Architetture Dati

UniShare

Davide Cozzi
@dlcgold

Indice

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/dlcgold/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 necessità 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*:

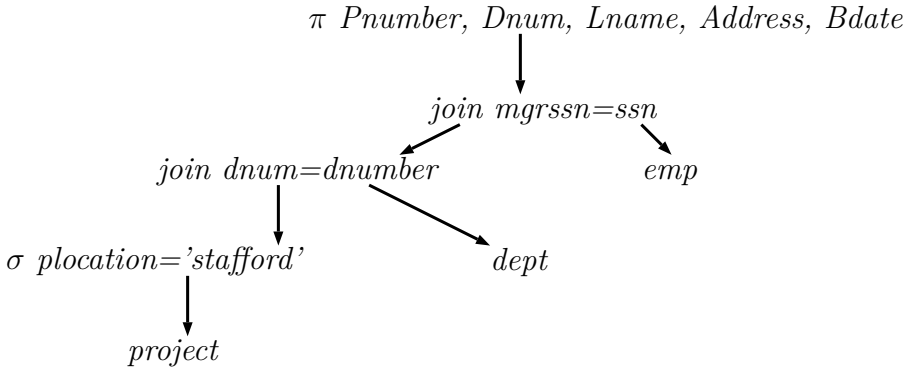
- un **query compiler** che prende una query in SQL e la traduce con un compilatore
- un **gestore di interrogazioni e aggiornamenti** che trasforma le query in SQL in algebra relazionale facendo operazioni di ottimizzazione
- un **gestore dei metodi di accesso** per permettere il passaggio tra file 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 l a 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 e 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:*
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):*

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

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 tutto 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 se una transazione termina prima che la successiva iniziale, 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 unisco 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 è una descrizione architetturale abilitante il protocollo di esecuzione di transazioni distribuite), 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 prestazione 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 interazione distribuita tra server rappresentata 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, transizioni, 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 $R_1 \cup R_2 \cup \ldots R_n = R$
   - l'unione garantisce la ricostruibilità
2. **verticale**:
   - $schema(R) = L = (A_1, \ldots, A_m)$ e $schema(R_i) = L_i = (A_{i1}, \ldots, A_{ik})$
   - garantisce la completezza, infatti $L_1 \cup L_2 \cup \ldots L_n = L$, 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 ottimizzatone 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 decisione 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 *rappresentata il tipo di responsabilità*

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

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

$$\pi_{ename}(EMP >\!<_{eno} (\sigma_{resp=''manager''}(ASG)))$$

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. $EMP1 = \sigma_{eno} \leq' E3'(EMP)$
4. $EMP2 = \sigma_{eno} >' E3'(EMP)$

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 la 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 dell'operazione di *semijoin*.

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

$$R \ semijoin_A \ S \equiv \pi_{R^*}(R \ join_A \ A)$$

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 $semijoin_A$ 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):

- $R \ join_\theta \ S \iff (R \ semijoin_\theta \ S) \ join_\theta \ S$
- $R \ join_\theta \ S \iff R \ join_\theta \ (S \ semijoin_\theta \ R)$
- $R \ join_\theta \ S \iff (R \ semijoin_\theta \ S) \ join_\theta \ (A \ semijoin_\theta \ R)$

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 **distribuited requests**, dove ogni singola operazione SQL si può riferire, grazie ad un *ottimizzatore distribuito*, a qualunque insieme di server, e di **distribuited 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 valore 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 transizione 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 necessità 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

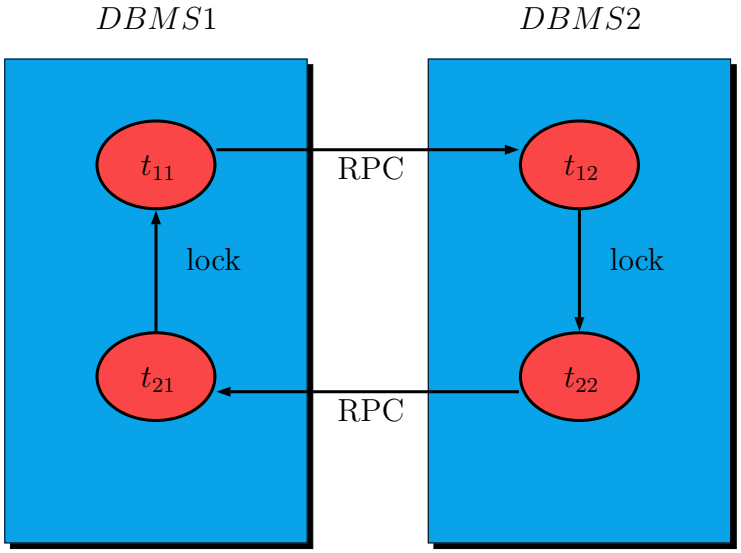


Figura 3.1: Grafico dell'esempio di deadlock distribuito

- $x < y$ 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 *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:

