



UNIVERSITÀ
DEGLI STUDI
FIRENZE

Scuola di Scienze Matematiche, Fisiche e Naturali
Corso di Laurea in Informatica

Tesi di Laurea

IMPLEMENTAZIONE E TESTING DI UN
CLUSTER DI DATABASE SU UNA RETE DI
PARI

IMPLEMENTATION AND TESTING OF A
PEER NETWORK DATABASE CLUSTER

LINDA LUCIANO

Relatore: *Lorenzo Bettini*
Correlatore: *Correlatore*

Anno Accademico 2016-2017

INDICE

1	Definizione del problema	7
1.1.1	Cluster di database	7
1.1.2	Rete di pari	10
1.1.3	Sistemi di ridondanza disco (RAID)	11
1.1.4	Codice di correzione errore (Erasure Coding)	17
1.2	Software utilizzato per gli esperimenti	19
1.2.1	PosgreSQL	19
1.2.2	Pglogical	40
1.3	Hardware utilizzato per gli esperimenti	45
2	Definizione del progetto	49
2.1.1	Architettura del progetto	49
2.1.2	Simulazione di un filesystem distribuito (Dati e Metadati)	51
2.2	Considerazioni statistiche sulla ridondanza sul dato	52
2.3	Considerazioni statistiche sulla ridondanza del metadato	52
2.4	Resilienza ai cambiamenti di rete	52
3	Definizione del quadro sperimentale	53
3.1	Lancio in configurazione 1	53
3.2	Lancio in configurazione 2	53
3.3	Lancio in configurazione 3	53
4	Conclusioni e possibili evoluzioni	55
4.1	Utilizzo di dischi SSD	55
4.2	Utilizzo di processori Dual Core	55
5	esercizi	57

ELENCO DELLE FIGURE

Figura 1	Architecture Shared Nothing [4]	8
Figura 2	Architecture Shared Disk - Shared Everything [4]	9
Figura 3	Sezionamento senza ridondanza - Questa configurazione ha sezionamento, ma nessuna ridondanza dei dati. Offre le migliori prestazioni, ma nessuna tolleranza agli errori.[8]	13
Figura 4	Replicazione - Questa configurazione è costituita da almeno due unità che duplicano la memorizzazione dei dati. Non c'è sezionamento. [8]	14
Figura 5	Sezionamento a livello di bit - Questa configurazione ha alcuni dischi che memorizzano le informazioni di errore di verifica e correzione (ECC). [8]	15
Figura 6	Sezionamento a livello di byte con disco di parità - Questa configurazione utilizza la rigatura e dedica un'unità a memorizzare informazioni di parità. [8]	15
Figura 7	Sezionamento a livello di blocco con disco di parità [8]	16
Figura 8	Sezionamento a livello di blocco con parità distribuita [8]	16
Figura 9	Sezionamento a livello di blocco con doppia parità distribuita [8]	17
Figura 10	Migrazione e aggiornamenti PostgreSQL [14]	42
Figura 11	Aggregazione [14]	42
Figura 12	A cascata e distribuzione dati [14]	43
Figura 13	A cascata e distribuzione dati [14]	43
Figura 14	immagine1 [8]	45
Figura 15	immagine2 [8]	46
Figura 16	rack [8]	47
Figura 17	server [8]	48
Figura 18	Semplice esempio di un Provider e tre Subscriber abbonati	49
Figura 19	Set di replica o (<i>replication set</i>)	51

"Inserire citazione"

— Inserire autore citazione

1

DEFINIZIONE DEL PROBLEMA

(replicazione dei dati - introduzione)

1.1.1 *Cluster di database*

Un cluster è una raccolta di componenti che garantisce scalabilità e disponibilità distribuendone i costi. Un cluster di database (SQL usa il termine cluster di catalogo) è una collezione di database gestiti da una singola istanza di un server database in esecuzione. Un'istanza è la raccolta di memoria e processi che interagiscono con un database, cioè l'insieme di file fisici che effettivamente memorizzano i dati.[1] A tal fine, è possibile creare un cluster di database per applicazioni enterprise high-end, memorizzando e elaborando informazioni sui nodi.

L'architettura per un cluster di database è distinta da come le responsabilità dei dati sono condivise tra i nodi di calcolo.

Seguono due dei vantaggi principali offerti dal clustering, specialmente in un ambiente di database di alto volume:

- *Fault tolerance* (tolleranza di guasti): in caso di guasto del singolo server, il cluster offre un'alternativa, poiché esiste più di un server o istanza per gli utenti a cui connettersi.
- *Load balancing* (bilanciamento del carico): la funzionalità di clustering è generalmente impostata per consentire agli utenti di essere assegnati automaticamente al server con il minor carico.[1]

Ci sono differenti tipi di architetture clustering che si diversificano da come vengono memorizzati i dati e allocate le risorse. La prima modalità di clustering è conosciuta come architettura "*shared-nothing*" (SN). È un'architettura di elaborazione distribuita in cui ogni nodo/server è totalmente indipendente e autonomo, pertanto nessuno dei nodi condivide memoria o archiviazione del disco. Più generalmente, non esiste un unico punto di

contesa nel sistema.[3] Il partizionamento è tale che ogni nodo possiede un sottoinsieme dei dati, ovvero ogni nodo ha accesso esclusivo su quel particolare sottoinsieme. [2]))

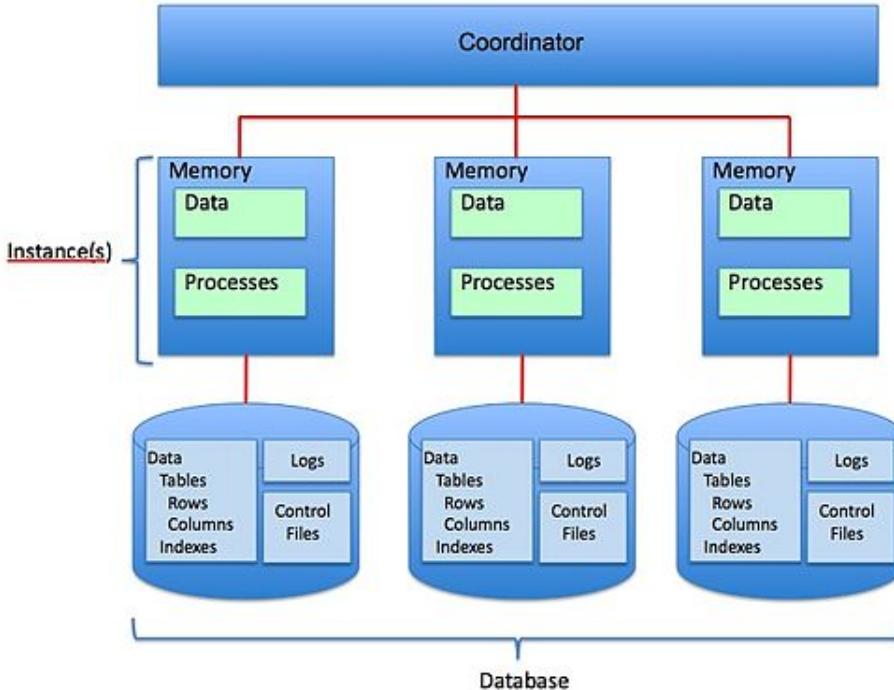


Figura 1: Architecture Shared Nothing [4]

I vantaggi dell'architettura SN rispetto a un'entità centrale che controlla la rete (un'architettura basata su controller) riguarda l'eliminazione di qualsiasi singolo punto di guasto, consentendo funzionalità di auto-riparazione (*self-healing*) e fornendo un vantaggio nell'offrire aggiornamenti non distruttivi.[5] *Shared-nothing* è anche noto come "*database sharding*". In generale, un sistema SN divide i suoi dati in vari nodi su database diversi o può richiedere a ciascun nodo di mantenere la propria copia dei dati dell'applicazione utilizzando un qualche tipo di protocollo di coordinamento.[3]

Si oppone a quest'ultima, l'architettura nota come "*shared-disk*" (disco condiviso), in cui tutti i dati vengono memorizzati centralmente in un unico disco e sono accessibili da tutti i nodi di cluster.[4] In questo tipo di struttura quindi più istanze di database vengono raggruppate in un singolo database sul disco. Nei sistemi di dischi condivisi, i blocchi (o

pagine) di dati su disco possono avere un solo proprietario.((L'architettura *shared-disk* è un esempio di *Synchronous multi-master*, ovvero ogni istanza del database può scrivere (cioè è un master) in modo sincrono.[4] **DA CHIEDERE E DA CONTROLLARE ULTIME FRASI. GUARDA COMMENTO per il sito**

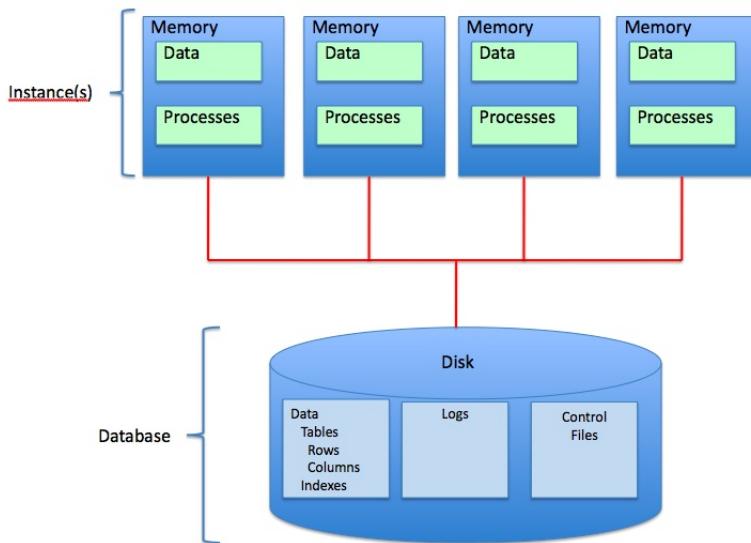


Figura 2: Architecture Shared Disk - Shared Everything [4]

In un'architettura SD, grandi reti di computer possono operare su un singolo set di dati senza la necessità di replicare o bloccare quel set di dati.[4] *Shared disk* ha due vantaggi:

- ogni processore ha la propria memoria,
- il bus di memoria non è un collo di bottiglia (contrariamente all'architettura "*shared-everything*"),
- il sistema offre un modo semplice per fornire un certo grado di tolleranza agli errori.

La distinzione tra i due tipi è diventata confusa di recente con l'introduzione di distribuzione della cache. In questa configurazione, i dati sono ancora gestiti centralmente, ma controllati da un potente "server virtuale" composto da molti server che lavorano insieme come uno.[2]

1.1.2 Rete di pari

Peer-to-peer networking (P2P) è un modello di comunicazione decentralizzato in cui ciascuna parte ha la stessa responsabilità per l'elaborazione dei dati e ciascuna parte può avviare una sessione di comunicazione. A differenza del modello *client/server*, in cui il client effettua una richiesta di servizio e il server soddisfa la richiesta, il modello di rete P2P, noto anche come *peer networking*, consente a ciascun nodo di funzionare sia come client che come server.[6]

Quando una rete P2P viene stabilita su Internet, è possibile utilizzare un server centrale per indicizzare i file oppure stabilire una rete distribuita in cui la condivisione dei file viene suddivisa tra tutti gli utenti della rete che memorizzano un determinato file. Le dimensioni della rete e i file disponibili consentono di condividere enormi quantità di dati.

