IL DILUVIO DIGITALE

di Federico Ruggieri

L'arrivo di Internet e del Web ha reso globale il "mercato" dell'informazione digitale, ed ha anche messo a nudo una serie di problematiche legate all'informazione digitalizzata. Fino all'avvento di questi sistemi le problematiche classiche di archiviazione, accesso e mantenimento delle informazioni erano gestite da istituzioni (Archivi di Stato, Biblioteche Nazionali) oppure dalle organizzazioni stesse che producevano l'informazione più disparata e la custodivano finché lo reputavano necessario (Cineteche, Editori, Centri di Ricerca, Laboratori). La carta è stata per molti secoli il veicolo quasi esclusivo di trasporto ed archiviazione della conoscenza ed ha svolto il proprio compito con ammirevole longevità. Ancora oggi, documenti vecchi di secoli, possono essere esaminati, anche se in condizioni di estrema cautela.

L'avvento dei calcolatori elettronici e dei sistemi di archiviazioni digitali basati su file ha permesso un accesso all'informazione rapido e simultaneo per più persone, ha facilitato la riutilizzabilità dell'informazione e la sua archiviazione su svariati mezzi fisici magnetici ed ottici.

Nel panorama dell'archiviazione di grandi quantità di dati digitali la Fisica delle Alte Energie si è reputata a lungo uno dei campi con maggiori esigenze, sia in termini di quantità di dati da archiviare, che in LHC ammontano a decine o centinaia di PetaBytes (PB = 10^{15} Bytes), sia in termini di numero di persone che devono accedervi in maniera facile e contemporanea: migliaia di ricercatori sparsi per il mondo.

La disponibilità di grandi quantità di dati digitali è, in realtà, un fenomeno molto vasto che riguarda molti campi: Beni Culturali (Archivi e Biblioteche); Scienza (Genomica, Medicina, Scienze della Terra); Industria (Petrolifera, Assicurativa, Commercio).

Inoltre, con l'aumentare delle comunicazioni umane per via telematica nei vari campi, dall'istruzione al commercio, si sono venute anche a creare in rete delle tracce digitali numerosissime che costituiscono un diluvio di dati comportamentali che possono essere preziosi, sia per la modellazione e comprensione del comportamento umano, sia per usi di ricerca, che commerciali. Questa enorme quantità di dati oltre ad offrire nuove opportunità, crea però preoccupazioni inerenti ad un ampio spettro di questioni legate alla gestione, conservazione, accesso, proprietà intellettuale e protezione di dati, anche personali.

La realtà del mondo digitale in cui viviamo, presenta molti elementi di sfida tecnologica e svariati campi in cui queste sfide devono essere raccolte, pena la perdita di informazioni o la loro inutilizzabilità. Il problema che si deve affrontare è della stessa importanza di quello che vide lo sforzo dei monaci amanuensi per il salvataggio della cultura nello scorso millennio. Alcune delle domande a cui si cerca di rispondere sono: quanti dati digitali produciamo che "meritano" di essere salvati? Ed ancora: chi decide cosa va salvato? E poi: chi paga perché questi dati restino disponibili per la "società globale" di Internet e del Web?

