Ultime novità sull'archiviazione digitale dal SUN PASIG Meeting di Parigi

Roberto Gibellini
CILEA, Segrate

Abstract
Il mondo delle biblioteche ed in generale delle organizzazioni che si occupano di informazione necessita di soluzioni per gestire, conservare e rendere accessibili le diverse e sempre più rapidamente crescenti moli di dati digitali che vengono poste sotto la loro cura. Attraverso l'uso di nuove "open solutions" che comprendono: applicazione software, hardware, sistemi operativi, storage e servizi di consulenza, si è ora in grado di far fronte a queste sfide (Sun PASIG Meeting, Parigi, 14-16 novembre 2007).

Libraries and information organisations are in need of solutions to manage and preserve the diverse and rapidly increasing digital content that is being placed in their care. Through the use of new open solutions including application software, hardware, operating systems, storage solutions and consulting services, these organisations are now able to rapidly deploy a solution that meets the challenges (Sun PASIG Meeting, Paris, November 14-16, 2007).

Keyword: Preservation, Digital Archiving, Open Solutions Architecture.

Introduzione

Lo scopo del "Sun Preservation and Archiving Special Interest Group" [1] (Sun PASIG), tenutosi a Parigi lo scorso novembre (14, 15 e 16), è stato quello di offrire uno spazio di discussione, aperto a organizzazioni commerciali (SUN Solaris [2] "in primis") e istituzioni quali biblioteche nazionali, comunità open source, università, consorzi, etc., coinvolte nella complessa realtà dell'archiviazione e della conservazione digitale.

Considerando il campo dell'archiviazione digitale come una delle più grandi sfide tecnologiche e culturali del prossimo futuro, l'evento si è focalizzato sulla condivisione di "Open Computing Solutions" e "Best Practices" nei seguenti campi di interesse:

- confronto tra diversi modelli di architetture OAIS [3], architetture "services-oriented" e casi d'uso di interesse pratico;
- condivisione di codice software e soluzioni implementative;
- cooperazione in materia di "Open Standards";
- "storage architectures" in relazione alla conservazione, archiviazione e la gestione di dati;
- interazione tra soluzioni sviluppate da "open-community" di settore e organizzazioni commerciali.

Panoramica sul meeting

Nei tre giorni di meeting si sono affrontati diversi aspetti di questa complessa, e per certi versi ancora da definire, nuova realtà che va via via assumendo un peso sempre maggiore.

Si è passati da una panoramica di carattere generale durante, le presentazioni del primo giorno, a un analisi più dettagliata con la costituzione di diversi gruppi di lavoro in cui i vari partecipanti hanno condiviso le loro esperienze e conoscenze al fine di creare modelli e linee guide che potessero in futuro diventare degli standard.

Primo giorno

Oltre alle presentazioni che avevano l'obiettivo di descrivere, anche numericamente, la crescente necessità di affrontare con urgenza e completezza il fenomeno dell'archiviazione digitale, durante questa prima giornata ha avuto luogo anche un interessante incontro durante il quale è stato introdotto il nuovo *storage* della SUN. Lo StorageTek 5800 (Honeycomb) [4] è un

sistema di ultima generazione ottimizzato per l'archiviazione di dati statici su vasta scala. È stato definito dai tecnici SUN come "il primo di una nuova serie di sistemi storage a oggetti di terza generazione e semplifica il modo in cui gli oggetti vengono memorizzati, recuperati e trattati, fornendo nel contempo una garanzia di protezione inerente dei dati stessi".

L'Honeycomb System si basa su Solaris 10 ed è il primo sistema storage per contenuti statici commercialmente disponibile il cui codice è accessibile in ambito open source. Tra le caratteristiche principali di questo sistema possiamo citare:
– integrazione dello storage, del server, del sistema operativo e del sistema di management;
– garanzia di "failure tolerance" e integrità dei dati;
– design simmetrico in modalità "clustered" che permette una facile scalabilità orizzontale;
– protezione dei dati grazie alla codifica Reed-Solomon usata in configurazione RAID 6;
– gestione dei metadati e "query capability" integrate nel sistema;
– bit rot protection, real-time checksums;
– possibilità di amministrare fino a un petabyte di dati.

