

# Architetture Dati

UniShare

Davide Cozzi  
@dlcgold

# Indice

<b>1</b>	<b>Introduzione</b>	<b>2</b>
<b>2</b>	<b>Sistemi centralizzati</b>	<b>3</b>
2.1	Ottimizzazione delle query . . . . .	6
2.2	Transazioni . . . . .	7
2.3	Gestore della concorrenza . . . . .	9
<b>3</b>	<b>Sistemi distribuiti relazionali</b>	<b>11</b>
3.1	DDBMS . . . . .	14
3.1.1	Caratteristiche dei DDBMS . . . . .	16
3.1.2	Frammentazione e replicazione . . . . .	18
3.2	Query distribuite . . . . .	21
3.2.1	Accesso in lettura . . . . .	21
3.2.2	Accesso in scrittura e controllo di concorrenza . . . . .	27
3.2.3	2 phase locking . . . . .	28
3.2.4	Gestione dei deadlock . . . . .	29
3.2.5	Recovery management . . . . .	31

# Capitolo 1

## Introduzione

Questi appunti sono presi a lezione. Per quanto sia stata fatta una revisione è altamente probabile (praticamente certo) che possano contenere errori, sia di stampa che di vero e proprio contenuto. Per eventuali proposte di correzione effettuare una pull request. Link: <https://github.com/dlccgold/Appunti>.

# Capitolo 2

## Sistemi centralizzati

**Definizione 1.** Un **DBMS (DataBase Management System)** è un sistema, ovvero un software, in grado di gestire collezioni di dati che siano:

- *grandi, ovvero di dimensioni maggiori della memoria centrale dei sistemi di calcolo usati (se ho a che fare con una quantità di dati non così grande e con un uso personale posso affidarmi ad una hashmap piuttosto che ad un db)*
- *persistenti, ovvero con un periodo di vita indipendente dalle singole esecuzioni dei programmi che le utilizzano e per molto tempo*
- *condivise, ovvero usate da diversi applicativi e diversi utenti (fattore che porta anche allo studio del carico di lavoro, workload). L'accesso può essere sia in scrittura che in lettura (ovviamente anche entrambi) a seconda del caso. SI pongono quindi problemi di concorrenza e sicurezza*
- *affidabili, sia resistente dal punto di vista hardware (un guasto non deve farmi perdere i dati) che dal punto di vista della sicurezza informatica. Le transazioni devono essere quindi **atomiche** (o tutto o niente) e **definitive** (che non verranno più dimenticate). Il software può cambiare mentre i dati no*

A livello di architettura per un sistema centralizzati si hanno:

- uno o più *storage* per memorizzare i dati, a loro volta su uno o più file del *file system*
- il **DBMS**, il componente software che funge da componente logico

- diverse applicazioni che elaborano i dati provenienti dal db (*lettura*) ed eventualmente scrivono dati sullo stesso (*scrittura*)
- il **DBA** (*DataBase Administrator*) che tramite riga di comando o GUI si occupa di manutenzione, sicurezza, ottimizzazione etc. . . del DBMS

L'*architettura dati* di un DBMS è definita dall'ente *ANSI/SPARC* e è a tre livelli:

1. diversi **schemi esterni**, porzioni di db messi a disposizione per le varie applicazioni
2. uno **schema logico (o concettuale)**, che fa riferimento al *modello relazionale* dei dati ed è indipendente dalla tecnologia usata. Avendo un unico schema logico si ha un'unica semantica (perlomeno a livello astratto). Si ha unica base di dati, quindi un unico insieme di record interrogati e aggiornati da tutti gli utenti. Non si ha nessuna forma di eterogeneità concettuale
3. uno **schema fisico**, che fa riferimento alla tecnologia usata per implementare le tabelle per salvare i dati. Si ha un'unica rappresentazione fisica dei dati e quindi nessuna distribuzione e nessuna eterogeneità fisica

*Un unico schema fisico è collegato ad un unico schema logico.*

Inoltre si hanno:

- un **unico linguaggio di interrogazione** e quindi un'unica modalità di accesso ai dati
- un unico sistema di gestione per accesso, aggiornamento e gestione per la transazioni e le interrogazioni
- un'unica modalità di ripristino in caso d'emergenza
- un unico amministratore dei dati
- **nessuna autonomia gestionale**

Per il discorso della persistenza dei dati si ha necessità di una memoria secondaria dove il DBMS salva le strutture dati, studiando un modo efficiente di trasferimento dei dati nel *buffer* in memoria centrale. Il *buffer* è un'area di memoria (o meglio un componente software) nella memoria centrale che

cerca, tramite una logica di “vicinanza”, di mettere i dati della memoria secondaria in quella centrale. Si usa il **principio di località**.

A causa degli accessi condivisi al db si hanno problemi di **concorrenza**, avendo accesso multi-utente alla stessa dei dati condivisa, accesso che necessita anche di meccanismi di **autorizzazione**. In merito alla concorrenza si ha che le transazioni sono corrette se **seriali** (ordinate temporalmente) ma questo non è sempre applicabile e quindi si deve stabilire un *controllo della concorrenza*.

Per accedere ai dati di un db si hanno le **query (interrogazioni)** che fanno parte del modello logico a cui si interfaccia l'utente. Essendo i dati nelle memorie secondarie bisogna cercare un modo di rendere gli accessi performanti, in primis tramite opportune strutture fisiche in quanto e strutture logiche non sarebbero efficienti in memoria secondaria. Bisogna fare in modo che gli accessi alla memoria secondaria siano il più limitati possibili e quindi bisogna ottimizzare l'esecuzione delle query. Ovviamente una scansione lineare delle tabelle sarebbe troppo dispendiosa con tabelle grosse, ricordando che i file sono ad accesso sequenziale. Inoltre un ipotetico *join* tra tabelle renderebbe ancora più complesso l'accesso, soprattutto se *full-join*.

Per poter garantire tutto ciò che è stato detto l'architettura del DBMS deve essere organizzata in termini di *funzionalità cooperanti*:

- un **query compiler** che prende una query in SQL e la traduce con un compilatore
- un **gestore di interrogazioni e aggiornamenti** che trasforma le query in SQL in algebra relazionale facendo operazioni di ottimizzazione
- un **gestore dei metodi di accesso** per permettere il passaggio tra file e tabelle passando dal **gestore del buffer** e il **gestore della memoria secondaria** dove i dati non sono in forma tabellare ma di file e pagine
- un **DDL compiler**, dove DDL sta per Data Description Language, che si occupa dei comandi del DBA
- un **gestore della concorrenza**, che garantisce il controllo della concorrenza
- un **gestore dell'affidabilità**, che garantisce che un dato non vada perso
- un **gestore delle transazioni**

Gli ultimi quattro entrano in uso specialmente in fase di scrittura.

**Tutto deve essere veloce!**

In un sistema distribuito la parte di query compiler, gestore delle interrogazioni, gestore delle transazioni e gestore della concorrenza resta invariato mentre il resto cambia drasticamente (in quanto i dati sono distribuiti) dovendo gestire diversamente l'accesso ai dati e la sua sicurezza. Bisogna gestire anche come i vari nodi devono interagire coi dati.

## 2.1 Ottimizzazione delle query

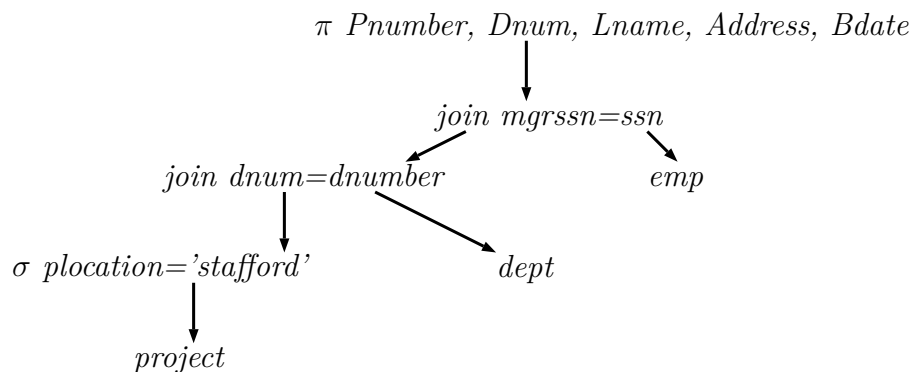
Ottimizzare le query è tutt'altro che banale.

Il primo step è il **parsing**, che stabilisce se la query è sensata dal punto di vista sintattico e se i vari nomi di tabelle e attributi sono coerenti con lo schema. Per questo ultimo aspetta ci si appoggia al **Data Catalog**, un particolare db che contiene informazioni sui vari database, in primis sui vari nomi delle tabelle e per ciascuna sui nomi di ogni attributo. SI ha quindi una soluzione per gestire i *metadati*. Il parser effettua un'analisi lessicale, per la sintattica e la semantica, usando il dizionario e la traduzione in algebra relazionale, producendo un **query tree**. Si calcola anche un **query plan logico**, utilizzando regole sintattiche di buon senso, per capire cosa fare prima (per esempio se fare prima una *select* o un *join*) per ottenere il risultato corretto nel minor tempo possibile (prima di fare una *join* magari seleziono prima una sottotabella con i dati potenzialmente utili, togliendo quelli sicuramente inutili... magari quel *join* può anche essere evitato). La query viene quindi rappresentata come un albero dove le foglie corrispondono alle strutture dati logiche, ovvero le tabelle. I nodi interni sono invece le varie operazioni algebriche (*select*, *join*, *proiezione*, *prodotto cartesiano* e *operazioni insiemistiche*).