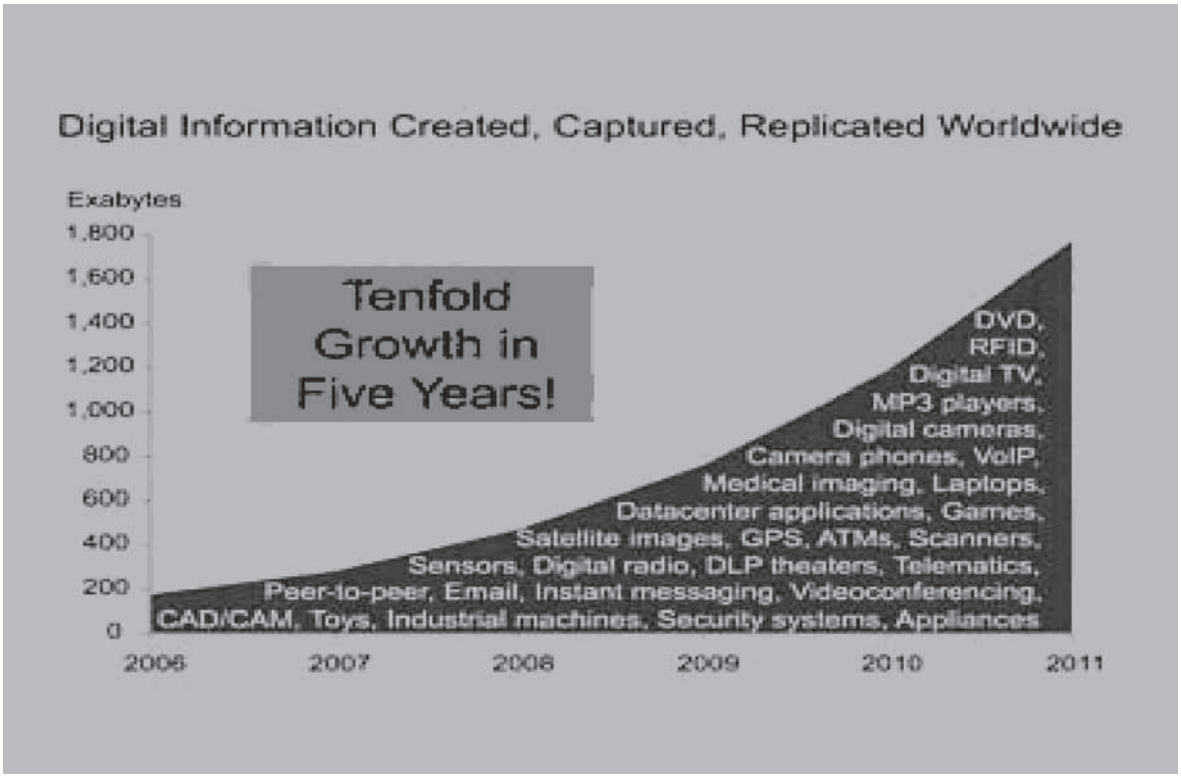
L'attenzione al Diluvio dei dati Digitali è recente, ma le prime discussioni erano iniziate già alla fine dello scorso millennio. In particolare, nel settore della ricerca ci si era, da tempo, posto il problema della riutilizzabilità dei dati provenienti da esperimenti ed apparecchiature scientifiche. Un esempio evidente sono i dati di varia natura provenienti dai satelliti che possono essere riutilizzati in vari contesti e con finalità diverse: dalla protezione civile a Google Maps.

La problematica è talmente sentita in molti paesi, che ci si è posti già il problema di come cambiare l'insegnamento per preparare meglio gli studenti, fin dai primi anni, ad affrontare queste problematiche con le metodologie necessarie Fostering Learning in the Networked World: The Cyberlearning Opportunity and Challenge [1].

La Commissione Europea ha avviato nell'ambito delle attività sulle eInfrastructures una riflessione su queste problematiche nel-

l'ambito dei dati scientifici [2]. Questo però e stato solo uno degli ultimi sforzi di supportare le decisioni dei prossimi anni sull'argomento. Molte nazioni infatti hanno già da tempo prodotto studi e linee guida sulla materia delle grandi infrastrutture di ricerca che includono anche archivi di dati [5].

Le problematiche relative al "Diluvio" dei Dati Digitali presentano almeno due aspetti che meritano approfondimento: l'aumento esponenziale dei dati digitali e la conservazione a lungo termine dei dati ritenuti rilevanti.



Come sale la Marea dei Dati (Fonte IDC)

COME SALE LA MAREA DEI DATI

Una valutazione riportata in un articolo su The Economist [3], valutava in 150 ExaBytes (EB = 10^{18} Bytes) la quantità di dati creata nel 2005 in tutto il mondo e che nel 2010 tale quantità era prevista in 1200 ExaBytes. Un'altra stima riportata da IDC [4] quota in 281 ExaBytes il totale dei dati nel 2007 e, per il 2011 una previsione di un fattore 10 rispetto al 2006. Nel 2007 si sarebbe verificato, per la prima volta, il superamento dello spazio di archiviazione disponibile nel mondo. Da quel momento non tutta l'informazione creata può più essere conservata e dal 2011 è previsto che il 50% di tale informazione non avrà spazio per essere registrata.

Questa inarrestabile marea di dati è prodotta in vari modi e si presenta in varie fogge (o tipi di Files), dai film registrati con dimensioni di alcuni GBytes (GB = 10^9 Bytes), alle foto archiviate, fino a dei semplici numeri telefonici o indirizzi IP. Inaspettatamente neanche la metà dei dati è prodotta intenzionalmente dalle persone, mentre la maggioranza è costituita da dati di foto di sorveglianza, da tracce di navigazione Web, archivi storici di transazioni e liste di mail, ecc.. Insomma buona parte della marea di dati che sommerge il mondo è un sottoprodotto delle attività primarie.

