Scuola Superiore
dell'Università degli Studi di Udine

Classe Scientifico-Economica

Colloquio di fine anno

CREAZIONE DI UN FILE SYSTEM DISTRIBUITO

Relatore:
Prof. Scagnetto Ivan

Allievo:
Corrado Alessio

Anno Accademico 2019-20

Contents

1 Introduzione

Lo scopo del progetto è la realizzazione di un **file system distribuito**[1].

Le caratteristiche che deve avere sono:

- *Efficienza*: I tempi di risposta devono essere ragionevoli e non degradare con l'aumento del numero di nodi. L'utilizzo delle risorse a disposizione deve essere massimizzato, ovvero non devono esserci risorse inutilizzate o carichi di lavoro troppo sbilanciati.
- *Accesso concorrente*: Più client possono effettuare richieste in modo concorrente. Deve essere tenuta in considerazione qualche politica di *fairness*, in modo che nessun client sia eccessivamente privilegiato o venga escluso. Deve essere garantita la *consistenza* specificamente in caso di accessi concorrenti alla stessa risorsa.
- *Replicazione dei dati (files)*: Per garantire un buon livello di *fault tolerance*, tutti i dati devono poter essere replicati. In questo modo problematiche su un nodo non compromettono il contenuto di un file.
- *Replicazione dei metadati (struttura del fs)*: La struttura del *file system* deve essere sempre replicata, in modo che non possa essere persa o corrotta .
- *Disponibilità e consistenza*: deve essere massimizzata la disponibilità delle risorse, mantenendo al contempo una consistenza forte. In particolare, ogni richiesta deve riferirsi all'ultima versione della risorsa (se non diversamente indicato).
- *Sicurezza*: Deve essere curata la gestione della sicurezza: sia a livello di connessione (utilizzando opportunamente cifratura e autenticazione) che di accesso (autenticazione dell'utente, privilegi sul fs). Deve essere minimizzata l'esposizione ad attacchi malevoli. Deve essere possibile ripristinare il sistema a seguito di un attacco o errore.

Numero di repliche dei dati Il numero di repliche di un file deve essere proporzionale sia all'importanza che al numero di accessi di esso: documenti con *dati critici* hanno numero di copie maggiore, per agevolarne la *disponibilità* e diminuire le *probabilità di perdita*.

Bilanciamento del carico Per migliorare l'efficienza viene effettuato un bilanciamento del carico: in ogni momento il sistema può copiare o spostare dati in modo da non sovraccaricare un singolo nodo. Inoltre, per il trasferimento in uscita di un documento si cerca di utilizzare il nodo con maggiore *banda disponibile*.

2 Architettura del sistema

Verranno di seguito analizzate diverse possibili architetture del sistema, comparandone pregi e difetti.

2.1 Single-server

La prima idea è quella di avere un *singolo server*, il quale memorizza *sia la struttura del fs che i dati*. Il client interagisce direttamente con esso per effettuare qualsiasi operazione.

Il fs può essere memorizzato a partire da una *directory nel filesystem del server*, oppure in una *partizione separata*. Con entrambe le soluzioni è il sistema operativo ad occuparsi della sua gestione.

La *replicazione dei dati* avviene mediante il salvataggio in diversi dispositivi di memorizzazione (es HDD). Per far ciò la scelta più consona è di utilizzare un sistema **RAID 10**[2]: i dati vengono salvati in copia (RAID 1) per avere *tolleranza ai guasti* e in striping (RAID 0) per aumentare la *velocità di lettura*.

Con schede di rete (e connessioni) multiple è possibile migliorare la tolleranza sia ai *guasti* che alle *congestioni di rete*.

Vantaggi

- Semplice da implementare.
- Modello più economico.
- Il fs è gestito direttamente dal sistema operativo.
- Minima esposizione contro attacchi hacker.
- Con gli accorgimenti sopra descritti si ha già una buona fault tolerance.

Svantaggi

- *Single point of failure* per quanto riguarda il server (i dati sono parzialmente protetti dal meccanismo RAID).
- Nessuna protezione per *eventi eccezionali* (eg. catastrofe naturale, incendio nell'edificio, blocco della rete...).

.

Figure 1: Single-server

2.2 Virtual server

La soluzione single-server può essere migliorata aggiungendo un *livello di virtualizzazione*. Il sistema non viene quindi eseguito direttamente sull'hardware ma all'interno di una macchina virtuale (VM)[3].

La macchina virtuale viene periodicamente *copiata* (interamente o in modo incrementale[4]) e salvata in un dispositivo di memorizzazione dedicato.

Nel caso ci fosse un *crash* o *errore critico del sistema operativo*, basta ricaricare la macchina virtuale più recente. La *perdita di dati* è limitata all'età dell'ultimo backup.

Vantaggi

- *Recovery* buona e in tempi brevi in caso di crash.
- Naturale implementazione in ambienti di loro natura virtualizzati (es *container*).

Svantaggi

- Durante il *periodo di transizione* il sistema non risponde.
- *Prestazioni* leggermente ridotte a causa della virtualizzazione.

Figure 2: Single-server con virtualizzazione

2.3 Cloud backup

Il backup può essere effettuato nel cloud, piuttosto che in un dispositivo dedicato all'interno del datacenter. Inoltre il backup può essere esteso anche ai dati, oltre che al fs.

Vantaggi

- Protezione in caso di *eventi eccezionali*: i dati non risiedono in un unico luogo fisico.
- Il provider può fornire ulteriori garanzie di replicazione.

Svantaggi

- In generale effettuare un backup nel cloud, anche se solo incrementale, richiede un *elevato utilizzo di banda* e *tempi maggiori*.
- I *costi* possono essere maggiori.

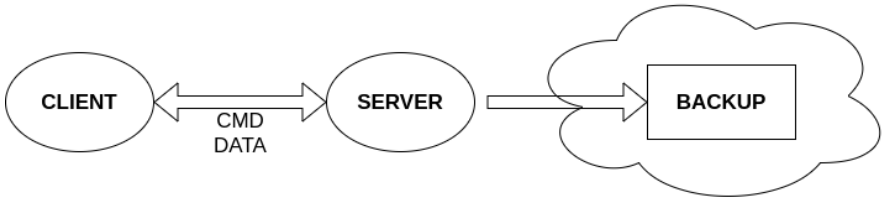


Figure 3: Aggiunta del backup nel cloud

2.4 Server as distributed database

Al posto di utilizzare il filesystem del sistema operativo, è possibile utilizzare un database distribuito[5]. Sia metadati che dati vengono *distribuiti automaticamente* nei nodi, garantendo consistenza e fault tolerance.

Vantaggi

- I dati vengono automaticamente replicati e gestiti dal dbms.
- Utilizzando un dbms in commercio, si beneficia dall'avere tutto già pronto e costantemente aggiornato.

Svantaggi

- Maggiore complessità.
- La connessione avviene tra il client ed un nodo, il quale poi distribuisce i dati bidirezionalmente agli altri nodi. Inoltre l'utilizzo di banda complessiva può essere doppio rispetto al necessario: nel caso i dati vadano trasferiti dal client al server (gateway) al server di destinazione o vice versa.

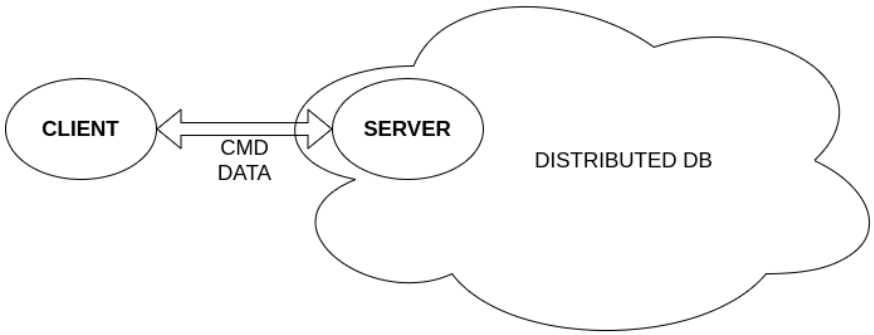


Figure 4: Utilizzo di un database distribuito

2.5 Meta+Data server

L'idea è di far *comunicare direttamente il client con i server in cui verranno memorizzati i dati*. Così facendo si elimina il bottleneck del nodo che prima doveva fungere da "gateway".

Un primo sviluppo è quello di *separare i compiti*: si hanno quindi un **meta server** e uno o più **data server**. Il meta server si occupa di memorizzare e gestire il filesystem (metadati), mentre i data server memorizzano soltanto i dati conenuti nei documenti.

Per effettuare una qualsiasi *operazione di trasferimento* il client si rivolge inizialmente al meta server, dal quale ottiene la configurazione per contattare il giusto data server e trasferire i dati. Le *operazioni sul filesystem* sono eseguite all'interno del meta server. In questo modo ci sono *due comunicazioni bidirezionali*: client-meta, per la gestione del fs; client-data, per il trasferimento dei dati.

Per migliorare le prestazioni, soprattutto nel caso in cui il client abbia una banda più ampia rispetto ai data server, i documenti di grandi dimensioni vengono spezzati in chunks e distribuiti.

Come nel caso single-sever, il meta server è unico, virtualizzato e sottoposto a backup periodico.

Vantaggi

- Trasferimenti concorrenti con data server multipli permettono di migliorare le prestazioni nel caso in cui il client abbia una banda elevata rispetto ai dataserver.
- Un *unico meta server* permette di mantenere in modo semplice la consistenza del filesystem.

Svantaggi

- La frequenza dei *backup* del meta server è fondamentale per minimizzare la perdita di dati in caso di fault.
- Tutti i data server devono essere *raggiungibili* dal client (maggiore attenzione alla sicurezza e struttura della rete).

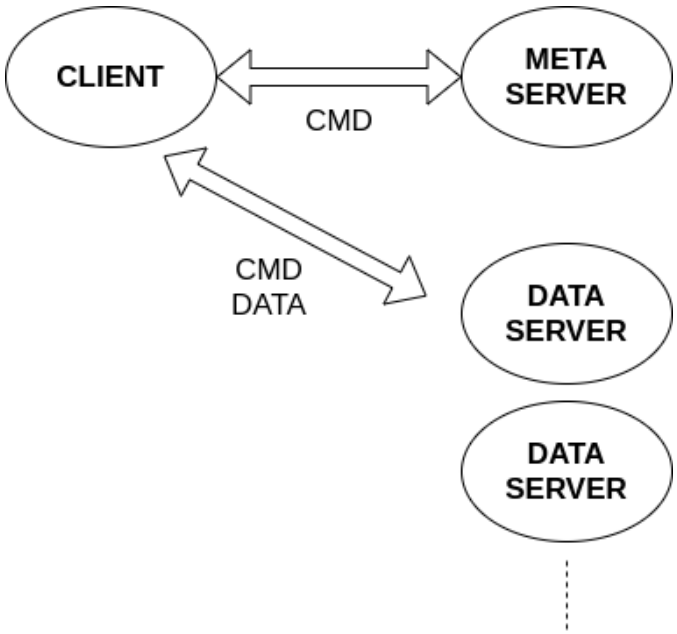


Figure 5: Separazione di meta e data server

2.6 Meta+Data server 2

Facendo dialogare meta e data server, è possibile:

1. Spostare l'*onere di configurazione dei data server* da client a meta server.
2. Gestire in qualsiasi momento la *replicazione dei dati*, ad esempio per bilanciare il carico o far fronte alla perdita di un nodo.
3. Il meta server può monitorare *prestazioni* e *stato di salute* dei data sever.

Il client quindi si limita a trasferire i dati da/verso data server utilizzando token forniti dal meta server.

Vantaggi

- Il sistema può *riconfigurarsi* in ogni momento per far fronte a qualsiasi evenienza.
- Ruolo del client semplificato.
- Maggiore sicurezza (minore esposizione) perché il client non può inviare comandi ai data server.

Svantaggi

- Maggiore complessità nel gestire operazioni *concorrenti* ed *asincrone* in nodi diversi.

