Prototipo per un servizio di cloud storage federato per il mondo accademico e della ricerca

Simon Vocella, Andrea Biancini, Cristiano Valli, Mario Reale, Fabio Farina

Consortium GARR

Abstract. Lo storage personale è una delle categorie di servizio di maggior successo del paradigma cloud. GARR sta svolgendo un progetto di sperimentazione in questa direzione a seguito delle richieste di una parte della propria comunità. In particolare si stanno effettuando studi in merito alle problematiche legate alla creazione di uno storage personale cloud con autenticazione e autorizzazione federata, che abbia elevati standard di resilienza e confidenzialità dei dati e sfrutti al massimo i benefici offerti dalla Federazione IDEM per la gestione delle identità. Il risultato di queste attività è riassunto nell'architettura del servizio GARRbox e nella produzione di un'istanza prototipale. Questo documento discute gli aspetti salienti previsti per il servizio, le lezioni imparate fino ad oggi e disegna le possibili evoluzioni che GARRbox dovrà considerare al fine di rispondere nel miglior modo possibile alle esigenze della Comunità GARR.

1. Introduzione

L'approccio Cloud è al momento la risposta di maggiore successo alla richiesta degli utenti di accedere a servizi complessi in modo semplice e trasparente tramite Internet. Nella definizione proposta dal NIST[1] i servizi Cloud sono fortemente caratterizzati dai seguenti benefici:

- Percezione di risorse illimitate: caratteristica principale del cloud storage è l'espandibilità dinamica delle risorse assegnabili a ciascun utente al variare delle necessità nel tempo.
- Elasticità: le risorse sono assegnate e rilasciate automaticamente in base al carico di lavoro istantaneo, garantendo la continuità del servizio.
- Accesso multi-modale e ubiquo: gli utenti interagiscono con la Cloud utilizzando diversi protocolli e dispositivi di accesso in mobilità, in modo trasparente rispetto ai tecnicismi della soluzione.
- Un modello di business definito: il modello economico cloud è proporzionale alla quantità di risorse consumate.
- Ottimizzazione d'uso delle risorse: il cloud storage model è indipendente dall'hardware utilizzato, permettendo pertanto il riutilizzo di ciò che già si ha (aumento dell'efficienza aziendale sul piano costi/benefici).

La prospettiva di accedere a insiemi di grandi risorse, controllandole direttamente con un costo di utilizzo contenuto, se non marginale, è ovviamente allettante per la comunità della Ricerca Italiana. Per venire incontro a tale richiesta, GARR ha progettato un prototipo di servizio cloud storage chiamato GARRbox, per la sincronizzazione e la condivisione dei dati personali di ricerca in modo semplice, rapido, sicuro, controllabile e indipendente dalla tipologia di risorse hardware a disposizione.

GARRbox rientra nella categoria delle cloud infrastrutturali IaaS e adotta un modello di servizio a cloud pubblica di tipo federato (Community). Il modello Community Cloud, come Helix Nebula [2], rappresenta un'alternativa valida per ridurre sia lo sforzo della messa in opera dei servizi, sia l'impatto finanziario che un unico partner dovrebbe affrontare nel creare un nuovo servizio, velocizzando l'adozione di soluzioni condivise dalla comunità e raggiungendo quindi un bacino di potenziali utenti più ampio.

Una Community Cloud necessita di accordi di federazione, di standard, e di strumenti per l'interoperabilità sia dei piani di controllo delle ri-

Workshop GARR Calcolo e Storage Distribuito - Selected Papers. Roma 29-30 novembre 2012
ISBN: 978-88-905077-4-8 © 2013 Associazione Consortium GARR

sorse sia dei dati. La scelta naturale in questo senso è adottare l'esperienza della federazione IDEM nella gestione delle identità, estendendola alle necessità dello storage cloud federato.

Il contributo descrive l'approccio e i principi che GARR ha seguito per creare un proto-servizio di GARRbox per la propria Direzione, al fine di validare i principi architetturali che saranno estesi per un servizio su scala nazionale.