Le prime reti P2P come Napster utilizzavano il software client e un server centrale, mentre reti successive come Kazaa e BitTorrent eliminavano il server centrale e dividevano i compiti di condivisione tra più nodi per liberare la larghezza di banda.

Seguono i vantaggi di una rete *peer-to-peer*:

- se un dispositivo collegato interrompe la connessione, il servizio non termina a differenza del modello *client-server*,
- è possibile configurare i computer in gruppi di lavoro *peer-to-peer* per consentire la condivisione di file e altre risorse su tutti i dispositivi. *Peer networking* consente di condividere facilmente i dati in entrambe le direzioni, sia per i *download* sul computer che per gli *upload* dal computer,
- su Internet, le reti peer-to-peer gestiscono un volume elevato di traffico di condivisione file distribuendo il carico su più computer. Poiché non si basano esclusivamente sui server centrali, le reti P2P possono scalare meglio e sono più resistenti delle reti *client-server* in caso di guasti o colli di bottiglia del traffico,
- Le reti *peer-to-peer* sono relativamente facili da espandere. Con l'aumentare del numero di dispositivi nella rete, aumenta la potenza della rete P2P, poiché ogni computer aggiuntivo è disponibile per l'elaborazione dei dati.[7]

Le reti *peer-to-peer* sono vulnerabili agli attacchi di sicurezza. Poiché ogni dispositivo partecipa al traffico di routing attraverso la rete, gli

hacker possono facilmente lanciare attacchi *denial of service*. Il software P2P funge da server e client, il che rende le *peer networking* più vulnerabili agli attacchi remoti rispetto alle reti client-server. I dati corrotti possono essere condivisi su reti P2P modificando i file già presenti in rete per introdurre codice dannoso.[7]

1.1.3 Sistemi di ridondanza disco (RAID)

RAID, acronimo di *redundant array of independent disks*, insieme ridondante di dischi indipendenti (originariamente *redundant array of inexpensive disks*), è una tecnologia che permette di memorizzare dati su più dischi rigidi in un computer (o collegati ad esso) in modo da garantire una gestione sicura dei dati[8]. I dispositivi RAID sono convenienti per sistemi che abbiano necessità di grandi quantità di dati continuamente disponibili.

Il RAID, con modalità differenti a seconda del tipo di configurazione, trae vantaggio dai principi di ridondanza dei dati e di parallelismo in modo da ottenere:

- incrementi di prestazioni (in lettura/scrittura);
- aumenti nella capacità di memorizzazione disponibile;
- miglioramenti nella tolleranza ai guasti, ne segue migliore affidabilità[10].

Il RAID rende il sistema resiliente alla perdita di uno o più hard disk, permettendo di sostituirli senza l'interruzione del servizio.

I volumi RAID vengono percepiti dal sistema operativo come una singola unità, indipendentemente dal numero di componenti che li costituiscono.

Il RAID funziona mettendo i dati su più dischi e consentendo operazioni di input/output (I/O) di sovrapporsi in modo equilibrato. Poiché l'utilizzo di più dischi aumenta il tempo medio tra i guasti, memorizzare i dati ridondantemente aumenta la tolleranza agli errori.

I dati vengono suddivisi in "*stripes*", ovvero in sezioni di stessa lunghezza, detta l'unità del sezionamento e scritti su differenti dischi. Quando si richiede una lettura di dimensione superiore all'unità di sezionamento, diverse implementazioni di diversi sistemi RAID distribuiscono l'operazione su più dischi in parallelo, aumentando le prestazioni. Ad esempio, se abbiamo sezioni da 1 bit e un array di D dischi, le sequenze di dati lunghe almeno D bit sfruttano tutti i dischi. **preso tutto da wikipedia**

RAID hardware e software

Il RAID può essere implementato sia con hardware dedicato che con software specifico.

Nel primo caso si tratta di unità di controllo che gestiscono tutto autonomamente, facendo in modo che il sistema operativo veda un disco normale. Nel secondo caso, è il sistema operativo che associa i dischi e li gestisce usando una forma di ridondanza attraverso un normale controller (ATA, SCSI, Fibre Channel o altro).

Le unità di controllo RAID sono più costose di quelle normali; tuttavia, se non si creano altri tipi di problemi, hanno il vantaggio di non creare difficoltà al sistema operativo.

Controllore RAID

Un controller RAID è un dispositivo hardware o un programma software utilizzato per gestire unità disco fisso (HDD) o unità SSD (*Solid State Drive, SSD*) in un computer o un array di archiviazione in modo da funzionare come unità logica.

Un controller offre un livello di astrazione tra un sistema operativo e le unità fisiche. Un controller RAID presenta gruppi a applicazioni e sistemi operativi come unità logiche per le quali è possibile definire schemi di protezione dei dati. Poiché il controller ha la possibilità di accedere a più copie di dati su più dispositivi fisici, ha la capacità di migliorare le prestazioni e proteggere i dati in caso di crash di sistema.

Nel RAID hardware, un controller fisico viene utilizzato per gestire l'array RAID. Il controller può assumere la forma di una scheda PCI o PCI Express (PCIe), progettata per supportare un formato di unità specifico come SATA o SCSI (alcuni controller RAID possono anche essere integrati con la scheda madre.)

Un controller RAID può anche essere solo software, utilizzando le risorse hardware del sistema host. Il RAID basata su software generalmente fornisce funzionalità simili a RAID *hardware-based*, ma la sua prestazione è tipicamente inferiore a quella delle versioni hardware.[13]

Livelli RAID

La caratteristica fondamentale che identifica una configurazione RAID è, come citato in precedenza, l'array, che rappresenta il tipo di collegamento logico che c'è tra i vari dischi. Con tale criterio viene determinato il livello RAID, ovvero la configurazione della tipologia di RAID e stabilito il

numero minimo di hard disk che sono necessari per attivarlo. A seconda del livello RAID sono implementate diverse caratteristiche operative per ottenere maggiori prestazioni o una maggiore sicurezza dei propri dati oppure entrambe le condizioni.

Si distinguono sei livelli, da 0 a 5. Questo sistema numerato consente di differenziare le versioni e di scegliere come diffondere i dati attraverso l'array ed è stato suddiviso in tre categorie: livelli RAID standard, nidificati e non standard[8] (segue la descrizione dei livelli standard e qualche nidificato).

LIVELLI RAID STANDARD

- RAID 0: livello privo di ridondanza. Si occupa di unire due o più dischi, all'interno dei quali i dati vengono suddivisi equamente (tramite striping o sezionamento), in modo da bilanciare anche il carico di operazioni di lettura e scrittura che li riguardano. Livello che consente di realizzare un disco virtuale di grandi dimensioni, più efficiente, ma la rottura di uno dei dischi porta alla perdita di tutti i dati.

RAID 0 è noto anche con il nome di *block striping*.[bibliografia commentata](#)

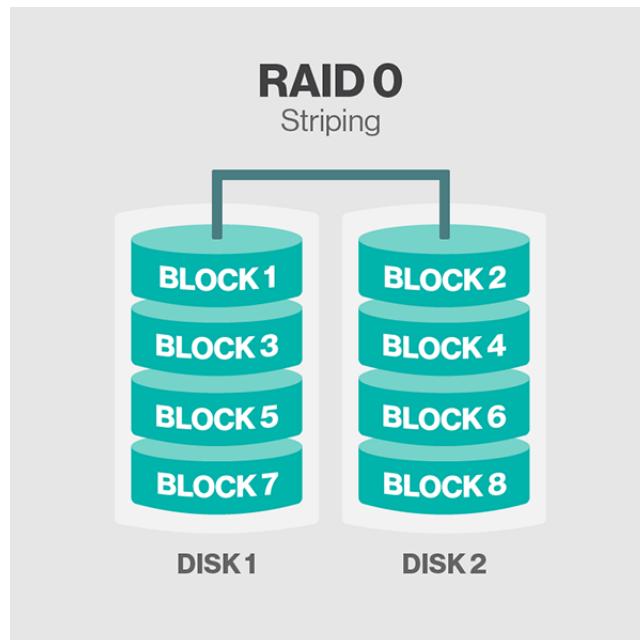


Figura 3: Sezionamento senza ridondanza - Questa configurazione ha sezionamento, ma nessuna ridondanza dei dati. Offre le migliori prestazioni, ma nessuna tolleranza agli errori.[8]

- RAID 1: livello che si occupa di unire assieme due o più dischi riproducendo fedelmente gli stessi dati. Questa configurazione mantiene quindi almeno una copia esatta di tutti i dati, detta "mirror". In questo caso, la rottura di un disco non pregiudica l'utilizzo dei dati che sono disponibili nel disco o nei dischi rimanenti. Più precisamente, l'affidabilità aumenta linearmente al numero di dischi presenti: un sistema con N dischi è in grado di resistere alla rottura di $N-1$ componenti.

La lettura delle prestazioni è migliorata poiché entrambi i dischi possono essere letti contemporaneamente. La scrittura delle prestazioni è la stessa di quella per il singolo disco.[8]

RAID 1 è conosciuto anche come *disk mirroring*.