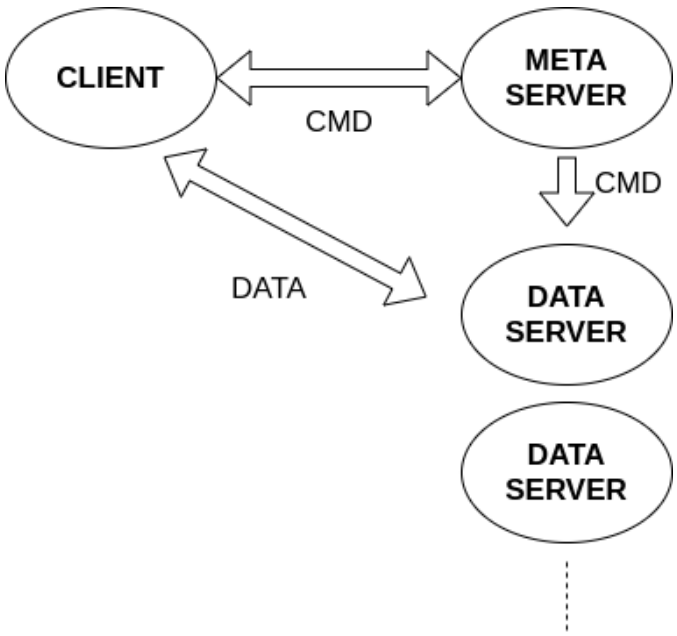


Figure 6: Meta + data server, il meta server dialoga con i data server

2.7 Meta server as distributed system

In questa soluzione non è presente un unico meta server, ma viene utilizzato un sistema distribuito. In generale è preferibile utilizzare un dbms distribuito presente in commercio per gestire i dati a basso livello.

Come dimostrato dal teorema CAP bisogna rinunciare alla disponibilità per garantire la coerenza del fs.

Vantaggi

- Possibilità di utilizzo di software in commercio.
- Maggiore fault tolerance dei metadati.
- Minore tempo di down in caso di malfunzionamenti.

Svantaggi

- Maggiore difficoltà nell'implementazione dei meta server e dei protocolli di comunicazione tra i vari componenti.
- Maggiore overhead (e quindi minori prestazioni) nel caso di sistemi di piccole/medie dimensioni.

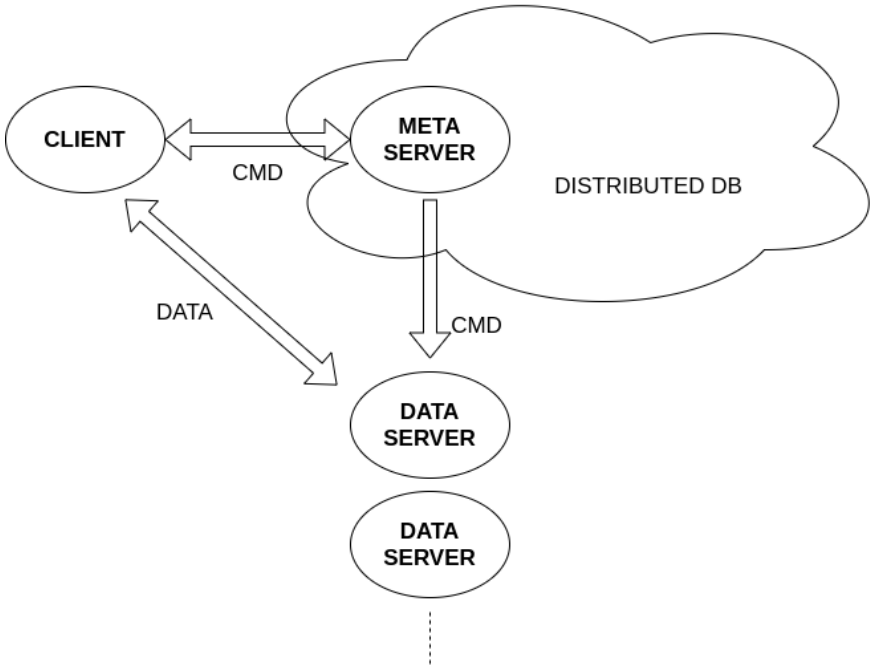


Figure 7: Il meta server è implementato come database distribuito

3 Struttura del filesytem

Il filesystem[6] ha una struttura non convenzionale. Descriveremo quindi la semantica dei path e i metadati ad essi associati.

3.1 Semantica dei path

Nei *sistemi tradizionali* il filesystem è formato da *cartelle, file e link*. La *struttura* è ad albero oppure grafo, con o senza la possibilità di cicli. La cartella "base" è root (nei sistemi unix-like denotata con la stringa "/"). Un *path*[7] può essere associato ad uno delle tre tipologie sopra menzionate; il filesystem memorizza questa associazione oltre che ad altri metadati.

Il simbolo '/' viene usato per separare i *componenti di un path*: ad esempio il path "/dir1/dir2/file1" identifica il file "file1" figlio della directory "dir2" che a sua volta è figlia della directory "dir1" che è figlia di root. Per questo '/' compare tra i caratteri vietati per la nominazione di un oggetto.

Nel *sistema presentato* esistono soltanto i **documenti**. Essi sono *contenitori di dati*, ed hanno associati alcuni *metadati* (in versione semplificata rispetto a UNIX).

Un path che termina con un carattere diverso da '%' rappresenta un *singolo documento* (es: "documento1", "home/lavoro#foto/mare-foto1"). Un path che termina con '%' rappresenta l'*insieme di documenti* che ha come prefisso i caratteri antecedenti al simbolo finale (es. "home/lavoro#foto%" include tutti i path che rispettano l'espressione regolare "home/lavoro#foto*"). Il carattere '%' può essere usato solo come terminatore. Il carattere ';' è vietato in quanto usato come separatore degli argomenti.

Questa organizzazione possiede intrinsecamente una *struttura ad albero*. I path che rappresentano insiemi di documenti sono chiamati *repository*. Le operazioni effettuate su repository hanno effetto su tutti i documenti (anche non ancora creati) appartenenti alla repository. La repository root è naturalmente identificata con il path "%". "" è un documento valido: si consiglia di usarlo per salvare le informazioni sul sistema (es proprietario, descrizione, contatti) in formato testuale con codifica UTF.

3.2 Metadati di un documento

I metadati associati ad un documento sono:

- uid: identificatore univoco dell'oggetto, in formato UUID Type 1[8];
- path: deve rappresentare un documento, stringa UTF[9];
- created: timestamp di creazione in formato UTC[10];
- size: dimensione in byte, 0 per le directory, formato unsigned int;
- owner: utente proprietario, stringa UTF;
- priority: numero minimo di copie, coincide con la priorità, formato unsigned int;

- checksum: checksum sha1, formato esadecimale;
- deleted: timestamp di eliminazione (vuoto se non eliminato), formato UTC.

4 Comandi gestiti dal client

Di seguito l'elenco e la descrizione dei comandi gestiti dal client, su cui deve basarsi l'implementazione del sistema.

- list: lista i metadati di un path (documento o documenti attualmente in una repository);
- get: download di un documento;
- push: upload di un documento;
- rem: eliminazione di path (documento o repository);
- lock: lock e unlock di un path (documento o repository);
- setPriority: aggiornare la priorità di un path (documento o repository).

5 Protocollo di comunicazione

Il protocollo di comunicazione[11] tra nodi diversi si basa sullo scambio di messaggi. Il flusso di dati è binario. Viene utilizzata una connessione di tipo TCP per garantire ordine di consegna, affidabilità, gestione della connessione.

Il formato dei comandi è testuale (codifica UTF). Un comando è formato dalla parola chiave che lo identifica e dalla lista degli argomenti. Il separatore utilizzato è il punto e virgola; i comandi sono terminati dal newline.

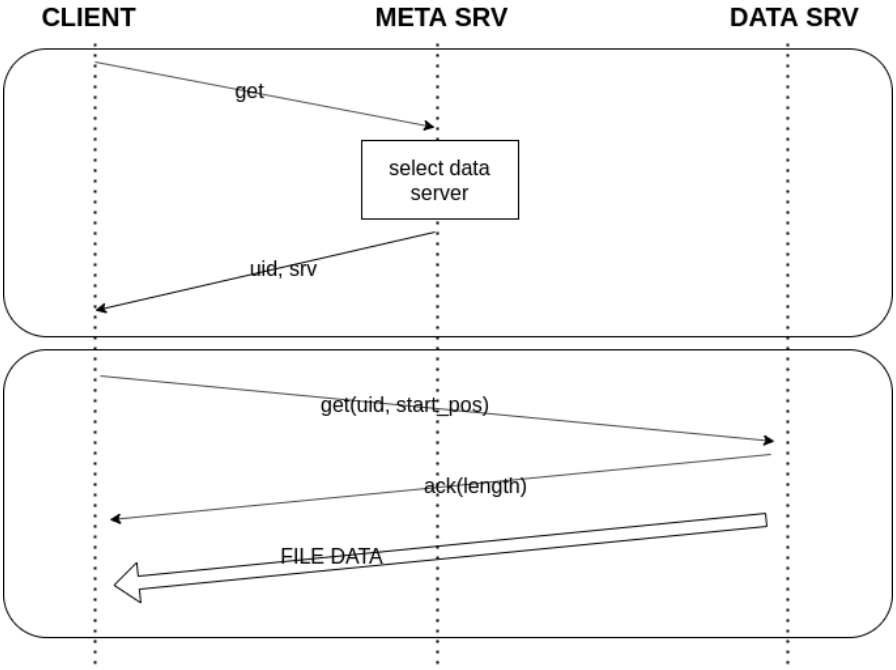
Di seguito i flussi di dati coinvolti nelle diverse tipologie di richieste.

5.1 get

La richiesta di tipo *get* richiede il trasferimento in entrata di un documento.

Inizialmente il *client* contatta il *meta server*, richiedendo un particolare *documento* (mediante path o uid). Il *meta server* seleziona il *data server* ottimale per gestire la richiesta (deve essere online, contenere la risorsa e non essere sovraccarico). Il *meta server* risponde quindi al client specificando l'uid della risorsa (richiesto per identificarla) e l'indirizzo del *data server* a cui collegarsi. Il *client* si disconnette dal *meta server*.

Il *client* si connette al *data server* inviando una richiesta *get(uid,start_pos)*. L'argomento *start_pos* indica da che byte iniziare a trasferire il documento (indice 0-based). Se la risorsa è disponibile (come dovrebbe essere) il *data server* invia un *ack* seguito dal flusso di dati del documento. Altrimenti risponde con *err* seguito dai dettagli.



5.2 push

La richiesta di tipo *push* richiede il trasferimento in uscita di un documento.

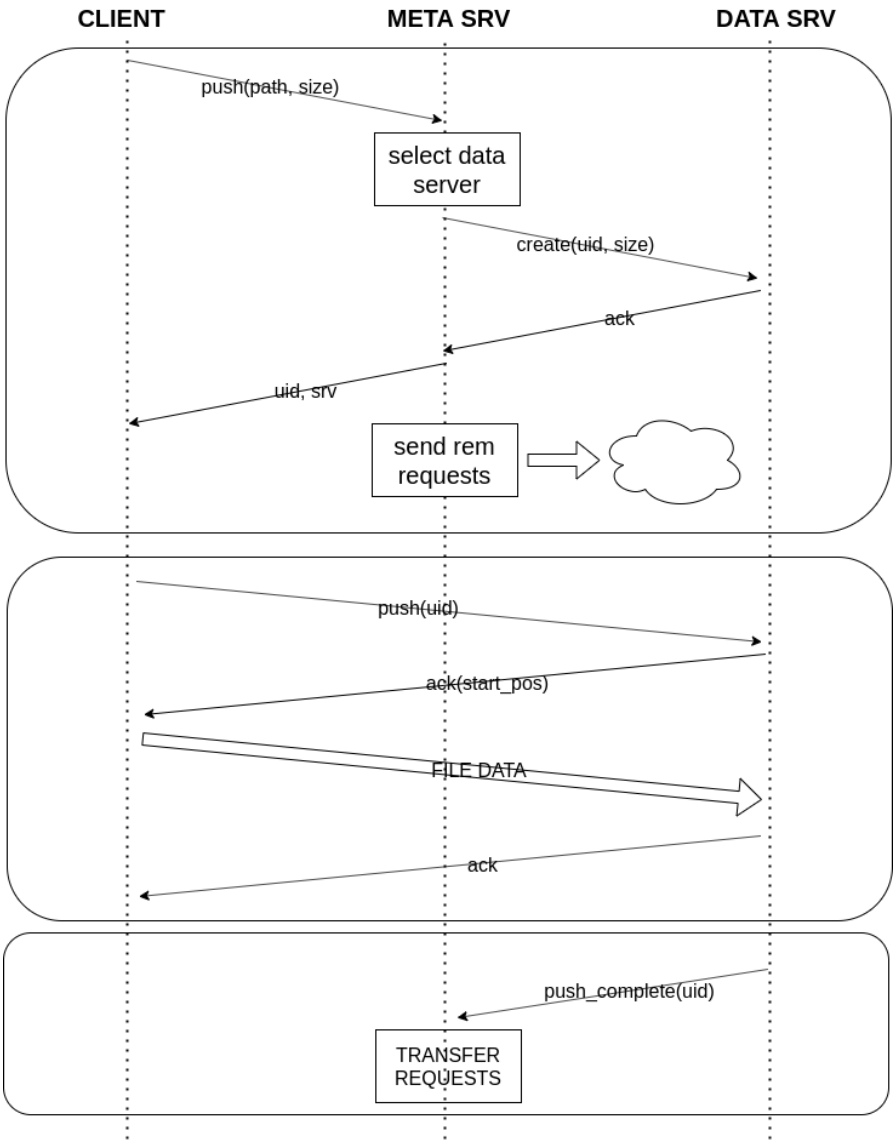
Inizialmente il *client* contatta il *meta server*, inviando *path* e *size*. Il *meta server* seleziona il *data server* ottimale per gestire la richiesta (deve essere online, poter contenere la risorsa e non essere sovraccarico). Il *meta server* si occupa anche di generare un nuovo *uid* da assegnare alla risorsa.