**Esempio 1.** Vediamo una query:

```
SELECT Pnumber, Dnum, Lname, Address, Bdate
FROM Project P, Dept D, Emp E
Where P.dnum=D.dnumber and E.ssn = D.mgrssn
and P.location = 'Stafford'
```

che produce:



ma l'ottimizzatore va oltre (magari invertendo i where etc...) con un query plan più efficiente che permette di cambiare automaticamente le query in altre più efficienti.

Si ha un db chiamato **Statistics** che contiene statistiche sulla storia delle query nonché altre informazioni sui dati. L'uso di tale db permette di ottimizzare le query.

Solo dopo questo processo si ha la trasformazione delle tabelle logiche in strutture fisiche e metodi di accesso alla memoria e la trasformazione delle operazioni algebriche nelle loro implementazioni sulle strutture fisiche. Per la trasformazione si usano proprietà algebriche e una stima dei costi delle operazioni fondamentali per diversi metodi di accesso (in poche parole le regole della ricerca operativa). L'ottimizzazione ha complessità **esponenziale** e quindi si introducono approssimazioni basate su euristiche, usando un'alberatura di costi usando la tecnica del **Branch&Bound**.

## 2.2 Transazioni

Una **transazione** è l'insieme di istruzioni di accesso in lettura e scrittura ai dati, istruzioni eventualmente inserite in un linguaggio di programmazione. Una transazione gode di proprietà che garantiscono la corretta esecuzione anche in ambito di concorrenza e sicurezza, tanto che sono paradigmatiche



del modello relazionale. Le transazioni iniziano con un **begin-transaction** (a volte finiscono con *end-transaction*, opzionale) e all'interno deve essere eseguito tra:

- **commit work**, per terminare correttamente la lettura e/o scrittura
- **rollback work**, per abortire la transazione

Un sistema transazionale OLTP (*OnLine Transaction Processing*) è in grado di definire ed eseguire transazioni per conto di un certo numero di applicazioni concorrenti anche alto.

**Esempio 2.** Vediamo un esempio di transazione (esempio di addebito su un conto corrente e accredito su un altro):

```
start transaction;
update ContoCorrente
  set Saldo = Saldo + 10 where
  NumConto = 12202;
update ContoCorrente
  set Saldo = Saldo - 10 where
  NumConto = 42177;
commit work;
```

oppure, con anche la verifica che ci siano ancora soldi dopo il prelievo (con eventuale aborto):

```
start transaction;
update ContoCorrente
  set Saldo = Saldo + 10 where
  NumConto = 12202;
update ContoCorrente
  set Saldo = Saldo - 10 where
  NumConto = 42177;
select Saldo into A
  from ContoCorrente
  where NumConto = 42177;
if (A >= 0) then commit work
else rollback work;
```

Il controllo può essere fatto a posteriori grazie al rollback che permette di “dimenticare” tutte le operazioni precedenti

Le istruzioni commit work e rollback work possono comparire più volte all'interno del programma ma esattamente una delle due deve essere eseguita. Si ha un **approccio binario**.

Bisogna approfondire quindi le **unità di elaborazione** che hanno le proprietà cosiddette *ACID*:

- **Atomicità**, ovvero una transazione è un'unità atomica di elaborazione. Non si può lasciare il db in uno "stato intermedio". Un problema prima del commit cancella tutte le operazioni svolte (*UNDO*) e un problema dopo il commit non deve avere conseguenze, se necessario vanno ripetute le operazioni (*REDO*)
- **Consistenza**, ovvero la transazione rispetta i vincoli di integrità (se lo stato iniziale è corretto lo è anche quello finale). Quindi se ci sono violazioni non devono restare alla fine (nel caso *rollback*)
- **Isolamento**, ovvero la transazione non risente delle altre transazioni concorrenti. Una transazione non espone i suoi stati intermedi evitando l'*effetto domino* (si evita che il rollback di una transazione vada in cascata con le altre). L'esecuzione concorrente di una collezione di transazioni deve produrre un risultato che si potrebbe ottenere con una esecuzione sequenziale
- **Durabilità** (ovvero persistenza), ovvero gli effetti di una transazione andata in commit non vanno persi anche in presenza di guasti (a tal fine si sfrutta il **recovery manager**, che garantisce l'affidabilità, del DBMS)

## 2.3 Gestore della concorrenza

Il **gestore della concorrenza** permette di eseguire in parallelo più operazioni.

Definiamo **schedule** come una sequenza di esecuzione di un insieme di transazioni. Uno schedule è **seriale** se una transazione termina prima che la successiva inizi, altrimenti è **non seriale**. Qualora non sia seriale si potrebbero avere problemi.

Si sfrutta quindi la **proprietà di isolamento** facendo in modo che ogni transazione esegua come se non ci fosse concorrenza: *un insieme di transazioni eseguite concorrentemente produce lo stesso risultato che produrrebbe una (qualsiasi) delle possibili esecuzioni sequenziali delle stesse transazioni allora si ha la proprietà di isolamento*.

Si ha quindi che uno schedule è serializzabile se l'esito della sua esecuzione è lo stesso che si avrebbe con una qualsiasi sequenza seriale delle transazioni contenute.

Si hanno quindi diversi algoritmi per il controllo della concorrenza secondo varie tipologie:

- controllo basato su *conflict equivalence*
- controllo di concorrenza basato su *locks* (*protocollo 2PL o two phase locking, shared locks e gestione dei deadlock*). Il protocollo 2PL è usato nei DBMS dove per costruzione si hanno schedule serializzabili usando i lock per bloccare l'accesso alla risorse da parte di una transazione fino a che una risorsa non sia rilasciata. Si hanno quindi i concetti di *lock* e *unlock* che garantiscono l'uso esclusivo di una risorsa e l'autorizzazione esclusiva dell'uso di una risorsa viene dato dal gestore delle transazioni. Si hanno delle **tabelle di lock**. Si ha che, in ogni transazione, tutte le richieste di *lock* precedono tutti gli *unlock* (che comunque devono essere fatti dopo l'operazione di *commit*)
- controllo di concorrenza basato su *timestamps*

## Capitolo 3

# Sistemi distribuiti relazionali

Abbiamo visto nei sistemi centralizzati come ci fosse una sola base dati. In un sistema distribuito abbiamo diversi **basi dati locali**, diverse applicazioni su ogni nodo di elaborazione (dove ogni nodo condivide varie informazioni) con gli utenti che accedono alle varie applicazioni. Questo tipo di architettura prende il nome di **architettura shared nothing**, in quanto i DBMS di ogni singola macchina sono autonome (anche di vendor diversi) ma che lavorano insieme.

Un sistema distribuito permette non solo di avere dati “distribuiti” tra vari nodi ma anche di “duplicarne” alcuni per diversi scopi, coi nodi collegati in rete (addirittura si hanno soluzioni interamente distribuite nel cloud).

Confrontando un db distribuito con un multi-database (ovvero vari database completi da “unificare”) notiamo come entrambi abbiano un’alta distribuzione, il primo una bassa eterogeneità (a differenza del secondo, dove nei vari db potrei avere forte differenza di tipologia dei dati contenuti). Si ha anche bassa autonomia nel caso si db distribuiti a differenza del multi-database (dove ogni db è singolarmente autonomo).

Bisogna capire cosa distribuire. Si hanno diverse condizioni (che possono essere presenti simultaneamente):

- le applicazioni, fra loro cooperanti, risiedono su più nodi elaborativi (**elaborazione distribuita**)
- l’archivio informativo è distribuito su più nodi (**base di dati distribuita**)

La distribuzione si dice essere **ortogonale e trasparente** agli altri.

Capire cosa distribuire è una parte consistente dello studio di come costruire un’architettura distribuita (magari frutto di situazioni particolari come la “fusione” di due sistemi a causa di un’acquisizione aziendale etc... dove

diverse logiche applicative e diverse strutture dati possono creare situazioni molto pericolose).

Possiamo classificare i db distribuiti. Si ha innanzitutto che un **DBMS Distribuito Eterogeneo Autonomo** è in generale una federazione di DBMS che collaborano nel fornire servizi di accesso ai dati con livelli di *trasparenza* definiti (infatti le diversità tra db nei nodi vengono “nascosti” a vari *livelli di trasparenza* per distribuzione, eterogeneità e autonomia). Come abbiamo visto esiste l’esigenza di integrare a posteriori vari db preesistenti (anche a causa di integrazione di nuovi applicativi o nuove cooperazioni di processi) e questa situazione è spinta dallo sviluppo della rete.