1.1 Le caratteristiche di GARRbox

GARRbox offre funzionalità simili agli strumenti di sincronizzazione dei dati commerciali come Dropbox, Google Drive e SugarSync. In aggiunta presenta i seguenti benefici:

- È pensato per essere gestito da enti nazionali e sul territorio nazionale, garantendo quindi conformità alle leggi sulla gestione dei dati in materia di privacy, resilienza e copyright.
- È pensato come sforzo comune della comunità R&I Italiana, garantendo economie di scala e condivisione equa dei benefici tra tutti i membri.
- Si basa su un'infrastruttura ad hoc, ma potrà federare eventuali risorse esterne sottoutilizzate, aumentandone l'efficienza e riducendo i costi di messa in opera del servizio, supportando la sostenibilità del servizio sul lungo periodo.
- È sotto l'egida di GARR, a garanzia di equità, dell'apertura agli standard, della certezza che i dati siano preservati e di sostenibilità del servizio per parecchi anni a venire.
- Garantisce confidenzialità e riservatezza dei dati dell'utente finale, tramite l'utilizzo di meccanismi di cifratura.

Come già accennato, GARRbox offre le caratteristiche del paradigma cloud: tramite un piano di controllo unico è possibile accedere in modo semplice ad uno storage distribuito con resilienza e replica dei dati in modo trasparente. Gli utenti percepiscono il servizio come un disco virtuale aggiuntivo o un tool di sincronizzazione, e quando sarà necessario più spazio, potranno ottenerlo in modo elastico.

Da progetto, GARRbox supporta le seguenti tecnologie:

- Autenticazione e autorizzazione federate tramite IDEM.
- Accesso ai dati tramite canali multipli: da browser, cellulare, da riga di comando.