Vista la chiara svolta open source della SUN, almeno per quanto riguarda questo progetto, vogliamo sottolineare come anche i clienti appartenenti ai settori universitario, sanitario e scientifico come The Alberta Library, CareStream Healthcare, Johns Hopkins University, Max Planck Institute, Oxford University [5], Purdue University, Southampton University, Stanford University [6], University of Calgary e University of Michigan, siano impegnati nell'implementazione del nuovo sistema, migliorando così i livelli di affidabilità, efficienza e garanzia dell'integrità dei dati e riducendo nel contempo i costi associati allo storage. Anche un crescente numero di partner, organizzazioni open source e ISV (Independent Software Vendor) sta lavorando insieme a SUN per sviluppare prodotti ed applicazioni da integrare nel sistema StorageTek 5800: per esempio BackBone, EPrints Services [7], Fedora Commons [8], General Atomics - Nirvana, StorageSwitch, Tiani Spirit e VTLS [9].

Secondo – Terzo giorno

Un ruolo di primo piano lo hanno avuto i gruppi di lavoro che si sono costituiti il secondo giorno del meeting e che hanno poi collaborato al fine di elaborare modelli comuni per tutta la sua durata. Gli aspetti analizzati all'interno di questi gruppi sono stati molteplici, anche se in questo articolo ci si soffermerà su quelli più importanti che l'architettura di un "Repository Documentale" dovrebbe considerare.

Partendo dall'assioma che la gestione di un sistema di archiviazione moderno non può prescindere dall'utilizzo dei metadati [10], vediamo qui di seguito elencate alcune delle problematiche che si devono affrontare:
– modellizzazione e astrazione dei contenuti:
  1. metadati strutturati;
  2. diverse politiche di storage per la prevenzione di diverse tipologie di rischio (perdita di dati sensibili, perdite di dati di valore storico/culturale, etc.);
  3. virtualizzazione/astrazione dei dati;
– storage:
  1. astrazione del livello di storage, per le applicazioni la via di accesso ai dati deve essere univoca e indipendente dal tipo di storage utilizzato;
  2. copie multiple degli stessi dati, anche in formati diversi, per diversi tipi di utilizzo del dato stesso (per esempio fruizione via web, archiviazione, etc.);
– accesso al sistema: sicurezza, autenticazione, autorizzazioni, controlli, polizie, etc.;
– servizi e policy per la conservazione dei dati;
– creazione dei metadati (entry: descrittive, tecniche, relative ad eventi, diritti sui documenti, provenienza, etc.);
– gestione dei metadati (acquisizione, *delivery*, standardizzazione, semantica, etc.);
– sistema di ricerca e indicizzazione dei dati;
– servizi di trasformazione/conversione dei dati (gestione dei diversi formati anche in proiezione futura);
– servizi di verifica dell'autenticità dei dati (firma digitale);
– reporting (log, statistiche, gestione degli errori);
– Workflow Engine (inserimento dati, migrazione, QoS, duplicazione, etc.);
– Business Continuity/Disaster Recovery;
– infrastuttura hardware (*clustering*, architettura piattaforma);
– gestione ambienti: di test, di sviluppo, pre-produzione, etc.;

Conclusioni

Le considerazioni finali che si possono trarre da questo interessante spazio di confronto organizzato dalla SUN sono le seguenti:
– *i metadati sono al centro dell'intero sistema.* Solo grazie a essi, infatti, si possono realizzare tutte quelle operazioni sull'oggetto digitale

(immagini, pdf, file multimediali, etc.) che sono proprie di un sistema di archiviazione documentale, quali la gestione, la visualizzazione, la ricerca, etc. Sfortunatamente *non esiste uno standard internazionale per la generazione dei metadati*, solitamente nascono da un accordo tra le parti in gioco;

– *non esistono soluzioni "open source" o commerciali in grado di coprire l'intero progetto*. L'approccio vincente è quello di studiare a tavolino l'architettura nel suo complesso, avendo ben chiari gli obbiettivi da raggiungere e il tipo di servizio che si vuole offrire, specialmente in termini di QoS. Una volta scelta la strada da intraprendere, trovare i giusti "partner/community" per realizzare il lavoro sembra essere il logico passo successivo. A tal proposito, si vuole segnalare l'organizzazione non-profit Fedora Commons [8], che gioca sicuramente un ruolo di primo piano quando si parla di progetti di questo tipo.

Bibliografia

[1] URL: http://sun-pasig.org/index.html
[2] URL: http://www.sun.com
[3] URL: http://public.ccsds.org/publications/archive/650x0b1.pdf
[4] URL: http://www.sun.com/storagetek/disk_systems/enterprise/5800/
[5] URL: http://www.ouls.ox.ac.uk/
[6] URL: http://www-diglib.stanford.edu/
[7] URL: http://www.eprints.org/
[8] URL: http://www.fedora-commons.org/
[9] URL: http://www.vtls.com/
[10] URL: http://it.wikipedia.org/wiki/Metadato