ADVANCE STUDY ON DATA MINING AND ITS PRIVACY PROCESS

[1]Pradeep Agrawal & [2]Yudhisthir Sharma

Research Scholar, Monad University, Hapur (India)

ABSTRACT

The efficient database management systems have been very important assets for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection whenever needed. The proliferation of database management systems has also contributed to recent massive gathering of all sorts of information. Today, we have far more information than we can handle: from business transaction and scientific data, to satellite pictures, text reports and military intelligence. Information retrieval is simply not enough anymore for decision making.

I INTRODUCTION TO DATA MINING

Confronted with huge collections of data, we have now created new needs to help us make better managerial choices. These needs are automatic summarization of data, extraction of the "essence" of information stored, and the discovery of patterns in raw data. We are in an age often referred to as the information age. In this information age, because we believe that information leads to power and success, and thanks to sophisticated technologies such as computers, satellites, etc., we have been collecting tremendous amounts of information. Initially, with the advent of computers and means for amass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. Unfortunately, these massive collections of data stored on disparate structures very rapidly became over shelling.

II PRIVACY POLICY AND STANDARDS

The privacy policy describes both what is allowed as well as not allowed in the system. Privacy standards should be prescriptive guidance for people building and operating systems, and should be backed by reusable services wherever practical. This is very important, it is no longer acceptable for enterprise privacy to exclusively function as an arbiter; privacy in the enterprise needs architecture and design advocates, and backing at runtime.

III PRIVACY ARCHITECTURE

The privacy architecture is a strategic framework that allows the development and operations staff to align efforts, in addition the privacy architecture can drive platform improvements which are not possible to make at a project level. A given software development project may not be able to make a business case to purchase an XML Privacy
Risk management, privacy policy and standards, and privacy architecture govern the privacy processes and defense in depth architecture through design guidance, runtime support, and assurance services. Privacy metrics are used for decision support for risk management, privacy policy and standards, and privacy architecture. The privacy architecture should have a reference implementation for developers and other IT staff to review what functions the privacy mechanisms performs, and how they do it.

IV PRIVACY PROCESSES

Privacy processes carry out the intent of the enterprise risk management, privacy policy and standards, and privacy architecture. They are broken into discrete domains because they solve very different problems, and require different staffing, support models, and success criteria.

4.1 SDL: Privacy functions as a collaborative design partner in the software development lifecycle (SDL), from requirements, architecture, design, coding, deployment, and withdrawal from service. Privacy adds value to the software development lifecycle through prescriptive and proscriptive guidance and expertise in building secure software. Privacy can play a role in all phases of the SDL, but an iterative, phased-based integration of privacy into the SDL is the wisest path, each additional privacy process improvement must fit with the overall SDL approach in the enterprise, which varies widely.

4.2 Identity Management: deals with the creation, communication, recognition, and usage of identity in the enterprise. Identity management includes provisioning services, directories, multi-factor authentication, federation, and so on. All access control is predicated on identity, a central concern to privacy architecture, the quality of the system's authentication and authorization cannot be stronger than the identity management process. The utility of the identity management architecture comes through mapping the subject request's claims (or assertions) to policy enforcement decision workflow; and the object's protection model, often in the form of group and/or role membership.

4.3 Threat Management

Threats differ from vulnerabilities in that threats are the actors that breach or attempt to breach privacy policies and mechanisms. The privacy gaps that are exploited by threats are called vulnerabilities. Threat Management tools and processes include: Privacy Monitoring, Web Application Firewall, Privacy Incident Management Processes, Privacy Event Management System, Incident Response Planning Processes, cryptography, and Forensic Analysis Process and Tools. The threat environment is inherently unpredictable and in large part out of control of the enterprise.

4.4 Vulnerability Management

The vulnerabilities may reside at any system layer – database, operating system, servers, and so on; specialized tools probe for known vulnerabilities. It is important to differentiate threat management and vulnerability management.