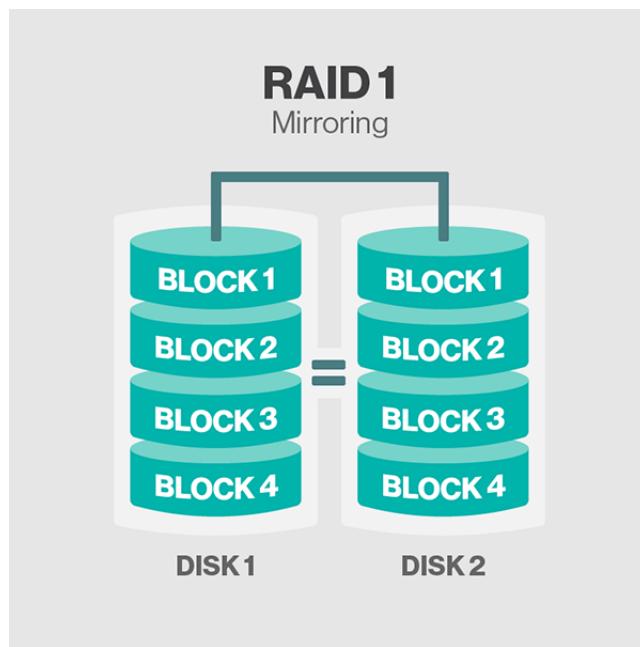


Figura 4: Replicazione - Questa configurazione è costituita da almeno due unità che duplicano la memorizzazione dei dati. Non c'è sezionamento. [8]

- RAID 2: livello che divide i dati al livello di bit (invece che di blocco) e usa un *codice di Hamming* per la correzione d'errore che permette di correggere errori su singoli bit e di rilevare errori doppi. Questi dischi sono sincronizzati dal controllore, in modo tale che la testina di ciascun disco sia nella stessa posizione in ogni disco.[10]

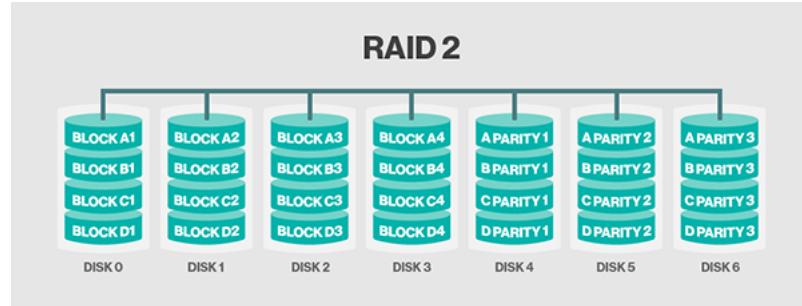


Figura 5: Sezionamento a livello di bit - Questa configurazione ha alcuni dischi che memorizzano le informazioni di errore di verifica e correzione (ECC). [8]

- RAID 3: livello che si occupa di unire assieme almeno tre o più dischi, all'interno dei quali i dati vengono suddivisi equamente, in modo da bilanciare anche il carico di operazioni di lettura e scrittura che li riguardano. Dedicano uno di questi dischi al contenimento di un sistema di codici di controllo, che permettono di ricostruire i dati nel caso in cui uno degli altri dischi si rompa. Le informazioni ECC vengono utilizzate per rilevare gli errori. Il recupero dei dati viene effettuato calcolando l'esclusiva OR (XOR) delle informazioni registrate sulle altre unità.[8]

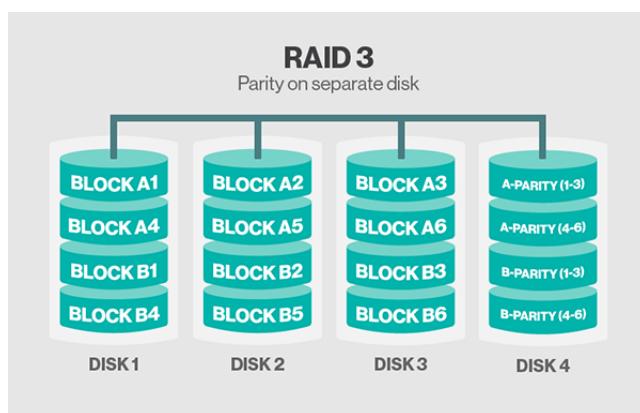


Figura 6: Sezionamento a livello di byte con disco di parità - Questa configurazione utilizza la rigatura e dedica un'unità a memorizzare informazioni di parità. [8]

- RAID 4: livello simile al livello tre, con la differenza che i dati vengono distribuiti in modo più efficiente tra i dischi, ma rimane compito di un disco separato il sistema di codici di controllo che permette la ricostruzione dei dati, chiamati "blocchi di parità".
Questo livello utilizza grandi sezionamenti, il che significa che è possibile leggere i record da un'unica unità.[8]

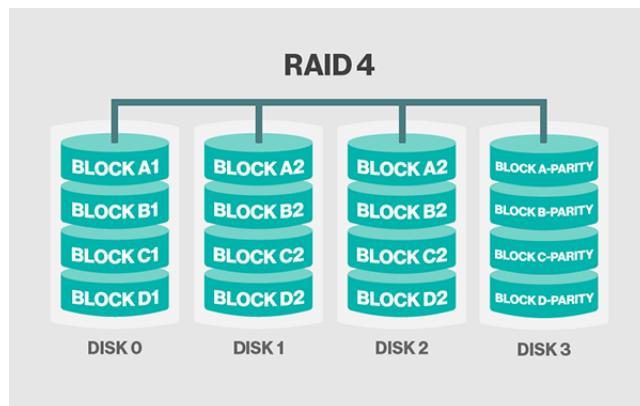


Figura 7: Sezionamento a livello di blocco con disco di parità [8]

- RAID 5: livello basato su livello di blocco con parità che risiedono su ciascuna unità. L'architettura dell'array consente alle operazioni di lettura e scrittura di coprire più unità. Ciò determina prestazioni migliori di quelle di un'unità singola, ma non altrettanto elevate di quella di un array RAID 0.[8]

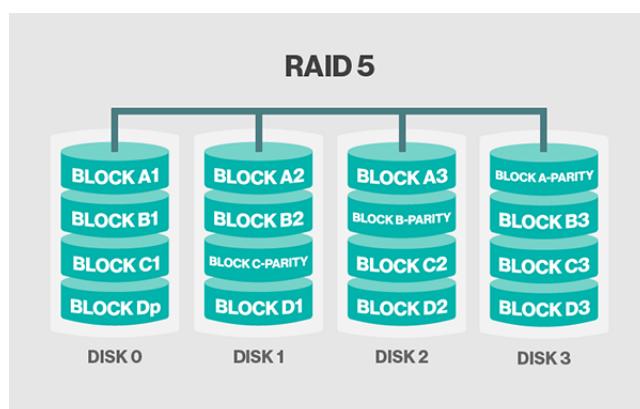


Figura 8: Sezionamento a livello di blocco con parità distribuita [8]

- RAID 6: livello simile a RAID 5, ma include un secondo schema di parità distribuito attraverso le unità nell'array. L'utilizzo di una parità aggiuntiva consente all'array di continuare a funzionare anche se due dischi non funzionano contemporaneamente. Tuttavia, questa protezione supplementare è più costosa. Le matrici RAID 6 hanno un costo superiore a gigabyte (GB) e spesso hanno prestazioni di scrittura più lente degli array RAID 5.[8]

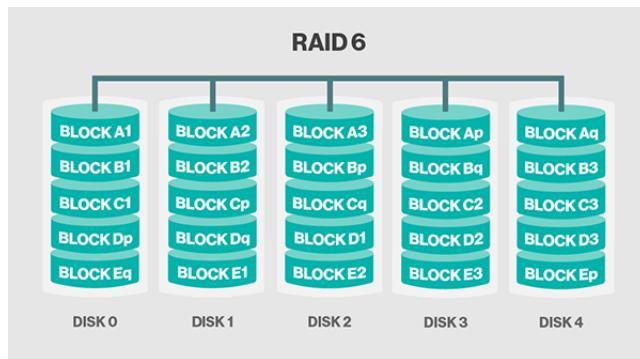


Figura 9: Sezionamento a livello di blocco con doppia parità distribuita [8]

LIVELLI RAID NIDIFICATI ANNIDATI I livelli Annidati sono dei tipi di livelli più complessi ottenuti dalla combinazione di alcuni livelli RAID Standard. Esempi classici sono le configurazioni RAID 0+1 o 10.

1.1.4 Codice di correzione errore (*Erasure Coding*)

La codifica di cancellazione, noto come *Erasure Coding* (EC) è un metodo di protezione dei dati, i quali vengono suddivisi in frammenti, estesi e codificati con pezzi di dati ridondanti e memorizzati su un insieme di posizioni o supporti di memorizzazione diversi.

L'obiettivo della codifica di cancellazione è quello di consentire di ricostruire i dati che vengono danneggiati utilizzando le informazioni sui dati memorizzati altrove nell'array. Lo svantaggio della codifica di cancellazione è che può essere più intenso della CPU e che può tradursi in una maggiore latenza.[11]

La codifica di cancellazione è utile con la presenza di grandi quantità di dati e tutte le applicazioni o sistemi che devono tollerare i guasti, come sistemi di array a dischi, griglie di dati, applicazioni di archiviazione distribuite. Un caso comune di utilizzo corrente per la codifica di cancellazione è in un sistema object-based cloud storage[11].

Come funziona

La codifica di cancellazione crea una funzione matematica per descrivere un insieme di numeri in modo che possano essere controllati per l'accuracy e recuperati in caso di perdita. Questo è il concetto fondamentale dei metodi di codifica di cancellazione, implementati più frequentemente utilizzando i codici *Reed-Solomon*[11].

In termini matematici, la protezione offerta dalla codifica di cancellazione può essere rappresentata in forma semplice dalla seguente equazione:

$$n = k + m$$

dove:

- la variabile k è la quantità originale di dati o simboli
- la variabile m indica i simboli aggiuntivi o ridondanti che vengono aggiunti per fornire protezione dai guasti
- la variabile n è il numero totale di simboli creati dopo il processo di codifica di cancellazione[11]
- la variabile r , chiamata velocità di codice, è definita nel seguente modo:

$$r = \sqrt{\frac{k}{n}}$$

Ad esempio, in una configurazione 10 di 16, o EC 10/16, sei simboli supplementari (m) saranno aggiunti ai 10 simboli di base (k). I 16 frammenti di dati (n) saranno diffusi su 16 unità, nodi o posizioni geografiche. Il file originale potrebbe essere ricostruito da 10 frammenti verificati.[11]

I codici di cancellazione, noti anche come codici di correzione degli errori di avanzamento (FEC), sono stati sviluppati più di 50 anni fa. Da quel momento sono emersi diversi tipi. In uno dei tipi più comuni, *Reed-Solomon*, i dati possono essere ricostruiti utilizzando qualsiasi combinazione di simboli k o pezzi di dati, anche se i simboli m sono persi o non sono disponibili. Ad esempio, in EC 10/16, sei unità, nodi o posizioni geografiche potrebbero essere persi o non disponibili e il file originale sarà ancora recuperabile.[11]

1.2 SOFTWARE UTILIZZATO PER GLI ESPERIMENTI

1.2.1 PosgreSQL

PostgreSQL è un potente sistema *Open Source* di database relazionale (DBMS, *Database Management System*) cioè è un sistema software progettato per consentire la creazione e manipolazione efficiente di database, ovvero di collezioni di dati strutturati.

Ha più di 15 anni di sviluppo attivo e un'architettura collaudata che ha guadagnato una notevole reputazione per l'affidabilità, l'integrità e la salvaguardia dei dati allocati e la correttezza di archiviazione.

PostgreSQL è un sistema di gestione dei database relazionale (*Object-Relational*, acronimo ORDBMS) basato su POSTGRES, Versione 4.2, sviluppato presso l'Università della California presso il Dipartimento di Informatica di Berkeley.[12]

Funziona su tutti i principali sistemi operativi, tra cui Linux, UNIX (AIX, BSD, HP-UX, SGI IRIX, Mac OS X, Solaris, Tru64), e Windows.[12]

Supporta gran parte dello standard SQL e offre molte funzionalità moderne:

- *queries* complesse
- foreign keys
- triggers
- views aggiornabili
- integrità transazionale
- controllo della concorrenza multiversione.

