

Conservazione delle basi di dati: la sfida internazionale e la soluzione dell' Archivio Federale Svizzero

Una gran parte delle informazioni amministrative è registrata in basi di dati. La sfida di oggi consiste nel conservare l'informazione e renderla accessibile per gli anni a venire, garantendo la trasmissione della conoscenza, ma anche la sostenibilità per le amministrazioni. La mancanza di standardizzazione ha reso fino ad ora molto complessa l'archiviazione del contenuto delle basi di dati. L'Archivio Federale Svizzero ha sviluppato un nuovo formato basato sul linguaggio XML che consente la conservazione a lungo termine del contenuto delle basi di dati relazionali. SIARD (Software-Independent Archiving of Relational Databases) offre una soluzione unica per la conservazione dei dati, dei metadati e delle relazioni tra di essi in un formato conforme allo standard ISO.

Perché archiviare le basi di dati?

La conservazione a lungo termine dei dati è sempre stata di importanza cruciale per le amministrazioni, garantendo la pianificazione e la stabilità. Oggi è opinione comune che i dati elettronici siano già al sicuro e di conseguenza l'archiviazione è spesso considerata superflua, poiché l'abitudine ad accedere ai contenuti con un semplice doppio click ha influenzato il modo in cui noi pensiamo alla conservazione dei documenti digitali.

Nondimeno la conservazione delle basi di dati ha un senso. In primo luogo, in un contesto tecnologico in continua evoluzione solo l'archiviazione può veramente garantire l'accesso ai dati e prevenirne la perdita. In secondo luogo, circa l'85% dei dati immagazzinati è inattivo, rendendo le basi di dati troppo complesse e costose da mantenere. [1] Infine l'archiviazione è spesso richiesta dalla legge, per garantire la libertà d'informazione (es. la legge svizzera sulla trasparenza amministrativa) o per documentare le attività del governo (es. il francese Code du patrimoine). L'archiviazione dunque offre una buona risposta alle nostre necessità, adempiendo agli obblighi di legge, facilitando la gestione dei dati e abbattendo i costi delle operazioni. Si tratta tuttavia di un lavoro difficile e che nasconde alcune insidie.

Criticità dell'archiviazione: cosa archiviare?

Una breve storia delle basi di dati ci aiuterà a far luce sul principale aspetto critico della loro archiviazione. Le prime basi di dati degli anni '60 erano strutturate secondo una chiara gerarchia (relazioni 1:1 o 1:n). Questa struttura ad albero era incline a ridondanze, necessarie per consentire relazioni complesse (n:m). Il modello gerarchico fu dunque sostituito dal modello reticolare che consente di rappresentare relazioni multiple senza ripetizioni. Più tardi veniva introdotto un altro modello, il modello a oggetti (object-oriented), che consiste in classi di informazioni che rappresentano i dati. Sebbene sia possibile accedere velocemente ai dati, in questo modello il numero delle query è limitato. Oltre a tutte le particolari differenze, i modelli di dati sopra descritti hanno una cosa in comune, il rapporto di dipendenza tra dati e codice, ossia il linguaggio software con il quale è stata creata la base di dati. Questa stretta relazione complica l'estrazione dei dati dal codice. Se non conosciamo il codice l'archiviazione diventa impossibile.

C'è però un'eccezione alla regola: le basi di dati relazionali. Questo modello introdotto intorno al 1970 da Edgar Codd, risolve il problema della dipendenza tra dati e codice, immagazzinando i dati in tabelle. Questa serie di tabelle correlate consente relazioni multiple (n:m) e un numero indefinito di query. L'uso delle chiavi primarie (identificatori univoci per ciascuna riga) e delle chiavi esterne (riferimenti ad altre tabelle) elimina la necessità di ripetere le informazioni. E anche se il software cambia, il contenuto dei dati rimane immutato. Tutto ciò che dobbiamo fare per archiviare i dati è estrarre le tabelle e memorizzarle. Archiviare le basi di dati relazionali è più semplice, in termini di lavoro e di costi. Dal momento che più del 90% delle basi di dati è costituito da basi di dati relazionali, la migliore strategia è probabilmente quella di concentrare i nostri sforzi sulla loro conservazione. Il modello relazionale rappresenta la soluzione al principale fattore di criticità dell'archiviazione. Ma questo è solo il primo passo. Il secondo, e forse il più complicato, consiste nel trovare un formato idoneo a garantire l'accesso futuro ai dati immagazzinati. Questo è esattamente ciò che l'Archivio Federale Svizzero ha cercato di fare.

Note

[1] Yuhanna, Noel "Database Archiving Remains an Important Part of Enterprise DBMS Strategy", Information & Knowledge Management Professionals (2007): <ftp://ftp.software.ibm.com/software/data/sw-library/data-management/optim/reports/forrester-archiving.pdf>

[2] ARELDA è o acronimo per ARchiving of ELectronic Data.

[3] <http://www.planets-proejct.eu/>

[4] O SFA attualmente utilizza una aplicação baseada no sistema JAVA, o SIARD Suite, que permite navegar através do arquivo SIARD e adicionar ou actualizar metadados.

Bibliografia

1 - Estratégia de aquisição e disposições dos Arquivos Nacionais (2007)

http://www.nationalarchives.gov.uk/documents/acquisitions_strategy.pdf

2 - Codd, E.F., "A Relational Model of Data for Large Shared Data Banks", Communications of the ACM, vol. 13, n.º 6 (1970), 377-387.

3 - Code du Patrimoine, July 30, 2008.