The threat environment contains many unknown mysteries around attacker techniques and goals, attackers will identify currently unknown vulnerabilities (zero day attacks), but there are many known vulnerabilities that the privacy team can act on, while the threat landscape is inherently less predictable meaning privacy is reactive to threats and can be generally proactive towards dealing with known vulnerabilities.

4.5 Risk Metrics

Measure the overall assets, and their attendant countermeasures, threats, and vulnerabilities. Since risk metrics are focused on assets, they allow the privacy architecture to be measured in business terms. Risk metrics inform stakeholders on privacy posture based on information that is harvested from the privacy processes, especially vulnerability management and threat management, and the defense in depth stack.

The four main phases in the process are: Architecture Risk Assessment, Privacy Architecture& Design, Implementation, and Operation & Monitoring.

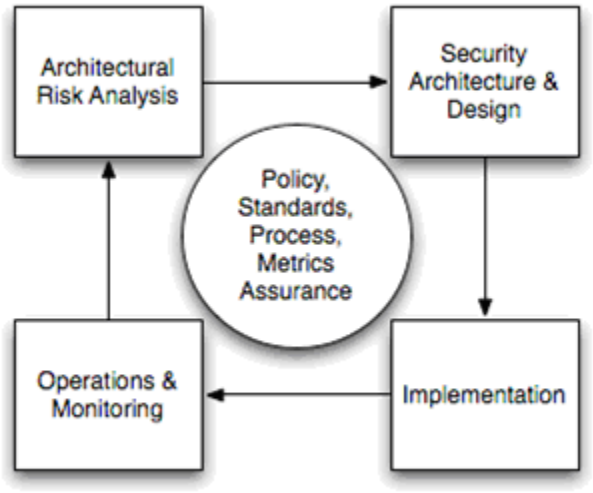

Figure 1: Privacy Architecture Lifecycle

V PRIVACY ARCHITECTURE AND DESIGN

Architecture and design of privacy services that enable business risk exposure targets to be met. The policies and standards, and risk management decisions drive the privacy architecture and the design of the privacy processes and defense in depth stack.

5.1 Implementation

Privacy processes and services implemented, operational, and managed. Assurance services are targeted at verifying that the Risk Management, Privacy Policy and Standards, Privacy Architecture decisions are reflected in the actual runtime implementation.

5.2 Operations and Monitoring

Ongoing processes, such as vulnerability management and threat management that monitor and manage the operational state as well as the breadth and depth of systems privacy. Operational and monitoring processes should be instrumented with privacy metrics to better measure the runtime environment.

5.3 Dashboard Reporting

The information privacy dashboard provides a way to track progress over time across the privacy architecture and processes. Given the many moving parts in a distributed enterprise, tracking and alignment of efforts is a challenge. The example dashboard below shows one way to roll up across multiple efforts and report on progress at an executive level.