Seguono le caratteristiche principali di PostgreSQL; le funzionalità introdotte nella recente versione 9.1 :

- elevata aderenza agli standard SQL;
- architettura client-server con una gamma completa di driver e di client;
- progettato in modo altamente concorrente, evitando che i processi in scrittura blocchino i processi in lettura;

- altamente configurabile ed estendibile, consentendo svariati tipi di applicazioni;
- elevate scalabilità e prestazioni, unite a un ampio spettro di possibilità per tarare la configurazione;
- sofisticato ottimizzatore delle query, adeguato per la business intelligence;
- supporto completo per Java, Python, Perl, PHP e molti altri linguaggi, sia per le procedure interne al server di database che per l'accesso da parte di client;
- elevata affidabilità, con una vasta serie di caratteristiche per durabilità e alta disponibilità;
- tipi di dati avanzati, come ad esempio GIS, full text search, e molti altri;
- internazionalizzazione, codifiche multibyte e collation. [13]

PostgreSQL è progettato per essere estensibile, poiché è possibile definire i propri tipi di dati, tipi di indici e lingue funzionali. Offre un ricco set di strumenti per gli sviluppatori in modo da gestire l'accesso simultaneo ai dati. Inoltre vi è la possibilità di ottimizzarlo per soddisfare le proprie esigenze, tramite uno sviluppo di un plugin personalizzato.

Un database di classe enterprise, PostgreSQL vanta funzionalità sofisticate come il controllo della concorrenza multiversione, il ripristino in tempo reale (*point in time recovery*), *tablespaces*, replica asincrona, transazioni nidificate (*savepoints*), backup in linea/a caldo, un sofisticato query planner/optimizer, ed il *write ahead logging* per una maggiore tolleranza ai guasti.

È estremamente scalabile sia nella quantità pura di dati che può gestire sia nel numero di utenti concorrenti che può ospitare. Esistono sistemi Active PostgreSQL in ambienti di produzione che gestiscono oltre 4 terabyte di dati.[12]

Alcuni limiti generali di PostgreSQL sono inclusi nei punti riportati di seguito:

Dimensione massima del database	Illimitato
Dimensione massima Tabella	32 TB
Dimensione massima Riga	1,6 TB
Dimensione Massima campo	1 GB
Numero Massimo Righe per tabella	Illimitato
Numero Massimo colonne per la tabella	250 - 1600
Numero Massimo indici per la tabella	Illimitato[13]

MVCC (*Multi-Version Concurrency Control*) è una tecnica avanzata per migliorare le prestazioni del database in un ambiente multiutente, mantenendo la coerenza dei dati. A differenza della maggior parte degli altri sistemi di database che utilizzano i *locks* per il controllo della concorrenza, Postgres mantiene la coerenza dei dati utilizzando questo modello. Ciò significa che durante l'interrogazione di un database ogni transazione vede un'istantanea di dati (una versione del database), indipendentemente dallo stato corrente dei dati sottostanti. Questo protegge la transazione dalla visualizzazione di dati incoerenti che potrebbero essere causati da altri eventuali aggiornamenti simultanei delle transazioni sulle stesse righe di dati, fornendo l'isolamento della transazione per ciascuna sessione del database.

La differenza principale tra i modelli multiversione e di blocco è che nei blocchi MVCC acquisiti per la query (lettura) i dati non sono in conflitto con i blocchi acquisiti per la scrittura dei dati e quindi la lettura non blocca mai la scrittura e la scrittura non blocca mai la lettura.[12]

Caratteristiche PostgreSQL

PostgreSQL ha il pieno supporto per le subquery (incluse sottotitoli nella clausola FROM), i livelli di isolamento delle transazioni di lettura e serializzabili. E mentre PostgreSQL ha un catalogo di sistema completamente relazionale che supporta più schemi per database, il suo catalogo è accessibile anche attraverso lo schema di informazioni definito nello standard SQL.

Le funzionalità di integrità dei dati includono le chiavi primarie (combinate), le chiavi estranee con aggiornamenti e cancellazioni a cascata, i controlli dei vincoli, i vincoli unici e non i vincoli nullo.

Ha anche una serie di estensioni e funzionalità avanzate. PostgreSQL supporta indici composti, unici, parziali e funzionali che possono utilizzare qualsiasi metodo di archiviazione B-tree, R-tree, hash o GiST.

Altre funzionalità avanzate includono l'ereditarietà delle tabelle, i sistemi di regole e gli eventi del database. L'ereditarietà di tabella mette un orientamento orientato all'oggetto sulla creazione della tabella, consentendo ai progettisti di database di derivare nuove tabelle da altre tabelle, trattandole come classi di base. Ancora meglio, PostgreSQL supporta sia l'ereditarietà singola che quella multipla in questo modo.

Il sistema di regole, chiamato anche il sistema di riscrittura delle query, consente al progettista del database di creare regole che identificano operazioni specifiche per una determinata tabella o vista e le trasformano dinamicamente in operazioni alternative quando vengono elaborate.

Il sistema eventi è un sistema di comunicazione interprocesso in cui i messaggi e gli eventi possono essere trasmessi tra i client utilizzando i comandi LISTEN e NOTIFY, consentendo sia la semplice comunicazione peer to peer sia un coordinamento avanzato sugli eventi del database. Poiché le notifiche possono essere rilasciate da trigger e stored procedure, i client PostgreSQL possono monitorare eventi di database come gli aggiornamenti, gli insert o le eliminazioni di tabella quando vengono eseguiti.[12]

Elevata personalizzazione

PostgreSQL esegue procedure memorizzate in più di una dozzina di linguaggi di programmazione, tra cui Java, Perl, Python, Ruby, Tcl, C / C ++ e il proprio PL / pgSQL, simile a PL / SQL di Oracle. I trigger e le stored procedure possono essere scritti in C e caricati nel database come libreria, permettendo una grande flessibilità nell'estensione delle sue funzionalità. Allo stesso modo, PostgreSQL include un *framework* che consente agli sviluppatori di definire e creare i propri tipi di dati personalizzati insieme a funzioni di supporto e operatori che definiscono il loro comportamento.

Il codice sorgente di PostgreSQL è disponibile sotto una licenza libera open source: la licenza PostgreSQL. Questa licenza dà la libertà di utilizzare, modificare e distribuire PostgreSQL in qualsiasi forma, sorgente aperta o chiusa.[12]

Non esiste un unico formato per tutti i software di replica. È necessario capire le proprie esigenze e come si adattino diversi approcci. Ad esempio, ecco due estremi nello spazio del problema di replica: Hai alcuni server collegati a una rete locale che vuoi mantenere sempre la corrente per scopi di failover e di bilanciamento del carico. Qui si considererebbero soluzioni sincrone, desiderose e quindi prive di conflitti. I tuoi utenti prendono una

copia locale del database con loro sui computer portatili quando lasciano l'ufficio, apportano modifiche mentre sono lontani e hanno bisogno di unire quelli con il database principale quando tornano. Qui si desidera un approccio asincrono e pigro di replica e sarà costretto a considerare come gestire i conflitti nei casi in cui lo stesso record sia stato modificato sia sul server master che su una copia locale. Questi sono problemi di replica del database, ma il modo migliore per risolverli sarà molto diverso. E come si può vedere da questi esempi, la replica ha molte terminologie specifiche che dovrà capire.

Hot Standby / Streaming Replication sarà disponibile a partire da PostgreSQL 9.0 e fornisce una replica binaria asincrona a uno o più standby. Gli standby poss

(spiegazione cinci -audio-)

La replica delle modifiche dello schema è un problema frequentemente discusso e solo pochi sistemi di database forniscono le estensioni necessarie per implementarlo. PostgreSQL non fornisce la possibilità di definire i trigger richiamati sulle modifiche dello schema, quindi un modo trasparente per replicare le modifiche dello schema non è possibile senza un sostanziale lavoro nel sistema centrale di PostgreSQL. Invece tendono ad essere gruppi di istruzioni DDL e DML che modificano più oggetti di database e fanno manipolazioni di dati di massa come l'aggiornamento di una nuova colonna al suo valore iniziale. Il sistema di replica Pglogical avrà un meccanismo per eseguire script SQL in modo controllato come parte del processo di replica.

Write-Ahead Logging (WAL)

Write-Ahead Logging (WAL) è un metodo standard per garantire l'integrità dei dati. Il concetto centrale di WAL è che le modifiche ai file di dati (dove risiedono tabelle e indici) devono essere scritte solo dopo che tali modifiche sono state registrate, ovvero dopo che i record di registro che descrivono le modifiche sono stati scaricati nella memoria permanente. Se viene seguita questa procedura, non abbiamo bisogno di svuotare le pagine di dati sul disco su ogni *commit* di transazione, perché sappiamo che in caso di crash saremo in grado di recuperare il database usando il log: eventuali modifiche che non sono state applicate alle pagine di dati possono essere rifatte dai record del registro. Questo è il recupero *roll-forward*, noto anche come REDO.

Poiché WAL ripristina il contenuto del file di database dopo un arresto anomalo, i *filesystem* registrati non sono necessari per l'archiviazione affidabile dei file di dati o dei file WAL.

L'utilizzo di WAL determina un numero notevolmente ridotto di scritture su disco, poiché è necessario scaricare il file di registro sul disco per garantire che una transazione venga eseguita, piuttosto che ogni file di dati modificato dalla transazione.

Il file di registro viene scritto in modo sequenziale e pertanto il costo della sincronizzazione del log è molto inferiore al costo dello svuotamento delle pagine di dati. Ciò è particolarmente vero per i server che gestiscono molte piccole transazioni toccando diverse parti dell'archivio dati. Inoltre, quando il server sta elaborando molte piccole transazioni simultanee, un `fsync` del file di log può essere sufficiente per il *commit* di molte transazioni.[12]

Archiviando i dati WAL possiamo supportare il ripristino in qualsiasi istante dei dati WAL disponibili: installiamo semplicemente un backup fisico preliminare del database e riproduciamo il registro WAL fino al momento desiderato. Inoltre, il backup fisico non deve essere un'istantanea istantanea dello stato del database: se viene eseguito per un certo periodo di tempo, la riproduzione del log WAL per quel periodo risolverà eventuali incoerenze interne.**vedere se metterlo**