Possiamo quindi dividere i livello di federazione su tre categorie tra loro ortogonali (ovvero indipendenti):

- autonomia
- distribuzione
- eterogeneità

### Autonomia

L’**autonomia** fa riferimento al grado di indipendenza tra i nodi e si hanno diverse forme:

- **autonomia di progetto**, il livello “massimo” dove ogni nodo ha un proprio modello dei dati e di gestione delle transazioni
- **autonomia di condivisione**, dove ogni nodo sceglie la porzione di dati da condividere ma condividendo con gli altri nodi lo schema comune
- **autonomia di esecuzione**, dove ogni nodo sceglie in che modo eseguire le transazioni

Si hanno quindi:

- **DBMS Strettamente integrati** con nessuna autonomia, con dati logicamente centralizzati, un unico data manager per le transazioni applicative e vari data manager locali che non operano in modo autonomo ma eseguono le direttive centrali
- **DBMS semi-autonomi**, dove ogni data manager è autonomo ma partecipa a transazioni globali, dove una parte dei dati è condivisa e dove sono richieste modifiche architetturali per poter fare parte della federazione

- **DBMS Peer to Peer** completamente autonomi, dove ogni DBMS lavora in completa autonomia ed è inconsapevole dell'esistenza degli altri

### Distribuzione

Per la **distribuzione** dei dati si hanno 3 livelli classici:

- **distribuzione client/server**, in cui la gestione dei dati è concentrata nei server, mentre i client forniscono l'ambiente applicativo e la presentazione
- **distribuzione Peer to Peer**, in cui non c'è distinzione tra client e server, e tutti i nodi del sistema hanno identiche funzionalità DBMS
- **nessuna distribuzione**

Le prime due possono anche non essere distinte.

### Eterogeneità

L'**eterogeneità** può invece riguardare vari aspetti:

- **modello dei dati** (relazionale, XML, object oriented (OO), json)
- **linguaggio di query** (diversi dialetti SQL, query by example, linguaggi di interrogazione OO o XML)
- **gestione delle transazione** (protocolli diversi per il gestore della concorrenza o per il recovery)
- **schema concettuale e logico** (concetti rappresentati in uno schema come attributo e in altri come entità)

Quindi si hanno vari tipi di DBMS:

- **DBMS distribuito omogeneo (DDBMS)** quando si ha alta distribuzione ma non si hanno autonomia ed eterogeneità (gestiti solitamente dallo stesso vendor)
- **DBMS eterogeneo logicamente integrato (data warehouse)** quando si ha alta eterogeneità ma non si hanno distribuzione e autonomia

- **DBMS distribuiti eterogenei** quando si ha alta eterogeneità e distribuzione ma non autonomia
- **DBMS federati distribuiti** quando si ha alta distribuzione, semi autonomia e non eterogeneità
- **DBMS distribuiti federati eterogenei** quando si ha alta distribuzione ed eterogeneità e semi autonomia
- **multi db MS**, totalmente autonomi ed eventualmente omogenei o eterogenei

*Si hanno molti altri sistemi in base alle 3 categorie.*

## 3.1 DDBMS

Parliamo di **DBMS distribuito omogeneo (DDBMS)**.

Studiamo uno schema in cui si passa da un sistema centralizzato ad un sistema distribuito.

Si hanno due architetture di riferimento:

- **l'architettura dati**
- **l'architettura funzionale**, ovvero l'insieme di tecnologie a supporto dell'architettura dati

Non avendo eterogeneità mantengo lo stesso schema di un DBMS centralizzato ma distribuisco dati bisogna prendere lo schema centralizzato e aggiungere componenti tra lo schema logico e lo schema fisico. Infatti non si avrà più un solo schema logico e un unico schema fisico ma tanti schemi logici e fisici locali (ad ogni logico corrisponde un fisico). I vari schemi logici inoltre si interfacciano con uno **schema logico globale**, i vari schemi logici locali non sono quindi altro che delle *viste* dello schema logico globale. Questa organizzazione tra schemi logici locali e schema logico globale è la cosiddetta **organizzazione LAV (Local As View)**. In ogni caso il progettista interroga lo schema logico globale e saranno varie tecnologie ad interrogare gli schemi logici locali (si fa una sorta di routing delle query).

Per ciascuna funzione (come query processing, transaction manager etc...) si possono avere vari tipi di gestione:

- centralizzata/gerarchica o distribuita

- con assegnazione statica o dinamica dei ruoli

*Lo schema globale viene progettato prima degli schemi locali.*

Ovviamente cambia il **processo di progettazione** nel caso dei DDBMS. Normalmente si ha un approccio *top-down* per la progettazione, con:

1. analisi dei requisiti
2. progettazione concettuale
3. progettazione logica
4. progettazione fisica

ma questo tipo di progettazione va cambiato e quindi si introduce una nuova fase e si cambiano le ultime due:

1. analisi dei requisiti
2. progettazione concettuale
3. **progettazione della distribuzione**, per capire dove mettere i dati
4. progettazione logica **locale**, che traduce dallo schema concettuale globale allo schema logico locale solo alcuni concetti
5. progettazione fisica **locale**

Si introduce il concetto di **portabilità**, ovvero la capacità di eseguire le stesse applicazioni DB su ambienti runtime diversi (anche con SQL diversi e differenti dallo standard). La portabilità è a *compile-time*

Si ha anche il concetto di **interoperabilità** (tra vendors diversi), ovvero la capacità di eseguire applicazioni che coinvolgono contemporaneamente sistemi diversi ed eterogenei (con zero autonomia). A tal fine sono stati introdotti dei *middleware*, tra cui **ODBC** che si occupa dell'accesso a dati di diversi vendor. ODBC, a livello architetturale, si pone sopra il DBMS e da un'immagine indipendente da ciò che c'è sotto (funziona come una sorta di *driver*), trasformando tutto in una sorta di SQL standard. Si hanno anche dei protocolli, come **X-Open Distributed Transaction Processing (DTP)**, che consentono di eseguire delle transazioni secondo una logica diversa. Questo protocollo stabilisce una serie di API che vengono implementate da ogni singolo DBMS per offrire una connettività standard (approccio molto usato per transazioni con vendor diversi). Il protocollo funziona sia se si ha che fare con omogeneità che con eterogeneità.

Si hanno altri approcci:



- **basi dati parallele**, con incremento delle prestazioni mediante parallelismo sia di storage devices che di processore (scalabilità orizzontale). Un esempio sono le **basi dati GRID**
- **basi dati replicate** dove si ha la replicazione della stessa informazione su diversi server per motivi di performance. Importanti per i temi della consistenza e della sicurezza
- **Data warehouses**, ovvero DBMS centralizzati, risultato dell'integrazione di fonti eterogenee, dedicati nel dettaglio alla gestione di dati per il supporto alle decisioni. Prevede la *cristallizzazione* dei dati, acquisiti da varie sorgenti, creando un nuovo schema con la memorizzazione dei dati in formato nuovo (solitamente relazionale). Non usa un approccio LAV

### 3.1.1 Caratteristiche dei DDBMS

Si hanno vari tipi di architetture DDBMS:

- **shared-everything**, ad esempio *SMP server*, dove il db management system e il disco sono in un unico nodo
- **shared-disk**, ad esempio *Oracle RAC*, dove diversi db management systems agiscono su una stessa **SAN (Storage Area Network)**, ovvero un'architettura dati di puro storage (con tanti dischi in raid). I vari db accedono ai dati secondo una certa regolazione. Viene distribuito il carico sui db ma si hanno problemi di concorrenza e hanno grandi problemi di scalabilità e costo economico
- **shared-nothing**, sempre più usati, dove ogni db management system ha il suo disco. È molto scalabile e, a patto di gestire la complessità, posso aggiungere nodi in modo illimitato (**scalabilità orizzontale**). Si presta molto all'ambiente cloud. Sono *architetture federate*.

Vediamo quindi le proprietà generali di un DDBMS (facendo esplicito riferimento alle architetture *shared-nothing* per la loro scalabilità):

- **località**, secondo il *principio di località*, che garantisce un aumento di performances (nonché di sicurezza) tenendo i dati si trovano “vicino” alle applicazioni che li utilizzano più frequentemente

- **modularità**, permettendo di scalare orizzontalmente e permettendo modifiche a dati ed applicazioni a basso costo
- **resistenza ai guasti**
- **prestazioni ed efficienza**

Concentrandoci sulla **località** si ha che la partizione dei dati corrisponde spesso ad una partizione naturale delle applicazioni e degli utenti. I dati risiedono più vicino a dove vengono più usati ma possono comunque essere raggiunti anche da lontano (*globalmente*). Si cerca inoltre sempre di più di spostare i dati verso le applicazioni (paradigma ribaltato nel caso di *big data*).

In merito alla **modularità** si nota come la distribuzione dinamica dei dati si adatta meglio alle esigenze delle applicazioni (magari spostando solo sottotabelle verso alcuni nodi etc. ..., sia in modo trasparente rispetto all'utente che altrimenti).

Parlando di **resistenza ai guasti** si ha una maggior fragilità a causa delle unità che aumentano di numero ma si ha **ridondanza** e quindi maggiore resistenza ai guasti di dati e applicazioni ridondate (*fail soft*).

Discorso più interessante è da farsi sulle **prestazioni**. Ogni nodo in un sistema shared-nothing gestisce db di dimensioni ridotte. Inoltre ogni nodo può essere ottimizzato ad hoc ed è più semplice gestire e ottimizzare applicazioni locali. Si ha inoltre distribuzione del carico totale e parallelismo tra transazioni locali che fanno parte di una stessa transazione distribuita (anche se questo aspetta obbliga soluzioni di coordinamento e appesantisce il carico sulla rete, che rischia di diventare un "collo di bottiglia").

I DDBMS hanno ovviamente **funzionalità** specifiche.