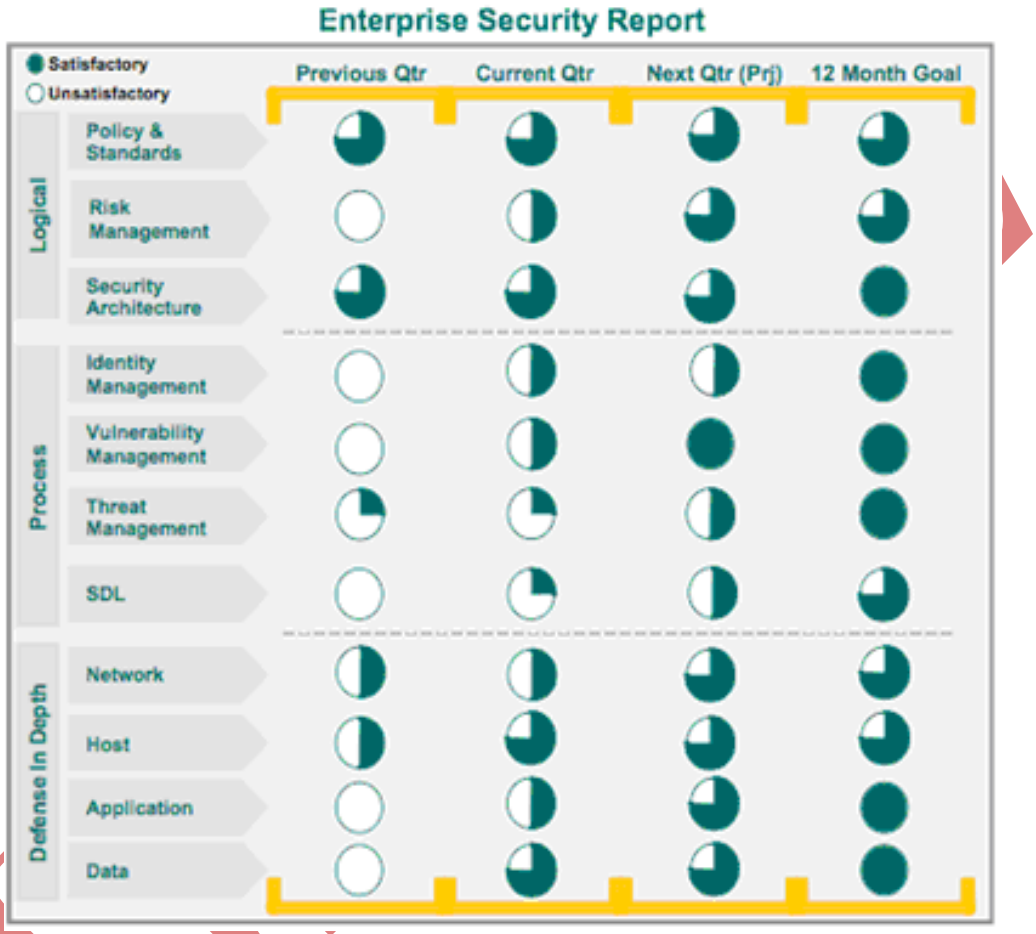

Figure 2: Enterprise Privacy Executive Report

The privacy architecture blueprint describes the key decisions, building an Enterprise Privacy Executive Report helps for senior management to understand the domains, the progress in those domains, and the key investment areas. Example: Applying the Enterprise Privacy Architecture

5.4 Blueprint

It describes a brief example of activities that enable applying the blueprint in the context of a static analysis project. Static analysis is the process of scanning and analyzing source code to identity privacy vulnerabilities. As with many privacy projects these efforts are typically treated as one off projects driven by a single goal, such as compliance, and not ordinarily mapped into a strategic context.

VI WHAT ARE DATA MINING AND KNOWLEDGE DISCOVERY

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, of not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for extraction of interesting knowledge that could help in decision-making. Data Mining, also popularly known as Knowledge Discovery in Data bases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. The iterative process consists of the following steps:

Data cleaning: also know as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

Data Selection: At this step, the data relevant to the analysis is decided on and retrieved from the data collection.

Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

Data mining: it is crucial step in which clever techniques are applied to extract patterns potentially useful.

Pattern evaluation: in this step, strictly interesting representing knowledge is identified based on given measure.