Il *meta server* si collega al *data server* dove verrà inizialmente salvata la risorsa, inviando un comando *create(uid, size)*. Questo comando predispone il *data server* per ospitare un nuovo documento con tale uid e dimensione. Se non ci sono errori, il *data server* risponde al *meta server* con *ack*. Altrimenti con *err* seguito dai dettagli.

Il *meta server* risponde quindi al client specificando l'uid della risorsa (richiesto per identificarla) e l'indirizzo del *data server* a cui collegarsi. Il *meta server* si disconnette dal *client* e invia le eventuali richieste di eliminazione ai data server (se il documento è una nuova versione di uno esistente).

Inizia quindi il caricamento del documento. Il *client* si collega al *data server* e invia una *push(uid)*. Se non ci sono errori, il *data server* risponde con *ack* e invia la posizione (indice 0-based) *last_pos* del primo byte non trasferito (equivalente al numero di byte già trasferiti). A quel punto il *client* trasferisce la porzione rimanente di documento. Al termine del trasferimento, se non ci sono errori, il *data server* risponde con *ack*. La connessione viene chiusa.

Quando il trasferimento è completato con successo, inizia la terza fase. Il *data server* invia un messaggio del tipo *push_complete(uid)* al *meta server*. Il meta server inizia quindi ad inviare ai *data server* interessati le richieste di trasferimento, per raggiungere il grado di replicazione voluto.

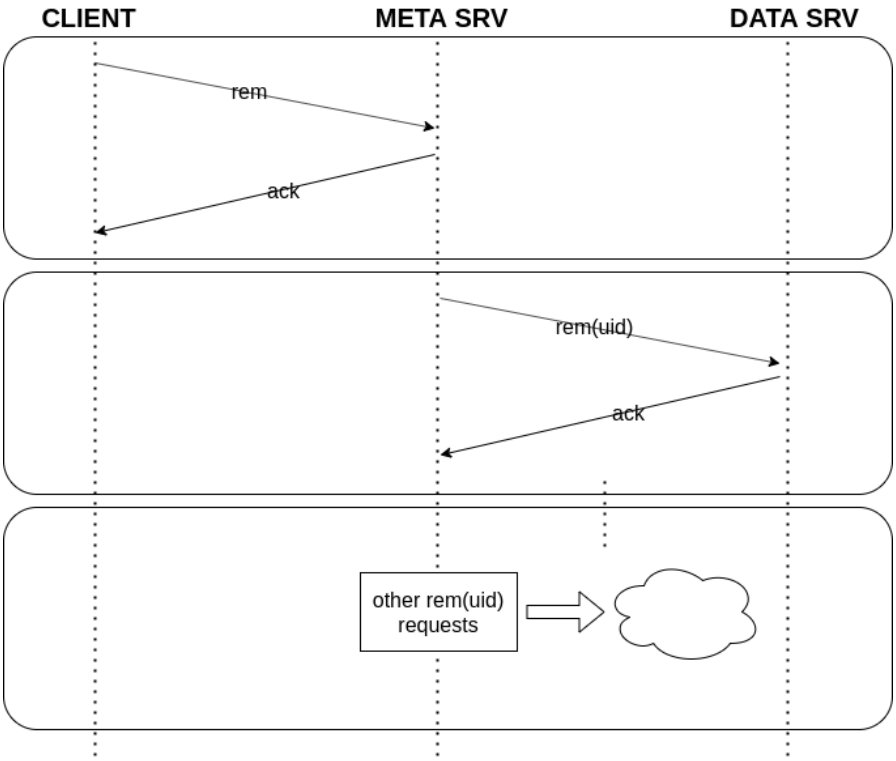


5.3 rem

La richiesta di tipo *rem* richiede l'eliminazione di un intero path o singolo uid dallo storage.

Il *client* invia al *meta server* una richiesta *rem*. Se non ci sono errori, il *meta server* risponde con *ack* e chiude la connessione.

Nel caso la richiesta di eliminazione sia di tipo fisico, il *meta server* contatta separatamente tutti i *data server* contenenti dati relativi a quel particolare *uid* e invia una richiesta *rem(uid)*.

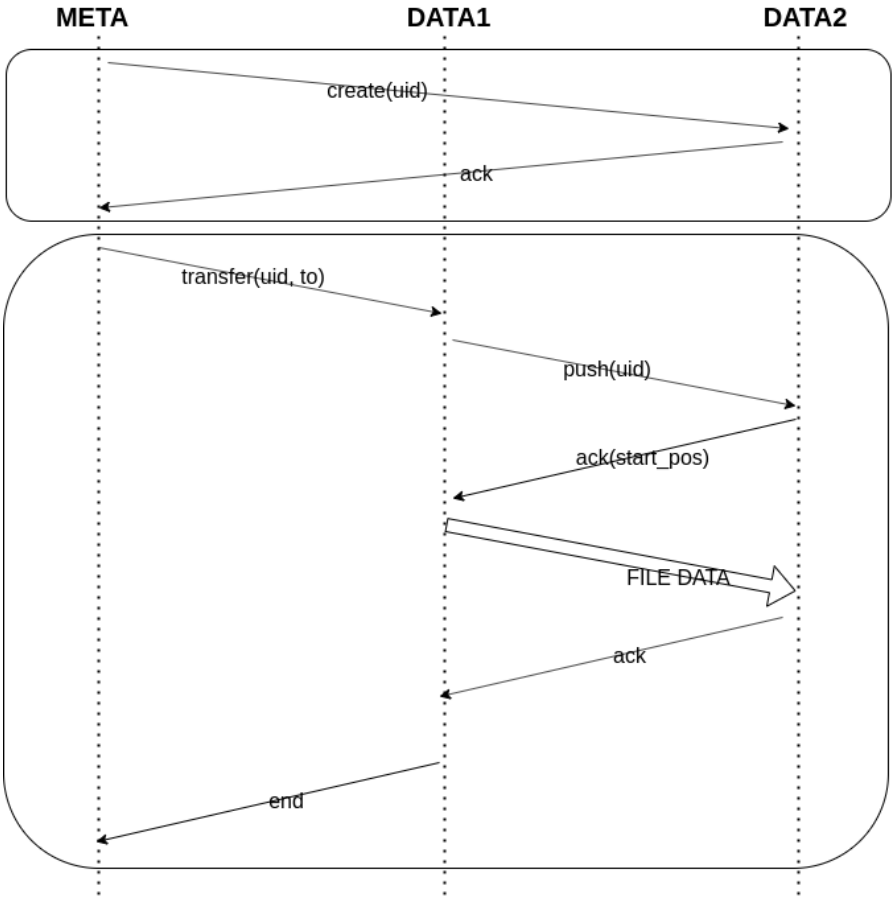


5.4 transfer

La richiesta di tipo *transfer*, effettuata esclusivamente dal metaserver, richiede la copia di un documento da un *data server* ad un altro. Viene usata per trasferire i dati tra i nodi in modo da gestire la ridondanza.

Inizialmente, come nel caso di una *push*, il meta server contatta il *data server* di destinazione (*DATA2*) tramite una richiesta *create(uid)*, per inizializzarlo a ricevere il documento. Se non ci sono errori il *data server* risponde con *ack*. La connessione viene chiusa.

A questo punto il *meta server* invia al *data server* di origine (*DATA1*) una richiesta *transfer(uid, to)* per ordinare il trasferimento. Se *DATA1* riesce a connettersi con successo a *DATA2*, risponde a *META* con *begin_transfer(uid)*. *META1* invia a *META2* una richiesta di tipo *push(uid)* e da luogo al trasferimento. Al termine *DATA1* invia un messaggio *end_transfer(uid)* a *META* per notificare l'avvenuto trasferimento.



6 Logging

Tutte le operazioni prima di essere eseguite vengono registrate in appositi database di log[12]. In questo modo, se un'operazione viene interrotta per qualsiasi motivo, può essere analizzata e rieseguita, in modo da non lasciare spazzatura.

Ciò è molto utile anche in fase di debugging/testing: analizzare lo stacktrace ed il flusso di operazioni avvenute precedentemente ad un problema è il modo più efficace per comprenderne le cause. Il modulo *logging*, parte della libreria standard di Python3, permette di organizzare le operazioni di log in modo compatto ed efficace.

7 Sicurezza ed autenticazione

Tutte le connessioni vengono cifrate a livello intermedio tramite protocollo TLS [13] over TCP. In questo modo la cifratura è invisibile alla logica dell'application e può in ogni momento essere modificata dal programmatore. Inoltre, essendo una tecnologia ampiamente diffusa e collaudata, presenta una vulnerabilità minima.

L'autenticazione da client a server avviene tramite token all'inizio della connessione. La password di ciascun utente può essere cambiata in ogni momento dall'utente stesso, dopo essersi loggato.

L'autenticazione da server a server avviene tramite certificato digitale.

Per garantire sia l'integrità che la sicurezza (per evitare modifiche volute) viene conservato e verificato il checksum (sha1 [14]) di tutti i documenti in ingresso.

8 Generazione di una chiave identificativa

Il problema della generazione di una chiave identificativa[15] è spesso sottovalutato. Consiste nella creazione di un codice che identifichi univocamente una risorsa, un oggetto o un sistema. Questa chiave deve avere la caratteristica di unicità, ovvero non devono esistere codici duplicati: altrimenti non sarebbe più possibile identificare correttamente la risorsa. Inoltre, per ragioni di efficienza, la chiava deve avere lunghezza limitata, in modo da limitarne il tempo di comparazione e lo spazio occupato. In genere le chiavi non hanno un significato, quindi possono essere interpretate indifferentemente come interi o stringhe esadecimali.

Strategie di generazione di una chiave Alcune possibili strategie di generazione di chiavi sono:

- Contatore intero.
- Funzione di hash applicata ad un insieme di proprietà variabili con il tempo.
- Numero casuale.
- UUID.

Il caso più semplice è quello del contatore intero. Solitamente si implementa con un contatore che parte da 0 e viene incrementato di un unità ad ogni generazione. I vantaggi sono: semplicità di implementazione e garanzia di unicità nel contesto del generatore, ovvero il generatore garantisce una sequenza di valori senza duplicati. Questa soluzione può essere implementata efficacemente soltanto all'interno di un sistema che disponga di un solo generatore, quindi di un solo nodo. Inoltre ciò espone all'esterno il numero

di chiavi generate sin ora, informazione che potrebbe essere sfruttata per attacchi al sistema.

Per superare il problema dell'unicità del nodo una soluzione è prependere al codice l'identificativo del nodo. Ciò garantisce l'unicità della chiave nell'insieme generato dall'intero cluster. È necessario che gli identificativi dei nodi siano univoci, pena il fallimento del sistema. Questo potrebbe essere un problema nel caso venga effettuato un merge tra cluster diversi: se due nodi avevano lo stesso identificativo, almeno uno di essi deve essere rinominato, assieme a tutte le chiavi da esso generate. Ciò potrebbe non essere possibile.