Ogni server ha buona capacità di gestire transazioni indipendentemente, anche se le interazioni distribuite tra server rappresentano un carico supplementare. Per le interrogazioni si ha che le query arrivano dalle applicazioni e i risultati dai server mentre per le transazioni le richieste transazionali arrivano dalle applicazioni ma sono richiesti **dati di controllo** per il coordinamento. La gestione della rete deve essere ottimizzata e serve uno studio sulla distribuzione locale dei dati.

Ricapitolando si hanno le seguenti funzionalità specifiche:

- **trasmissione** di query, transazioni, frammenti di db e dati di controllo tra i nodi
- **frammentazione, replicazione e trasparenza** (secondo vari livelli), fattori legati alla natura distribuita dei dati

- un **query processor** e un **query plan** per la previsione di una strategia globale accanto a strategie per le query locali. Si gestisce il passaggio tra schema logico globale e quelli locali. Chi esegue la query lo fa senza pensare alla frammentazione dei dati
- **controllo di concorrenza** tramite algoritmi distribuiti, fondamentale per gli accessi *in scrittura*
- **strategie di recovery e gestione dei guasti**, sia in merito alla rete che all'hardware stesso

### 3.1.2 Frammentazione e replicazione

Si definisce **frammentazione** come la possibilità di allocare porzioni (*chunk*) diverse del db su nodi diversi.

Si definisce **replicazione** come la possibilità di allocare stesse porzioni del db su nodi diversi.

Si definisce **trasparenza** come la possibilità per l'applicazione di accedere ai dati senza sapere dove sono allocati (serve qualcosa che instradi le query).

#### frammentazione

Esistono due tipi di frammentazione:

1. **frammentazione orizzontale**, che prevede di prendere una tabella e frammentare in base alle righe (le prime  $n$  da una parte, le seconde  $m$  dall'altra etc...). Si mantiene quindi inalterato lo schema in quanto ottengo solamente delle tabelle più piccole in quanto pezzi. Per spezzare uso una *select* (per la **selezione**) che selezioni ogni volta un certo "blocco" di tabella
2. **frammentazione verticale**, che consente di ridurre la dimensionalità della tabelle spezzandola in base alle colonne. In ogni nuova tabella però la prima colonna deve essere uguale alla prima della tabella originale (ovvero dove si ha la chiave primaria), questo per garantire che si possa ricomporre la tabella (e lo schema) originale (con operazioni di *join*, o meglio un *natural join*) e garantire la trasparenza. Anche in questo caso uso una *select* (per la **proiezione**) che selezioni ogni volta un certo numero di colonne da mettere nella nuova tabella

Bisogna quindi garantire:

- **completezza**, ovvero ogni record della relazione  $R$  di partenza deve poter essere ritrovato in almeno uno dei frammenti
- **ricostruibilità**, ovvero la relazione  $R$  di partenza deve poter essere ricostruita senza perdita di informazione a partire dai frammenti
- **disgiunzione**, ovvero ogni record della relazione  $R$  deve essere rappresentato in uno solo dei frammenti
- **replicazione**, l'opposto della disgiunzione

Quindi possiamo definire meglio le proprietà dei due tipi di frammentazione per la relazione  $R$ , frammentata in diversi  $R_i$ :

1. **orizzontale**:

- $schema(R_i) = schema(R), \forall i$
- ogni  $R_i$  contiene un sottoinsieme dei record di  $R$
- è definita da una proiezione su una condizione  $ci$ :  $\sigma_{ci}(R)$
- garantisce la completezza, infatti  $R_1 \cup R_2 \cup \dots R_n = R$
- l'unione garantisce la ricostruibilità

2. **verticale**:

- $schema(R) = L = (A_1, \dots, A_m)$  e  $schema(R_i) = L_i = (A_{i1}, \dots, A_{ik})$
- garantisce la completezza, infatti  $L_1 \cup L_2 \cup \dots L_n = L$ , dove i vari  $L_i$  sono i frammenti verticali ed  $L$  è la tabella originale
- si garantisce la ricostruibilità in quanto  $L_i \cap L_j \supseteq chiave\ primaria(R), \forall i \neq j$  (ovvero ogni frammento deve contenere la chiave primaria)

## Replicazione

Approfondiamo ora la **replicazione**. Si hanno diversi aspetti positivi per l'accesso *in lettura*, come il miglioramento delle prestazioni in quanto consente la coesistenza di applicazioni con requisiti operazionali diversi sugli stessi dati e aumenta la *località dei dati* usati da ogni applicazioni. Nel momento in cui si ha l'accesso *in scrittura* si hanno però diversi aspetti negativi. Si hanno diverse complicazioni architetturali, tra cui la gestione della transazioni e

l'updates di copie multiple, che devono essere tutte aggiornate. Inoltre bisogna studiare dal punto di vista progettuale cosa replicare, quanto replicare (ovvero capire quante copie mantenere), dove allocare le copie e le politiche per gestirle.

In merito all'allocazione studiamo anche gli **schemi di allocazione**. Ogni frammento può essere allocato su un nodo diverso. Lo schema globale quindi è solo *virtuale* (in quanto non materializzato in un solo nodo) e lo **schema di allocazione** definisce il *mapping* tra un frammento e un nodo. Si ha quindi una tabella, un **catalogo**, che ci da informazioni sul partizionamento, associando ogni frammento al nodo in cui è allocato.

### Trasparenza

Con la **trasparenza** si ha la separazione della semantica di alto livello dalle modalità di frammentazione e allocazione. Si separa quindi la *logica applicativa* dalla *logica dei dati* ma per farlo serve uno strato software che gestisca la traduzione dallo schema unico ai sottoschemi, comportando un aumento di complessità del sistema e una perdita di prestazioni (problemi che si riducono con un *mapping* integrato del DDBMS).

Le applicazioni (transazioni, interrogazioni) non devono essere modificate a seguito di cambiamenti nella definizione e organizzazione dei dati e si hanno due tipi di trasparenza, che si applicano agli schemi ANSI-SPARC nel modello distribuito (schema logico globale e schemi logici/fisici locali):

1. **trasparenza logica (o indipendenza logica)**, ovvero in dipendenza dell'applicazione da modifiche dello schema logico. Un'applicazione che usa un frammento non viene modificata se vengono modificati altri frammenti
2. **trasparenza fisica (o indipendenza fisica)**, ovvero in dipendenza dell'applicazione da modifiche dello schema fisico

Frammentazione e allocazione sono tra lo schema logico globale e ogni schema logico locale.

Si hanno quindi tre livelli di trasparenza:

- **trasparenza di frammentazione**, che permette di ignorare l'esistenza dei frammenti ed è lo scenario migliore per la programmazione applicativa con un'applicazione scritta in SQL standard. Il sistema si occupa di convertire query globali in locali e relazioni in sotto-relazioni. La scomposizione delle query per ogni sotto-relazione è detta **query rewriting**

- **trasparenza di replicazione/allocazione**, dove l'applicazione è consapevole dei frammenti ma non dei nodi in cui si trovano. In questo caso la query è già spezzata in quanto si sa di avere a che fare con un sistema frammentato
- **trasparenza di linguaggio**, dove l'applicazione specifica sia i frammenti che i nodi, nodi che possono offrire interfacce che non sono SQL standard. Tuttavia l'applicazione sarà scritta in SQL standard a prescindere dai linguaggi locali dei nodi. Le query vengono quindi tradotte ottimizzate di query. *Questo è il livello di trasparenza più basso*

## 3.2 Query distribuite

Analizzeremo prevalentemente DDBMS distribuiti *shared-nothing* e, in seguito, *architetture di replica* (con un *replication server* atto a gestire al replica). Le query sono ovviamente le operazioni più importanti. Possono essere di sola *lettura* (tramite operazioni come la *select*) o anche di *scrittura*. Le due tipologie di operazioni vengono gestite in modo molto differente (la lettura sincrona non è un problema, se non hardware risolvibile con una distribuzione del carico, a differenza della scrittura sincrona). Le operazioni devono essere eseguite **velocemente**.

### 3.2.1 Accesso in lettura

Studiamo prima le **query in scrittura**.

Esistono, in un sistema relazionale, una serie di attività che convertono la query in SQL in algebra relazionale e solo dopo si ha la distribuzione. L'utente, ignaro dello schema distribuito, interroga lo schema logico globale e il DDBMS decompone la query secondo una localizzazione specifica in base ai singoli frammenti (ovvero deve distribuire la query in modo sensato). Si ha anche un'ottimizzazione globale della query prima della distribuzione in modo che anche la distribuzione stessa sia ottimizzabile correttamente, infatti il gestore delle interrogazioni manda ai singoli nodi i giusti frammenti di query che verranno ottimizzati localmente. Ho quindi nel complesso 4 fasi che compongono il **query processor**:

1. **query decomposition**
2. **data localization**
3. **global query optimization**

#### 4. local optimization

##### Query decomposition

La *query decomposition* opera sullo schema logico globale non tenendo conto della distribuzione. In questo caso si hanno tecniche di ottimizzazione algebrica (usando quindi l'algebra relazionale indipendentemente dalla distribuzione) analoghe a quelle usati in sistemi centralizzati e si ha come output un **query tree** non ottimizzato rispetto ai **costi di comunicazione**. Il costo di comunicazione riguarda il costo di uso della **rete** e dipende da vari fattori. Il costo di comunicazione è il vero “collo di bottiglia” in sistemi distribuiti.

