continuity aspects that are important for the continuous provision of data in business processes and for decision-making purposes. The aspects in the second class enable innovative, better, more efficient and more effective data usage. The problems of data overload, data integration, data quality, data degradation, data ownership or stewardship, data privacy, and data visualization are described. The last aspect is of managerial nature. Data governance is important for planning, supervising and controlling of all management activities exercised to improve organizational data and information.

It is been shown that data management challenges are numerous. A data management analysis in a typical organization will most probably reveal that some data management aspects are more problematic than others and resolving data management issues can be dynamic. Consequently, data governance will constantly need to discover novel and innovative ways to deal with data management problems.

REFERENCES

W3C. (2008). Platform for privacy preferences (P3P) project. Retrieved on September 2, 2008, from www.w3.org/P3P/

Allen, J. H. (2001). *The CERT guide to system and network security practices*. Boston: Addison-Wesley.

Bocij, P., Chaffey, D., Greasley, A., & Hickie, S. (2006). *Business information systems*. Harlow, UK: Prentice Hall Financial Times.

Brumec, J. (1997). A contribution to IS general taxonomy. *Zbornik radova, 21*(22), 1-14. Cali, A., De Giacomo, G., & Lenzerini, M. (2001). Models for information integration: Turning local-as-view into global-as-view. In *Proc. of Int. Workshop on Foundations of Models for Information Integration, 10th Workshop in the Series Foundations of Models and Languages for Data and Objects*. Retrieved on August 16, 2008, from www.dis.uniroma1.it/~degiacom/papers/2001/CaDL01fmii.ps.gz

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (1999). *Readings in information visualization: Using vision to think*. San Francisco: Morgan Kaufmann.

DAMA. (2008). *DAMA-DMBOK: Functional framework*, version. 3. Retrieved on August 16, 2008, from http://www.dama.org/files/public/DMBOK/DI_DAMA_DMBOK_en_v3.pdf

Davenport, T. H. (1993). *Process innovation: Reengineering work through information technology*. Boston: Harvard Business School Press.

EU. (1995). *Directive 95/46/EC of the European Parliament and the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data*. Retrieved on September 2, 2008, from http://ec.europa.eu/justice_home/fsj/privacy/docs/95-46-ce/dir1995-46_part1_en.pdf

Gartner. (2005). *Business drivers and issues in enterprise information management.* Retrieved on September 2, 2008, from http://www.avanade.com/_uploaded/pdf/avanadearticle4124441.pdf

Gartner Inc. (2006). *Defining, cultivating, and measuring enterprise agility*. Retrieved on September 15, 2008, from http://www.gartner.com/resources/139700/139734/defining_cultivating_and_mea_139734.pdf

Halevy, A. Y. (2001). Answering queries using views: A survey. *The VLDB Journal*, *10*(4), 270–294. doi:10.1007/s007780100054

Hall, D., & McMullen, S. H. (2004). *Mathematical techniques in multisensor data fusion*. Norwood, USA: Artech House.

Healy, M. (2005). Enterprise data at risk: The 5 danger signs of data integration disaster (White paper). Pittsburgh, PA: Innovative Systems. Retrieved on August 22, 2008, from http://www.dmreview.com/white_papers/2230223-1.html

IBM. (2007). *The IBM data governance council maturity model: Building a roadmap for effective data governance*. Retrieved on August 22, 2008, from ftp://ftp.software.ibm.com/software/tivoli/whitepapers/LO11960-USEN-00_10.12.pdf

IBM. (2008). RPO/RTO defined. Retrieved on September 15, 2008, from http://www.ibmsystemsmag.com/mainframe/julyaugust07/ittoday/16497p1.aspx

IDS. (2008). *IDS white paper: The diverse and exploding digital universe*. Framingham: IDC.

Juran, J. M., & Godfrey, A. B. (1999). *Juran's quality handbook*. McGraw-Hill.

Lakshmanan, L. V., Sadri, F., & Subramanian, S. N. (2001). SchemaSQL: An extension to SQL for multidatabase interoperability. *ACM Transactions on Database Systems*, *26*(4), 476–519. doi:10.1145/503099.503102

Langseth, J. (2004). *Real-time data warehousing: Challenges and solutions*. Retrieved on August 29, 2008, from http://DSSResources.com/papers/features/langseth/langseth02082004.html

Lu, J. J. (2006). *A data model for data integration*. (ENTCS 150, pp. 3–19). Retrieved on August 16, 2008, from http://www.sciencedirect.com

Motro, A., & Anokhin, P. (2006). Fusionplex: Resolution of data inconsistencies in the integration of heterogeneous information sources. [from http://www.sciencedirect.com]. *Information Fusion*, *7*, 176–196. Retrieved on August 16, 2008. doi:10.1016/j.inffus.2004.10.001

Oracle. (2008). *Oracle® database high availability overview 11g release 1 (11.1)*. Retrieved on September 8, 2008, from http://download.oracle.com/docs/cd/B28359_01/server.111/b28281/overview.htm#i1006492

Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal*, *10*(4), 334–350. doi:10.1007/s007780100057

Schmidt, K. (2006). *High avalability and disaster recovery*. Berlin: Springer.

Scofield, M. (1998). Issues of data ownership. *DM Review Magazine*. Retrieved on August 29, 2008, from http://www.dmreview.com/issues/19981101/296-1.html

Ullman, J. D. (1997). Information integration using local views. In F. N. Afrati & P. Kolaitis (Eds.), *Proc. of the 6th Int. Conf. On Database Theory* (ICDT'97). (LNCS 1186, pp. 19-40). Delphi.

VisualLiteracy. (2008). Visual literacy: An e-learning tutorial on visualization for communication, engineering, and business. Retrieved on September 3, 2008, from http://www.visual-literacy.org/

Woods, D. D., Patterson, E. S., & Roth, E. M. (2002). Can we ever escape from data overload? A cognitive systems diagnosis. *Cognition Technology and Work*, *4*, 22–36. doi:10.1007/s101110200002

Xu, L., & Embley, D. W. (2004). Combining the best of global-as-view and local-as-view for data integration. *Information Systems Technology and Its Applications, 3rd International Conference ISTA'2004* (pp. 123-136). Retrieved on September 2, 2008, from www.deg.byu.edu/papers/PODS.integration.pdf