Una delle soluzioni più comunemente usate per la generazione di identificativi è l'utilizzo degli *UUID*[8]. Gli **UUID** (Universally Unique Identifier) sono stringhe binarie di 128 cifre generate con precise regole, dipendenti dalla versione. Spesso vengono rappresentati in esadecimale ed compaiono tra i tipi primitivi di molti linguaggi (tra cui CQL).

Gli UUID di tipo 4 sono generati casualmente. Se viene utilizzato un generatore con distibuzione uniforme, per avere il 50% di probabilità di registrare almeno una collisione è necessario generare $2.71 * 10^{18}$ valori. Questa probabilità, seppur non nulla, quasi sempre viene trascurata. Quando però la generazione di un duplicato potrebbe portare il sistema in uno stato inconsistente, è necessario adottare accorgimenti specifici per riportare il sistema ad uno stato safe.

Gli UUID di tipo 1, o *TimeUUID*, sono invece generati a partire da istante temporale (UTC), un identificativo del nodo e un numero di sequenza. Come identificativo del nodo si adotta l'indirizzo MAC della scheda di rete del pc, il quale è garantito essere univoco dal costruttore. La presenza dell'istante temporale permette di ordinare gli identificativi per ordine cronologico. L'identificativo del nodo impedisce collisioni tra nodi diversi. Il numero di sequenza elimina le collisioni (fino a 2^{12}) per elementi generati dallo stesso nodo nello stesso millisecondo: la capacità di generazione garantita è quindi di 4096 UUID tipo 1 al millisecondo.

Il problema principale degli UUID tipo 1 è causata dalla presenza di due nodi associati alla medesima scheda di rete, per i quali la probabilità di conflitto è estremamente elevata. Ciò si risolve imponendo l'esecuzione di una sola istanza del software per host, dato un cluster.
https://tools.ietf.org/html/rfc4122.html

9 Problema della consistenza in un sistema distribuito

I sistemi distribuiti sono per loro natura soggetti a frequenti errori, ritardi di comunicazione e situazioni di concorrenza. È quindi fondamentale assicurarsi di mantenere la

consistenza[16] dello stato globale in modo che eventuali vincoli di integrità non siano violati.

Un esempio lampante è quello delle richieste del tipo "INSERT .. IF NOT EXISTS". Tale operazione non può essere valutata in un solo nodo, in quanto l'identificativo cercato potrebbe essere stato inserito in un altro nodo, ma non ancora propagato all'intero sistema.

Per verificare l'esito di tale richiesta è quindi necessario adottare una politica specifica, quale potrebbe essere il meccanismo del **QUORUM**[17]. Essa stabisce che se in un istante la maggioranza assoluta ($Q = floor(N/2) + 1$, con N numero di nodi) dei nodi concorda su una proprietà, allora essa in quell'istante è vera. Capiamo meglio con un esempio: viene inserito il valore 1 all'interno di un nodo. Esso propaga il valore ad almeno $Q - 1$ nodi distinti. Al termine dell'operazione almeno la Q nodi contengono il valore 1. Per verificare la presenza del valore 1, viene effettuata una richiesta ad almeno Q nodi distinti qualsiasi. Per il principio della piccionaia, almeno uno di essi conterrà il valore specificato e quindi risponderà "yes", dando il risultato corretto.

Il nuovo problema da affrontare è che l'algoritmo precedente vale se e solo se tutte le richieste vengono valutate allo stesso istante, e il risultato ottenuto è garantito solo per lo stesso istante. Questo però è impossibile, in quanto comunicazioni e computazioni non sono istantanee. La soluzione adottata da Cassandra[1] è l'utilizzo del protocollo *PAXOS*[18] per effettuare *Lightweight transactions*[19]. Le LWT sono utilizzate in particolare nell'esecuzione di comandi "INSERT ... IF NOT EXISTS" e "UPDATE ... IF CONDITION", in modo da garantire la consistenza.

Inoltre, Cassandra permette di stabilire il livello di consistenza[20] per ogni query di tipo select. Esso può essere ANY, ONE, TWO, THREE, ..., QUORUM, ..., ALL, in base alle necessità del sistema. Questa scelta consente di bilanciare il tradeoff tra *disponibilità* e *consistenza*. ANY fornisce la maggiore disponibilità (basta che un nodo qualsiasi risponda), ALL la maggiore consistenza (tutti i nodi devono concordare). Per questo progetto è stata scelta l'opzione QUORUM, in quanto garantisce forte consistenza ma tolleranza ai guasti fino a $floor(N)$ nodi.

10 Implementazione

Di seguito verranno mostrati i punti salienti dell'implementazione demo da me sviluppata. Questa implementazione non è commercializzabile, in quanto incompleta, ma ha lo scopo di provare tutte le funzionalità del sistema sopra descritto.

[1] Cassandra è il *distributed dbms* impiegato nella mia implementazione, uno tra i più attualmente utilizzati

10.1 Strumenti software

Il linguaggio scelto per implementare client, dataserver e metaserver è **Python 3**[2]. Questo linguaggio ha numerosi vantaggi: è *platform-independent*, è *semplice*, *efficace* da utilizzare, permette la scrittura di codice abbastanza *compatto*, possiede una collezione di *librerie* praticamente illimitata, si adatta bene alla programmazione *concorrente*. Unico punto debole è l'*efficienza*, minore rispetto al C++, ma comunque più che sufficiente per lo scopo.

Come dbms per il metaserver ho scelto **Apache Cassandra**[3], uno dei più famosi sistemi *NoSQL* distribuiti in commercio. Gestisce automaticamente la *replicazione* e *concorrenza* all'interno del cluster, con efficacia, in modo da assicurare *disponibilità* e *affidabilità*. Le *prestazioni* sono ottime, come il supporto della *community* e la *documentazione*. *Cql* (Cassandra Query Language) una sintassi molto simile a SQL, cosa che agevola molto l'intervento di programmatori non specializzati.

Dato che i dataserver devono essere più leggeri possibile, come dbms per i dataserver ho scelto **SQLite3**[4]. E' completamente integrato nella libreria standard di Python 3, quindi non necessita dell'installazione di software aggiuntivo. Tutti i dati vengono memorizzati all'interno di un *unico file compresso*. Le *performances* sono più che adatte per le operazioni richieste. Sia buona *documentazione* che *facilità* di utilizzo sono punti a favore. Il linguaggio è un *dialetto SQL*, molto simile a quello degli altri sistemi (l'unica differenza importante sta nei tipi di dato, che sono semplificati).

I *file di configurazione* sono in formato **YAML**[21]: un formato *open*, *human-readable* tra i più diffusi al mondo e intuitivi.

10.2 Struttura del codice

Il codice sia di dataserver che metaserver si articola in 5 parti:

- Caricamento della cofigurazione.
- Interfaccia di connessione e scambio dati.
- Interfaccia al database.
- Operazioni periodiche.
- ServerSocket e relativo handler.

Il client dispone soltanto dell'interfaccia di connessione/scambio dati e dell'implementazione delle operazioni.

[2] `https://www.python.org/download/releases/3.0/`
[3] `https://cassandra.apache.org/`
[4] `https://www.sqlite.org/index.html`

10.3 Caricamento della configurazione

Sia per metaserver che dataserver è obbligatorio fornire il percorso del file di configurazione come primo argomento.

10.3.1 Metaserver

Codice

```
if len(sys.argv) > 1:
    configFile = sys.argv[1]
    print("load config", configFile)
    config = yaml.full_load(open(configFile, "r"))
    print("config", config)
else:
    logging.error("Please provide config file")
    exit(1)

database = Database()

for dataserver in config["dataservers"]:
    database.addDataServer(dataserver)
    print("add dataserver", dataserver)
```

Esempio di configurazione

```
dataservers:
    - localhost:10010
    - localhost:10011
    - localhost:10012
```

10.3.2 Dataserver

Codice

```
if len(sys.argv) > 1:
    configFile = sys.argv[1]
    print("load config", configFile)
    config = yaml.full_load(open(configFile, "r"))
    print("config", config)
    NAME = config["name"]
    HOST = config["host"]
    PORT = config["port"]
    METASERVER = config["metaserver"]
    workingDir = config["workingDir"]
else:
    logging.error("Please specify config file")
    exit(1)

if not os.path.exists(workingDir):
    os.makedirs(workingDir)
```

```
17      os.chdir(workingDir)
18      logging.info("work on %s", os.getcwd())
19
20  #create db if not exists
21  if not os.path.exists("database.db"):
22      with sqlite3.connect('database.db') as conn:
23          with open("../../dataserver/create_db.sqlite3", "r") as sql:
24              conn.executescript(sql.read())
25
26  database = Database()
27  database.setStats(config["storage"], config["downspeed"], config["upspeed"
        ])
28
29  SERVER = str(HOST) + ":" + str(PORT)
```

Esempio di configurazione

```
1  name: dataserver10012
2  workingDir: /home/ale/Dropbox/tesina iot/code/data/dataserver10012
3  host: localhost
4  port: 10012
5  storage: 100000000
6  downspeed: 50
7  upspeed: 10
8  metaserver: localhost:10000
```

10.4 Interfaccia di connessione e scambio dati

L'interfaccia è unica per client, metaserver e dataserver. E' implementata come *adapter*[22] per la classe socket e StreamRequestHandler. Utilizza gli oggetti *rfile* (lettura) e *wfile* (scrittura), ottenuti dal socket, per effettuare i trasferimenti.

Il trasferimento dati avviene in *modalità binaria*, per ragioni di efficienza.

I file vengono scritti dal metodo ad alta efficienza *socket.sendfile*, il quale permette di evitare il *double buffering*. Vengono letti a blocchi di 1024 byte con il metodo *socket.read*.

Codice

```
1  def CustomConnection(cls):
2      def addrFromString(self, addr):
3          ip, port = addr.split(":")
4          return (ip, int(port))
5
6      def write(self, *args):
7          res = ";".join([str(i) for i in args]).strip() + "\n"
8          logging.info("write %s", res)
9          self.wfile.write(res.encode())
10
11     def readline(self):
12         line = self.rfile.readline().strip().decode().split(";")
```

```
13          logging.info("readline %s", line)
14          return line
15
16      def readFile(self, size, outFile):
17          size = int(size)
18          pos = 0
19          while pos != size:
20              chunk = min(1024, size - pos)
21              # print("read", chunk, "bytes")
22              data = self.rfile.read(chunk)
23              outFile.write(data)
24              pos += len(data)
25
26      def writeFile(self, filepath, startIndex):
27          if type(startIndex) != int:
28              startIndex = int(startIndex)
29          with open(filepath, "rb") as file:
30              self.sendfile(file, startIndex)
31
32      setattr(cls, "writeFile", writeFile)
33      setattr(cls, "addrFromString", addrFromString)
34      setattr(cls, "readFile", readFile)
35      setattr(cls, "write", write)
36      setattr(cls, "readline", readline)
37
38      return cls
39
40
41  @CustomConnection
42  class Connection(socket.socket):
43      def __init__(self, endpoint):
44          super(Connection, self).__init__(socket.AF_INET, socket.SOCK_STREAM
                )
45          self.connect(self.addrFromString(endpoint))
46          self.wfile = self.makefile('wb', 0)
47          self.rfile = self.makefile('rb', -1)
48
49
50  @CustomConnection
51  class DataServerHandler(StreamRequestHandler):
52      def sendfile(self, *args, **kwargs):
53          self.connection.sendfile(*args, **kwargs)
54
55      ...code...
```

10.5 Interfaccia al database

Per interfacciarsi al database metaserver e dataserver hanno degli oggetti dedicati, chiamati *Database*. Essi espongono tutte le query sotto forma di *API*: in questo modo gli utilizzatori del database sono svincolati dall'effettiva implementazione sottostante. Questo è un vantaggio sia in termini di *incapsulamento*, che rende più semplice la

progettazione e manutenzione del software, che nel caso in cui si voglia cambiare dbms, perché l'utilizzatore non deve essere modificato.

10.5.1 Metaserver

Di seguito la struttura del database utilizzato dal metaserver assieme ad una parte di implementazione dell'oggetto Database.