##### Data localization

La *data localization* considera la frammentazione delle tabelle e la distribuzione, capendo ad esempio dove effettuare le *select* etc. . . . Si procede quindi all'ottimizzazione delle operazioni rispetto alla frammentazione, tramite **tecniche di riduzione**. Viene quindi prodotta una query efficiente per la frammentazione ma non ottimizzata. Supponiamo per esempio di avere una tabella su 3 nodi (distribuita tramite frammentazione orizzontalmente) e che di base la query faccia la richiesta a tutti e tre (facendo l'unione dei risultati). Usando la tecnica di riduzione, qualora, per esempio, effettivamente sia necessaria solo in un nodo, si avrà che la query sarà distribuita unicamente nel nodo corretto.

##### Global query optimization

La *global query optimization* si basa sulle statistiche sui frammenti per effettuare l'ottimizzazione. Viene arricchito il **query tree**, creato con gli operatori dell'algebra relazionale, tramite gli **operatori di comunicazione** (ovvero *send* e *receive*), che vengono effettuati tra nodi. Alcune query, dopo aver tenuto conto dei tempi di comunicazione, potranno essere eseguite in parallelo (grazie all'indipendenza data dallo *shared-nothing*). In questo caso le decisioni più rilevanti riguardano le operazioni di *join* (che è uno degli operatori più complicati) e, in particolare (come vedremo più avanti), l'ordine tra i *join* n-ari e la scelta tra *join* e *semijoin* (un operatore particolare per i sistemi distribuiti). L'operatore di *join* infatti “ingrandisce” i dati “fondendo” tabelle, che magari sono frammentate in più tabelle su vari nodi. L'uso di *send* e *receive* permette la comunicazione dei dati tra i nodi, anche se questo rischia di diventare troppo esoso in termini di prestazioni. Si hanno quindi degli **algoritmi di calcolo del costo adattivi** e si deve studiare la rete e i suoi

ritardi, che dipendono dalla *topologia* della rete stessa e dal carico applicativo. Si ha quindi una fase di **ottimizzazione a runtime**, dove si riadatta il **query plan**. Per riadattarlo si fa in primis monitoring sull'esecuzione della query, si procede adattando il modello di costo (eventualmente con software automatici) e, eventualmente, riadattando la query se si calcola uno scarto di costo troppo elevato. È il DBA che stabilisce delle soglie temporali entro le quali ottenere una risposta. Per questo conta la trasparenza, in quanto non è l'applicazione che deve interessarsi di questo aspetto. Si introduce un nuovo *layer* di complessità.

In alcuni casi è impossibile ad avere un DDBMS che si occupi di questo tipo di ottimizzazioni.

**La distribuzione non è predicibile a priori.** Vediamo un semplice esempio:

**Esempio 3.** *Si supponga di avere il seguente schema:*

- *Employee (eno, ename, title), di cardinalità 400*
- *AssiGN(eno, projectno, resp, dur), di cardinalità 1000 e dove resp rappresenta il tipo di responsabilità*

*Si ha la seguente query: trovare i nomi dei dipendenti che sono anche manager di progetti:*

```
SELECT ename
FROM Employee E JOIN AssiGN A on E.eno=A.eno
WHERE resp='manager'
```

*Che, in algebra relazionale già abbastanza ottimizzata ipotizzando che ci siano pochi manager, sarebbe:*

$$\pi_{ename}(EMP \bowtie_{eno} (\sigma_{resp="manager"}(ASG)))$$

*Supponiamo di avere poi 5 nodi uguali, il quinto per il risultato e i primi 4 frammentati orizzontalmente secondo questo schema di divisione per nodo:*

1.  $ASG1 = \sigma_{eno \leq 'E3'}(ASG)$
2.  $ASG2 = \sigma_{eno > 'E3'}(ASG)$
3.  $EMP1 = \sigma_{eno \leq 'EMP'}(EMP)$
4.  $EMP2 = \sigma_{eno > 'EMP'}(EMP)$

*Vediamo quindi una prima esecuzione:*



- chiedo al nodo 1 i manager e sposto i risultati di  $\sigma_{resp="manager"}(ASG1)$  sul nodo 3 (dove sono descritti in modo più completo) come  $(ASG'1)$ . Il risultato di  $EMP1 ><_{eno} (ASG1)$ , calcolato sul nodo 4, lo porto sul nodo 5  $EMP'1$
- chiedo al nodo 2 i manager e sposto i risultati di  $\sigma_{resp="manager"}(ASG'2)$  sul nodo 4 (dove sono descritti in modo più completo) come  $(ASG'2)$ . Il risultato di  $EMP2 ><_{eno} (ASG'2)$ , calcolato sul nodo 4, lo porto sul nodo 5 come  $EMP'2$
- sul nodo 5 il risultato sarà  $EMP'1 \cup EMP'2$

Vediamo una seconda soluzione:

- contemporaneamente chiedo al nodo 1 e al nodo 3 di mandare al nodo 5 tutti i manager. Sempre contemporaneamente a queste due operazioni chiedo al nodo 2 e al nodo 4 di mandare al nodo 5 tutte le informazioni. I tempi saranno basati sul più lento dei quattro nodi, che determinerà il tempo massimo dell'operazione (che sono circa calcolabili a priori tramite la tabella delle statistiche)
- nel nodo 5 calcolo il risultato:

$$(EMP1 \cup EMP2) ><_{eno} \sigma_{resp="manager"}(ASG1 \cup ASG2)$$

I tempi di trasporto detteranno quale soluzione tra le due è la più performante ma, viste le cardinalità esigue di dati, probabilmente vince la seconda (dove si ha un solo spostamento globale). Questa seconda strategia costringe a pensare a particolari strutture di accesso secondarie dette **indici**, che permettono interrogazioni efficaci ma che **non possono essere “portati”** in sistemi distribuiti. Quindi nel nodo 5 non ho gli **indici** e quindi devo fare l'intero **prodotto cartesiano** per il join (che però in questo caso ha un tempo trascurabile, grazie alla bassa cardinalità dei dati, rispetto ai costi di trasferimento, generalmente non trascurabili rispetto ai costi delle operazioni interne ad un nodo).

Riprendendo l'esempio definiamo:

- **costo di messaggio** come il costo fisso di spedizione o ricezione di un messaggio (detto *setup*)
- **costo di trasmissione** come il costo, fisso rispetto alla topologia, di trasmissione dati

- **costo di comunicazione** come la somma tra il costo di messaggio, moltiplicato per il numero di messaggi, più il costo di trasmissione, moltiplicato per il numero di *bytes* trasmessi
- **costo totale** come la somma dei costi delle operazioni (*I/O* e *CPU*) più i costi di comunicazione (*comunicazione*)
- **response time** come la somma dei costi qualora si tenga conto del *parallelismo delle trasmissioni*, quindi come la somma tra il costo di messaggio, moltiplicato per il numero di messaggi comunicati in modo sequenziale, più il costo di trasmissione, moltiplicato per il numero di *bytes* trasmessi in modo sequenziale. In questo conto volendo posso usare dei **pesi** basati sulla cardinalità delle unità da trasferire e tenere conto del massimo tempo di risposta che si ottiene

Si ha quindi che:

- nelle **grandi reti geografiche** i costi di *comunicazione* sono molto maggiori del costo di *I/O*, circa di 10 volte
- nelle **reti locali** i costi di *comunicazione* e *I/O* sono paragonabili, grazie alle reti *gigabit* in locale

Tendenzialmente il costo di comunicazione è ancora il **fattore critico** ma sempre meno.

Bisogna scegliere cosa **minimizzare**:

- il *response time*, aumentando il parallelismo che però può portare ad un aumento del *costo totale*, con un maggior numero di trasmissione e un maggior processing locale. Nell'esempio 3 potrebbe sembrare la seconda soluzione, che effettivamente parallelizza di più ma non minimizza i costi di risposta
- il *costo totale*, senza tener conto del parallelismo utilizzando meglio le risorse e aumentando il *throughput* ma peggiorando così il *response time*. Nell'esempio 3 è la prima soluzione

### Join e Semijoin

Il **join** presenta il problema di portare alla perdita dell'**indice**. Bisogna quindi studiare come effettuare l'operazione tra due tabelle su due nodi diversi. Una prima operazione è data dall'operazione di *semijoin*.

**Definizione 2.** Definiamo, in algebra relazionale, l'operazione *semijoin*, tra due tabelle  $R$  e  $S$ , sull'attributo  $A$ , come:

$$R \text{ semijoin}_A S \equiv \pi_{R^*}(R \text{ join}_A S)$$

dove  $R^*$  è l'insieme degli attributi di  $R$ .

In altre parole scelgo esplicitamente di tenere solo gli attributi di  $R$  dopo il *semijoin*.

Quindi con  $R \text{ semijoin}_A S$  ho la proiezione sugli attributi di  $R$  operazione di *join* e quindi ho che il *semijoin* non è **commutativo**.

Dalla seconda tabella porto solo la serie di attributi che mi servono esplicitamente ( $\pi_A(S)$ ) riducendo il carico di lavoro.

**Alla fine il nostro  $R'$  con i risultati del *semijoin* sarà trasportato nel nodo di  $S$**

Prese due tabelle allocate su nodi differenti, il *join* tra di esse può quindi essere calcolato tramite operazioni di *semijoin*, valgono infatti le seguenti equivalenze (che portano a diverse strategie a seconda della stima dei costi):

- $R \text{ join}_\theta S \iff (R \text{ semijoin}_\theta S) \text{ join}_\theta S$
- $R \text{ join}_\theta S \iff R \text{ join}_\theta (S \text{ semijoin}_\theta R)$
- $R \text{ join}_\theta S \iff (R \text{ semijoin}_\theta S) \text{ join}_\theta (A \text{ semijoin}_\theta R)$

In tutti i casi si riduce lo spostamento dei dati.