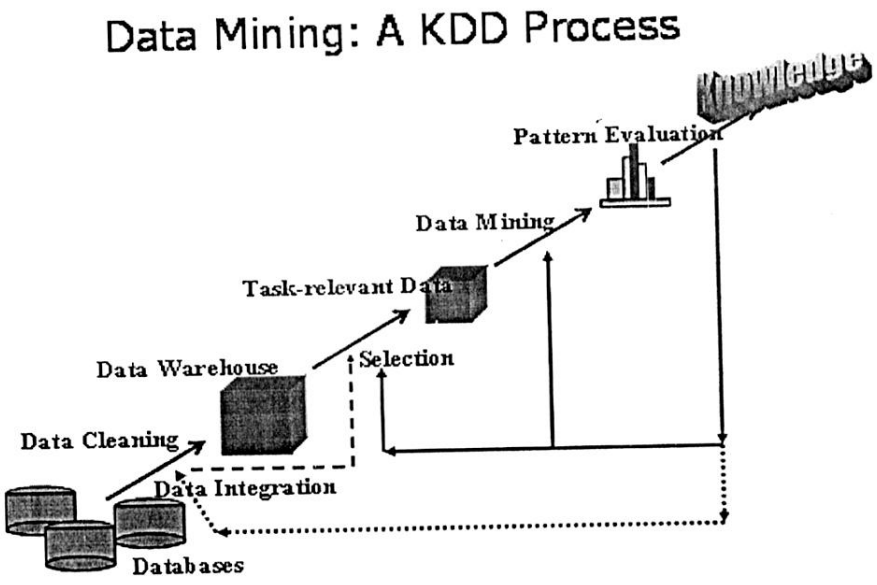

Figure 3: Data Mining is the core of Knowledge Discovery process

Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

It is common to combine some of these steps together. For instance, data cleaning and data integration can be performed together as a pre processing phase to generate a data warehouse.

Date selection and data transformation can also be combined where the consolidation of the date is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data. The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measure can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results. Data mining derives its name from the similarities between searching for valuable information in a large database and mining. Both imply either sifting through a large amount of material the material

to exactly pinpoint where the values reside. Other similar terms referring to data mining are: data dredging, knowledge extraction and pattern discovery.

VII APPLICATIONS OF DATA MINING

7.1 General Data Mining Applications

These applications are typically of value to any type of organization that perform the functions discussed here, and is not limited to only certain organizations operating in certain industries. This section highlights some of the applications of data mining in organizations in general typical business functions that these data mining applications fall under are:

(1) E-commerce and E-business Functions

Data mining can help organizations to identify the information that would be most suitable to put on the Web by conducting an analysis of competitors and potential customers, and determining the expertise of the organization (Thuraisingham, 2003:276) [23] In order for any organization to get onto the World Wide Web, it needs to decide what information to put on its web site. Data mining can also be used to assist organizations waning to conduct e-business with other organizations by providing information with regard to selecting the best partners, identifying competitors and determining the best pricing policies for its products of services. Summarizes the e-commerce and e-business applications of data mining as being the following:

- Selecting partner organization.
- Analyzing customer profiles.
- Determining which products to market online.
- Assessing the similarity of user browsing patterns.
- Identifying web pages that are viewed together.
- Assessment of the similarity of web page contents.

- Categorization of web pages based on content.

The web site will assist online users in their decision making by listing for instance other books that have been bought by people who also bought the particular book that the online user is currently looking at. One of the most useful applications of data mining in an e-commerce environment can be found on the Amazon.com web site. The web site also makes use of data mining to bundle books that are often bought in pairs, together and to sell or market these at reduced price if they are bought together. Amazon.com also uses data mining to profile its customers, and online users that return to the web site after having bought a book for instance, has been provided with lists of other book titles that they might be interested in based on the category of book that they just bought.

(ii) Business Intelligence Functions

In order to be able to run a business effectively, management needs intelligence relating to competitors, customer, partners and employees, as well as intelligence relating to market conditions, future trends, government policies and more lists the applications of data mining in business intelligence as follows:

- Analyzing competitor strategies.
- Analyzing customer profiles.
- Determining you won business strategies.

(iii) Customer Relationship Management Functions

This information is then used to serve the customers in the best possible way. The CRM is a key application of business intelligence and deals with mining information about customers from public as well as private databases in order to build customer profiles. This mined information is then used to provide customers with easy access to other similar products that they might be interested in. Amazon.com once again provides a good example of how this mined information can be put to good use by gathering information such as, customers who bought this book also bought those books. Typical applications of data mining in customer relationship management are therefore:

- Building and analyzing customer profiles.
- Developing customer-specific products.