File di configurazione di PostgreSQL

Segue parte del file di configurazione di PostgreSQL chiamato `postgresql.conf` ottenuta dall'ultima versione 10.0, ponendo attenzione ai parametri dei lanci di configurazione:

```
# -----
# PostgreSQL configuration file
# -----
#
# This file consists of lines of the form:
#
#     name = value
#
# (The "=" is optional.) Whitespace may be used. Comments are introduced with
# "#" anywhere on a line. The complete list of parameter names and allowed
# values can be found in the PostgreSQL documentation.
#
# The commented-out settings shown in this file represent the default values.
# Re-commenting a setting is NOT sufficient to revert it to the default value;
# you need to reload the server.
```

```

#
# This file is read on server startup and when the server receives a SIGHUP
# signal. If you edit the file on a running system, you have to SIGHUP the
# server for the changes to take effect, run "pg_ctl reload", or execute
# "SELECT pg_reload_conf()". Some parameters, which are marked below,
# require a server shutdown and restart to take effect.
#
# Any parameter can also be given as a command-line option to the server, e.g.
# "postgres -c log_connections=on". Some parameters can be changed at run time
# with the "SET" SQL command.
#
# Memory units: kB = kilobytes           Time units: ms  = milliseconds
#                 MB = megabytes          s    = seconds
#                 GB = gigabytes         min = minutes
#                 TB = terabytes        h    = hours
#                                         d    = days

#-----
# FILE LOCATIONS
#-----

# The default values of these variables are driven from the -D command-line
# option or PGDATA environment variable, represented here as ConfigDir.

#data_directory = 'ConfigDir'# use data in another directory
# (change requires restart)
#hba_file = 'ConfigDir/pg_hba.conf'# host-based authentication file
# (change requires restart)
#ident_file = 'ConfigDir/pg_ident.conf'# ident configuration file
# (change requires restart)

# If external_pid_file is not explicitly set, no extra PID file is written.
#external_pid_file = ''# write an extra PID file
# (change requires restart)

#-----
# CONNECTIONS AND AUTHENTICATION
#-----
```

```
# - Connection Settings -  
  
listen_addresses = '*'# what IP address(es) to listen on;  
# comma-separated list of addresses;  
# defaults to 'localhost'; use '*' for all  
# (change requires restart)  
port = 5432 # (change requires restart)  
max_connections = 100 # (change requires restart)  
#superuser_reserved_connections = 3 # (change requires restart)  
#unix_socket_directories = '/tmp'# comma-separated list of directories  
# (change requires restart)  
#unix_socket_group = ''# (change requires restart)  
#unix_socket_permissions = 0777 # begin with 0 to use octal notation  
# (change requires restart)  
#bonjour = off # advertise server via Bonjour  
# (change requires restart)  
#bonjour_name = ''# defaults to the computer name  
# (change requires restart)  
  
# - Security and Authentication -  
  
#authentication_timeout = 1min # 1s-600s  
#ssl = off  
#ssl_ciphers = 'HIGH:MEDIUM:+3DES:!aNULL' # allowed SSL ciphers  
#ssl_prefer_server_ciphers = on  
#ssl_ecdh_curve = 'prime256v1'  
#ssl_dh_params_file = ''  
#ssl_cert_file = 'server.crt'  
#ssl_key_file = 'server.key'  
#ssl_ca_file = ''  
#ssl_crl_file = ''  
#password_encryption = md5 # md5 or scram-sha-256  
#db_user_namespace = off  
#row_security = on  
  
# GSSAPI using Kerberos  
#krb_server_keyfile = ''  
#krb_caseins_users = off
```

```
# - TCP Keepalives -
# see "man 7 tcp" for details

#tcp_keepalives_idle = 0 # TCP_KEEPIDLE, in seconds;
# 0 selects the system default
#tcp_keepalives_interval = 0 # TCP_KEEPINTVL, in seconds;
# 0 selects the system default
#tcp_keepalives_count = 0 # TCP_KEEPCNT;
# 0 selects the system default

#-----
# RESOURCE USAGE (except WAL)
#-----

# - Memory -

shared_buffers = 256MB # min 128kB
# (change requires restart)
#huge_pages = try # on, off, or try
# (change requires restart)
#temp_buffers = 8MB # min 800kB
max_prepared_transactions = 100 # zero disables the feature
# (change requires restart)
# Caution: it is not advisable to set max_prepared_transactions nonzero unless
# you actively intend to use prepared transactions.
#work_mem = 32MB # min 64kB
#maintenance_work_mem = 128MB # min 1MB
#replacement_sort_tuples = 150000 # limits use of replacement selection sort
#autovacuum_work_mem = -1 # min 1MB, or -1 to use maintenance_work_mem
#max_stack_depth = 2MB # min 100kB
dynamic_shared_memory_type = posix # the default is the first option
# supported by the operating system:
#   posix
#   sysv
#   windows
#   mmap
# use none to disable dynamic shared memory
# (change requires restart)
```

```

# - Disk -

#temp_file_limit = -1 # limits per-process temp file space
# in kB, or -1 for no limit

# - Kernel Resource Usage -

#max_files_per_process = 1000 # min 25
# (change requires restart)
shared_preload_libraries = 'pglogical'# (change requires restart)

# - Cost-Based Vacuum Delay -

#vacuum_cost_delay = 0 # 0-100 milliseconds
#vacuum_cost_page_hit = 1 # 0-10000 credits
#vacuum_cost_page_miss = 10 # 0-10000 credits
#vacuum_cost_page_dirty = 20 # 0-10000 credits
#vacuum_cost_limit = 200 # 1-10000 credits

# - Background Writer -

#bgwriter_delay = 200ms # 10-10000ms between rounds
#bgwriter_lru_maxpages = 100 # 0-1000 max buffers written/round
#bgwriter_lru_multiplier = 2.0 # 0-10.0 multiplier on buffers scanned/round
#bgwriter_flush_after = 512kB # measured in pages, 0 disables

# - Asynchronous Behavior -

#effective_io_concurrency = 1 # 1-1000; 0 disables prefetching
max_worker_processes = 64 # (change requires restart)
#max_parallel_workers_per_gather = 2 # taken from max_parallel_workers
#max_parallel_workers = 8 # maximum number of max_worker_processes that
# can be used in parallel queries
#old_snapshot_threshold = -1 # 1min-60d; -1 disables; 0 is immediate
# (change requires restart)
#backend_flush_after = 0 # measured in pages, 0 disables

#-----
# WRITE AHEAD LOG

```

```

#-----



# - Settings -



wal_level = logical # minimal, archive, hot_standby, or logical
# (change requires restart)
#fsync = on # turns forced synchronization on or off
synchronous_commit = off # synchronization level;
# off, local, remote_write, or on
#wal_sync_method = fsync # the default is the first option
# supported by the operating system:
#   open_datasync
#   fdatasync (default on Linux)
#   fsync
#   fsync_writethrough
#   open_sync
#full_page_writes = on # recover from partial page writes
#wal_compression = off # enable compression of full-page writes
#wal_log_hints = off # also do full page writes of non-critical updates
# (change requires restart)
#wal_buffers = -1 # min 32kB, -1 sets based on shared_buffers
# (change requires restart)
#wal_writer_delay = 200ms # 1-10000 milliseconds
#wal_writer_flush_after = 1MB # measured in pages, 0 disables

#commit_delay = 0 # range 0-100000, in microseconds
#commit_siblings = 5 # range 1-1000

# - Checkpoints -



#checkpoint_timeout = 5min # range 30s-1d
#max_wal_size = 1GB
#min_wal_size = 80MB
#checkpoint_completion_target = 0.5 # checkpoint target duration, 0.0 - 1.0
#checkpoint_flush_after = 256kB # measured in pages, 0 disables
#checkpoint_warning = 30s # 0 disables

# - Archiving -



#archive_mode = off # enables archiving; off, on, or always

```

```

# (change requires restart)
#archive_command = ''# command to use to archive a logfile segment
# placeholders: %p = path of file to archive
#           %f = file name only
# e.g. 'test ! -f /mnt/server/archivedir/%f && cp %p /mnt/server/archivedir/%f'
#archive_timeout = 0 # force a logfile segment switch after this
# number of seconds; 0 disables

#-----
# REPLICATION
#-----

# - Sending Server(s) -

# Set these on the master and on any standby that will send replication data.

max_wal_senders = 64 # max number of walwriter processes
# (change requires restart)
#wal_keep_segments = 0 # in logfile segments, 16MB each; 0 disables
#wal_writer_timeout = 60s # in milliseconds; 0 disables

max_replication_slots = 64 # max number of replication slots
# (change requires restart)
track_commit_timestamp = on # collect timestamp of transaction commit
# (change requires restart)

# - Master Server -

# These settings are ignored on a standby server.

synchronous_standby_names = '1 (r1, r2, r3, r4)'
# standby servers that provide sync rep
# method to choose sync standbys, number of sync standbys,
# and comma-separated list of application_name
# from standby(s); '*' = all
#vacuum_defer_cleanup_age = 0 # number of xacts by which cleanup is delayed

# - Standby Servers -

```

```

# These settings are ignored on a master server.

#hot_standby = on # "off" disallows queries during recovery
# (change requires restart)
#max_standby_archive_delay = 30s # max delay before canceling queries
# when reading WAL from archive;
# -1 allows indefinite delay
#max_standby_streaming_delay = 30s # max delay before canceling queries
# when reading streaming WAL;
# -1 allows indefinite delay
#wal_receiver_status_interval = 10s # send replies at least this often
# 0 disables
#hot_standby_feedback = off # send info from standby to prevent
# query conflicts
#wal_receiver_timeout = 60s # time that receiver waits for
# communication from master
# in milliseconds; 0 disables
#wal_retrieve_retry_interval = 5s # time to wait before retrying to
# retrieve WAL after a failed attempt

# - Subscribers -

# These settings are ignored on a publisher.

max_logical_replication_workers = 16 #4 # taken from max_worker_processes
# (change requires restart)
#max_sync_workers_per_subscription = 2 # taken from max_logical_replication_w

#-----
# QUERY TUNING
#-----

# - Planner Method Configuration -

#enable_bitmapscan = on
#enable_hashagg = on
#enable_hashjoin = on
#enable_indexscan = on
#enable_indexonlyscan = on

```

```

#enable_material = on
#enable_mergejoin = on
#enable_nestloop = on
#enable_seqscan = on
#enable_sort = on
#enable_tidscan = on

# - Planner Cost Constants -

#seq_page_cost = 1.0 # measured on an arbitrary scale
#random_page_cost = 4.0 # same scale as above
#cpu_tuple_cost = 0.01 # same scale as above
#cpu_index_tuple_cost = 0.005 # same scale as above
#cpu_operator_cost = 0.0025 # same scale as above
#parallel_tuple_cost = 0.1 # same scale as above
#parallel_setup_cost = 1000.0 # same scale as above
#min_parallel_table_scan_size = 8MB
#min_parallel_index_scan_size = 512kB
#effective_cache_size = 4GB

# - Genetic Query Optimizer -

#gqo = on
#gqo_threshold = 12
#gqo_effort = 5 # range 1-10
#gqo_pool_size = 0 # selects default based on effort
#gqo_generations = 0 # selects default based on effort
#gqo_selection_bias = 2.0 # range 1.5-2.0
#gqo_seed = 0.0 # range 0.0-1.0

# - Other Planner Options -

#default_statistics_target = 100 # range 1-10000
#constraint_exclusion = partition # on, off, or partition
#cursor_tuple_fraction = 0.1 # range 0.0-1.0
#fromCollapse_limit = 8
#joinCollapse_limit = 8 # 1 disables collapsing of explicit
# JOIN clauses
#force_parallel_mode = off