L'uso del *semijoin* è conveniente sse il costo del suo calcolo e del trasferimento del risultato è inferiore al costo del trasferimento dell'intera relazione e del costo dell'intero *join* (e questo dipende dal numero di attributi coinvolti).

*Avere più di un join complica la situazione, anche solo per la scelta dell'ordine in cui eseguirli.*

### Local optimization

La *local optimization* si occupa dell'ottimizzazione degli schemi locali. Ogni nodo riceve una *fragment query* e la ottimizza, con tecniche analoghe ai sistemi centralizzati, in modo completamente indipendente. Si hanno comunque operazioni di ottimizzazione locale a priori sul fatto che il *global query optimization* punti a ridurre i costi di comunicazione (nel caso di un DDBMS in rete geografica) o ad aumentare il parallelismo (in caso di DDBMS in rete locale).

In ogni caso nella progettazione di sistemi di gestione dati distribuiti bisogna tener conto di:

- tipologie di query distribuite
- stime o statistiche sullo storico query distribuite già eseguite (fatto periodicamente dal DDBMS)
- topologia della rete
- carico aspettato e workload previsto

### 3.2.2 Accesso in scrittura e controllo di concorrenza

In questo caso la situazione si complica. Un conto è avere delle **remote requests (read-only)**, che possono essere un numero arbitrario di query SQL in sola lettura, un altro è avere delle **remote transactions (read-write)**, ovvero un numero arbitrario di operazioni SQL che prevedono anche *insert* e *update*. Ragionando in un'ottica in cui si ha un numero arbitrario di server si parla di **distributed requests**, dove ogni singola operazione SQL si può riferire, grazie ad un *ottimizzatore distribuito*, a qualunque insieme di server, e di **distributed transactions**, dove ogni operazione è diretta ad un unico server e dove le transazioni possono modificare più di un db, tutto ciò grazie ad un *protocollo transazionale di coordinamento distribuito*, detto **two-phase commit** (in questo caso si ha spesso a che fare con sistemi *eterogenei* e *federati*). Deve valere la **proprietà di atomicità** di *ACID*. Sempre riguardo *ACID* si ha che la **proprietà di consistenza** non dipende dalla distribuzione, in quanto le proprietà sono indipendenti dall'allocazione, e si ha che la **proprietà di durabilità** viene garantita localmente. Vanno invece rivisti architetturalmente la **proprietà di atomicità**, tramite componenti come il *reliability control* e il *recovery manager* (in caso di guasti), e la **proprietà di isolamento**, tramite il *concurrency control* (senza il quale si può incorrere in *update fantasma*, dove un *update* viene cancellato da uno seguente).

Rivedendo i **principi del controllo di concorrenza**.

**Definizione 3.** Data una transazione  $t_i$  si ha che essa viene scomposta in sotto-transazioni  $t_{ij}$  a seconda del nodo  $j$ -simo su cui viene eseguita. A sua volta una transazione  $t_{ij}$  viene nominata  $r_{ij}$  per le operazioni di lettura e  $w_{ij}$  per le operazioni di scrittura. L'uso di una risorsa  $x$  viene indicata, per esempio, con  $r_{ij}(x)$  o  $w_{ij}(x)$ .

Ogni sotto-transazione viene schedulata in modo indipendente di server di ciascun nodo e quindi la **schedule globale** (dove *schedules* indica le sequenze delle transazioni da eseguire) dipende dalle **schedules locali** di ogni nodo.

Purtroppo la **serializzabilità locale di ogni schedule non garantisce**

la sua serializzabilità globale. Si viene a creare un **grafo globale dei conflitti** in quanto i conflitti non sono a livello locale ma a livello globale (infatti si hanno risorse occupate tra i vari nodi, si arriva, in certi casi, ad una **situazione di deadlock**).

Si ha quindi che lo *schedule globale* è serializzabile sse **gli ordini di serializzazione sono gli stessi per tutti i nodi coinvolti** (nel caso in cui il db non sia replicato).

Qualora si abbia un db replicato si aggiunge un altro problema qualora le scritture riguardino due repliche diverse. In tal caso si può violare la **mutua consistenza** (che dice che al termine della transazione tutte le copie devono avere lo stesso valore) dei due db locali, anche con due schedule localmente seriali. Si introduce quindi un **protocollo di controllo delle repliche**. Il **protocollo di controllo delle repliche** viene chiamato **ROWA (*Read Once Write All*)**. In base a questo protocollo, dato un item logico  $X$  (con  $x_1 \dots x_n$  items fisici), si ha che le transazioni vedono solamente  $X$  ed è il protocollo che si occupa di mappare  $read(X)$  su una copia qualunque e  $write(X)$  su tutte le copie. Questo meccanismo torna utile nell'uso standard delle repliche, permettendo al client di leggere dal nodo più vicino ma imponendo, eventualmente, la scrittura su tutte le copie e bloccando le operazioni fino a quando l'ultima scrittura non è avvenuta. Si hanno purtroppo problemi di perdita di prestazioni a causa di questa scrittura “di massa”. Per lo stesso problema si hanno anche condizioni di rilascio tramite *protocolli asincroni*.

### 3.2.3 2 phase locking

Anche l'**algoritmo 2PL (*2 Phase Locking*)** viene esteso al caso di schemi distribuiti. Si hanno due possibili strategie:

1. **primary site**, *centralized*, in quanto basata sui siti
2. **primary copy**, in quanto basata sulle copie

2PL prevede che prima di rilasciare un *lock* debba averli richiesti tutti, e nello *stricted 2PL* solo dopo che sia anche stato effettuato il *commit*.

Nel caso di *centralized 2PL* si ha un **lock manager (LM)** per ogni nodo, è un'architettura “master-slave”. Il “master” è appunto il *lock manager* che gestisce i *lock* per l'intero db distribuito. Gli “slave” sono invece i *data processor*, che seguono quanto fa il *lock manager coordinatore* (che se non è disponibile per problemi tecnici del nodo comporta seri problemi in quanto la scelta di un nuovo *lock manager* tra quelli di ogni nodo è parecchio complicata).

Si ha inoltre che il **transaction manager (TM)** del nodo in cui inizia la

transazione sarà ritenuto il *TM coordinatore* dei transaction manager. La transazione anche in questo caso sarà eseguita dai *data processor* nei vari nodi. Il TM coordinatore formula all'LM coordinatore le richieste di *lock*, che vengono concesse tramite l'*algoritmo 2PL*. Una volta concesse il TM coordinatore le comunica ai vari *data processor*, assegnando ad essi i vari lock e l'accesso ai dati. Al termine delle operazioni i *data processor* comunicheranno il termine al TM coordinatore che a sua volta lo comunicherà all'LM coordinatore, che rilascerà i lock. Si ha però un effetto “collo di bottiglia” sul nodo del LM che deve gestire moltissime richieste e fino a che non risponde sistema va in *wait*. Una soluzione a questo problema è individuata nella tecnica della **copia primaria**. Prima dell'assegnazione del *lock*, viene individuata per ogni risorsa una *copia primaria*. Inoltre si ha che i diversi nodi hanno diversi *lock manager* attivi, ognuno che gestisce una partizione dei lock complessivi, relativi alle risorse primarie residenti nel nodo. Inoltre, per ogni risorsa nella transazione, il TM comunica le richieste di lock al LM responsabile della copia primaria, che assegna i *lock*. Si evita quindi il “collo di bottiglia” ma è necessario determinare a priori il LM per ogni risorsa. Inoltre si necessita di una **directory globale** dove tutti i nodi “vedono tutto”.

### 3.2.4 Gestione dei deadlock

Indipendentemente da quanto appena discusso si può creare un'**attesa circolare** tra transazioni di due o più nodi. Bisogna quindi applicare un algoritmo distribuito. Per costruire l'algoritmo dobbiamo ragionare che siamo in una “rete tra pari” *Peer-to-Peer* (e non “master-slave”) e quindi bisogna definire un protocollo su cui costruire l'algoritmo *asincrono e distribuito*. L'algoritmo potrebbe partire su uno qualsiasi dei nodi.