Struttura

```
1  --il fattore di replicazione deve essere <= al numero di server a
       disposizione
2  create keyspace metaserver with replication = {'class' : 'SimpleStrategy',
       'replication_factor' : 1};
3
4
5  use metaserver;
6
7
8  /*un metaserver*/
9  create table dataserver(
10     server text,       --indirizzo in formato host:port del dataserver
11     online boolean,      --se in questo momento (all'ultima rilevazione) e'
           online
12     capacity bigint,      --capacita' in byte totale
13     remaining_capacity bigint,      --capacita' in byte disponibile (totale-
           utilizzata)
14     available_down float,      --banda in Mb/s disponibile per il download
15     available_up float,        --banda in Mb/s disponibile per l'upload
16     primary key(server)
17 );
18
19 /*un documento*/
20 create table object(
21     uid uuid,
22     path text,         --path associato al documento
23     created timestamp,      --istante di creazione (nel metaserver)
24     owner text,       --proprietario (colui che l'ha creato)
25     size bigint,       --dimensione in byte
26     priority int,      --livello di priorita' (di replicazione)
27     checksum text,
28     deleted timestamp,        --istante di eliminazione (NULL se non
           eliminato)
29     primary key(uid)
30 );
31
32 /*lock attivi*/
33 create table object_lock(
34     path text,     --path del documento
35     user text,       --utente che detiene il lock
36     primary key(path)
37 );
38
```

```
39 /*log delle performance dei dataserver*/
40 create table performance_log(
41     server text,           --indirizzo del dataserver nel formato host:port
42     time timestamp,      --istante di registrazione del record
43     online boolean,
44     capacity bigint,
45     remaining_capacity bigint,
46     available_down float,
47     available_up float,
48     primary key(server, time)
49 );
50
51 /*oggetti che vanno verificati*/
52 create table pending_object(
53     uid uuid,            --uid dell'oggetto
54     enabled boolean,       --se false, l'oggetto non puo essere attualmente
        processato perche' non e' disponibile o non sono disponibili
       server in cui replicarlo
55     primary key(uid)
56 );
57
58 /*documenti salvati nei dataserver*/
59 create table stored_object(
60     uid uuid,        --documento
61     server text,       --indirizzo del server in cui e' salvato, nel
       formato host:port
62     created timestamp,
63     complete boolean,
64     primary key(uid, server)
65 );
66
67 /*utilizzato per le ricerche di prefissi del tipo "where path like '<prefix
    >%'"*/
68 create custom index object_path_idx on object(path)
69 using 'org.apache.cassandra.index.sasi.SASIIndex';
70
71 /*utilizzato per selezionare l'ultima versione di un documento, dato il
    path (con cql non e' possibile ordinare nella query) */
72 create materialized view pathToObject as select * from object where path is
     not null
73 primary key (path, uid)
74 with clustering order by (uid desc);
```

Codice

```
1 class Database:
2     def __init__(self):
3         #livello di consistenza delle query
4         profile = ExecutionProfile(
5             consistency_level=ConsistencyLevel.QUORUM,
6             serial_consistency_level=ConsistencyLevel.SERIAL  #LWT
7         )
8
```

```
9           #connessione al cluster
10          self.cluster = Cluster(protocol_version=4, execution_profiles={
                EXEC_PROFILE_DEFAULT: profile})
11          self.session = self.cluster.connect("metaserver")
12
13      #aggiunta di un nuovo dataserver. inizialmente stato=disconnesso
14      def addDataServer(self, addr):
15          self.session.execute("insert into dataserver(server, online) values
                (%s, false)", (addr, ))
16
17      #creazione di un nuovo documento
18      def addObject(self, uid, path, owner, size, priority, checksum):
19          created = datetime_from_uuid1(uid)  #l'istante di creazione del
                documento coincide con l'istante di creazione del suo uid
20
21          self.session.execute("insert into object(uid, path, created, owner,
                size, priority, checksum) "
22                               "values (%s, %s, %s, %s, %s, %s, %s)"
23                               , (uid, path, created, owner, size, priority,
                                    checksum))
24          return uid
25
26      #una nuova copia fisica di un documento viene creata in un dataserver
27      def addStoredObject(self, uid, server, complete=False, created=None):
28          if created is None: created = datetime.now() #istante di creazione
                della replica, non del documento
29
30          self.session.execute("insert into stored_object(uid, server,
                created, complete) values (%s, %s, %s, %s)",
31                               (uid, server, created, complete))
32          return id
33
34      #effettua il lock di un path (singolo documento)
35      def lockPath(self, path, user="root"):
36          #se il path che si intende bloccare e' gia' presente nella tabella,
                il constraint INSERT IF NOT EXISTS fallisce e applied viene
                settato a False (altrimenti True)
37          return self.session.execute("insert into object_lock(path, user)
                values (%s, %s) if not exists",
38                                      (path, user)).one().applied
39
40      #unlock di un path
41      def unlockPath(self, path):
42          self.session.execute("delete from object_lock where path = %s", (
                path, ))
43
44      #ottiene lo stato di lock di un path e l'identificativo dell'utente che
                detiene il lock (se bloccato, altrimenti "")
45      def getPathLock(self, path):
46          r = self.session.execute("select * from object_lock where path = %s
                ", (path, )).one()
47
48          if r is None:    #none se non presente nel db l'entry relativa al
```

```
            path specificato (quindi e' sbloccato)
49              return False, ""
50          else:
51              return True, r.user
52
53      #ottiene i metadati di un documento dato l'uid
54      def getObjectByUid(self, uid):
55          return self.session.execute("select * from object where uid = %s",
                (uid, )).one()
56
57      #lista di tutti i documenti appartenenti ad un determinato path (
            singolo o repository)
58      def list(self, path):
59          if path == "" or path == '%':
60              return self.session.execute("select * from object").all()
61          else:
62              return self.session.execute("select * from object where path
                    like %s", (path, )).all()
63
64      #lista dei dataserver
65      def getDataServers(self, onlyOnline=True):
66          if onlyOnline:   #solo i dataserver online (secondo l'ultima
                rilevazione)
67              return self.session.execute("select * from dataserver where
                    online=true allow filtering").all()
68          return self.session.execute("select * from dataserver").all()
69
70      #update dello status di un dataserver (ottenuto a seguito di una
            rilevazione)
71      def updateDataServerStatus(self, addr, online, remaining_capacity=0,
            capacity=0, available_down=.0, available_up=.0):
72          self.session.execute("update dataserver "
73                               "set capacity=%s, remaining_capacity=%s,
                                    available_down=%s, "
74                               "available_up=%s, online=%s where server=%s",
75                               (capacity, remaining_capacity,
76                                available_down, available_up, online, addr))
77          #performances dei dataserver loggate per eventuali analisi
78          self.session.execute("insert into performance_log(server, time,
                capacity, remaining_capacity, available_down,"
79                               "available_up, online) values (%s, %s, %s, %s,
                                    %s, %s, %s)",
80                               (addr, datetime.now(), capacity,
                                    remaining_capacity,
81                                available_up, available_down, online))
82
83      #rimuove una replica fisica di un documento
84      def removeStoredObject(self, uid, server):
85          self.session.execute("delete from stored_object where uid = %s and
                server = %s", (uid, server))
86
87      #rimuove un documento dato l'uid
88      def removeObject(self, uid):
```

```
89          self.session.execute("delete from object where uid = %s", (uid, ))
90
91      #ottiene la lista di dataserver che contengono l'uid specificato
92      def getServersForUid(self, uid, complete=True, online=True):
93          if complete:   #solo dataserver la cui replica e' completa (
                trasferita completamente e validata)
94              res = self.session.execute(
95                  "select server from stored_object where uid = %s and
                        complete = true allow filtering", (uid, )).all()
96          else:
97              res = self.session.execute(
98                  "select server from stored_object where uid = %s allow
                        filtering", (uid, )).all()
99          if online:  #solo dataserver online (secondo ultima rilevazione)
100             def f(srv):
101                 return self.session.execute("select online from dataserver
                        where server = %s", (srv.server, )).one().online
102             res = list(filter(f, res))
103         return res
104
105     #testa se un dataserver e' online
106     def isServerOnline(self, server):
107         return self.session.execute("select online from dataserver where
                server = %s", (server, )).one().online
108
109     #setta una replica di un documento come completa
110     def setComplete(self, uid, server):
111         self.session.execute("update stored_object set complete = true
                where uid = %s and server = %s", (uid, server))
112
113     #marca come eliminato l'ultimo uid (in ordine di creazione) relativo ad
             un path singolo specificato.
114     def markDeleted(self, path):
115         res = self.getUidForPath(path)
116         timestamp = datetime.now()
117         if res is not None:
118             self.session.execute("update object set deleted = %s where uid
                    = %s",
119                                  (timestamp, res.uid))
120
121     #marca come eliminato uno specifico uid
122     def markUidDeleted(self, uid):
123         timestamp = datetime.now()
124         self.session.execute("update object set deleted = %s where uid = %s
                ",
125                              (timestamp, uid))
126
127     #ottiene l'uid dell'ultima versione del documento relativo ad un path
            singolo
128     def getUidForPath(self, path):
129         path = str(path)
130         return self.session.execute("select * from pathToObject where path
                = %s", (path, )).one()
```

```
131
132     #ottiene la lista di tutti gli uid relativi ad un path singolo (tutte
            le versioni di un documento)
133     def getUidsForPath(self, path):
134         return self.session.execute("select * from object where path like %
                s", (path,)).all()
135
136     #gli uid di tutti i documenti che hanno una replica all'interno del
            dataserver specificato
137     def getUidsForServer(self, server):
138         return self.session.execute("select uid from stored_object where
                server = %s allow filtering", (server,))
139
140     #aggiunge un uid alla lista di quelli da verificare (per azioni
            periodiche)
141     def addPendingUid(self, uid):
142         self.session.execute("insert into pending_object(uid, enabled)
                values (%s, True)", (uid,))
143
144     #aggiorna la priorita' di un documento (dato l'uid)
145     def updatePriority(self, uid, priority):
146         self.session.execute("update object set priority=%s where uid=%s",
                (priority, uid))
147
148     #lista di tutti gli uid pendenti
149     def getPendingUids(self, onlyEnabled=True):
150         if onlyEnabled:  #solo quelli attivi
151             return self.session.execute("select uid from pending_object
                    where enabled=True allow filtering").all()
152         return self.session.execute("select uid from pending_object").all()
153
154     #disabilita un uid pendente. utilizzato ad esempio quando un documento
            va replicato ma non ci sono dataserver disponibili.
155     def disablePendingUid(self, uid):
156         self.session.execute("update pending_object set enabled=False where
                uid = %s", (uid,))
157
158     #rimuove un uid pendente
159     def removePendingUid(self, uid):
160         self.session.execute("delete from pending_object where uid = %s", (
                uid,))
161
162
163 database = Database()
```

10.5.2 Dataserver

Di seguito struttura e implementazione del database relativo ai dataserver.