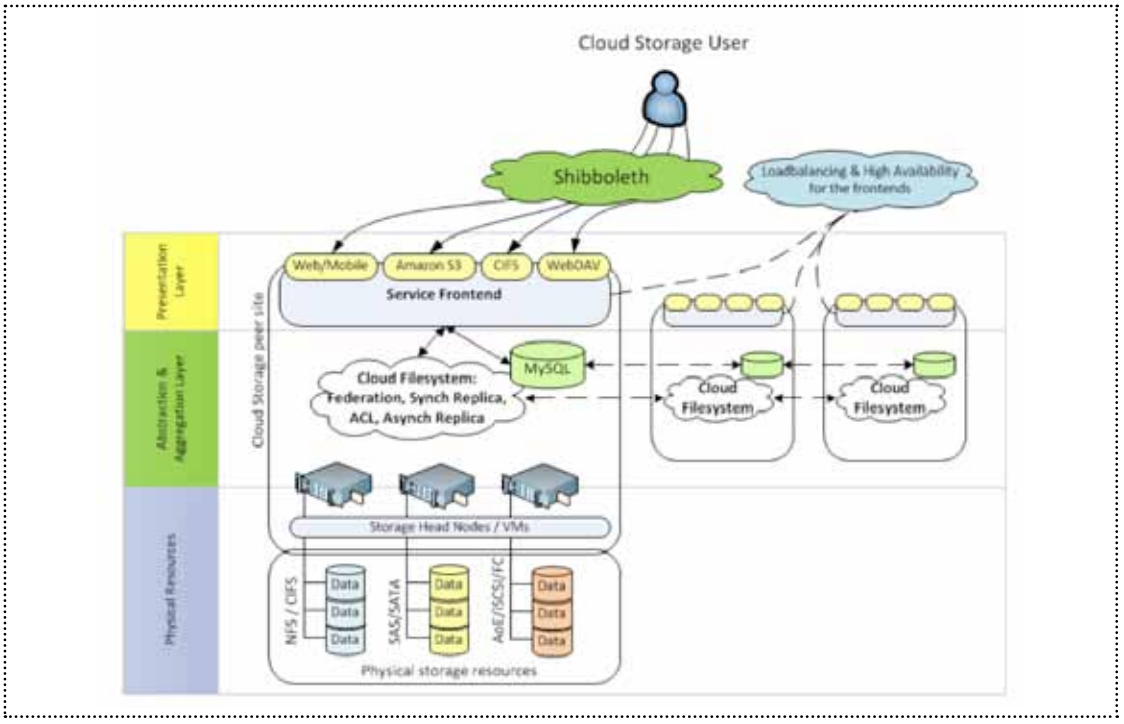

Fig 1 - L'architettura a strati del servizio di cloud storage

- Ridondanza e crittografia dei dati e dei metadati, se richiesto.
- *Logging* e *audit* capillare delle operazioni sui dati.

L'architettura di GARRbox segue il modello *multi-tier* per avere scalabilità, resilienza e supporto di protocolli aperti a ogni strato, semplificando l'adozione e il coinvolgimento di nuovi partner nella federazione. I tre livelli mostrati in Figura 1 sono:

- Le risorse fisiche di storage. GARRbox si basa su risorse di GARR e sulla cooperazione con i membri della Comunità, al fine di mettere a fattor comune risorse distribuite, per garantire replica e bilanciamento.
- Lo strato d'integrazione e federazione, basato su tecnologie di file system cloud distribuiti. Lo strato implementa gli aspetti cloud e di virtualizzazione per garantire l'elasticità delle risorse. Inoltre è responsabile della replicazione dei dati e della loro disponibilità in caso di fallimento.
- I *front-end* con interfaccia multi-canale. Le interfacce spazieranno da portali web a client di sincronizzazione, da applicazioni mobili a protocolli di basso livello con le relative API. Ognuno dei moduli di accesso è integrato con la Federazione IDEM.

Quest'approccio permette di integrare tecnologie eterogenee nei diversi strati, permettendo di restare sempre al passo con le evoluzioni ICT, senza dover mutare il quadro generare del servizio e assicurando in ogni momento ridondanza, alta disponibilità e affidabilità.

1.2 Le differenze tra GARRbox e le offerte commerciali

Molte istituzioni e aziende private sono ancora scettiche nel conservare dati sensibili sulle cloud commerciali. Le ragioni di questa cautela sono principalmente legate ai problemi oggettivi di sicurezza che le cloud pongono. In particolare, i provider commerciali sono in grado di assicurare la persistenza dei dati, ma non sono in grado di garantire il livello di privacy e, in generale, di fiducia necessario per conservare dati sensibili, ad esempio immagini mediche per scopi di ricerca. In aggiunta, i provider commerciali non possono garantire che i dati degli utenti restino dentro i domini imposti dalla giurisdizione Europea .

GARR assicura un alto livello di fiducia alla propria comunità: gestendo la connettività per la Comunità ha già un'ampia esperienza nell'affrontare le problematiche di sicurezza degli utenti, e l'adozione di GARRbox non introduce un nuovo soggetto nella gestione delle informazioni.

GARRbox supera queste limitazioni grazie al ruolo istituzionale di GARR, assicurando che i dati siano ospitati su data center della comunità, con tutti i criteri ottimali di ridondanza a failover. GARRbox assicura che la privacy e le policy di accesso siano regolate tramite IDEM, lasciando agli utenti pieno controllo su come e a chi sono concessi gli accessi ai dati. In aggiunta, i dati potranno essere protetti con i migliori metodi di crittografia, rendendoli inaccessibili anche ai gestori dello storage fisico. Grazie alla Comunità, il servizio non sarà mai soggetto a improvvisi cambiamenti unilaterali degli accordi tra gli utenti e il servizio stesso. Analogamente, GARRbox non sarà soggetto a *vendor lock* a nessun livello, favorendo sempre protocolli e soluzioni aperte.