Si ipotizza innanzitutto che tutte le sotto-transazioni siano attivate in modo sincrono, tramite *Remote Procedure Call (RPC)* bloccante, ovvero una transazione chiede di fare un'operazione su un certo nodo facendo una RPC ad un altro nodo, mettendosi in attesa fino a che non finisce. Si possono generare due tipi di attesa:

1. **attesa da RPC**, una sotto-transazione su un nodo attende un'altra sotto-transazione, della stessa transazione, su un altro nodo
2. **attesa da rilascio di risorsa**, una sotto-transazione su un nodo attende un'altra sotto-transazione, della stessa transazione, sullo stesso nodo a causa del rilascio di una risorsa (che normalmente è la tipica situazione che porta ad un deadlock in un sistema centralizzato)

La composizione dei due tipi di attesa può generare un **deadlock globale**. E' possibile caratterizzare le condizioni di attesa su ciascun nodo tramite condizioni di precedenza e serve quindi specificare qualche notazione, per rappresentare il fatto che ogni nodo deve capire quali sono le transazioni sono in attesa per una chiamata esterna o per l'accesso ad una risorsa interna:

- $EXT_i$  per una chiamata all'esecuzione di una transazione sul nodo  $i$
- $x < y$  per indicare che  $x$  sta aspettando il rilascio di una risorsa da parte di  $y$  (che può essere anche  $EXT$ )
- indichiamo quindi la **sequenza di attesa generale** al nodo  $k$  come:

$$EXT < T_{ik} < T_{jk} < EXT$$

**Esempio 4.** *Sul DBMS1 si ha:*

$$EXT2 < T_{21} < T_{11} < EXT2$$

*e sul DBMS2:*

$$EXT1 < T_{12} < T_{22} < EXT1$$

*ovvero sul nodo 2 c'è la transazione 1 che sta aspettando che finisca. La transazione 1 sul nodo 1 sta aspettando che la transazione 2 sul nodo 1 finisca, che a sua volta sta aspettando che la transazione sul nodo 2 finisca. Però sul DBMS2 scopriamo che la transazione 1 sul nodo 1 è in attesa della transazione 2 sul nodo 2 finisca. Quest'ultima sta aspettando che finisca la transazione 1 sul nodo 2 che a sua volta attende la transazione sul nodo 1. Si ha quindi un **deadlock distribuito**, rappresentato nell'immagine 3.1.*

Per risolvere il problema del **deadlock distribuito** ogni nodo, ad un certo punto con un suo ordine temporale, prende la sua sequenza di attesa e la aggiunge ad alle condizioni di attesa locale degli altri nodi legati da  $EXT$ . Dopodiché analizza la situazione, rilevando potenziali **deadlock locali**, e comunica le sequenze di attesa alle altre istanze dell'algoritmo, ovvero agli altri nodi. Qualora si abbia un *deadlock locale* si crea un grafo dedicato allo stesso dove sarà possibile notare un ciclo, che rappresenta il deadlock. Ovviamente è possibile evitare che due nodi scoprano lo stesso deadlock, rendendo così quindi più efficiente l'algoritmo che invia le sequenze di attesa solo in alcuni modi:

1. **in avanti**, verso il nodo dove è attiva la sotto-transizione attesa (nodo nel quale vede che non ci siano deadlock (???)). Nei sistemi

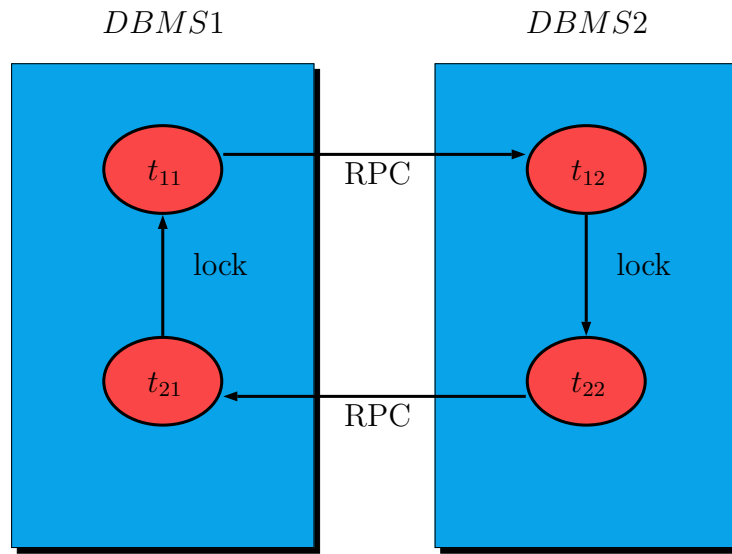


Figura 3.1: Grafico dell'esempio di deadlock distribuito

*Peer-to-Peer* questi sono meccanismi *coreografati*, decisi a priori. Esempio alla slide 68 del quarto PDF della costruzione di un grafo delle transazioni con presenza di ciclo

2. solamente quando l'identificatore del secondo nodo attende il rilascio della risorsa identificatore del primo nodo

### 3.2.5 Recovery management

Approfondiamo quindi come garantire la **proprietà di atomicità**. Abbiamo visto come uno dei guasti possibili in un sistema distribuito sia quello legato alla perdita di messaggi sulla rete, nonché al partizionamento delle stessa (comportando magari l'isolamento dei nodi). Ricapitolando abbiamo diversi tipi di guasti:

- **guasti nei nodi**, sia *soft* che *hard*
- **perdita di messaggi**, che lascia l'esecuzione di un protocollo in uno stato di *indecisione*, in quanto ogni messaggio del protocollo è seguito da un *ack* e la perdita o del messaggio o dell'*ack* stesso genera incertezza (non potendo decidere se il messaggio sia arrivato o meno)
- **partizionamento della rete**, dove una transazione distribuita può essere attiva contemporaneamente su più sotto-reti temporaneamente isolate. In questa situazione i singoli nodi non riescono



a capire bene chi sia isolato e la cosa può portare i nodi a fare scelte contraddittorie

Si è quindi studiato il **protocollo two phase commit (2PC)** che cerca di funzionare in presenza di guasti di rete. Questo tipo di protocollo consente ad una transizione di giungere ad un eventuale *commit* o *abort* su ciascuno dei nodi che partecipano alla transazione. In questo protocollo la decisione di *commit* o *abort* tra due o più **resource managers (RM)** (i server) viene certificata da un **transaction manager (TM)** (il coordinatore). Lo scambio dei messaggi (e il salvataggio di un log per ciascuno) tra TM e RMs è ciò su cui si basa il protocollo 2PC. Si ha quindi sempre un'architettura "master-slave" (in modo metaforico si può rappresentare come il prete che unisce in matrimonio i due sposi). Si ha quindi il TM che interroga i RMs riguardo allo stato della loro esecuzione.

### 2PC in assenza di guasti

Si hanno diverse fasi:

1. durante la prima fase il TM interroga (con un *prepare* o *ready\_to\_commit*) tutti i nodi per capire come ciascun nodo intenda terminare la transazione, autonomamente o irrevocabilmente *commit* o *abort* (magari per violazione della concorrenza locale o per violazione di qualche vincolo di consistenza etc...). I nodi risponderanno quindi o con *ready\_to\_commit* o con *not\_ready\_to\_commit*
2. nella seconda fase il TM prende la decisione globale. Si ha che se anche solo un nodo richiede un *abort* allora si avrà *abort* per tutti i nodi, altrimenti *commit*, chiudendo la transazione. Il TM si occupa anche di comunicare ai RMs la decisione finale per poter procedere con le azioni locali

Le fasi sono schematizzate in figura 3.2.

Come abbiamo detto precedentemente si ha la raccolta di *log* nei quali compaiono due tipi di record:

1. **record di transazione**, con le informazioni sulle operazioni effettuate
2. **record di sistema**, con l'evento di *checkpoint* e di *dump* (ovvero la copia esatta del db in un certo stato)

Le scritture sui log avvengono prima della decisione delle operazioni, che a loro volta si suddividono in *prepare* e *global decision*. Ai log del TM vengono aggiunti ulteriori dettagli:

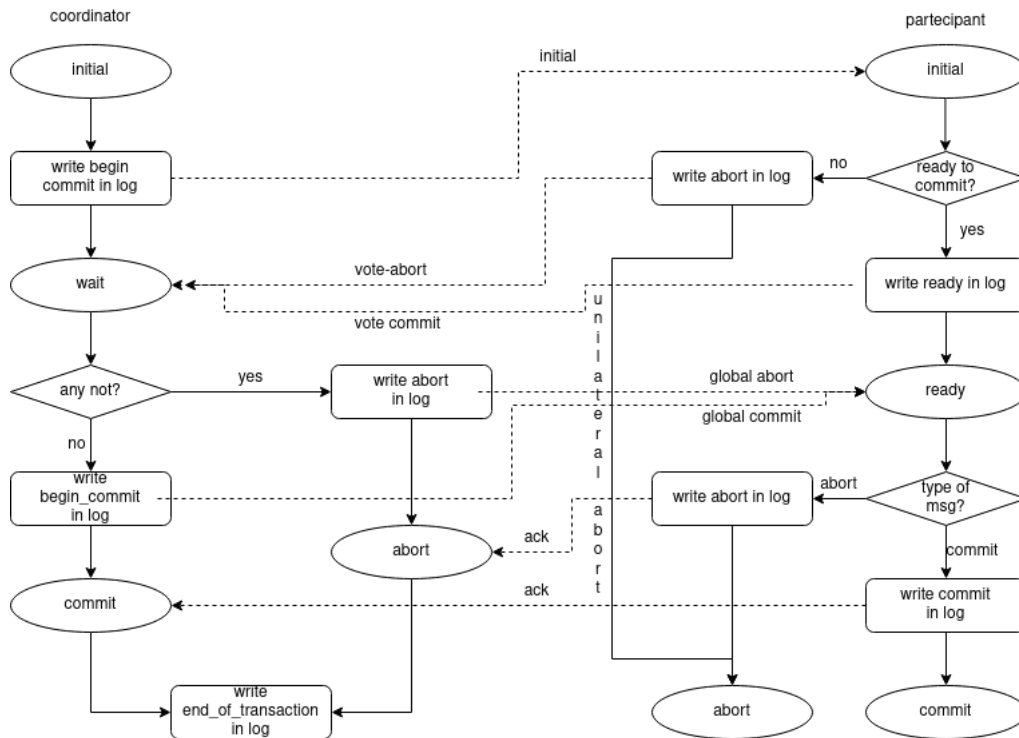


Figura 3.2: Diagramma del protocollo 2PC, dove negli ovali abbiamo le *decisioni* e nei rettangoli le *azioni sui log*. Entrambi i tipi di frecce indicano un *messaggio*. La prima fase è fino alle decisioni globali incluse.