- **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**

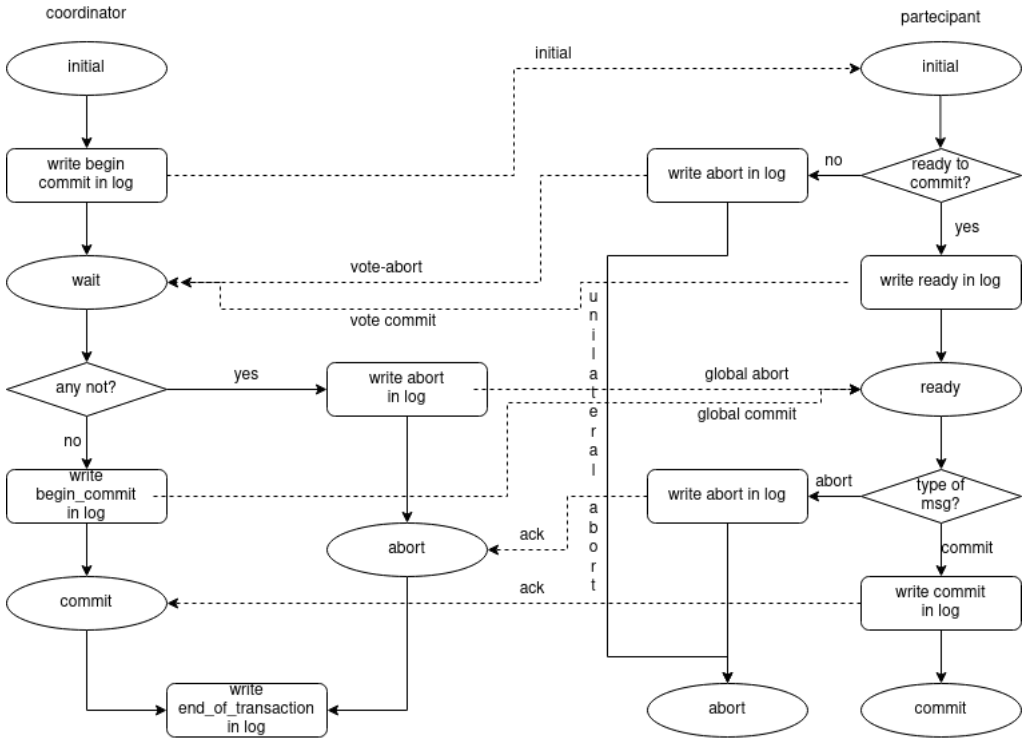


Figura 3.2: Diagramma del protocollo 2PC, dove negli ovali abbiamo le *decisioni* e nei rettangoli le *azioni sui log*. Entrambi i tipi di frecce indicano un *messaggio*. La prima fase è fino alle decisioni globali incluse.

- **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 riposte. 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 ripete 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 di procede con un **warm restart**:
  - nel caso di *abort* o *azione* si procede con l'*undo* dell'operazione
  - nel caso di *commit* di 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*.
Se durante lo svolgimento del protocollo 2pc si partiziona la rete avendo due sottoreti una con TM e RM1 che e la seconda sottorete con RM2, RM3. In questo caso il TM continua a mandare il messaggio della global decision a RM2 RM3 dopo lo scadere del timeout. Si concluderà comunque, con alte probabilità, con un *abort* da parte del TM.

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 ricordasi 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 scriver e*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.

3.3 Repliche

Parliamo ora delle tecnologie per la **replicazione di dati**.
Tra le principali soluzioni architetturale troviamo **IBM replication technologies** e **Microsoft SQL Server replication technologies** (*i dettagli delle implementazioni non saranno oggetto d'esame*).

Definizione 5. *La **replica** è il processo di cerare e mantenere istanze dello stesso db allineate tra loro, consentendo la condivisione di dati ma anche comportando cambiamenti architetturali. Si ha che le eventuali modifiche devono essere viste da tutti i nodi.*

Definizione 6. *Definiamo **sincronizzazione** è il processo che m i consente di allineare le copie, prima o poi.*

In base all'ultima definizione si capisce che spesso le copie non sono aggiornate istantaneamente. Si ha quindi la **replica sincrona** o la **replica asincrona** (non avendo allineamento *realtime*). IBM preferisce un approccio *based and mode based* mentre Microsoft uno basato su *snapshot, transactional* e *merge*.
Vediamo le differenze tra le due repliche:

- le **repliche sincrone** cercano di far si che tutte le repliche vengano aggiornate contemporaneamente (ad esempio si ha il protocollo ROWA). Scrivo in modo sincrono su tutti nodi e solo quando tutte le repliche confermano la scrittura avanzo con le transazioni (che fallisce se ho nodi non disponibili). Nelle repliche sincrone si necessitano molti scambi di messaggi. Di fatto si obbliga due o più *storage* ad aggiornarsi e a fare *rollback* in caso di fallimento. Si hanno quindi alte disponibilità, un auto *fail-over* (bloccando la transazioni in caso di guasti sui nodi) e un *data loss* minimo. Le repliche sincrone vengono soprattutto usate nei *disaster recovery*, ovvero in situazioni *mission critical* (ad esempio sistemi bancari dove i db di backup, almeno 3, devono stare a centinaia di kilometri di distanza, in zone sismiche tra loro diverse). Gli svantaggi delle repliche sincrone sono la necessità di una rete valida, si hanno problemi di scalabilità, di costi e minor flessibilità

- nelle **repliche asincrone** prima si aggiorna il db *target* e poi le repliche (normalmente dopo pochi secondi ma anche dopo giorni). Si hanno evidenti vantaggi di costo, scalabilità e flessibilità (perché in caso di problema lavoro in primis sul db principale) ma a rischio di *data loss* (nell'intervallo di tempo tra la scrittura del db principale e delle repliche). Normalmente si usano soluzioni asincrone per accessi online e la loro efficienza, per bilanciamento del calcolo etc...

In caso di perdita di dati bisogna analizzare i singoli contratti per capire legalmente come rispondere di dati che non verranno recuperati probabilmente.
Ci sono vari contesti in cui pensare alla replica dei dati:

- condivisione di dati da utenti tra loro scollegati. Un esempio è una copia su un portatile con una replica usata da un commerciale. Si possono avere conflitti nel momento in cui più utenti con db replicati lavorano offline. Si ha il *merge conflict*

- *data consolidation*, ovvero quando un'azienda vuole tenere più copie dei dati in vari punti e alla fine bisogna riportare i dati

a livello centrale a cadenza periodica. Può servire per fare data warehousing o anche solo semplicemente per monitorare le vendite delle varie filiali o per aggiornare il catalogo

- *data distribution* che è il caso degli *e-commerce*. È un caso tipicamente *mission-critical* e bisogna aumentare l'accesso ai dati e si ha una costante sincronizzazione realtime bidirezionale per evitare problemi. Un altro caso è la distribuzione tra diversi uffici, dove le repliche locali hanno magari dati non presenti nel db globale