```

```
#-----
# ERROR REPORTING AND LOGGING
#-----

# - Where to Log -

#log_destination = 'stderr'# Valid values are combinations of
# stderr, csvlog, syslog, and eventlog,
# depending on platform. csvlog
# requires logging_collector to be on.

# This is used when logging to stderr:
#logging_collector = off # Enable capturing of stderr and csvlog
# into log files. Required to be on for
# csvlogs.
# (change requires restart)

# These are only used if logging_collector is on:
#log_directory = 'log'# directory where log files are written,
# can be absolute or relative to PGDATA
#log_filename = 'postgresql-%Y-%m-%d_%H%M%S.log'# log file name pattern,
# can include strftime() escapes
#log_file_mode = 0600 # creation mode for log files,
# begin with 0 to use octal notation
#log_truncate_on_rotation = off # If on, an existing log file with the
# same name as the new log file will be
# truncated rather than appended to.
# But such truncation only occurs on
# time-driven rotation, not on restarts
# or size-driven rotation. Default is
# off, meaning append to existing files
# in all cases.
#log_rotation_age = 1d # Automatic rotation of logfiles will
# happen after that time. 0 disables.
#log_rotation_size = 10MB # Automatic rotation of logfiles will
# happen after that much log output.
# 0 disables.

# These are relevant when logging to syslog:
```

```
#syslog_facility = 'LOCAL0'
#syslog_ident = 'postgres'
#syslog_sequence_numbers = on
#syslog_split_messages = on

# This is only relevant when logging to eventlog (win32):
# (change requires restart)
#event_source = 'PostgreSQL'

# - When to Log -

#client_min_messages = notice # values in order of decreasing detail:
#    debug5
#    debug4
#    debug3
#    debug2
#    debug1
#    log
#    notice
#    warning
#    error

#log_min_messages = warning # values in order of decreasing detail:
#    debug5
#    debug4
#    debug3
#    debug2
#    debug1
#    info
#    notice
#    warning
#    error
#    log
#    fatal
#    panic

#log_min_error_statement = error # values in order of decreasing detail:
#    debug5
#    debug4
#    debug3
```

```
# debug2
# debug1
# info
# notice
# warning
# error
# log
# fatal
# panic (effectively off)

#log_min_duration_statement = -1 # -1 is disabled, 0 logs all statements
# and their durations, > 0 logs only
# statements running at least this number
# of milliseconds

# - What to Log -

#debug_print_parse = off
#debug_print_rewritten = off
#debug_print_plan = off
#debug_pretty_print = on
#log_checkpoints = off
#log_connections = off
#log_disconnections = off
#log_duration = off
#log_error_verbosity = default # terse, default, or verbose messages
#log_hostname = off
#log_line_prefix = '%m [%p]' # special values:
#  %a = application name
#  %u = user name
#  %d = database name
#  %r = remote host and port
#  %h = remote host
#  %p = process ID
#  %t = timestamp without milliseconds
#  %m = timestamp with milliseconds
#  %n = timestamp with milliseconds (as a Unix epoch)
#  %i = command tag
#  %e = SQL state
```

```
#      %c = session ID
#      %l = session line number
#      %s = session start timestamp
#      %v = virtual transaction ID
#      %x = transaction ID (0 if none)
#      %q = stop here in non-session
#          processes
#      %% = '%'
# e.g. '<%u%%%d> '
#log_lock_waits = off # log lock waits >= deadlock_timeout
#log_statement = 'none'# none, ddl, mod, all
#log_replication_commands = off
#log_temp_files = -1 # log temporary files equal or larger
# than the specified size in kilobytes;
# -1 disables, 0 logs all temp files
log_timezone = 'UTC'

# - Process Title -

#cluster_name = ''# added to process titles if nonempty
# (change requires restart)
#update_process_title = on

#-----
# RUNTIME STATISTICS
#-----

# - Query/Index Statistics Collector -

#track_activities = on
#track_counts = on
#track_io_timing = off
#track_functions = none # none, pl, all
#track_activity_query_size = 1024 # (change requires restart)
#update_process_title = on
#stats_temp_directory = 'pg_stat_tmp'
```

```
# - Statistics Monitoring -  
  
#log_parser_stats = off  
#log_planner_stats = off  
#log_executor_stats = off  
#log_statement_stats = off  
  
#-----  
# AUTOVACUUM PARAMETERS  
#-----  
  
#autovacuum = on # Enable autovacuum subprocess? 'on'  
# requires track_counts to also be on.  
#log_autovacuum_min_duration = -1 # -1 disables, 0 logs all actions and  
# their durations, > 0 logs only  
# actions running at least this number  
# of milliseconds.  
#autovacuum_max_workers = 3 # max number of autovacuum subprocesses  
# (change requires restart)  
#autovacuum_naptime = 1min # time between autovacuum runs  
  
# - Statement Behavior -  
  
#search_path = '"$user", public'# schema names  
#default_tablespace = ''# a tablespace name, '' uses the default  
#temp_tablespaces = ''# a list of tablespace names, '' uses  
# only default tablespace  
#check_function_bodies = on  
#default_transaction_isolation = 'read committed'  
#default_transaction_read_only = off  
#default_transaction_deferrable = off  
#session_replication_role = 'origin'  
#statement_timeout = 0 # in milliseconds, 0 is disabled  
#lock_timeout = 0 # in milliseconds, 0 is disabled  
#idle_in_transaction_session_timeout = 0 # in milliseconds, 0 is disabled  
#vacuum_freeze_min_age = 50000000  
#vacuum_freeze_table_age = 150000000  
#vacuum_multixact_freeze_min_age = 5000000  
#vacuum_multixact_freeze_table_age = 150000000
```

```
#bytea_output = 'hex'# hex, escape
#xmlbinary = 'base64'
#xmloption = 'content'
#gin_fuzzy_search_limit = 0
#gin_pending_list_limit = 4MB

# - Locale and Formatting -

datestyle = 'iso, mdy'
#intervalstyle = 'postgres'
timezone = 'UTC'
#timezone_abbreviations = 'Default'      # Select the set of available time zone
# abbreviations. Currently, there are
#   Default
#   Australia (historical usage)
#   India
# You can create your own file in
# share/timezonesets/.
#extra_float_digits = 0 # min -15, max 3
#client_encoding = sql_ascii # actually, defaults to database
# encoding

# These settings are initialized by initdb, but they can be changed.
lc_messages = 'C'# locale for system error message
# strings
lc_monetary = 'C'# locale for monetary formatting
lc_numeric = 'C'# locale for number formatting
lc_time = 'C'# locale for time formatting

# default configuration for text search
default_text_search_config = 'pg_catalog.english'

# - Other Defaults -

#dynamic_library_path = '/cyone1/postgres10/usr/lib/postgresql/'
#local_preload_libraries =
#session_preload_libraries = ''

#-----
```

```
# LOCK MANAGEMENT
#-----

#deadlock_timeout = 1s
#max_locks_per_transaction = 64 # min 10
# (change requires restart)
#max_pred_locks_per_transaction = 64 # min 10
# (change requires restart)
#max_pred_locks_per_relation = -2 # negative values mean
# (max_pred_locks_per_transaction
# / -max_pred_locks_per_relation) - 1
#max_pred_locks_per_page = 2           # min 0

#-----
# VERSION/PLATFORM COMPATIBILITY
#-----


# - Previous PostgreSQL Versions -

#array_nulls = on
#backslash_quote = safe_encoding # on, off, or safe_encoding
#default_with_oids = off
#escape_string_warning = on
#lo_compat_privileges = off
#operator_precedence_warning = off
#quote_all_identifiers = off
#sql_inheritance = on
#standard_conforming_strings = on
#synchronize_seqscans = on

# - Other Platforms and Clients -

#transform_null_equals = off

#-----
# ERROR HANDLING
#-----
```

```

#exit_on_error = off # terminate session on any error?
#restart_after_crash = on # reinitialize after backend crash?

#-----
# CONFIG FILE INCLUDES
#-----

# These options allow settings to be loaded from files other than the
# default postgresql.conf.

#include_dir = 'conf.d'# include files ending in '.conf' from
# directory 'conf.d'
#include_if_exists = 'exists.conf'# include file only if it exists
#include = 'special.conf'# include file

#-----
# CUSTOMIZED OPTIONS
#-----


# Add settings for extensions here

pglogical.conflict_resolution = 'last_update_wins'

```

1.2.2 Pglogical

Pglogical è un sistema logico di replica implementato come estensione di PostgreSQL. Completamente integrato, non richiede alcun triggers o programmi esterni. Questa alternativa alla replica fisica è un metodo altamente efficiente per replicare i dati utilizzando un modello di *publish/subscribe* per la replica selettiva.[14]

Vantaggi

I vantaggi offerti da Pglogical sono i seguenti:

- Replica sincrona

- Replica ritardata
- Risoluzione dei conflitti configurabili
- Capacità di convertire lo standby fisico in una replica logica
- Può pubblicare i dati da PostgreSQL a un abbonato Postgres-XL
- Le sequenze possono essere replicate
- Nessun trigger significa ridurre il carico di scrittura sul Provider
- Nessuna re-esecuzione di SQL significa overhead e latenza ridotti per il Sottoscrittore
- Il sottoscrittore non è in ripristino di riposo caldo, in modo da poter utilizzare tavoli temp, non sbloccati o normali
- Non è necessario annullare le query per consentire alla replica di continuare la riproduzione
- Il sottoscrittore (*subscriber*) può avere diversi utenti e protezione, indici diversi, impostazioni di parametri diversi
- Replica solo un database o un sottoinsieme di tabelle, noto come set di replica (*Replication Sets*)
- Replicare in versioni o architetture di PostgreSQL, consentendo aggiornamenti a bassa o zero-downtime
- Più server a monte in un singolo subscriber per l'accumulo di cambiamenti.[14]

Casi di uso

I diagrammi che seguono descrivono i gestori di database delle funzioni che sono in grado di eseguire con pglogical:

Come funziona pglogical?

Pglogical utilizza le funzioni di Decodifica Logica aggiunte da 2ndQuadrant (e disponibili da PostgreSQL 9.4). Pglogical funziona ancora più veloce con PostgreSQL 9.5 e successive, con bassi overhead su entrambi i provider e abbonati.