Struttura

```
1
2 --un oggetto
```

```
3  create table object(
4      uid text primary key,
5      local_path text,    --percorso del file system dov'e' salvato
6      size integer,    --dimensione in byte (totale)
7      complete integer,     --true se e' completo, false se deve ancora essere
           totalmente trasferito
8      created text,        --istante di creazione (nel dataserver)
9      checksum text
10 );
11
12 --configurazione del dataserver
13 create table stats(
14     storage int,         --storage totale in byte
15     downspeed int,       --banda totale in download, in Mb/s
16     upspeed int          --banda totale in upload, in Mb/s
17 );
18
19 --documenti che vanno eliminati
20 create table toBeDeleted(
21     uid text primary key         --uid del documento
22 );
```

Codice

```
1  class Database:
2      def __init__(self):
3          try:
4              #connessione al database
5              self.connection = sqlite3.connect('database.db')
6              self.cursor = self.connection.cursor()
7
8          except sqlite3.Error as error:
9              print("error while connecting to database", error)
10
11     #ottiene i metadati relativi ad un documento, dato l'uid
12     def getObject(self, uid):
13         return self.cursor.execute("select * from object where uid = ?", (
               str(uid), )).fetchone()
14
15     #elimina i dati relativi ad un documento
16     def deleteUid(self, uid):
17         self.cursor.execute("delete from object where uid = ?", (uid, ))
18         self.cursor.execute("delete from transfer where uid = ?", (uid, ))
19         self.cursor.execute("delete from toBeDeleted where uid = ?", (uid,
               ))
20         self.connection.commit()
21
22     #lista di tutti i documenti (anche non completi)
23     def getStoredData(self):
24         return self.cursor.execute("select uid, created, complete from
               object").fetchall()
25
26     #spazio complessivo riservato ai documenti attualmente salvati (per i
```

```
        documenti parziali si conta comunque la dimensione totale)
27      def reservedSpace(self):
28          res = self.cursor.execute("select sum(size) from object").fetchone
                ()[0]
29          if res is None:
30              res = 0
31          return res
32
33      #aggiunge un documento alla coda di eliminazione
34      def addToBeDeleted(self, uid):
35          self.cursor.execute("insert into toBeDeleted(uid) values (?)", (uid
                , ))
36          self.connection.commit()
37
38      #ottiene tutti i documenti in coda di eliminazione
39      def getToBeDeleted(self):
40          return self.cursor.execute("select * from toBeDeleted").fetchall()
41
42      #crea una nuova replica di un documento
43      def addObject(self, uid, local_path, size, checksum):
44          complete = 0
45          created = datetime.now()
46
47          self.cursor.execute("insert into object(uid, local_path, size,
                complete, checksum, created) values (?, ?, ?, ?, ?, ?)",
48                              (uid, local_path, size, complete, checksum,
                                  created))
49          self.connection.commit()
50
51          return self.cursor.lastrowid   #primary key dell'entry inserita
52
53      #setta un documento come completo
54      def setComplete(self, uid):
55          self.cursor.execute("update object set complete=1 where uid = ?", (
                uid, ))
56          self.connection.commit()
57
58      #ottiene le impostazioni del nodo (banda, capacita', ecc)
59      def nodeStats(self):
60          return self.cursor.execute("select * from stats").fetchone()
61
62      #modifica le impostazioni del nodo
63      def setStats(self, storage, downspeed, upspeed):
64          self.cursor.execute("delete from stats")
65          self.cursor.execute("insert into stats(storage, downspeed, upspeed)
                values (?, ?, ?)",
66                              (storage, downspeed, upspeed))
67          self.connection.commit()
68
69  database = Database()
```

10.6 Operazioni periodiche

Le operazioni che non possono essere eseguite subito (es: eliminazione di un documento mentre sta venendo trasferito) oppure che vanno eseguite regolarmente (es: monitoraggio delle prestazioni del dataserver) sono valutate *periodicamente* ed eseguite nel primo momento utile.

Lo schema di esecuzione è molto semplice: ogni operazione periodica è *schedulata ad intervalli regolari* (es 10 secondi) e viene eseguita da un *looper thread*.

10.6.1 Metaserver

Nel metaserver le operazioni periodiche sono:

- Controllo dello stato dei dataserver.
- Mantenimento del grado di replicazione dei documenti.

Controllo dello stato dei dataserver

```
1  def onDataServerConnect(addr):
2      print("server", addr, "connected")
3
4      for elem in database.getUidsForServer(addr):
5          database.addPendingUid(elem.uid)
6
7      for elem in database.getPendingUids(onlyEnabled=False):
8          database.addPendingUid(elem.uid)
9
10 def onDataServerDisconnect(addr):
11     database.updateDataServerStatus(addr, False)
12
13     for elem in database.getUidsForServer(addr):
14         database.addPendingUid(elem.uid)
15
16     print("server", addr, "disconnected")
17
18
19 def checkDataServerStatus(addr):
20     wasOnline = database.isServerOnline(addr)   #status dell'ultima
           rilevazione
21
22     try:
23         with Connection(addr) as conn:  #richiesta dello status al
               dataserver
24             conn.write("status")
25             res = conn.readline()
26
27         status, reservedCapacity, totCapacity, downSpeed, bandDown, upspeed
               , bandUp = res
28
```

```
29          database.updateDataServerStatus(addr, True, totCapacity -
                reservedCapacity, totCapacity,
30                                           downSpeed - bandDown, upspeed -
                                                bandUp)
31
32          if not wasOnline:
33              onDataServerConnect(addr)
34      except ConnectionRefusedError as err:  #accade quando il dataserver non
            e' raggiungibile
35          if wasOnline:
36              onDataServerDisconnect(addr)
37
38  def monitorDataServers():
39      for server in database.getDataServers(onlyOnline=False):
40          checkDataServerStatus(server.server)
```

Mantenimento del grado di replicazione dei documenti

```
1   def processPendingUids():
2       for elem in database.getPendingUids():
3           processPendingUid(database, elem.uid)
4
5   def processPendingUid(database, uid):
6       #metadati dell'uid
7       res = database.getObjectByUid(uid)
8
9       priority = res.priority
10      size = res.size
11      checksum = res.checksum
12
13      #repliche attualmente disponibili (solo online)
14      serversContaining = {x.server for x in database.getServersForUid(uid)}
15      numCopies = len(serversContaining)
16
17      if numCopies == priority:  #se la priorita' coincide con il numero di
            repliche disponibili non serve fare nulla
18          database.removePendingUid(uid)
19          return
20
21      #server liberi per ricevere una nuova replica
22      availableServers = [x.server for x in database.getDataServers() if x.
            server not in serversContaining and x.remaining_capacity > size]
23
24      serversContaining = list(serversContaining)
25
26      if numCopies > priority:  #se il numero di copie disponibili supera la
            priorita' bisogna eliminarne una
27          random.shuffle(serversContaining)
28          target = serversContaining[0]
29
30          print("remove", uid, "target", target)
31
32          try:
```

```
33              with Connection(target) as conn:  #invia richiesta di
                    eliminazione al dataserver
34                  conn.write("deleteUid", uid)
35                  print(conn.readline())
36
37              database.removeStoredObject(uid, target)  #eseguito solo se la
                    richiesta di eliminazione (connessione) ha avuto successo
38          except:
39              pass
40
41          return
42
43      if numCopies < priority:   #non ci sono abbastanza copie disponibili
44
45          if len(serversContaining) == 0 or len(availableServers) == 0:   #se
                non ci sono sorgenti o destinazioni disponibili bisogna
                attendere
46              database.disablePendingUid(uid)
47              return
48
49          try:
50              #scegli una sorgente ed una destinazione
51              random.shuffle(availableServers)
52              random.shuffle(serversContaining)
53              source = serversContaining[0]
54              target = availableServers[0]
55
56              database.addStoredObject(uid, target) #crea una nuova replica
57
58              def process():
59                  try:
60                      with Connection(target) as conn:  #crea il documento
                            sul server di destinazione
61                          conn.write("createUid", uid, size, checksum)
62                          print(conn.readline())
63
64                      with Connection(source) as conn:   #ordina il
                            trasferimento
65                          conn.write("transfer", uid, target)
66                          print(conn.readline())
67                  except:
68                      database.removeStoredObject(uid, target) #la replica
                            non e' stata creata con successo
69                      database.addPendingUid(uid)
70
71              _thread.start_new_thread(process)  #l'operazione, essendo lunga
                    , viene eseguita su un thread separato
72
73          except:
74              print("ERROR")
```

Scheduling ed esecuzione

```
schedule.every(5).seconds.do(monitorDataServers)
schedule.every(5).seconds.do(processPendingUids)
schedule.run_all()

def repeatedActions(_):
    while True:
        schedule.run_pending()
        time.sleep(0.1)

_thread.start_new_thread(repeatedActions, (None,))
```

10.6.2 Dataserver

Nel dataserver le operazioni periodiche sono:

- Controllo dello *stato interno* (utilizzo della banda e della memoria di massa).
- Eliminazione dei documenti in coda.

Controllo dello stato interno

```
#bandup and banddown in MB/s
performances = {"sent": 0, "recv": 0, "lastTime" : 0, "bandup" : 0, "
    banddown" : 0}

def getPerformance():
    res = psutil.net_io_counters()
    curtime = time.time()
    delta = curtime - performances["lastTime"]
    performances["lastTime"] = curtime
    performances["bandup"] = (res.bytes_sent - performances["sent"]) /
        delta / 1000000
    performances["banddown"] = (res.bytes_recv - performances["recv"]) /
        delta / 1000000
    performances["sent"] = res.bytes_sent
    performances["recv"] = res.bytes_recv
```

Per tenere traccia dei documenti attualmente in trasferimento, che quindi non possono essere eliminati, viene utilizzato il set *processing*. Esso contiene in ogni momento tutti gli *uid* relativi a documenti in caricamento e in trasferimento.

Eliminazione dei documenti in coda

```
def processDeleteUids():

    database = Database()
    uids = database.getToBeDeleted()

    with processingLock:
        for uid, in uids:
            if uid not in processing:
```

```
9                   uid, localPath, size, complete, created, checksum =
                        database.getObject(uid)
10
11                  if os.path.exists(localPath): os.remove(localPath)
12                  database.deleteUid(uid)
```

Scheduling ed esecuzione

```
1  schedule.every(5).seconds.do(getPerformance)
2  schedule.every(15).seconds.do(processDeleteUids)
3
4  def runPeriodicTasks(_):
5      while True:
6          schedule.run_pending()
7          time.sleep(0.1)
8
9  _thread.start_new_thread(runPeriodicTasks, (None,))
```

10.7 ServerSocket e relativo handler

Il punto di contatto di metaserver è il metodo *handle*. Esso si occupa del *dispatching delle richieste*, con l'ausilio di una mappa delle operazioni. Si ricorda che, secondo il protocollo di comunicazione, una richiesta è formata dal nome ed eventuali parametri, separati dal punto e virgola e terminati da newline. Una soluzione analoga è presente per il dataserver.

Una eventuale *richiesta non riconosciuta* viene automaticamente scartata.

Metaserver e relativo handler

```
1      ...
2
3      def handle(self):
4          self.database = Database()
5
6          switcher = {
7              "getPath": self.getPath,
8              "pushPath": self.pushPath,
9              "list": self.list,
10             "pushComplete": self.pushComplete,
11             "test": self.test,
12             "deletePath": self.deletePath,
13             "addDataServer": self.addDataServer,
14             "getUid": self.getUid,
15             "lockPath": self.lockPath,
16             "getPathLock": self.getPathLock,
17             "unlockPath": self.unlockPath,
18             "updatePriorityForPath": self.updatePriorityForPath,
19             "updatePriorityForUid": self.updatePriorityForUid,
20             "permanentlyDeletePath": self.permanentlyDeletePath
21         }
```

```
22
23          print("handle request from " + str(self.client_address))
24
25          data = self.readline()
26
27          switcher[data[0]](data[1:])
28
29  with MetaServer((HOST, PORT), MetaServerHandler, bind_and_activate=True) as
        server:
30      server.serve_forever()
```

In questa sezione vengono analizzate le implementazioni di tutti i metodi/operazioni che possono essere eseguite. L'analisi è suddivisa per operazioni, in modo da poter apprezzare meglio le interazioni tra i componenti.

10.7.1 Caricamento di un documento

Questa azione viene usata dal client per caricare un documento.

Inizialmente il client si collega al dataserver inviando *path, dimensione, checksum, priorità, utente.*

Client

```
1  def sendFile(localPath = "small_file.txt", remotePath="testfile.txt",
       priority=1, user="default"):
2      with Connection(METASERVER) as conn:
3          size = getsize(localPath)
4
5          checksum = hashFile(localPath)
6
7          conn.write("pushPath", remotePath, size, checksum, priority, user)
8          res = conn.readline()
9
10         if res[0] != "ok":
11             print("ERR", res)
12             return False
13
14         state, uid, addr = res
15     ...
```

Il metaserver controlla che il path non sia bloccato da un altro utente, cerca un dataserver che possa ospitare il documento e gli assegna un nuovo uid.

Metaserver

```
1  def pushPath(self, args):
2      path, size, checksum, priority, user = args
3      size = int(size)
4
```

```
5       lock, lockUser = self.database.getPathLock(path)
6       if lock and lockUser != user:
7           self.write("err", "path locked by " + lockUser)
8           return False
9
10      dataservers = [x for x in self.database.getDataServers() if x.
            remaining_capacity > size]
11
12      if len(dataservers) == 0:
13          self.write("err", "no data servers available")
14          return False
15
16      random.shuffle(dataservers)
17      target = dataservers[0]
18
19      uid = uuid4()
20      addr = target.server
21      ...
```

Il metaserver contatta il dataserver di destinazione creando la risorsa.