- prestazioni, accesso efficiente, *load balancing* e accesso offline. Se non si hanno necessità di update immediati (tipo un sito vetrina) allora la replica garantisce la disponibilità e l'accessibilità a basso costo. Con il load balancing scarico gli utenti su diverse macchine replicate, in primis per le molte realtà di sola lettura o comunque con pochissime scritture (un esempio può anche essere un social network dove anche eventuali ritardi di qualche secondo non sono problematici). Per la disponibilità bisogna fare un forte testing dell'intera architettura, se cade un server bisogna puntare ad una replica e bisogna essere sicuri e spesso i costi sono troppo alti (*RIVEDERE QUESTA PARTE*)

- separazione tra *data entry* e *reporting*, se si usa lo stesso server per entrambi i compiti (che in pratica sono scrittura costante e lettura costante) può essere utile separare in due server. Si evitano così i rallentamenti dati dai *lock*. Bisogna studiare i tempi di sincronizzazione

- coesistenza di applicazioni, questo è un caso particolare. Qualora sia necessario cambiare applicazione devo, eventualmente, cambiare anche i sistemi. Bisogna quindi travasare i dati vecchi e durante il trasferimento bisogna comunque mantenere funzionanti le applicazioni. Quindi bisogna far coesistere i due database durante il trasferimento (facendo il travaso di notte bloccando le transazioni). I costi sono incredibili e si può avere anche coesistenza delle applicazioni e non solo dei dati, con migrazioni parziali (magari per area geografica) etc... questo comporta che magari due filiali devono collaborare con due applicazioni e due db diversi

Ci sono anche casi in cui non si dovrebbe replicare:

- quando ci sono frequenti update su più copie, portando le copie a possibili conflitti che devono essere scoperti e gestiti "manualmente"
- quando la consistenza è *critical* magari in contesti di trasferimento di fondi etc.... In questo caso solitamente si impone un protocollo ROWA (con transazioni *ACID compliant*), riducendo le prestazioni per avere l'autorizzazione dei *commit* da parte delle repliche

Ma spesso bisogna comunque replicare "scegliendo il male minore" (ad esempio nelle banche) e bisogna quindi analizzare il singolo caso, anche in base al budget. Inoltre non bastano le tecnologie serve un'ottima organizzazione.
Concludendo si ha che i benefici della replica sono:

- disponibilità
- affidabilità
- prestazioni
- riduzione di carico
- lavoro offline
- supporto a molti utenti

Distinguiamo anche delle classi di tipologie di replica:

- **data distribution**, di tipo *1:many*, con un *source* che distribuisce, in modo sincrono o asincrono, le varie copie passive ai *target*
- **Peer-to-Peer**, dove i vari nodi sono interconnessi e si aggiornano tra di loro. Si usa un approccio ROWA
- **data consolidation**, di tipo *many:1*, dove ho più *source* che aggiornano un *target* a livello centrale
- **bi-directional** (per il *conflict detention resolution*), dove una copia primaria e uno secondaria possono leggere e scrivere a vicenda tra loro (è quindi una versione semplificata del *Peer-to-Peer* con due Peer)
- **multi-Tier staging**, in cui si hanno meccanismi intermedi tra *source* e *target* con "aree di deposito" dette aree di *staging*

Per realizzare una replica posso fare in diversi modi:

- faccio letteralmente il backup del disco con una persona che stacca il disco dal server e lo copia, riattaccando infine copia e disco originale
- posso prima fare il backup attaccando un altro disco e poi mettere nella nuova macchina il disco copia
- posso fare una *replica incrementale*, ovvero faccio un *full backup* e sposto solo il file di *log* delle transazioni nel nuovo server, rieseguendo quanto fatto (essendo contenuto nel *log*). Quindi prima faccio un *full backup* e poi un *backup* del *log*, questo per ogni replica. Un'alternativa è l'**event publish**. Un **event publish** è una replica senza *apply* e leggo i file di log. Analizzando gli eventi tramite particolari meccanismi riscrivo quindi sull'architettura target. Da un *publisher* si passa ad un *distributor* e infine ai *subscriber*

Sulle slide *repliche* si ha un approfondimento delle architetture IBM e Microsoft opzionali per il corso.

Vengono qui aggiunte le cose dette in live.
Un **db parallelo** è studiato per le prestazioni. Si ha accesso parallelo ai dati, parallelismo *intra-query* (stessa query su frammenti diversi), parallelismo *inter-query* (tante query diverse) e sono fatti da elementi hardware posti vicini tra loro.
Parliamo di **persistenza dei dati**. Bisogna garantire la durabilità dei dati, bisogna quindi usare un db. Si hanno anche problemi per creare oggetti persistenti a partire da un linguaggio OOP, si ha il cosiddetto l'**object-relational paradigm mismatch**. Si separa quindi l'aspetto OOP e l'aspetto relazionale per risolvere il problema per poi "mappare" l'uno nell'altro, tramite i **data mapper**. Si hanno operazioni **CRUD**:

- Create
- Read
- Update
- Delete

Si ha quindi il cosiddetto **Object–relational mapping**, ovvero questa tecnica di mapping. Si hanno vari framework che lo implementano.
Storicamente si è provato a fare db ad oggetti ma con scarsi risultati.
Parliamo ora del *caso Gitlab*. Questo è un esempio paradigmatico di cosa succede in caso di mala gestione delle repliche. Il 31 Gennaio 2017 dei malintenzionati erano entrati nel db PosgreSQL (con due repliche) e stavano scrivendo in modo significativo sul nodo primario DB1 e sul secondario DB2. I tantissimi lock si bloccavano a vicenda rendendo non disponile il db. Alle 21 prendono il db secondario e cercano di riorganizzare tutto ma il db primario aveva fatto chiamate al secondario che però non aveva risposto. I tecnici iniziano a fare danni extra, svuotando la directory dei dati del secondario ma per errore ha "droppato" il database primario, cancellando 300GB di dati dal primario. C'erano ancora i backup, fatti a *snapshot* ogni 24h. C'era un backup manuale di 6 ore prima per altri motivi. Il problema è che non sono stati trovati i backup, una volta trovati c'erano problemi a causa di incompatibilità tra PosgreSQL 9.2 e 9.6. I dati su Azure non erano backuppati per scelta di Azure, che non faceva i backup dei database. Nel complesso 5 sistemi di backup in totale non funzionavo (alcuni, tipo S3, per errori di codice erano vuoti). Fortunatamente c'era il backup manuale di 6 ore prima e si è risolto il problema ma perdendo 5000 progetti, 700 user e migliaia di righe di codice.
Bisogna premiare la trasparenza di Gitlab che ha documentato tutto quello che è accaduto (senza specificare i nomi dei DBA).