2. Il prototipo di servizio per la Direzione

Per verificare le scelte architetturali descritte nel paragrafo precedente, abbiamo creato un prototipo di servizio aperto ai soli utenti della Direzione GARR. Il sistema è stato dimensionato per supportare circa trenta utenti, offrendo a ognuno 10 GB di spazio.

Le risorse fisiche sono costituite da dispositivi *off the shelf* accedute tramite POSIX. Per il livello di aggregazione che astrae le risorse fisiche, esponendo un file system resiliente omogeneo, è stato sperimentato il file system distribuito GlusterFS. I dati degli utenti sono replicati in triplice copia localmente in Direzione e ulteriormente mantenuti come backup su un'istanza secondaria ospitata dalla Sezione INFN di Milano-Bicocca. Nello strato di presentazione, ogni in-

terfaccia è realizzata come un modulo indipendente per scalabilità e personalizzazione, ed è implementata come un servizio web che accede ai dati del file system aggregato. Questo strato è anche responsabile delle funzioni di autenticazione e autorizzazione, implementato con Shibboleth SP.

Il prototipo fornisce funzioni di navigazione dei dati tramite un'interfaccia di *file-browsing* per caricare, scaricare e gestire i propri dati su Internet. I file memorizzati sul servizio sono quindi accessibili tramite:

- Un'interfaccia web basata su Ajaxplorer, per accedere ai *repository* e gestire le preferenze, la condivisione e le attività amministrative. Inoltre, il portale espone lo stato della federazione e il monitoraggio.
- Un gateway custom compatibile con il protocollo Amazon S3. L'interfaccia è realizzata per superare la limitazione che il web ha nella gestione delle cartelle annidate. L'interfaccia permette agli utenti di accedere ai propri dati ed eseguirne una migrazione verso provider cloud differenti, utilizzando gli strumenti client standard per l'accesso ai servizi di Amazon.
- Un'interfaccia WebDAV che permette di integrare facilmente le applicazioni preesistenti e i mount point nativi dei diversi sistemi operativi.
- Un'interfaccia di download BitTorrent, che sfruttando funzionalità *peer-to-peer* permette di migliorare la disponibilità di un file, ottimizzando l'utilizzo della banda. Questa interfaccia è stata pensata per aumentare la diffusione di materiale per la divulgazione e la didattica.

2.1 La gestione delle identità

La sicurezza è una questione fondamentale per i servizi Cloud. La certezza che i dati siano accessibili solo dagli utenti che ne hanno diritto è il presupposto fondamentale per le infrastrutture *multi-tenancy*, come i dispositivi fisici sui quali opera il servizio cloud. Il prototipo realizzato affronta sin d'ora queste problematiche adottando gli standard di sicurezza web e integrando i meccanismi di autenticazione e autorizzazione basati su SAML, secondo le disposizioni della Federazione IDEM. Gli attributi sono utilizzati per identificare non solo l'utente, ma anche per estrarre i gruppi cui l'utente appartiene e la sua istituzione. Il prototipo usa queste informazioni per determinare automaticamente la visibilità che l'utente ha sui dati e sui contenuti condivisi. Anche i limiti e le quote di spazio disponibili possono essere definite secondo regole basate su questi attributi. Le regole sono definite dagli amministratori di ogni sito. Le quote sono implementate con una doppia soglia: quando un utente supera la soft-quota riceve una notifica via e-mail. Quando la hard-quota è superata, l'utente non può caricare ulteriori contenuti.

2.2 Condivisione e versioning dei dati

Il prototipo supporta due modelli di condivisione:

- I file possono essere condivisi tramite URL dinamici. L'URL può essere protetto tramite password e può rimanere valido per un tempo limitato specificando una scadenza.
- Le cartelle possono essere condivise con altri utenti del servizio per creare spazi di collaborazione.