- **prepare record** (in figura 3.2 *begin\_commit*) che contiene l'identità (nodi e transazioni) di tutti i RMs
- **global commit** o **global abort** che descrive la decisione globale. La decisione del TM diventa esecutiva quando scrive nel proprio log **global commit** o **global abort**
- **complete record** (in figura 3.2 *begin\_of\_transaction*), che viene scritto alla fine del protocollo

Al log dei RMs vengono aggiunti ulteriori dettagli:

- **ready record**, per segnalare la disponibilità irrevocabile a partecipare alla fase di commit. Si hanno diverse politiche sul **protocollo 2PL** (*recoverable*, *2PL*, *ACR*, *strict 2PL*). Inoltre questo log contiene l'indicazione del TM e i records (come nel caso centralizzato), ovvero *begin*, *insert*, *delete*, *update*, *commit*

- **not ready record** (in figura 3.2 *abort*) per segnalare l'indisponibilità del RM al commit

In entrambe le fasi del 2PC possono avvenire guasti e in entrambe tutte le componenti devono poter decidere in base al loro stato. Viene quindi introdotto l'uso di **timers**, stabilendo un **timeout** entro il quale il TM aspetta una risposta (in entrambe le fasi, anche se nella prima è più importante). Se il timeout viene superato si lancia un *abort*. Vediamo distinte le due fasi

1. il TM scrive *prepare* nel suo log e invia *prepare* ai RMs, fissando un timeout massimo per le risposte. Gli RMs *recoverable*, ovvero pronti al *commit*, scrivono *ready* nel loro log e inviano *ready* al TM. Gli RMs non *recoverable*, ovvero non pronti al *commit* a causa di un deadlock, scrivono *not-ready* nel loro log e terminano il protocollo, con un *log unilaterale*. Il TM, come detto scrive nel suo log **global commit** o il **global abort**, il secondo nel caso in cui ci sia anche solo un *not-ready* o che scatti il timeout (assumendo che i nodi che non hanno risposto siano in *failure*).
2. il TM trasmette la decisione globale e fissa un secondo *timeout*. Gli RMs *ready* agiscono di conseguenza scrivendo *commit* o *abort* nei loro log e inviando un *ack* al TM. Solo dopo effettuano in locale *commit* o *abort*. Il TM raccoglie gli *ack* e, in assenza di qualche risposta, fissa un nuovo *timeout* ripetendo la trasmissione per gli RMs problematici. Questo fino a che non avrà ricevuto da tutti un *ack* e in quel momento scrive *complete* nel suo log

Abbiamo quindi appena visto un **paradigma centralizzato** “master-salve” dove i vari RMs non comunicano tra loro. Il TM è quindi il collo di bottiglia”. Si hanno altri paradigmi:

- **lineare**, dove gli RMs hanno un ordine prestabilito e comunicano sempre secondo un ordine prestabilito. Il TM è solo il primo di tale ordine. Questo paradigma è utile per reti senza possibilità di *broadcast*
- **distribuito**. In questo caso nella prima fase il TM comunica coi vari RMs, che però rispondono a tutti. I vari RMs decidono in base alle informazioni ricevute dagli altri RMs (comunicano tra loro in *broadcast*) e quindi non è più necessaria la seconda fase di *2PC*. Si genera una gran quantità di messaggi tra i vari RMs, avendo una situazione “tra pari” *Peer-to-Peer*, creando un problema di prestazioni, dovendo garantire la comunicazione tra tutti i nodi (anche tramite *ack*)

## 2PC in caso di guasto

In uno stato di guasto, nel caso **centralizzato**, un RM nello stato *ready* perde la sua autonomia e attende la decisione del TM. Nel caso di guasto del TM i vari RMs sono lasciati in *stato di incertezza* e le risorse allocate alla transazione restano bloccate.

**Definizione 4.** Definiamo **finestra di incertezza** come la finestra temporale fra la scrittura di *ready* nel log dei RMs e la scrittura di *commit* o *abort*. Questo intervallo è ridotto al minimo da **2PC**.

Durante la finestra di incertezza tutte le risorse acquisite tramite meccanismi di *lock* restano bloccate e, in caso di guasto durante la *finestra di incertezza*, TM e RMs usano i **protocolli di recovery**.

Si possono avere diversi guasti:

- **guasti di componenti**, ovvero guasti al TM o ai RMs.
- **perdita di messaggi o partizionamento della rete**

**Guasti di componenti** In questo caso devono essere usati protocolli con due compiti:

1. assicurare la terminazione delle procedure. Sono i **protocolli di terminazione**
2. assicurare il ripristino. Sono i **protocolli di recovery**

**In entrambi i casi questi protocolli funzionano sia che il guasto interessi un solo componente che più di uno.**

Partiamo dal caso in cui cada un RM. Può cadere prima di iniziare il protocollo, non rispondendo al *prepare* e portando all'*abort*. Può cadere dopo il *ready\_to\_commit* e se quando si riprende vede nei log lo stato *ready* si mette in *wait* non sapendo bene come fare (nel caso di *abort* semplicemente chiuderà il protocollo). Il TM provvederà a inviare a tale RM *commit* o *abort*. Non si hanno quindi problemi al protocollo, o si va diretti all'*abort* o si aspetta. Il TM capisce che un RM è caduto grazie al *timeout*.

Più complesso è il caso in cui cada il TM. Se cade prima di ricevere le risposte al *prepare*, quando si sarà ripreso, guarderà lo stato delle risposte al *prepare* ricevute, altrimenti si avrà un *abort unilaterale*.

Vengono introdotti gli **algoritmi bizantini** nel caso in cui il TM cada e il RMs debbano decidere insieme cosa fare. Per decidere serve visibilità.

Quindi, ricapitolando per le cadute del TM:

- cade quanto l'ultimo record del log è *prepare*, magari bloccando alcuni RMs. In tal caso si hanno 2 opzioni di recovery:
  1. decidere *global abort*, e procedere con la seconda fase di 2PC
  2. ripetere la prima fase, sperando di giungere a un *global commit*, richiedendo nuovamente lo stato dei RM
- cade quanto l'ultimo record del log è *global-commit* o *global-abort* il TM deve ripetere la seconda fase in quanto alcuni RMs potrebbero essere bloccati o comunque ignari della decisione presa prima della caduta
- l'ultimo record nel log è una *complete*, in tal caso non si hanno problemi

Ricapitoliamo anche le cadute dell'RM:

- l'ultimo record nei log è di *azione*, *abort* o *commit*, come nel caso centralizzato e in tal caso si procede con un **warm restart**:
  - nel caso di *abort* o *azione* si procede con l'*undo* dell'operazione
  - nel caso di *commit* si effettua nuovamente la transazione
- l'ultimo record nei log è *ready* e in tal caso l'RM si blocca non conoscendo la decisione del TM e si inseriscono, durante *warm restart*, nel *ready set* le transazioni dubbie. Si hanno quindi due alternative:
  1. **remote recovery**, ovvero l'RM chiede al TM cosa è accaduto
  2. il TM riesegue la seconda fase del protocollo

**Perdita di messaggi e partizionamento della rete** In questo caso il TM non riesce a distinguere tra perdita di messaggi *prepare* o *ready* nella prima fase e procede, scattando i *timeout*, con un **global abort**.

Durante la seconda fase la non distinzione tra perdita di *ack* o di decisioni dei gli RM porta alla ripetizione della seconda fase dopo il *timeout*.

Un partizionamento della rete non causa problemi ulteriori, dato che una transazione può avere successo solo se il TM e tutti i RM appartengono alla stessa partizione.

### Ottimizzazioni di 2PC

2PC può essere ottimizzato per:

- ridurre il numero di messaggi tra TM e RMs
- ridurre le scritture nei log

Si hanno due tipi di ottimizzazione:

1. **ottimizzazione read-only**, quando un RM sa che se la propria transazione è *read-only* allora non influenza l'esito finale della transazione. Al *prepare* risponde *read-only* e termina il protocollo. Il TM ignora i partecipanti *read-only* dalla seconda fase e, qualora si sapesse a priori, possono anche essere direttamente esclusi dal protocollo
2. **ottimizzazione presumed abort** che si basa sulla regola “scordarsi gli *abort* e ricordarsi i *commit*”. In questo caso il TM abbandona la transazione dopo la decisione di abort senza scrivere *global abort* nel log e senza aspettare risposta dai vari RMs. Se il TM riceve richiesta di un *remote recovery* il TM decide per il *global abort*. Non sarà più necessario quindi scrivere *global abort* o *prepare* nei log ma solo *global commit*, *ready* e *commit*

### Protocollo X/Open

Il protocollo 2PC è stato adottato nel protocollo **X/Open DTP (*Distributed Transaction Processing*)**, che è un consorzio di vendors che vogliono rendere portabile lo standard dell'ambiente UNIX. In questo protocollo si ha il *TX interface* per la comunicazione tra il TM e l'applicazione e si ha l'*XA interface* per la comunicazione tra TM e RMs. Ogni vendor ha la sua implementazione del modello.