3.4 Prima esercitazione

Esercizio 1. *Si consideri un db con le seguenti relazioni (e quindi tabelle):*

- *PRODUCTION (SerialNumber, PartType, Model, Quantity, Machine)*
- *PICKUP (SerialNumber, Lot,* ***Client, SalesPerson****, Amount)*
- *CLIENT (Name, City, Address)*
- *SALESPERSON (Name, City, Address)*

(con sottolineate le chiavi primarie e in grassetto le chiavi di integrità referenziale)

e ci poniamo l'obiettivo di partizionare il db secondo determinate specifiche.
Si assume che si abbiano le seguenti specifiche organizzative:

- *si hanno 4 centri di produzione (Dublino, San Jose, Zurigo* e *Taiwan, ciascuno responsabile, rispettivamente, di cpu, keyboard, screen* e *cable) e 3 centri di vendita (San Jose, Zurigo* e *Taiwan)*
- *le vendite sono distribuite secondo le località geografiche, i clienti a Zurigo sono serviti solo dai venditori di Zurigo etc.... Si ha però che i venditori di Zurigo servono anche Dublino*
- *ogni area geografica ha il proprio db (avremo quindi 4 db)*

Vogliamo studiare una ***frammentazione orizzontale*** *delle 4 tabelle.*
Ricordiamo quindi che abbiamo 4 centri di produzione (ciascuno responsabile di un prodotto e con un db ciascuno) e 3 punti di vendita.
Partiamo con la frammentazione della relazione PRODUCTION*, ottenendo 4 tabelle, una per componente prodotto, ottenendo (con σ abbiamo l'operazione di selezione nell'algebra relazionale):*

- $PRODUCTION_1 = \sigma_{partType=cpu}(PRODUCTION)$
- $PRODUCTION_2 = \sigma_{partType=keyboard}(PRODUCTION)$
- $PRODUCTION_3 = \sigma_{partType=screen}(PRODUCTION)$
- $PRODUCTION_4 = \sigma_{partType=cable}(PRODUCTION)$

Passiamo alla relazione PICKUP*. Anche in questo caso si frammenta per il prodotto facendo il join con la tabella* PRODUCTION*(ricordando che π è l'operazione di proiezione/join nell'algebra relazionale). Per comodità indichiamo tutti gli attributi di* PICKUP *con pick, avendo quindi:*

$$pick = SerialNumber, Lot, Client, SalesPerson, Amount$$

indico anche con SN SerialNumber

- $PICKUP_1 = \pi_{pick}(\sigma_{partType=cpu}(PICKUP\ SN = SN(PRODUCTION)))$
- $PICKUP_2 = \pi_{pick}(\sigma_{partType=keyboard}(PICKUP\ SN = SN(PRODUCTION)))$
- $PICKUP_3 = \pi_{pick}(\sigma_{partType=screen}(PICKUP\ SN = SN(PRODUCTION)))$
- $PICKUP_4 = \pi_{pick}(\sigma_{partType=cable}(PICKUP\ SN = SN(PRODUCTION)))$

Prendo quindi una proiezione di tutti gli elementi di PICKUP separando nei vari PICKUP in base al prodotto.
Passo a alle tabelle SALESPERSON*. Abbiamo 3 punti di vendita, quindi, circa come per* PRODUCTION*, frammento in base alle città di vendita:*

- $SALESPERSON_1 = \sigma_{City="SanJose"}(SALESPERSON)$
- $SALESPERSON_2 = \sigma_{City="Zurigo"}(SALESPERSON)$
- $SALESPERSON_3 = \sigma_{City="Taiwan"}(SALESPERSON)$

Manca solamente CLIENT*. Anche in questo caso divido in base alle città, ricordando che Zurigo e Dublino sono clienti entrambi di Zurigo:*

- $CLIENT_1 = \sigma_{City="SanJose"}(CLIENT)$
- $CLIENT_2 = \sigma_{City="Zurigo" \text{ or } City="Dublino"}(CLIENT)$
- $CLIENT_3 = \sigma_{City="Taiwan"}(CLIENT)$

Abbiamo finito la frammentazione e quindi dobbiamo solo distribuire tali tabelle:

- *le quattro tabelle con indice 1 andranno a San Jose, in* `company.sanjose.com`
- *le quattro tabelle con indice 2 andranno a Zurigo, in* `company.zurigo.com`
- *le quattro tabelle con indice 3 andranno a Taiwan, in* `company.taiwan.com`
- *le due tabelle con indice 4 andranno a Dublino (che quindi avrà solo parte di* PRODUCTION *e parte di* PICKUP *in quanto a Dublino non si ha un punto vendita), in* `company.dublino.com`

Esercizio 2. *Vediamo un esercizio in merito alla trasparenza, che ricordiamo essere a tre livelli:*

1. *di frammentazione*
2. *di replicazione/allocazione*
3. *di linguaggio*

Si chiede di fare delle interrogazioni, tenendo conto dei livelli di trasparenza, sul db costruito nell'esercizio precedente.
La prima query ci chiede di determinare la quantità dei prodotti che hanno

valore "77y6878" (abbiamo quindi a che fare con la trasparenza di frammentazione, infatti interroghiamo come se avessimo a che fare con un solo db):

(con `:Quan` *indichiamo il nome della tabella).*

Vediamo ora come fare nel caso di trasparenza di allocazione, quindi si sa di avere a che fare con un db distribuito. La query quindi si "sposterà" alla ricerca del giusto frammento:

Vediamo ora come funziona per la trasparenza di linguaggio. In tal caso dobbiamo considerare sia le frammentazioni che i vari indirizzi di allocazione, ovvero i `company.città.com`*:*

Nelle slide usa `union` *al posto di* `if :empty then`*.*

Esercizio 3. *Sempre sul db del primo esercizio effettuiamo la seguente query: determinare le macchine che utilizzano come componente "keyboard" e sono vendute al cliente "Brown".*

Per praticità vediamo solo la trasparenza di frammentazione e quella di allocazione.