(iv) Marketing and Sales Functions

Marketing and sales were two of the early business functions that drove the development of the data mining. The following as some of the data mining application relevant to these functions:

- Performing targeted marketing.
- Determining the marketing strategies of competitors.
- Prediction of sales trends.

- Market segmentation.
- Lifestyle behavior analysis.
- Online sales support.
- Analysing or predicting customer reaction to promotions.
- Marketing basket analysis.

(v) Enterprise Resource Management Function

The resources of an organization might include human, inventory, expertise or anything that is of value to the organization, and data mining has many applications in this area. Enterprise resource management involves the management of all the resources of an organization. Some of these applications are:

- Analyzing human resources for employee retention benefits and payroll.
- Analyzing enterprise resources or supply chain management.

One of the primary applications of data mining in enterprise resource management is to ensure that the organization's resources and capabilities are aligned with what is required in the business environment not only today, but also in the future.

(vi) Manufacturing and Planning Functions

Data mining has several applications in the manufacturing and planning functions of organizations. Some of these applications include the following:

- Analyzing and developing manufacturing schedules.
- Determining the best possible assembly routines.
- Monitoring plants to detect anomalies and unusual patterns.
- Assisting large manufacturing firms to predict production breakdowns and analyze product defects.

The main advantage data mining offers to manufacturing firms is its ability to assist in identifying the conditions that lead to critical situations. Engineers can then use this information to correct problems and prevent future failures.

(vii) Education and Training Functions

These applications related mainly to the identification of both current and future training needs. Data mining has several applications that are relevant to education and training. The following two main applications:

- Developing courses and schedules based on the needs identified by data mining.

- Predicting future trends in education.

VIII CONCLUSION

Usefulness of discovery: Data mining allows the discovery of knowledge potentially useful and unknown. Whether the knowledge discovered is new, useful or interesting, is very subjective and depends upon the application and the user. It is certain that data mining can generate, or discover, a very large number of patterns or rules. In some cases the number of rules can reach the millions. One can even think of a meta-mining phase to mine the oversized data mining results. To reduce the number of patterns or rules discovered that have a high probability to be non-interesting, one has to put a measurement on the patterns.

Discovered patterns can also be found interesting if they confirm or validate a hypothesis sought to be confirmed or unexpectedly contradict a common belief. This brings the issue of describing what is interesting to discover, such as meta-rule guided discovery that describes forms of rules before the discovery process, and interestingness refinement languages that interactively query the results for interesting patterns after the discovery phase. Typically, measurements for interestingness are based on thresholds set by the user. These thresholds define the completeness of patterns discovered. Identifying and measuring the interestingness of patterns and rules discovered, or to be discovered is essential for the evaluation of the mined knowledge and the KDD process as a whole. While some concrete measurements exist, assessing the interestingness of discovered knowledge is still an important research issue.

REFERENCE

[1]. Krishnaswamy, S., Zaslasvky, A., and Loke, S, W., "Internet Delivery of Distributed Data Mining Services: Architectures, Issues and Prospects ", Architectural Issues of Web-enabled Electronic Business, Murthy, V.K. and Shi, N. (eds.), 2003

[2]. D.B.Skillicorn and D. Talia, "Mining large data sets on grids: Issues and prospects ", Computing and Informatics, Special Issue on Grid Computing, 2002.

[3]. D.B.Skillicorn, "The case for datacentric grids", Workshop on Massively Parallel Programming, IPDPS 2002.

[4]. Ian Foster et aI., "The Physiology of the Grid: An Open Grid Services, Architecture for Distributed Systems Integration ", Technical Report, Globus Project, 2004.

[5]. Sue Spielman, "The Struts Framework - Practical Guide for Java Programming", Morgan Kaufmann Publishers, 2003.

[6]. Malcolm Davis, "Struts, an open-source MVC implementation ", IBM Technical article, 2001.

[7]. Rod Johnson, "J2EE Design and Development", Wrox, 2002

[8]. Apache Jakarta Project. http.//jakarta.apache.org/