<http://www.legifrance.gouv.fr/affichCode.do?sessionId=2FAA76FF7AE923389AC2146821608165.tp&cidTexte=LEGITEXT000006074236&dateTexte=20081001>

4 - Knowles, J.S. / Bell, D.M.R., "The Codasyl Model", in: Databases - Role and structure, P. M. Strocker, P. M. D. Gray, and M. P. Atkinson (eds) CUP, 1984. Swiss Federal Law on Archiving (BGA), June 26 1998: http://www.admin.ch/ch/d/sr/cl15_2_1.html

5 - Swiss Federal Law on the freedom of information in the federal administration (Offentlichkeitsgesetz, BGO), December 17, 2004: http://www.admin.ch/ch/d/sr/c152_3.html

La soluzione svizzera: il formato SIARD

L'Archivio Federale Svizzero (AFS) lavora sulla conservazione delle basi di dati sin dalla fine degli anni '90. L'AFS effettuò la scelta strategica di archiviare solo basi di dati relazionali. Come parte del progetto ARELDA, venne ideato e sviluppato un nuovo formato per l'archiviazione di basi di dati relazionali. [2] Il formato SIARD (Software-Independent Archiving of Relational Databases) fu presentato nel 2004. Da allora è stato elaborato e considerevolmente perfezionato nell'ambito del progetto PLANETS. [3] Verso la fine dell'estate 2008 l'Archivio Federale Svizzero presentò una versione definitiva del formato SIARD unitamente al software [4].

Cosa significa la conservazione a lungo termine con SIARD in termini pratici? Il software SIARD converte basi di dati proprietarie (MS Access, MS SQL, Oracle, etc.) nel formato non proprietario SIARD. L'archivio SIARD (con estensione di file .siard) rappresenta la base di dati nella sua struttura logica, mantenendo non solo i dati primari e i metadati, ma soprattutto le relazioni.

Un archivio SIARD consiste in un file strutturato, in formato ZIP-64 standard che accetta praticamente qualsiasi dimensione di file. Esso contiene due cartelle: la cartella header e la cartella content. La cartella header contiene la struttura della base di dati, i metadati. Un unico file, metadata.xml, ci consente di comprendere gli aspetti tecnici e il contesto originario della base di dati. In termini tecnici, al livello più alto (base di dati) SIARD memorizza l'identificatore, la versione del formato, il checksum crittografico (per verificare l'integrità dei dati primari), etc. A livello di schema, SIARD memorizza le liste di tabelle, le viste e le routine. A livello di tabella, registra infine i vincoli e i trigger. Andando ancora più in profondità, a livello di colonna SIARD specifica anche il tipo di SQL usato, i nomi dei LOBs (large objects) e soprattutto le chiavi esterne e le chiavi candidate con le relazioni associate. Allo stesso tempo SIARD contestualizza i dati. A livello di base di dati SIARD ci permette di registrare o aggiungere (mediante la suite SIARD) informazioni sulla provenienza dell'archivio, la sua descrizione, gli utenti, etc. Nei livelli inferiori ci permette di mantenere i dettagli delle tabelle, i nomi delle colonne e il contenuto. Queste informazioni descrittive rendono la base di dati intellegibile per gli utenti futuri, sia sotto l'aspetto tecnico che del contesto. La seconda cartella, content, memorizza i dati primari. I dati sono archiviati rispettando la struttura della base di dati. Per ciascuno schema SIARD genera automaticamente una cartella (schema 1, schema 2, etc.) contenente la corrispondente serie di tabelle in sottocartelle (tabella 1, tabella 2, etc.). I dati sono memorizzati in file XML (es. tabella1.xml). Lo schema così definito riflette lo schema di metadati delle tabelle SQL. Specifica inoltre che la tabella è memorizzata come una sequenza di linee contenenti i valori delle colonne, con differenti modelli XML. Vengono archiviati anche BLOBs e CLOBs (Binary or Character Large Objects), contenenti qualsiasi tipo di informazione, memorizzati in cartelle generate automaticamente (es. lob1, lob2, etc.) in file TXT o BIN (es. record1.text, oppure record1.bin, etc.).

SIARD costituisce dunque una copia fedele della base di dati archiviata. Usando SIARD infatti conserviamo sia i dati primari sia i metadati in un modo e in una forma che rendono la base di dati comprensibile e accessibile. Ma per quanto tempo?

SIARD e la conservazione a lungo termine

"L'eternità è molto lunga", dice Woody Allen, "specialmente verso la fine". Se parliamo di IT l'eternità invece potrebbe essere anche molto breve. La rapida obsolescenza dei formati minaccia l'accessibilità futura dei dati. Cosa potrebbe ridurre questo rischio? Una parola sola: standardizzazione.

L'uso degli standard ISO largamente accettati assicura in larga misura che i dati immagazzinati possano essere accessibili in futuro. Basandosi su questa affermazione SIARD memorizza automaticamente sia i dati primari che i metadati in formati standard ISO: SQL1999, UNICODE e, più importante di tutti, XML 1.0. Per garantire la standardizzazione SIARD converte tutta la struttura delle basi di dati proprietarie in un equivalente set di caratteri UNICODE. Inoltre SIARD non archivia sinonimi dato che essi non fanno parte dello standard SQL:1999. Aderire agli standard è una regola ferrea.

Conclusioni

SIARD è concepito come un formato open source. La sua descrizione è disponibile sul sito web dell'Archivio Federale Svizzero [5]. Non pretende di essere la soluzione per l'archiviazione di tutti i modelli di basi di dati, ma è comunque una soluzione fattibile e concreta per la conservazione a lungo termine delle basi di dati relazionali.