Metaserver

```
1       ...
2       with Connection(addr) as dataServer:
3           dataServer.write("createUid", uid, size, checksum)
4           response = dataServer.readline()
5
6       if response[0] != "ok":
7           self.write("err", response)
8           return False
9       ...
```

Dataserver

```
1   def createUid(self, args):
2       uid, size, checksum = args
3       local_path = os.getcwd() + "/" + uid
4
5       with processingLock:
6           totCapacity, downSpeed, upspeed = self.database.nodeStats()
7           reservedCapacity = self.database.reservedSpace()
8
9           if reservedCapacity + size > totCapacity:
10              self.write("err", "storage capacity exceeded")
11              return False
12
13          self.database.addObject(uid, local_path, size, checksum)
14
15      self.write("ok")
```

Il metaserver marca come eliminato l'oggetto associato al path (se esiste) e aggiunge il nuovo oggetto al database. Risponde al client indicando *uid, addr*.

Metaserver

```
    self.database.markDeleted(path)
    self.database.addObject(uid, path, "root", size, priority, checksum)
    self.database.addStoredObject(uid, addr)

    self.write("ok", uid, addr)
```

Il client si collega al dataserver target e invia il documento. Se il documento era stato già parzialmente caricato nel dataserver (es. file di grosse dimensioni), il caricamento riprende da dover era rimasto. Se il documento era già stato caricato completamente (ovvero il client tenta di caricare un documento già completo), viene generato un errore. Se il checksum al termine del caricamento non corrisponde, il documento viene subito eliminato per essere ricaricato in futuro.

Client

```
    ...
    with Connection(addr) as conn:
        conn.write("pushUid", uid)

        res = conn.readline()
        if res[0] != 'ok':
            print(res)
            return False

        status, startIndex = res

        print("send starting from", int(startIndex))

        conn.writeFile(localPath, startIndex)
        print(conn.readline())
```

Dataserver

```
def pushUid(self, args):
    uid, = args

    with processingLock:
        objinfo = self.database.getObject(uid)

        if objinfo is None:
            self.write("ERR", "uid info not found")
            return False

        uid, localpath, size, complete, created, checksum, = objinfo

        if complete:
            self.write("err", "file complete")
```

```
15              return False
16
17          processing.add(uid)
18
19      startIndex = 0
20      if os.path.exists(localpath):
21          startIndex = os.path.getsize(localpath)
22
23      assert startIndex < size    #altrimenti sarebbe complete
24
25      self.write("ok", startIndex)
26
27      with open(localpath, "a+b") as out:
28          self.readFile(size - int(startIndex), out)
29
30      file_checksum = hashFile(localpath)
31
32      with processingLock:
33          processing.remove(uid)
34
35          print("checksum", checksum, "file_checksum", file_checksum)
36
37          if file_checksum != checksum:
38              os.remove(localpath)
39              self.write("err", "checksum do not match")
40              return False
41
42          self.database.setComplete(uid)
```

Il dataserver invia al metaserver la notifica che il caricamento è completo, e quindi il documento può essere utilizzato.

Dataserver

```
1       ...
2       with Connection(METASERVER) as metaConn:
3           metaConn.write("pushComplete", uid, SERVER)
4
5       self.write("ok")
```

Il metaserver processa l'elemento per controllarne il grado di replicazione.

Metaserver

```
1   def pushComplete(self, args):
2       uid, addr = args
3
4       self.database.setComplete(uid, addr)
5       self.database.addPendingUid(uid)
6
7       self.write("ok")
```

10.7.2 Trasferimento di un documento

Il trasferimento viene usato dal metaserver per copiare un documento da un dataserver ad un altro, aumentandone così il grado di replicazione.

Il metaserver prima configura il server di destinazione per ricevere il documento, come nel caso del caricamento da parte del client. Poi invia la richiesta di trasferimento al server sorgente specificando *uid, indirizzo server destinazione*.

Metaserver

```
1 with Connection(target) as conn:
2     conn.write("createUid", uid, size, checksum)
3     print(conn.readline())
4
5 with Connection(source) as conn:
6     conn.write("transfer", uid, target)
7     print(conn.readline())
```

Il caricamento del documento dal server sorgente a destinazione avviene con lo stesso metodo usato nel caricamento da client a dataserver.

Dataserver sorgente

```
1  def transfer(self, args):
2      uid, server = args
3
4      with processingLock:
5          res = self.database.getObject(uid)
6          if res is None:
7              self.write("err", "uid not found")
8              return False
9
10         processing.add(uid)
11
12     uid, localPath, size, complete, created, checksum = res
13
14     with Connection(server) as target:
15         target.write("pushUid", uid)
16         status, startIndex = target.readline()
17         target.writeFile(localPath, startIndex)
18
19     with processing:
20         processing.remove(uid)
21
22     self.write("ok")
```

10.7.3 Scaricamento sul client di un documento (dato il path)

Questa è l'operazione con cui il client ottiene un documento dal sistema di storage.

Inizialmente il client contatta il metaserver indicando il path che vuole scaricare.

Client

```
1  def get(localPath = "testin.txt", remotePath = "ale/file1", newFile=True,
       user="default"):
2      if newFile and os.path.exists(localPath):
3          os.remove(localPath)
4
5      with Connection(METASERVER) as conn:
6          conn.write("getPath", remotePath, user)
7          res = conn.readline()
8
9      if res[0] != 'ok':
10         print("ERR", res)
11         return False
12
13     status, uid, addr, checksum = res
```

Il metaserver verifca che il path non sia bloccato da un altro utente, cerca l'uid associato al path (quello che corrisponde all'ultima versione del documento), cerca un dataserver che lo contenga (possibilmente seleziona il migliore) e risponde al client indicando *uid, indirizzo dataserver*.

Metaserver

```
1  def _getServerForUid(self, uid):
2      res = self.database.getServersForUid(uid)
3
4      if len(res) == 0:
5          self.write("err", "no copies available")
6          return False
7
8      random.shuffle(res)
9
10     addr = res[0].server
11     return addr
12
13 def getPath(self, args):
14     path, user = args
15
16     lock, lockUser = self.database.getPathLock(path)
17     if lock and lockUser != user:
18         self.write("err", "path locked by " + lockUser)
19         return False
20
21     res = self.database.getUidForPath(path)
22
23     if not res:
24         self.write("err", "path not found")
25         return False
26
27     if res.deleted is not None:
```

```
28          self.write("err", "path deleted")
29          return False
30
31      uid, checksum = res.uid, res.checksum
32
33      addr = self._getServerForUid(uid)
34
35      self.write("ok", uid, addr, checksum)
```

A quel punto il client si collega al dataserver sorgente, invia la posizione da cui vuole iniziare a leggere il documento (usato per riprendere il trasferimento interrotto di un documento di grandi dimensioni) e legge il documento. La chiamata usata in questo caso è *getUid*. Al termine viene verificato il checksum del documento per individuare eventuali errori o manomissioni.

Client

```
1       startIndex = 0
2       if os.path.exists(localPath):
3           startIndex = os.path.getsize(localPath)
4
5       with Connection(addr) as conn:
6           conn.write("getUid", uid, startIndex)
7           res = conn.readline()
8           status, size = res
9
10          with open(localPath, "a+b") as out:
11              conn.readFile(size, out)
12
13      if checksum != hashFile(localPath):
14          logging.error("HASH don't match")
15          return False
16
17      return True
```

Codice

```
1   def getUid(self, args):
2       uid, startIndex = args
3       startIndex = int(startIndex)
4
5       with processingLock:
6           res = self.database.getObject(uid)
7
8           if res is None:
9               self.write("err", "specified uid not present")
10              return False
11
12          uid, localPath, size, complete, created, checksum = res
13
14          if not complete:
15              self.write("err", "not complete")
```

```
16                 return False
17
18             processing.add(uid)
19
20         self.write("ok", size-startIndex)
21         self.writeFile(localPath, startIndex)
22
23         with processingLock:
24             processing.remove(uid)
```

10.7.4 Scaricamento di un documento dato l'uid

Nel caso in cui si voglia ottenere una particolare versione di un documento, si segue un meccanismo analogo al precedente. La differenza è che il client effettua una richiesta iniziale del tipo *getUid*, al posto che *getPath*, verso il metaserver.

Metaserver

```
1  def _getServerForUid(self, uid):
2      res = self.database.getServersForUid(uid)
3
4      if len(res) == 0:
5          self.write("err", "no copies available")
6          return False
7
8      random.shuffle(res)
9
10     addr = res[0].server
11     return addr
12
13 def getUid(self, args):
14     uid, user = args
15
16     res = self.database.getObjectByUid(uid)
17
18     if res is None:
19         self.write("err", "uid " + uid + " not found")
20         return False
21
22     checksum = res.checksum
23     addr = self._getServerForUid(uid)
24
25     self.write("ok", addr, checksum)
```

10.7.5 Eliminazione di un path

Questa operazione permette di eliminare logicamente un path. Tutti i documenti interessati vengono marcati come eliminati ma conservati, in modo che sia possibile reperirli tramite l'uid.

Il client effettua una richesta del tipo *deletePath* indicando *path, utente*. Remind: path può riferirsi ad un documento (es "doc1") o ad una repository (es "dir1%").

Client

```
1  def deletePath(path, user="default"):
2      conn = Connection(METASERVER)
3      conn.write("deletePath", path, user)
4      print(conn.readline())
5      conn.close()
```

Il metaserver ricerca tutti i documenti interessati e, se non sono bloccati, li marca come eliminati.

Metaserver

```
1  def deletePath(self, args):
2      path, user = args
3
4      res = self.database.getUidsForPath(path)
5
6      lockedPaths = []
7
8      for elem in res:
9          if elem.deleted is not None:
10             continue
11
12         lock, lockUser = self.database.getPathLock(elem.path)
13         if lock and lockUser != user:
14             lockedPaths.append(elem.path)
15         else:
16             uid = elem.uid
17             if elem.deleted is None:
18                 self.database.markUidDeleted(uid)
19
20     if len(lockedPaths):
21         self.write("err", "paths locked :", lockedPaths)
22     else:
23         self.write("ok")
```

10.7.6 Eliminazione permanente di un path

L'eliminazione permanente di un path causa la cancellazione logica e fisica di tutti i dati relativi ad un path. In questo modo tutto lo spazio verrà liberato e non sarà più possibile recuperare i dati. Questa operazione è simile all'eliminazione di un'intera repository di un sistema di versionamento.

Il client invia una richiesta del tipo *permanentlyDeletePath* al metaserver indicando *path, utente*.

Codice

```
def permanentlyDeletePath(path, user="default"):
    conn = Connection(METASERVER)
    conn.write("permanentlyDeletePath", path, user)
    print(conn.readline())
    conn.close()
```

Come nel caso precedente, il metaserver verifica che i documenti non siano bloccati da altri utenti. In seguito marca ogni documento non bloccato come eliminato e gli setta priorità nulla, di fatto schedulando l'eliminazione tutte le repliche.

Codice

```
def permanentlyDeletePath(self, args):
    path, user = args

    res = self.database.getUidsForPath(path)

    lockedPaths = []

    for elem in res:
        lock, lockUser = self.database.getPathLock(elem.path)
        if lock and lockUser != user and elem.priority > 0:
            lockedPaths.append(elem.path)
        else:
            uid = elem.uid
            if elem.deleted is None:
                self.database.markUidDeleted(uid)
            self.database.updatePriority(uid, 0)
            self.database.addPendingUid(uid)

    if len(lockedPaths):
        self.write("err", "paths locked :", lockedPaths)
    else:
        self.write("ok")
```

10.7.7 Lock

Mediante il meccanismo dei lock un client può "bloccare" un path (solo singolo documento). Quando un path è bloccato, tutte le richieste di altri utenti riguardanti quel path vengono respinte.

Per effettuare un lock, il client specifica la coppia *path, utente*.

Client

```
def lockPath(path, user="default"):
    conn = Connection(METASERVER)
    conn.write("lockPath", path, user)
    res, = conn.readline()
```

```
5       return res
```

Il metaserver risponde con *True/False* per indicare se il lock ha avuto successo. Fallisce quando si tenta di bloccare un path già bloccato da un altro utente.