QUANTO DURANO I DATI

Una pubblicità ben nota citava *"un diamante è per sempre"*; non è così per i dati archiviati. I mezzi di archiviazione informatici utilizzati sono basati su sistemi magnetici (hard disk, nastri) o ottici (CD, DVD, BluRay). Ognuno di questi sistemi ha capacità di archiviazione dif-

ferenti: un disco magnetico o un nastro viaggiano attualmente sui 2 TeraBytes (TB = 10^{12} Bytes) mentre un disco ottico BluRay può contenere fino a 25 GigaBytes (GB = 10^9 Bytes) per faccia. Queste capacità, enormi rispetto a pochi anni fa, sembrano quasi ridicole ed inadeguate a sostenere l'impatto della "marea" in arrivo.

L'aspetto più rilevante è però la durata dei dati archiviati su tali supporti: nessuno di essi garantisce il loro mantenimento per più di 20 o 30 anni, anche nelle migliori condizioni di conservazione.

Un altro aspetto da considerare è che, con la rapida evoluzione della tecnologia, è abbastanza improbabile che si possa ritrovare un lettore capace di rileggerli a distanza di decine di anni. Si pensi ai floppy disk: nessun computer acquistato negli ultimi anni ha più a bordo tale dispositivo "antidiluviano".

Dunque chi pensa di salvare i propri dati per decenni, ed oltre, ha la preoccupazione di travasarli regolarmente su supporti più moderni ogni 10 anni al massimo. Ma ci saranno ancora disponibili le applicazioni che sanno leggere ed interpretare quei dati? Anche questo aspetto ricade sulle spalle del creatore/utilizzatore dei dati stessi.

È dunque lecito porsi la domanda su quale debba essere la vita dei dati e quanti di essi debbano essere salvati perché non vada parsa informazione importante sotto il profilo culturale e storico, ma ci sia anche un limite alle risorse da impiegare.

LA SALVAGUARDIA DEI DATI

Salvaguardare i dati è un compito non banale e che, come si ricava dalle premesse, non può essere affrontato prescindendo dalla natura e dall'uso che viene fatto dei dati stessi. Non esiste, insomma, una soluzione per tutte le necessità.

Il Web Archive [6] è un esempio delle iniziative che tentano di salvaguardare il patrimonio culturale espresso via Web e che va irrimediabilmente perso quando un sito web aggiorna i suoi contenuti o viene spento. Il sito consente di rivedere pagine archiviate attraverso il servizio di "The Wayback Machine". Il contenuto delle pagine archiviate risale fino al 1996 e l'attività è frutto della collaborazione di più di 130 organizzazioni pubbliche e private, fra cui la Biblioteca Alessandrina di Alessandria in Egitto [7] che provvede a fornire un servizio alternativo di *back-up*. Nel 2007 è stato effettuata una scansione completa di 2 milioni di pagine web e tale archivio costituisce una fotografia del web per quell'anno.

Una certa quantità di dati deve essere conservata per anni per obblighi istituzionali o di legge: si pensi ad esempio ai dati delle transazioni finanziarie e bancarie oppure agli archivi di traffico degli operatori telefonici, mentre le pubbliche amministrazioni sono spinte all'utilizzo di archivi informatici ed alla eliminazione di quelli cartacei. La maggior parte di questi archivi hanno la caratteristica di essere scritti una volta e raramente consultati, per cui possono essere conservati con sistemi *"off-line"* come ad esempio armadi in cui conservare dischi ottici e/o nastri magnetici.

Questi dati che, in molti casi non verranno più consultati, si deterioreranno lentamente nel tempo e verranno mantenuti al sicuro per i periodi previsti (tipicamente almeno 5 anni), passati i quali, quando ci saranno esigenze di spazio, saranno distrutti per far posto a nuovi archivi. Nella sostanza subiranno la stessa fine della carta mandata al macero, con qualche preoccupazione in più per lo smaltimento.

In alcuni campi, come la medicina, i dati relativi alle persone fisiche (es. i pazienti) richiedono un trattamento che protegga la privacy, ma allo stesso tempo permetta un accesso ad applicazioni che li analizzano sotto il profilo statistico per l'individuazione dell'efficacia di trattamenti o individuazione di marcatori che permettano una diagnosi precoce di alcune malattie. Le tecniche in questi casi, oltre all'anonimizzazione dei dati, devono prevedere la non trasferibilità degli stessi al di fuori delle cliniche e/o ospedali dove sono stati raccolti. Le analisi sui dati devono, conseguentemente, essere effettuate su risorse di calcolo a disposizione *in situ* e solo i risultati delle analisi possono poi essere trasferiti altrove.