A differenza dei servizi di cloud storage pubblici, l'approccio federato consente agli utenti la condivisione delle cartelle anche secondo gli attributi previsti dalla federazione di identità. Nel dettaglio, gli utenti possono condividere le proprie cartelle con un'istituzione o con diversi criteri di raggruppamento, determinati dai valori dei diritti previsti da IDEM. Gli utenti possono specificare quali autorizzazioni hanno i partecipanti alla condivisione sui dati: sola lettura, sola scrittura, o pieno controllo.

Il prototipo prevede anche un sistema di controllo delle versioni per consentire agli utenti di tenere traccia delle modifiche sui file. Il sistema implementa un server GIT [3]: un sistema di controllo di revisione distribuito, che si distingue per efficienza e rapidità. Tramite l'interfaccia web, l'utente gestisce le versioni correnti e precedenti dei dati. Il numero totale di versioni per un file è limitato a un numero fisso e solo le ultime modifiche possono essere ripristina-

te. Questo vincolo è stato introdotto per impedire al numero di versioni di crescere troppo velocemente, penalizzando la reattività e l'efficienza del servizio.

3. Primi feedback degli utenti

Il prototipo è stato rilasciato agli utenti della Direzione il 25 settembre 2012. Ad oggi sono presenti 35 utenti registrati. L'analisi dei log indica che il sistema è acceduto principalmente dall'interfaccia web e che la maggior parte degli utenti usa il sistema con regolarità. Un numero esiguo di utenti (sei) lo utilizza saltuariamente. L'interfaccia compatibile con Amazon è utilizzata dagli utenti più esperti per spostare grandi quantitativi di file organizzati in cartelle annidate. La distribuzione dello spazio occupato mostra il classico andamento a coda lunga, con un utente che ha richiesto più di 10 GB, un ristretto gruppo di utenti che occupa buona parte della quota assegnata e la maggior parte delle utenze abbondantemente sotto la quota assegnata. Le aree di lavoro condiviso create dagli utenti sono 16: questo indica che i nostri utenti prediligono gli strumenti collaborativi offerti. L'architettura a strati ha permesso di minimizzare gli impatti dei malfunzionamenti, avendo un solo incidente risolto in meno di un'ora in sei mesi di sperimentazione (99,85% di uptime). I feedback utente guidano e guideranno le evoluzioni del servizio. In particolare gli utenti richiedono strumenti di sincronizzazione desktop e maggiori informazioni sulle modifiche ai documenti nelle aree condivise.

4. Conclusioni

Questo contributo presenta GARRbox, un'architettura di cloud storage federato per il mondo accademico e della ricerca, e il prototipo con cui GARR ne ha messo alla prova i principi. I risultati della sperimentazione indicano quali siano le tecnologie adatte per estendere il prototipo di servizio, creato per gli utenti della Direzione GARR, e quali aspetti debbano essere rivisti e rafforzati per offrire un servizio scalabile, sicuro ed efficiente alla Comunità.

Quando questi ultimi punti tecnologici saranno migliorati, si inizieranno a definire le politiche di gestione del funzionamento del servizio, in modo da trasformarlo in un servizio per la comunità e in una federazione di cloud storage. Gli sviluppi futuri faranno tesoro delle conoscenze acquisite con il prototipo e dei preziosi suggerimenti forniti dagli utenti. L'attività immediata che è in fase di sviluppo consiste nel miglioramento dell'interfaccia utente, comprensiva di client di sincronizzazione, con feedback più immediati e un migliore controllo delle versioni.

Riferimenti Bibliografici

[1] Mell, P., & Grance, T. (2009, August 8). National Institute of Standards and Technology - Cloud Computing. Retrieved September 4, 2009, from National Institute of Standards and Technology: http://csrc.nist.gov/groups/SNS/cloud-computing/index.html

[2] http://helix-nebula.eu

[3] http://git-scm.com

Simon Vocella
simon.vocella@garr.it
lavora da diversi anni come software developer. Interessato a nuove tecnologie emergenti, in particolare a sistemi distribuiti e tecnologie cloud. Ha collaborato con GARR lavorando nei progetti europei FEDERICA e NOVI e nel progetto cloud storage GARRbox.