Partiamo con la trasparenza di frammentazione:

Vediamo il caso di trasparenza di allocazione (e sappiamo che "keyboard" è solo in $PRODUCTION_2$ quindi interroghiamo un solo frammento e senza chiedere la specifica del partType):

Se avessimo voluto fare anche la trasparenza di linguaggio non sarebbe cambiato nulla dato che $PRODUCTION_2$ è solo a Zurigo.

Esercizio 4. *Sempre sul db del primo esercizio effettuiamo il cambiamento di indirizzo del cliente "Brown" che si sposta da "27 Church St.", Dublino, a "43 Park Hoi St.", Taiwan. Abbiamo quindi un cambio di allocazione nel db.*

Partiamo con la trasparenza di frammentazione:

La cosa si complica nel caso di trasparenza di allocazione, tenendo conto delle due città e dei loro db:

Se avessimo voluto fare anche la trasparenza di linguaggio non sarebbe cambiato nulla poiché le frammentazioni sono in locazioni diverse

Esercizio 5. *Sempre sul db del primo esercizio effettuiamo la seguente query: calcolare la somma di tutti gli ordini ricevuti a SanJose, Zurigo e Taiwan.*

Partiamo con la trasparenza di frammentazione:

Passiamo alla trasparenza di allocazione. Mi serviranno tutti i frammenti di `SALESPERSON`*. Inoltre devo considerare i vari* `PICKUP`*, tutti e quattro per il discorso delle vendite su Dublino:*

Esercizio 6. *Sempre sul db del primo esercizio cerchiamo di massimizzare il parallelismo delle inter-query.*

Prendiamo la seguente query: estrarre la somma delle quantità di produzione che sono raggruppate secondo i tipi e i modelli delle componenti.

Vogliamo però massimizzare il parallelismo, divido quindi tra le varie frammentazioni:

massimizziamo il parallelismo inter-query se si consideriamo che ogni partizione ha un DBMS diverso. Volendo potrei dividere la query ancora a seconda del modello, evitando il `group by` *(cosa utile nel caso in cui si abbia a che fare con un sistema fortemente multicore)*

Esercizio 7. *Vediamo un esempio di db replicato che può produrre inconsistenza.*
Prendiamo l'esempio del db prodotto nel primo esercizio. Supponiamo che ogni frammento di PRODUCTION *sia allocato a tutti i DBMS. Però ogni DBMS utilizza un frammento e trasmette i cambiamenti del frammento agli altri DBMS, permettendo di avere copie del db. In caso di fallimento di un DBMS il db sarebbe comunque accessibile dagli altri sistemi e non so avrebbe problemi con le query, avendo che il fallimento è trasparente sia ai client che al db stesso. Purtroppo quando si ha un fallimento si può generare un partizionamento di rete e questo può comportare delle inconsistenze. Se, per esempio, due transazioni tolgono 800 ad una certa quantità la seconda in ordine temporale fallirà ma se avvengono in due DBMS non connessi non falliranno, producendo inconsistenza.*

Esercizio 8. *Data una replicazione simmetrica dire quando produce inconsistenza. Un esempio è un db senza il concurrency control e quindi due transazioni come quelle dell'esercizio precedente possono causare inconsistenza anche senza fallimento della rete.*

Capitolo 4

Blockchains

Questa lezione è stata tenuta dal prof. Leporati.
Nonostante qualche ambiguità ed errore di scrittura, nel capitolo si userà il termine "nodo" quando si parla di rete P2P e il termine "blocco" quando ci si riferisce ai blocchi interni alla blockchain.

Definizione 7. *Una **blockchain**, in poche parole, è un registro pubblico, condiviso e decentralizzato che memorizza la proprietà di beni digitali.*

È quindi un registro in cui vengono memorizzate informazioni relative alla proprietà di qualcosa che può essere rappresentato tramite sequenze di bit (quindi qualcosa di digitale come i *bitcoin* o altre criptovalute).
Vengono memorizzate anche le **transazioni**, ovvero i cambi di proprietà.
Essendo pubblico tutti possono vedere cosa è stato registrato e in particolare "chi possiede cosa" e la storia di una certa proprietà.
Questo registro è condiviso in quanto gestito da più persone ed è decentralizzato in quanto non esiste un nucleo che abbia di poteri da amministratore rispetto agli altri, tutti sono allo stesso livello.
Questo registro è organizzato in blocchi. Si ha un primo blocco detto **genesis block** che da il via a tutto. Si hanno poi altri blocchi, in nero, collegati a questo e che formano una catena. Un blocco si lega al successivo tramite particolari funzioni crittografiche, dette **funzioni di hash crittografico**. Nel dettaglio il collegamento tra un blocco e il successivo è dato dal fatto che il valore di hash del blocco è contenuta all'interno del blocco successivo, facendo in modo che si estremamente difficile alterare il contenuto di un blocco, ovvero diverse transazioni, che sono organizzate secondo una precisa struttura dati che consente di verificare la validità in modo veloce ed efficiente. Per ogni blocco si calcola quindi l'hash e lo si salva nel blocco successivo. Per modificare una transazione quindi dovrei calcolare l'hash e dovrei fare il check con il blocco successivo. Quindi tutti possono verificare la validità

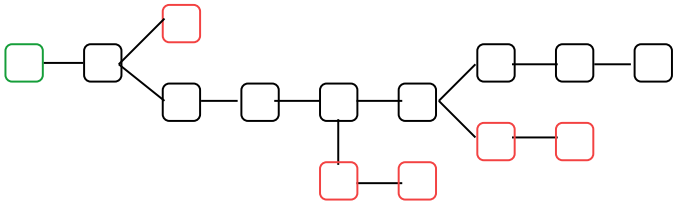


Figura 4.1: Esempio di schema di blockchain. In verde a sinistra troviamo il **genesis block**. In nero i blocchi che formano la catena. In rosso abbiamo i nodi *orfani*. Per comodità è stata rappresentata in orizzontale con la parte "alta" a destra

della blockchain. Modificare i dati in modo tale da ottenere lo stesso hash però è davvero difficile.