Metaserver

```
1  def lockPath(self, args):
2      path, user = args
3      res = self.database.lockPath(path, user)
4      self.write(res)
```

Per verificare se un path sia bloccato si può effettuare una richiesta del tipo *getPathLock*. Essa restituisce un booleano e, nel caso sia bloccato, anche l'utente che detiene il blocco.

Metaserver

```
1  def getPathLock(self, args):
2      path, = args
3      lock, user = self.database.getPathLock(path)
4      self.write(lock, user)
```

Per sbloccare un path è sufficiente invocare la richiesta *unlockPath*. Essa non effettua controlli e si affida alla correttezza di comportamento dei client: questo per permettere a chiunque di sbloccare un path bloccato da un utente che è andato offline (piuttosto che lasciarlo bloccato per tempo arbitrario).

Metaserver

```
1  def unlockPath(self, args):
2      path, = args
3      self.database.unlockPath(path)
4      self.write("ok")
```

Gestire i lock su repository, piuttosto che su singoli documenti, è tutt'altro che banale. Non ho trovato una soluzione efficiente per interrogare il database in modo da controllare se uno dei lock presenti sia prefisso del path che voglio controllare.

10.7.8 Modificare la priorità di un documento

È possibile in ogni momento modificare la priorità di un determinato path. Con *updatePriorityForPath* viene modificata la priorità di ogni documento relativo al path specificato, mentre con *updatePriorityForUid* è possibile effettuare l'operazione su una singola versione.

Il client contatta il metaserver indicando *path, utente, nuova priorità*. La nuova priorità deve essere >0.

Metaserver

```
1  def updatePriorityForPath(self, args):
2      path, priority, user = args
3      priority = int(priority)
4
5      if priority <= 0:
6          self.write("err", "priority must be >0")
7          return False
8
9      res = self.database.getUidsForPath(path)
10
11     lockedPaths = []
12
13     for elem in res:
14         if elem.priority <= 0:
15             continue
16
17         lock, lockUser = self.database.getPathLock(elem.path)
18         if lock and lockUser != user:
19             lockedPaths.append(elem.path)
20         else:
21             uid = elem.uid
22             self.database.updatePriority(uid, priority)
23             self.database.addPendingUid(uid)
24
25     if len(lockedPaths):
26         self.write("err", "paths locked :", lockedPaths)
27     else:
28         self.write("ok")
29
30 def updatePriorityForUid(self, args):
31     uid, priority = args
32     priority = int(priority)
33
34     res = self.database.getObjectByUid(uid)
35
36     self.database.updatePriority(uid, priority)
37     self.database.addPendingUid(uid)
38
39     self.write("ok")
```

10.7.9 Aggiungere un dataserver

È possibile aggiungere un nuovo dataserver utilizzando la richiesta *addDataServer*, a cui va specificato l'indirizzo del dataserver.

Al momento della connessione il metaserver richiede al dataserver la lista dei contenuti in modo da aggiornare le proprie tabelle.

Metaserver

```
def addDataServer(self, args):
    addr, = args

    self.database.addDataServer(addr)

    checkDataServerStatus(self.database, addr)

    with Connection(addr) as conn:
        conn.write("getStoredData")
        len, = conn.readline()

        for i in range(int(len)):
            uid, created, complete = conn.readline()
            created = dateutil.parser.parse(created)
            complete = complete == "1"
            self.database.addStoredObject(uid, addr, complete, created)
```

Dataserver

```
def getStoredData(self, args):
    data = self.database.getStoredData()
    self.write(len(data))
    for uid, created, complete in data:
        self.write(uid, created, complete)
```

10.8 Testing

Al termine dello sviluppo è stata eseguita una sessione di *testing*. Gli obiettivi di test sono:

- Correttezza di esecuzione
- Comportamento quando il protocollo viene violato (accidentalmente o in modo malevolo)
- Performances
- Scalabilità

10.8.1 Correttezza

Per effettuare il controllo di correttezza è stato creato un client (benevolo) che effettua tutte le operazioni disponibili, con diversi ordinamenti e anche in modo concorrente. Ad ogni passaggio sono stati controllati manualmente gli stati di tutti i database, nonché tutti i log.

I risultati, benché mostrino errori sporadici molto rari che solitamente non portano ad interruzioni del servizio, sono più che sufficienti per mostrare il funzionamento di quanto presentato nell'analisi.

10.8.2 Resistenza agli attacchi

In questo caso è stato ideato un client malevolo che intenzionalmente viola le regole di protocollo, tentando di portare il sistema in stato inconsistente o di errore. Un esempio è l'invio di richieste con argomenti di formato sbagliato.

I risultati non hanno messo in luce alcuna vulnerabilità specifica del protocollo e dell'implementazione.

10.8.3 Performance

La valutazione delle performance viene effettuata attraverso l'utilizzo di benchmark che portino il sistema ad un sovraccarico.

Dai dati rilevati una singola richiesta viene generalmente processata nell'arco del decimo di secondo. Effettuare richieste in modo concorrente riduce il tempo totale in modo rilevante: l'inserimento in parallelo di 100 file ha portato il tempo totale da 28 a 9 secondi.

Si è osservato che l'avvio del metaserver è particolarmente oneroso nel caso in cui sia presente una grande quantità di documenti (>10k), in quanto esso deve controllare lo stato di ogni replica all'interno dei dataserver (sincronizzazione). Sviluppi futuri dovrebbero migliorare questo passaggio in modo da poter supportare quantità di documenti maggiori.

10.8.4 Scalabilità

Il test di scalabilità consiste nel valutare il consumo di risorse e le performance del sistema all'aumentare del numero di nodi.

Il consumo di risorse è stato valutato eseguendo fino a 400 istanze di dataserver concorrenti nello stesso dispositivo. L'utilizzo di ram di ogni singola istanza è stimata a 8 MB (costanti). Il tempo di avvio del sistema è di 4 secondi ogni 100 nodi (ma questo dato dipende fortemente dall'hardware a disposizione). Il tempo di accesso al singolo documento rimane pressoché costante.

11 Considerazioni finali

Perchè questo approccio è migliore rispetto ad usare direttamente un database distribuito?

L'utilizzo di un database distribuito è senza dubbio più semplice: basta installare il dbms nei nodi voluti e dialogarci tramite uno dei tanti client in commercio.

Ci sono però numerosi aspetti negativi:

- Non è possibile stabilire il livello di replicazione per un singolo documento. Un livello di replicazione unitario per file temporanei (es. cache) riduce lo spreco di memoria, mentre un livello di replicazione elevato praticamente annulla la possibilità di perdita di dati di importanza inestimabile.
- Il dbms è generalmente dispendioso in termini di risorse. Mentre la semplicità di un dataserver ne permette persino l'installazione in un dispositivo come *Raspberry*. E' così possibile sfruttare ad esempio un impianto IoT per immagazzinare dati.
- Il client si collega ad un nodo il quale inoltra i dati alla destinazione. Con questa architettura invece invia direttamente i dati al dataserver di destinazione (e viceversa). Si ha quindi un risparmio di banda complessiva.
- Il metaserver fornisce funzionalità quali mantenimento di un filesystem, operazioni di lock, bilanciamento del carico personalizzato.

Sviluppi futuri?

Lo sviluppo principale consiste nell'implementazione dell'algoritmo che muove e replica i documenti per bilanciare il carico.

Altri sviluppi potrebbero essere la creazione di un client grafico, per l'utilizzo da parte del pubblico. Inoltre dovrebbe essere implementato un livello di protezione, sia dal punto di vista della cifratura che dei permessi (con eventuale login). Vanno effettuati test più approfonditi e migliorata la gestione degli errori, soprattutto per quanto riguarda i corner case. Una compressione lato client potrebbe permettere un risparmio di risorse (memoria e rete) non indifferente.

Come ultima cosa si potrebbe provare a modificare leggermente il metaserver in modo che possano essere eseguite concorrentemente più istanze. Usando già un dbms distribuito (Cassandra) ed essendo implementato in modo sessionless, la transizione dovrebbe essere estremamente semplice e veloce da implementare.

Sistema di versionamento

È interessante notare che il sistema può benissimo essere utilizzato come sistema di versionamento. Dato che i documenti non vengono mai eliminati, ma la modifica presuppone la creazione di un nuovo uid (quindi vecchia e nuova versione sono effettivamente documenti a se stanti che condividono solo il path), è possibile andare a selezionare una particolare versione di ogni documento basandosi sulla data di creazione. L'unico accorgimento è quello di effettuare dal client una richiesta del tipo *getUid*, al posto di *getPath*.

La creazione di una repository, a differenza di sistemi come *git*, è trasparente, in quanto ogni path è in sè una repository (oppure da un altro punto di vista l'intero filesystem è una grande repository). L'eliminazione di un documento tramite *deletePath* agisce soltanto sul fs, marcandolo come eliminato, ma non rimuove i dati dai dataserver. Quindi è sempre possibile recuperare i dati di un documento eliminato in questo modo. Invece *permanentlyDeletePath* elimina definitivamente un documento (o repository), permettendo la liberazione di spazio.

Cosa ho imparato con questo progetto?

Ho affinato la mia conoscenza del linguaggio Python 3. In particolare ho scoperto come utilizzare il pattern decorator per arricchire oggetti "di sistema" in modo da strutturare in modo efficace il codice.

Ho usato per la prima volta un database NoSQL, in particolare Apache Cassandra. Essendoci molti punti a favore, oltre che molte limitazioni, rispetto ai "classici" database SQL, non è stato facile capire come sfruttarne al meglio le potenzialità. Solo per citare alcuni esempi, operazioni come la generazione delle chiavi primarie ed effettuare filtri/join sono tutt'altro che scontate.

La gestione di un sistema distribuito coinvolge, per sua natura, il dialogo di più programmi. Va posta particolare cura allo sviluppo dei protocolli di comunicazione, sincronizzazione e fault tolerance, in un ambiente in cui errori e ritardi (sia interni che esterni al sistema) sono la norma. Tutte situazioni che in un sistema unitario non si presentano.

L'imprevedibilità elevata riduce drasticamente l'efficacia del testing. E' quindi ancora più fondamentale stabilire contratti, vincoli, invarianti, diagrammi di flusso, meccanismi di controllo, meccanismi di logging e criteri di scrittura del codice che aiutino a minimizzare sia la probabilità di errore che i danni a seguito di un errore.

Bibliography

[1] Distributed file system. https://en.wikipedia.org/wiki/Clustered_file_system#Distributed_file_system

[2] Raid. https://en.wikipedia.org/wiki/RAID.

[3] Virtual machine. https://www.vmware.com/topics/glossary/content/virtual-machine.

[4] Backup types. https://www.acronis.com/en-eu/articles/incremental-differential-backups/.

[5] Distributed database. https://www.tutorialspoint.com/distributed_dbms/distributed_dbms_databases

[6] File system. https://en.wikipedia.org/wiki/File_system.

[7] Path. https://en.wikipedia.org/wiki/Path_(computing).

[8] Uuid. https://en.wikipedia.org/wiki/Universally_unique_identifier.

[9] Utf. https://en.wikipedia.org/wiki/UTF-8.

[10] Utc. https://en.wikipedia.org/wiki/Coordinated_Universal_Time.

[11] Communication protocol. https://en.wikipedia.org/wiki/Communication_protocol.

[12] Log file. https://en.wikipedia.org/wiki/Log_file.

[13] Tls. https://hpbn.co/transport-layer-security-tls/.

[14] Sha-1. https://en.wikipedia.org/wiki/SHA-1.

[15] Surrogate key. https://en.wikipedia.org/wiki/Surrogate_key.

[16] Consistency. https://en.wikipedia.org/wiki/Consistency_(database_systems).

[17] Quorum. https://en.wikipedia.org/wiki/Quorum_(distributed_computing).

[18] Paxos. https://en.wikipedia.org/wiki/Paxos_(computer_science).

[19] Lightweight transactions (cassandra). https://blog.pythian.com/lightweight-transactions-cassandra/.

[20] Consistency level (cassandra). https://docs.datastax.com/en/cassandra-oss/3.0/cassandra/dml/dmlConfigConsistency.html.

[21] Yaml. https://en.wikipedia.org/wiki/YAML.

[22] Adapter design pattern. https://www.tutorialspoint.com/python_design_patterns/python_design_patt