Per tutte quelle applicazioni, invece, in cui i dati devono essere mantenuti a lungo "*on-line*", cioè a disposizione degli utilizzatori, nel passato si sono ampiamente utilizzate strategie di archiviazione gerarchica (Hierarchical Storage Management) in cui i dati ad accesso più frequente sono registrati sui supporti più veloci e con minor tempo di accesso (tipicamente dischi magnetici o a stato solido), mentre i dati scarsa-

mente utilizzati vengono archiviati su dispositivi più economici e con minore consumo energetico (generalmente librerie di nastri magnetici o supporti ottici).

Per i dati di LHC è stato studiato il problema che ha almeno due facce: da un lato la necessità di archiviare una grande quantità di dati (circa 15 PetaBytes/anno) e dall'altro lato permettere l'accesso a tali dati alla comunità dei ricercatori che è dell'ordine di diverse migliaia di persone sparse in tutto il mondo. Tutto questo rispettando dei ragionevoli limiti di costo e consentendo a tutti i ricercatori le stesse opportunità di accesso ed analisi dei dati.

Sono queste necessità che hanno portato la comunità scientifica di LHC a scegliere del 1999 di avviare un progetto per la realizzazione di una infrastruttura di Grid Computing che permettesse di affrontare con successo tutte le sfide elencate.