Si hanno vari nodi per l'uso della blockchain. I vari nodi sono nodi di una **rete Peer-to-Peer (*P2P*)** (come quelle per la condivisione di file come *torrent*, dove quindi ogni computer fa sia da *client* che da *server*). I vari nodi osservavano le proposte di transazione che vengono fatte dagli utenti (anche utenti che solo usano la blockchain), verificano che siano valide (ad esempio verificando che non avvenga il *double-spending*, ovvero, per esempio, una doppia concessione dello stesso bene), eseguono un **protocollo di consenso** (in quanto potrebbero esserci nodi "disonesti", con per esempio utenti che vorrebbero appropriarsi di beni altrui come bitcoin, si procede quindi a maggioranza secondo il *proof-of-work*) e infine procedono validando la transazione aggiungendo il blocco alla fine della blockchain ("in cima"). Una copia della blockchain viene memorizzata in ogni nodo della rete P2P alla fine della transazione, quindi quando i nodi sono d'accordo sull'aggiunta del blocco allora ciascuno lo aggiunge alla propria copia della blockchain (altrimenti qualsiasi proposta futura verrebbe bocciata).

In molte blockchain se si stabilisce che due blocchi possono essere attaccati ad un certo blocco e si inizia a lavorare su entrambi formando quindi due nuove catene. Alla fine "vince" la catena più lunga, fermando la continuazione dell'altro ramo. I blocchi del ramo "perdente" vengono resi *orfani* lasciando quindi una sola catena attiva. Le transazioni nei nodi *orfani* vengono dimenticate e bisognerà reinserirle. Questa cosa è molto inefficiente.

Ci sono reti, come quella *bitcoin*, in cui una transazione è valida sse vengono aggiunti 6 blocchi al di sopra di quello che contiene la transazione.

Le transazioni sono **pseudo-anonime** (comodo ambito bitcoin dove, come per i contanti, non si sa per quali mani sono passati quei soldi e per cosa sono stati usati prima). Ci sono vari meccanismi per rendere anonime le cose, alcuni migliori (come quelli di *monero* o *zcash*) e alcuni peggiori (come quelli

di *bitcoin* ed *ethereum*).
Si ha un ramo della *computer foreniscs* che si occupa di capire chi ha fatto certe transazioni, esplicitamente di criptovalute, sulla blockchain. Quindi si prova a raggiungere l'anonimato ma non sempre si riesce (basta un utente che paghi in *bitcoin* su un sito rivelando la propria identità).

4.1 Bitcoin

Bitcoin è una criptovaluta proposta da Satoshi Nakamoto (probabilmente un nome falso in quanto in molti stati creare valuta, anche digitale, è reato) nel 2008. Non è nato all'improvviso ma molte idee all'interno di *bitcoin* erano già presenti i diversi paper di crittografia precedenti (ad esempio la *proof-of-work* era già presente all'interno della gestione dello spam delle mail, dove veniva imposto un certo sforzo computazionale per mandare una mail in modo che l'invio non fosse completamente "gratuito", comportando che si possano mandare al massimo circa 2000 mail al giorno, al più di attrezzarsi con dei super computer). Ci sono stati tanti precedenti di tentativi di creazione di un sostituto digitale del denaro, con diverse difficoltà a causa di anonimato e *double-spending*. Inoltre rappresentare una moneta con una sequenza di bit consente la copia illimitata di tale sequenza, creando copie perfettamente identiche. Una soluzione per evitare il *double-spending* è quella di usare una *trust third party (TTP)*, ovvero tipicamente una banca che segna le transazioni monetarie, diventando però un "collo di bottiglia", venendo interpellata in tutte le transazioni. Inoltre la banca è conscia delle transazioni di denaro, che perdono così l'anonimato. L'idea di Satoshi Nakamoto e stata quella di unire vari protocolli crittografici usando al posto della banca un registro condiviso, una blockchain appunto, in cui vengono salvate le transazioni e dove vengono anche controllate, permettendo di evitare il *double-spending*. Si usa la blockchain quindi per creare un *trust*, sostituendo la funzione delle banche. Nel **genesis block** di *bitcoin* c'è infatti un messaggio (nel posto del blocco dedicato alla "causale") che fa riferimento al potere eccessivo delle banche.
Le proprietà memorizzate nella blockchain *Bitcoin* sono chiamate direttamente, come abbiamo già scritto, **bitcoins (*BTC*)** o frazioni di essi. Le transazioni sono parecchio complicate, con una struttura dati complessa, con diverse transazioni in ingresso (infatti una transazione può contenere transazioni), campi per generare nuovi bitcoin,0diversi *script di transazione* in output (scritti in un linguaggio "particolare").
Non verrà trattata nel dettaglio la conformazione delle transazioni.
Se un utente A vuole mandare un *bitcoin* ad un utente B, usando il pro-