Pglogical si basa molto sulle caratteristiche introdotte nell'ambito dello sviluppo BDR, tra cui:

Migrare e aggiornare PostgreSQL con tempi di inattività quasi a zero

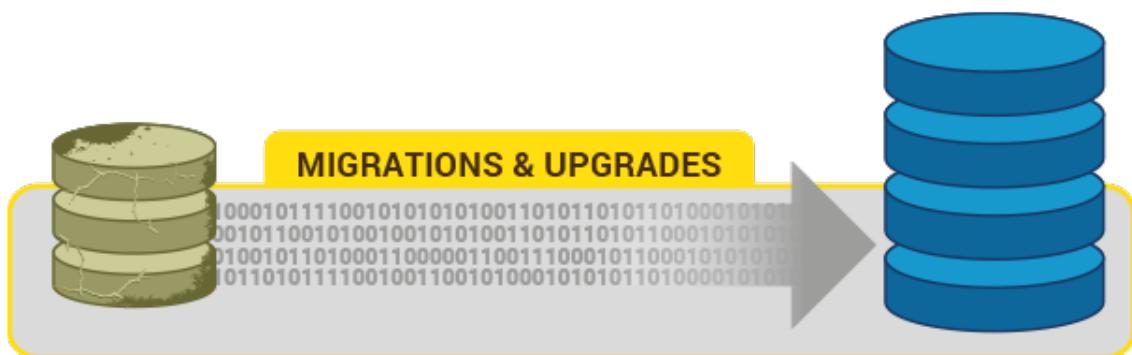


Figura 10: Migrazione e aggiornamenti PostgreSQL [14]
Accumulare le
modifiche provenienti da server di database scartati in un data warehouse

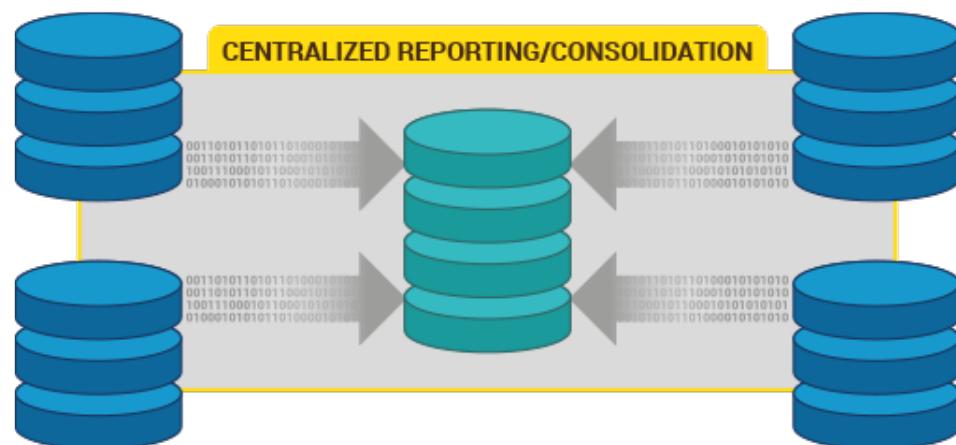


Figura 11: Aggregazione [14]

Copiare tutti o una selezione di tabelle di database ad altri nodi di un cluster

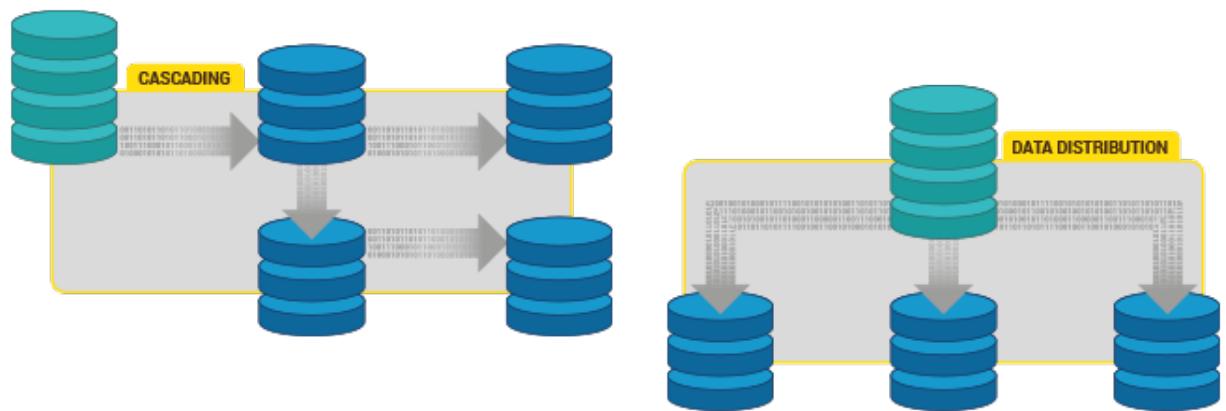


Figura 12: A cascata e distribuzione dati [14]
Le modifiche del database in tempo reale ad altri sistemi

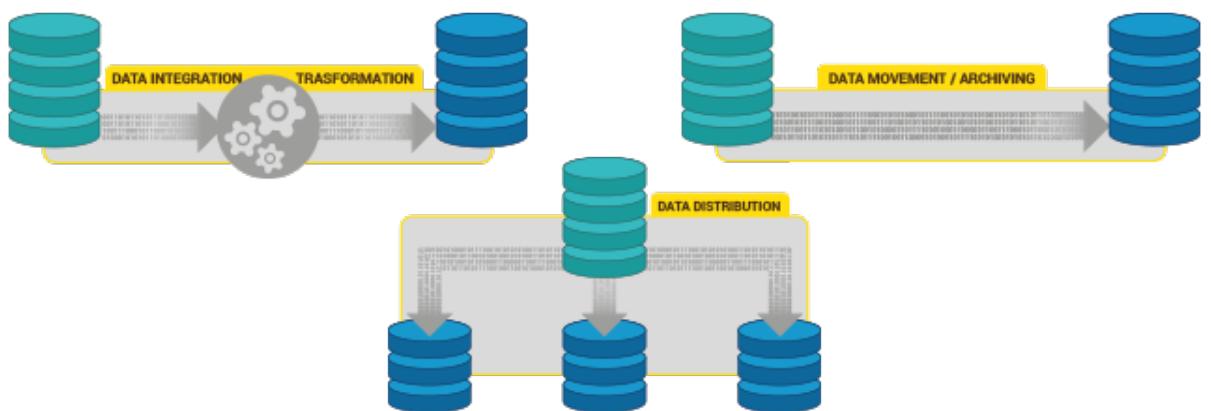


Figura 13: A cascata e distribuzione dati [14]

- Decodifica logica
- Slot di replica
- Lavoratori di sfondo statico
- Origini di replica
- Impegnano timestamp
- Messaggi WAL logici.[14]

Replica pglogical Bi-Directional Replication (BDR)?

No. pglogical non fornisce funzionalitÃ complete di replica multi-master e un supporto di modifica dello schema coerente, come fa la BDR. Pglogical incorpora le funzionalitÃ di BDR e le lezioni apprese da BDR per produrre una soluzione piÃ¹ semplice e piÃ¹ semplice a utilizzare per la replica unidirezionale, uti

Lo sviluppo di BDR continuerÃ per quelli che richiedono piena capacitÃ multi-master, riutilizzando gran parte del codice da pglogical.[14]

DA METTERE? DIFFERENZA TRA PG LOGICAL E BDR Replica pglogical Bi-Directional Replication (BDR)? No. pglogical non fornisce funzionalitÃ complete di replica multi-master e un supporto di modifica dello schema coerente, come fa la BDR. Pglogical incorpora le funzionalitÃ di BDR e le lezioni apprese da BDR per produrre una soluzione piÃ¹ semplice e piÃ¹ semplice a utilizzare per la replica unidirezionale, utilizzabile da piÃ¹ per

L'estensione pglogical fornisce la replica dello streaming logico per PostgreSQL, utilizzando il modello *Provider/Subscribe*. Si basa sulla tecnologia sviluppata come parte del Progetto BDR.

Utilizziamo i seguenti termini per descrivere i flussi di dati tra i nodi:

- **Nodi:** istanze del database PostgreSQL
- **Provider/Subscribers:** ruoli presi dai nodi
- **Set di replica:** una raccolta di tabelle che identificano i dati da replicare.

I casi d'uso supportati sono:

- Replica completa del database
- Replica selettiva di insiemi di tabelle mediante set di repliche

- Replica selettiva delle righe della tabella sul lato del publisher o del sottoscrittore (`row_filter`)
- Replica selettiva delle colonne della tabella sul lato dell'editore
- Raccolta dati / unione da più server upstream. [14]

1.3 HARDWARE UTILIZZATO PER GLI ESPERIMENTI

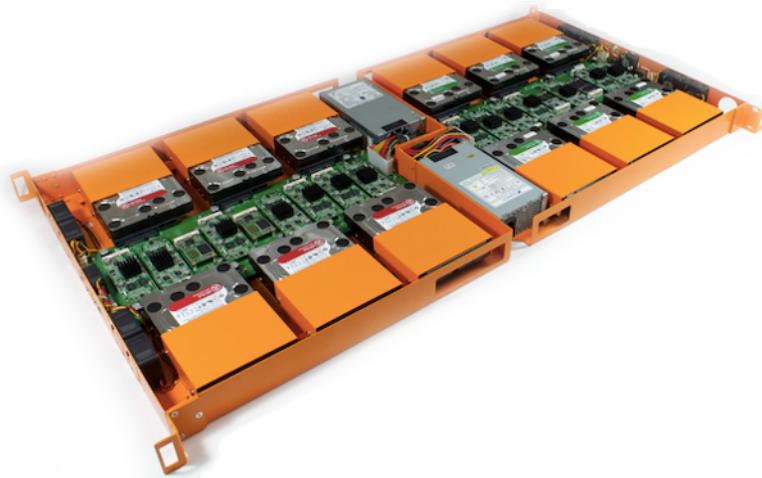


Figura 14: immagine1 [8]



Figura 15: immagine2 [8]



Figura 16: rack [8]



Figura 17: server [8]

2

DEFINIZIONE DEL PROGETTO

2.1.1 *Architettura del progetto*

Concetto Provider/Subscriber

Il modello di *provider/subscriber* definisce un flusso unidirezionale di informazioni da un oggetto *provider* a un numero qualsiasi di oggetti *subscribers*. Il concetto è che il *provider*, anche chiamato *publish*, ha informazioni o eventi utili che devono essere comunicati ad altri oggetti, ovvero tutti i suoi *subscriber*, che useranno tali informazioni per eseguire azioni aggiuntive o rimanere sincronizzati con il fornitore.

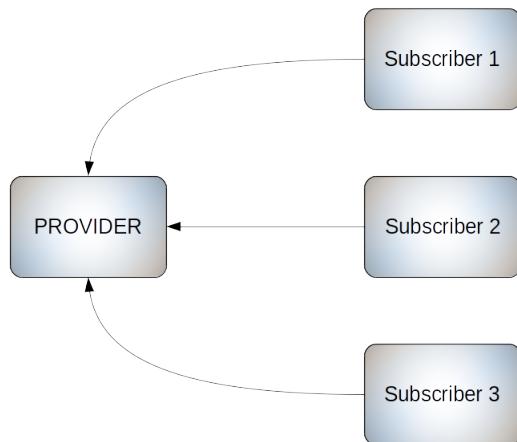


Figura 18: Semplice esempio di un Provider e tre Subscriber abbonati

Come anticipato nel capitolo precedente, la struttura base del sistema di replicazione Pglogical utilizza il modello *publish/subscriber* sopra descritto.

Al *provider* sono "abbonati" uno o più nodi *subscribers*. Ogni nodo che riceve i dati di replica da una fonte, quindi *provider*, può essere configurato per essere in grado di inoltrare tali dati agli altri nodi sottoscritti a sè.