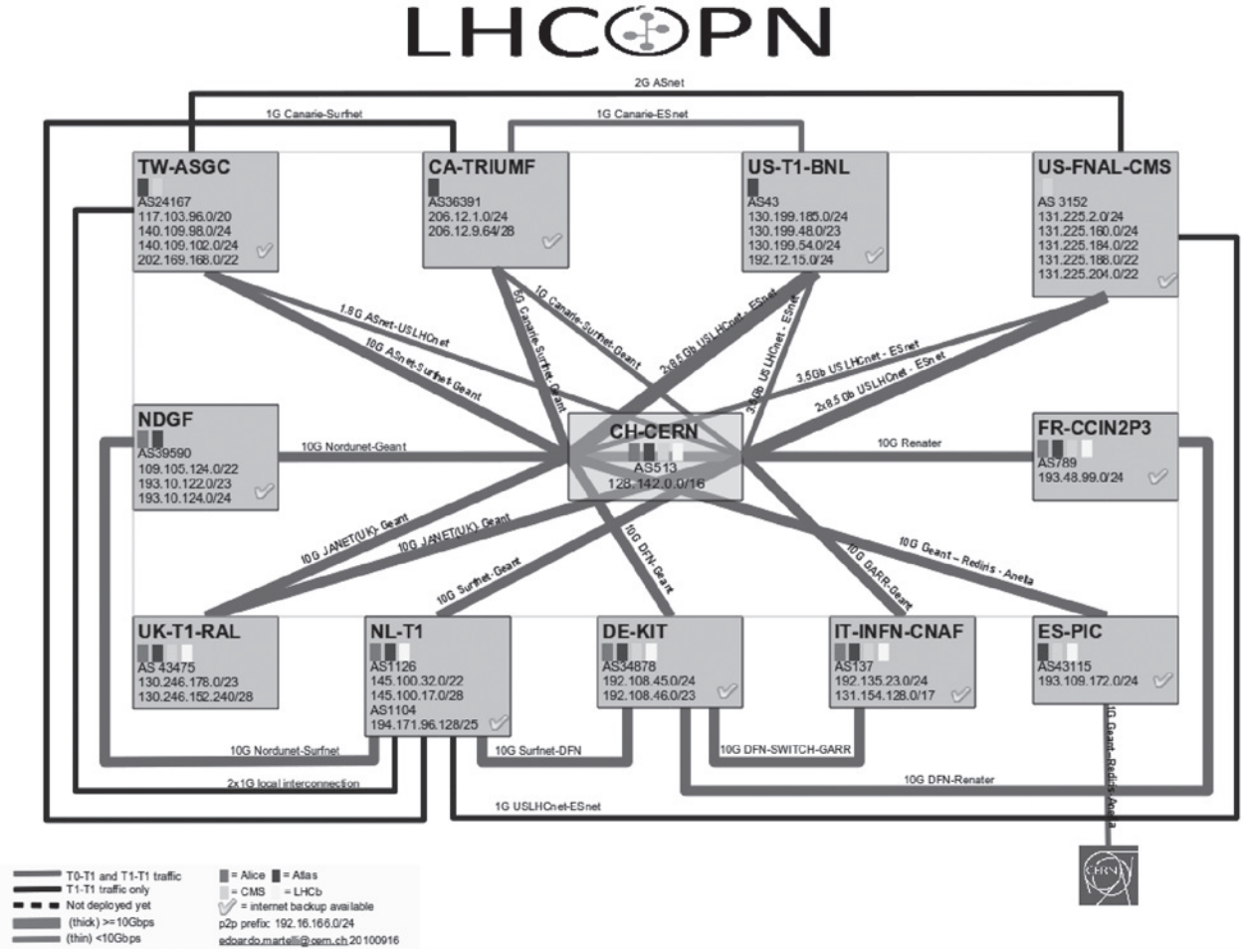


Figura 1 - La rete ottica privata di LHC

La struttura dei dati è distribuita fra il CERN ed i laboratori esterni in una struttura a livelli o "Tier" in cui il CERN svolge la funzione di sito dove i dati vengono creati (Tier0) e poi da esso i dati vengono distribuiti alle strutture principali (Tier1) nelle varie regioni del mondo e da questi i dati sono resi accessibili ai livelli successivi (Tier2 e Tier3) presso altri laboratori che fanno riferimento ciascuno al proprio Tier1.

Per realizzare il trasferimento continuo di una tale enorme quantità di dati fra il Tier0 ed i Tier1è stata realizzata una rete ottica dedicata "LHCOPN" [8] con connessioni da un minimo di 1 Gbps a 10 Gbps (Giga bit per secondo) il cui schema è mostrato in figura 1. L'origine riferimento non è stata trovata.

La realizzazione di una tale avanzatissima rete ad alte prestazioni è stata possibile grazie alle reti della ricerca: quella Europea GEANT e quelle delle nazioni come il GARR in Italia.

Bibliografia

[1] Fostering Learning in the Networked World: The Cyberlearning Opportunity and Challenge. A 21st Century Agenda for the National Science Foundation – Report of the NSF Task Force on Cyberlearning – June 24, 2008.
[2] Riding the wave – How Europe can gain from the rising tide of scientific data – October 2010. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf
[3] The Economist: "The Data Deluge" - http://www.economist.com/node/15579717?story_id=15579717
[4] The Diverse and Exploding Digital Universe - An IDC White Paper - sponsored by EMC, March 2008 - http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf
[5] Si vedano i documenti raccolti dal sito GDI2020 "A Vision for Global Research Data Infrastructures. http://www.grdi2020.eu/
[6] http://www.archive.org/
[7] http://archive.bibalex.org
[8] http://lcg.web.cern.ch/LCG/activities/networking/nw-grp.html

FEDERICO DE RUGGIERI

Laureato in Fisica presso l'Università di Bari è attualmente Dirigente di Ricerca dell'INFN presso la Sezione di Roma Tre.Ha svolto la sua attività professionale sperimentale principalmente nel campo della Fisica delle particelle. Ha partecipato alla costruzione, messa in opera ed analisi dei dati di diversi esperimenti al CERN ed a Frascati. Nel corso della sua carriera si è concentrato sugli aspetti del calcolo, delle reti di trasmissione dati e dei sistemi di acquisizione e processamento dei dati.Ha ricoperto molti incarichi: Presidente della Commissione Nazionale Calcolo INFN, Chairman del Comitato Europeo per il Calcolo nella Fisica delle Alte Energie (HEPCCC), Direttore del CNAF (Centro Nazionale per la Ricerca e Sviluppo nelle Tecnologie Informatiche e Telematiche dell'INFN), membro del Comitato Tecnico Scientifico della Rete GARR e del Consorzio CASPUR. In qualità di chairman dello HEPCCC ha avviato il primo Progetto Europeo DataGRID per lo sviluppo di Griglie Computazionali, co-finanziato dalla Commissione Europea ed è stato il Project Manager di successivi progetti Europei di estensione delle infrastrutture Grid nei paesi del Mediterraneo, del Medio Oriente ed in Cina. È anche docente di "Acquisizione Dati e Controllo di Esperimenti" nel corso di laurea magistrale in Fisica dell'Università di Roma Tre.

Contatti
INFN Sezione di Roma Tre Via della Vasca Navale, 84 00146 Roma Italy
Tel. +39 0657337232 Fax +39 0657337059 Email: Federico.Ruggieri@roma3.infn.it