Utilizzando la replica in cascata, ogni nodo *subscriber* è in contemporanea mittente e destinatario. Più nello specifico, ogni sottoscrittore è anche *provider* di altri *subscribers*.

Ci sono tre idee distinte dietro questa capacità:

1. La scalabilità: un database, in particolare il *publish* che riceve tutte le transazioni di aggiornamento dalle applicazioni client, ha solo una capacità limitata di soddisfare le query dei nodi sottoscritti durante il processo di replica.
2. Limitare la larghezza di banda network richiesta per un sito di backup mantenendo la possibilità di avere più slave nella posizione remota.
3. Essere in grado di configurare scenari di *failover*: in una configurazione da master a slave multipli, è improbabile che tutti i nodi slave siano esattamente nello stesso stato di sincronizzazione quando il master fallisce. Per garantire che uno slave possa essere promosso al master è necessario che tutti i sistemi rimanenti possano concordare lo stato dei dati. Poiché non è possibile eseguire il *rollback* di una transazione confermata, questo stato è indubbiamente lo stato di sincronizzazione più recente di tutti i nodi slave rimanenti.

Aggiunto alle funzionalità di PostgreSQL, che ci permette di replicare un intero database, il sistema di replicazione Pglogical può essere configurato per replicare in modo selettivo le righe di una tabella su entrambi i lati *publisher/subscriber*.

Le sequenze e le tabelle sono raggruppate in modo logico dentro un set di replica (*replication set*).

Ogni set corrisponde ai dati da replicare. È composto da una serie di flussi di dati di replica, che definiremo con R_n (dove n rappresenta la n -esimo set di replica), contenente un gruppo di oggetti da replicare indipendenti da altri oggetti provenienti dallo stesso master. In ogni caso,

tutte le tabelle che hanno relazioni che potrebbero essere espresse come vincoli di chiavi esterne e tutte le sequenze utilizzate per generare numeri di serie in queste tabelle dovrebbero essere contenute in uno stesso set.

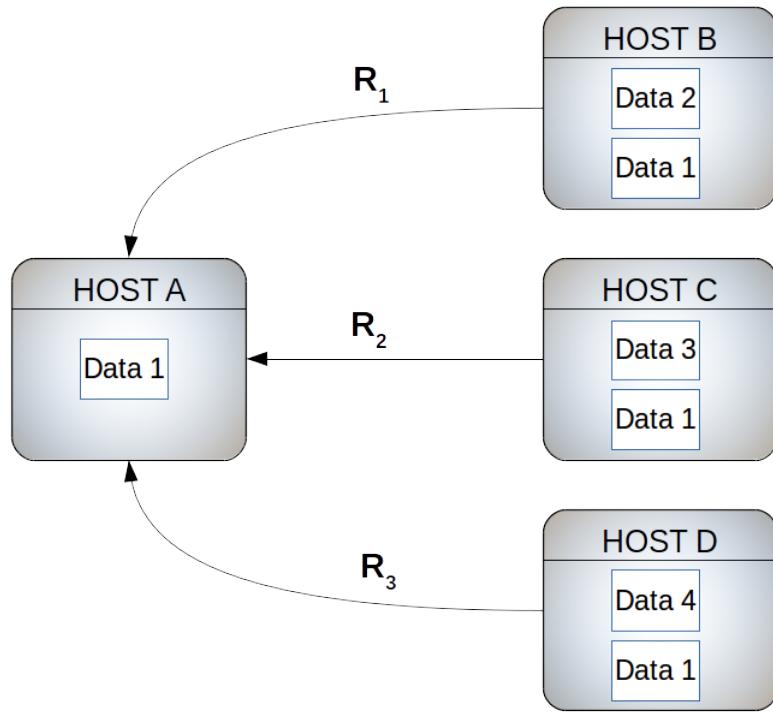


Figura 19: Set di replica o (*replication set*)

La figura illustra un semplice esempio di una configurazione di replica. Il set di replica è composto dal flusso di dati R_1 , R_2 e R_3 . Questo scenario raffigura quattro host (nel seguente caso quattro nodi) in cui il NODO A ha il ruolo di *publish* e i restanti sono i suoi sottoscrittori. Data 1 rappresenta il flusso di dati nativi di A da replicare. Ciascun *subscriber* ha esattamente i dati originali del NODO A, in quanto abbonati, in aggiunta ai propri nativi (Data 2, Data 3 e Data 4).

In questo modo se il *provider* fallisce (o per altri motivi non vi è più la possibilità di scriverci nuovi dati), l'HOST B, in quanto suo successivo, può essere promosso come *master* di entrambi i set.

Simulazione di un filesystem distribuito (Dati e Metadati)

inserisci organizzazione mappa - per far vedere come funziona il concetto di pub/sub, range, db Ciascun File viene inserito dentro un *bucket*,

che è un insieme di record all'interno di un database, ovvero un oggetto, contenente un insieme di campi o elementi, ciascuno dei quali è identificato da un nome univoco e da un tipo di dato.

Di ciascun File vengono scritte due informazioni:

1. file intero diviso in chunk,
2. una parte di metadato.

È necessario replicare i metadati sulle varie *Board* in modo tale che, in casi di fault, guasti o perdite, sia possibile ottenere nuovamente il dato originale.

Per i metadati la replica segue il concetto del modello *provider/subscriber*. Quando è scritto un File, per ogni chunk viene fatto un calcolo di un numero (**scia e qualcosa**) combinato con un numero pseudocasuale che restituisce un ID. Ogni Board gestisce un range di ID.

figura con le board e le replication sets

Sono replicate dalle 5 alle 8 Board (numero gestibile da un parametro configurabile).

Per quanto riguarda i dati è indispensabile scrivere i chunk. Il dato viene spartito in due connessioni e mandato in due *hosts* diversi. Più precisamente, è lanciata una *query* in parallelo che permette la scrittura in diversi database.

figura con due host e query in parallelo

CONSIDERAZIONI STATISTICHE SULLA RIDONDANZA SUL DATO

CONSIDERAZIONI STATISTICHE SULLA RIDONDANZA DEL METADATO

RESILIENZA AI CAMBIAMENTI DI RETE

3

DEFINIZIONE DEL QUADRO SPERIMENTALE

DA RIVEDERE - <http://www.slony.info/images/Slony-I-concept.pdf>

Failover: While it is relatively easy to tell in a master to multiple slave scenario which of the slaves is most recent at the time the master fails, it is nearly impossible to tell the actual row delta between two slaves. So in the case of a failing master, one slave can be promoted to the master, but all other slaves need to be re-synchronized with the new master. Performance: Storing the logging information in one or very few rotating log tables means that the replication engine can retrieve the actual data for one replication step with very few queries that select from one table only. In contrast to that a system that fetches the current values from the application tables at replication time needs to issue the same number of queries per replicated table and these queries will be joining the log table(s) with the application data table. It is obvious that this system's performance will be reverse proportional to the number of replicated tables. At some time the complete delta to be applied, which can not be split as pointed out already, will cause the PostgreSQL database system to require less optimal than in memory hash join query plans to deal with the number of rows returned by these queries and the replication system will be unable to ever catch up unless the workload on the master drops significantly.

LANCIO IN CONFIGURAZIONE 1

LANCIO IN CONFIGURAZIONE 2

LANCIO IN CONFIGURAZIONE 3

4

CONCLUSIONI E POSSIBILI EVOLUZIONI

UTILIZZO DI DISCHI SSD

UTILIZZO DI PROCESSORI DUAL CORE

5

ESERCIZI

1. Scrivere le possibili evoluzioni del programma

```
co X: = X+2 // X: = X+1 oc
```

assumendo che ciascun assegnamento è realizzato da tre azioni atomiche che caricano X in un registro (Load R X), incrementano il valore del registro (Add R v) e memorizzano il valore del registro (Store R X). Per ciascuna delle esecuzioni risultanti dall'interleaving delle azioni atomiche descrivere il contenuto dopo ogni passo della locazione condivisa X e dei registri privati, R₁ del processo che esegue il primo assegnamento ed R₂ per il processo che esegue il secondo assegnamento. Se assuma che il valore iniziale di X sia 50.

2. Si definisca il problema della *barrier synchronization* e si descrivano per sommi capi i differenti approcci alla sua soluzione. Se ne fornisca quindi una soluzione dettagliata utilizzando i semafori.
3. Considerare n api ed un orso che possono avere accesso ad una tazza di miele inizialmente vuota e con una capacità di k porzioni. L'orso dorme finchè la tazza è piena di k-porzioni, quindi mangia tutto il miele e si rimette a dormire. Le api riforniscono in continuazione la tazza con una porzione di miele finchè non si riempie; l'ape che aggiunge la k-esima porzione sveglia l'orso. Fornire una soluzione al problema modellando orso ed api come processi e utilizzando un monitor per gestire le loro operazioni sulla tazza. Prevedere che le api possano eseguire l'operazione *produce-honey* anche concorrentemente.
4. Descrivere le primitive di scambio messaggi send e receive sia sincrone che asincrone ed implementare

- `synch_send(v:int)`
- `send(v:int)`
- `receive(x:int)`

utilizzando le primitive di LINDA.

BIBLIOGRAFIA

- [1] Techopedia - *Definition - What does Clustering mean?* (Cited on page 7.)
- [2] Techopedia - *Techopedia explains Clustering* (Cited on pages 8 and 9.)
- [3] Wikipedia, the free encyclopedia - *Shared nothing architecture* (Cited on page 8.)
- [4] DA RIGUARDARE - https://en.wikibooks.org/wiki/Oracle_and_DB2,_Comparison_and_Comparison
- [5] Dave Wright - *The Advantages of a Shared Nothing Architecture for Truly Non-Disruptive Upgrades* solidfire.com. 2014-09-17. Retrieved 2015-04-21 (Cited on page 8.)
- [6] SearchNetworking - *peer-to-peer (P2P)* (Cited on page 10.)
- [7] Lifeware - *Introduction to Peer-to-Peer Networks* (Cited on pages 10 and 11.)
- [8] SearchStorage - *RAID (redundant array of independent disks)* (Cited on pages 3, 11, 13, 14, 15, 16, 17, 45, 46, 47, and 48.)
- [9] SearchStorage - *RAID controller* (Cited on pages 12, 20, and 21.)
- [10] Derek Vadala - *Managing RAID on Linux*, O'Reilly, 2002 (Cited on pages 11 and 15.)
- [11] SearchStorage - *erasure coding* (Cited on pages 17 and 18.)
- [12] PosgreSQL - *Documentation* (Cited on pages 19, 20, 21, 22, and 24.)
- [13] 2ndQuadrant Ltd - *PostgreSQL* (Cited on pages 12, 20, and 21.)
- [14] 2ndQuadrant Ltd - *pglogical* (Cited on pages 3, 40, 41, 42, 43, 44, and 45.)
- [15] Autore - *Titolo - altre informazioni*