prio client, che spesso viene chiamato **wallet**, specifica la quantità che vuole mandare e l'indirizzo di B. Ogni utente ha associato quindi un indirizzo, in qualità di lunga sequenza di numeri. Per ottenere l'indirizzo viene usata una **chiave pubblica**, usando quindi la *crittografia a chiave pubblica* in cui si usano algoritmi di cifratura e de-cifratura. Ciascuno crea con questi algoritmi una copia *chiave pubblica/chiave privata*, rende pubblica la *chiave pubblica* e comunica di usare quella chiave pubblica per risalire all'indirizzo a cui farsi spedire i *bitcoin*, infatti data la chiave pubblica all'utente che deve inviare i bitcoin la userà per cifrare il messaggio in modo che, tramite la chiave privata, solo il legittimo destinatario possa decifrare il messaggio.
Dalla chiave pubblica quindi ottiene l'indirizzo al quale A deve mandare i soldi per farli ricevere a B. Quindi l'indirizzo è una sorta di identità per B che però può generarsi tutte le coppie di chiavi che vuole, combinando poi i vari bitcoin ricevuti ai vari indirizzi che ha generato tramite la complessa struttura della transazione. Questa possibilità di generare infinite chiavi però è solo uno *pseudo-anonimato*.
Bisogna però anche dimostrare che A è proprietario dei *bitcoins* che vuole inviare a B. Per farlo A "firma" digitalmente la transazione tramite la sua *chiave segreta*. I nodi della rete P2P, che sono detti **miners**, verificano che la firma di A sia valida, verificano che non ci sia il *double-spending* e infine validano la transazione mettendola in un nuovo blocco della blockchain. Se B, che ha ricevuto i *bitcoins*, vuole a sua volta mandarli a C avvia una transazione esattamente come descritto sopra, firmando con a chiave privata che era accoppiata alla chiave pubblica con la quale A gli aveva mandato i *bitcoins*, permettendo che la firma sia verificata e validata. Quindi sulla blockchain il possedimento di un *bitcoin* è rappresentato dal fatto che si può inviare una transazione per cui quel *bitcoin* può essere dato a qualcun altro (ovviamente si è usato *bitcoin* ma si poteva parlare anche di più *bitcoins* o, vedremo in seguito, frazioni, che molto piccole, di esso). Non è segnato da nessuna parte quanti *bitcoins* possiede un certo utente ma solo le catene di transazioni che sono state fatte da ciascun *bitcoin*, vedendo quindi chi è l'ultimo proprietario. È quindi essenziale memorizzare le chiavi in quanto possedere equivale a conoscere una chiave segreta.
La blockchain registra ogni singola transazione (e sono circa un migliaio ogni 10 minuti, con un blocco che riesce a contenere circa un migliaio di transazioni e i blocchi vengono aggiunti uno ogni 10 minuti circa) e attualmente, in data 19 Ottobre 2020, si è arrivati a 290GB di blockchain.

4.1.1 Miners

Abbiamo parlato prima dei membri della rete P2P della blockchain *Bitcoin*, detti appunto **miners**.
I *miners* lavorano su un *pool* di transazioni proposte dai vari *wallet*. I miners scelgono circa un migliaio di queste transazioni alla volta e cercano di formare il nuovo blocco, validando le transazioni. Effettuano quindi i calcoli computazionali del *proof-of-work* per cui dimostrano di aver fatto un certo sforzo per avere il **diritto** di essere quelli che aggiungono il prossimo blocco alla catena. Avendo un'aggiunta ogni 10 minuti si ha una fortissima concorrenza tra i *miners*. Inoltre il carico di lavoro richiesto diventa sempre più difficile in quanto, grazie alla **legge di Moore**, diventa sempre più facile risolvere il "puzzle" crittografico, per il *proof-of-work,* che serve a risolvere il nuovo blocco e quindi il "puzzle" viene reso sempre più difficile (ovvero se un blocco arriva in meno di 10 minuti il blocco successivo sarà più difficile da produrre, in modo che nuovamente servano almeno 10 minuti, anche se equivalentemente verrà reso più facile se ci vogliono troppi minuti in più di 10, avendo così **autoregolazione**). Bisogna quindi parlare di questo "problema" crittografico da risolvere e per farlo bisogna un attimo specificare meglio le *funzioni di hash*.

Definizione 8. *Le **funzioni di hash** sono funzioni crittografiche che prendono in input una sequenza di bit, in teoria arbitrariamente lunga, anche se ogni funzione ha un limite teorico (ma praticamente irraggiungibile), e produce una sequenza che vorrebbe essere univoca (ma che non lo è) di poche centinaia di bit. Per esempio SHA1 produce in output una sequenza di 160 bit, MD5 di 128 bit, SHA256 di 256 bit (una di quelle usate in Bitcoin), RIPEMD di 160 bit (anch'essa usata in Bitcoin) etc...*
Le funzioni di hash devono essere sufficientemente facili da calcolare a partire da un certo input

L'idea è quindi è che prendo un file e, usando ad esempio SHA256, mi esce una sequenza di 256 bit, univoca per quel file. Purtroppo il dominio della funzione (ovvero i bit del file) è molto più grande del codominio (ovvero tutte le possibili sequenze di 256 bit) e quindi non si può avere davvero una **funzione iniettiva** e quindi si avranno sempre due sequenze in input che producono lo stesso output e quando questo avviene si ha una **collisione**. Le collisioni sono inevitabili ma le funzioni di hash sono fatte in modo tale che sia estremamente difficile trovare due input diversi che producano lo stesso output, motivo per cui si può anche pensare che le hash siano univoche. Si ha anche che è estremamente difficile cercare esplicitamente un input che abbai lo stesso hash di un altro, rendendo quindi molto difficile sostituire un

input dato con un altro ottenendo comunque lo stesso output, imbrogliando. Quest'ultimo problema è più difficile di quello della *collisione* (dove ho, nella pratica, un grado di liberà in più avendo due input).
Si hanno altre due proprietà:

1. dato un hash è estremamente difficile trovare un input, per questo si dice che la funzione di hash è *one-way* (essendo molto facile da calcolare ma difficilissimo da invertire)

2. anche se cambio un solo bit dell'input ottengo un output completamente diverso a quello ottenuto prima del cambiamento, rendendo impossibile capire la regola di calcolo o anche solo fare indagini statistiche. Infatti hash significa anche "polpettone", cosa che rappresenta bene l'azione di spezzettamento, calcolo e rimescolamento (con anche valori semi casuali) dell'input fatte dalle funzioni per calcolare l'output

Quindi nel *proof-of-work*, per dimostrare di aver svolto una certa quantità di lavoro viene preso il dato di cui devo calcolare l'hash, gli viene aggiunta una quantità casuale detta **nonce (si pronuncia "nons")** e calcolo l'hash del dato concatenato al *nonce* e vado a vedere se il risultato ha un certo numero prefissato di bit più significativi uguali a 0, riduco quindi il codominio, ovvero il numero di possibili output validi, dicendo che devono cominciare con un certo numero di zeri. Aumentando il numero di zeri riduco la probabilità di ottenere un risultato valido scegliendo un *nonce* a caso (e diminuendo ottengo l'opposto). Variando gli zeri ottengo quanto detto sopra in merito al variare del carico computazionale per restare sui 10 minuti.
Quindi, ricapitolando:

- il miner sceglie un migliaio di transazioni più o meno a caso
- il miner spara a caso un valore del *nonce*
- calcola l'hash di tutto il blocco:
  - se ottiene un valore con un numero di bit più significativi uguali a zero accettabile, prima degli altri, manda il blocco nella rete P2P per far validare il nuovo blocco e, se il controllo viene superato, il blocco viene accettato e messo in cima alla blockchain (quindi ogni nodo della P2P lo attacca in cima alla propria copia)
  - se non ottiene tale valore spara a caso un altro *nonce* e ci riprova

Se mentre si sta lavorando un nuovo blocco viene validato (anche dal nodo che stava lavorando al nuovo blocco) un altro blocco bisogna "buttare" quanto costruito anche se non tutto, butto infatti le transazioni che già compaiono nel nuovo blocco convalidato. Inoltre cambio l'hash del nuovo blocco a cui sto lavorando mettendo quello di quello appena convalidato (per il discorso spiegato all'inizio del capitolo).
L'importanza di essere colui che crea il blocco è data dal fatto che l'unico momento in cui si possono creare *bitcoins* è durante la creazione del nuovo blocco, chi aggiunge il blocco ha dei nuovi *bitcoins* che vengono creati insieme al blocco.
Il mining può essere fatto da macchine sempre più performanti (all'inizio bastava un portatile, poi si è passati ad usare i pc di interi uffici di notte (o qualche furbo anche, illegalmente, dell'università), poi si è passati a cluster "home-made" tramite hardware spesso dedicato solo al mining, mentre ora servono cluster estremamente potenti ma che comunque facciano ritornare le spese). Si hanno anche soluzioni di sub-affitto di hardware o di gente che condivide l'hardware per poi dividere i guadagni. A causa di queste *farm* professionali enormi, spesso collocate in paesi freddi per il risparmio del raffreddamento e dove la corrente costa meno (ovvero in zone disabitate di paesi spesso poveri e poco democratici), rischia di danneggiare l'idea di decentralizzazione della blockchain. **Bitcoin** ad un certo punto era in mano di pochissime persone (cinesi) con farm sparse nei monti asiatici. Si rischiano anche manipolazioni truffaldine del mercato, ad esempio proposte di *trading coi bitcoins* (cose comunque vietate nei mercati regolamentati ma non su quello delle criptovalute, non essendo regolamentato, e quindi se il broker di imbroglia sei fregato).
Un altro problema dell'*accentramento* è il cosiddetto **attacco del 51%**. Dato che ogni miner deve vincere contro tutti gli altri per poter essere quello che ha diritto di mettere il nuovo blocco allora se uno riesce a possedere il 51% dell potenza di calcolo dei miners, controllando il 51% dei nodi della rete P2P, praticamente vince sempre, validando sempre le transazioni, anche se non valide (ovviamente almeno il 51%).
"Tutti contro tutti" è anche molto inefficiente dal punto di vista energetico (la rete P2P che gestisce *bitcoin* consuma più di tutta l'Argentina). Per questo si cercano sempre nuovi algoritmi/alternative per il *proof-of-work*, come ad esempio *proof-of-stake*, dove stake sta per "quantità di soldi", ovvero l'idea in cui solo pochi nodi della P2P fanno il mining. Tali nodi vengono scelti in base alla quantità di soldi che ogni nodo ha deciso di rendere disponibili agli altri (su un conto speciale). Quindi chi mette più soldi ha più possibilità di essere scelto.
Al massimo si ha che si potranno costruire 21 milioni di *bitcoins* e questi sono

sempre più difficili da creare (è quindi paragonabile all'oro da un certo punto di vista).
Torniamo a parlare delle transazioni.
Nel caso in cui A voglia dare a B una certa porzione di *bitcoins* deve creare due transazioni:

- una in cui dal totale produce la frazione complementare a quella che deve dare a B, questa transazione è verso se stessi stessi
- una in cui da a B la frazione voluta

In poche parole se A ha 10 *bitcoins* deve fare una transazione a se stessa di 9 e una di 1 a B.
Questo può essere fatto anche avendo più indirizzi di partenza in cui si hanno i *bitcoins*, grazie al fatto che una transazione ha più input, e verso più destinatari, contemporaneamente, in quanto la transazione ammette anche più output nella sua struttura. **La somma totale degli input non deve essere minore a quella in output** (infatti la transazione non verrà validata). Può comunque essere maggiore in quanto la differenza, detta *fee*, può essere un compenso per il miner per far scegliere quella transazione da validare (che quindi non sceglie proprio a caso il migliaio di transazioni in quanto ordina le transazioni in base alle *fee*). Queste *fee* saranno l'unica fonte di guadagno per i miner quando non si potrà più produrre *bitcoins* per il limite visto precedentemente (in quel momento le *fee* aumenteranno di valore probabilmente).
Due blocchi possono essere aggiunti contemporaneamente perché magari due miner in luoghi distanti propagano insieme l'avviso di aver trovato un nuovo blocco valido. Può succedere quindi che vengano aggiunti entrambi anche se uno è destinato a diventare *orfano*.
Proporre una transazione con *fee* pari a 0 può portare al fatto che nessun miner la prenda in carico, comportando la creazione di *transazioni zombie* (che sono parecchie).
Un altro problema di *bitcoin* è che per essere sicuri che una transazione si andata a buon fine devo aspettare 6 blocchi, quindi un'ora di tempo, e questo limita i casi d'uso della moneta (di certo non va bene per prendere il caffè). Il numero di transazioni supportate è comunque minimo rispetto a quelle che si possono effettuare con la moneta fisica.

4.2 Ethereum

Il futuro delle blockchains non ha potenzialmente limiti. Ci sono molte applicazioni e molte varianti, una di queste è **Ethereum**.

Listing 1 Esempio di contratto in Solidity tratto dalla documentazione di Solidity `https://solidity.readthedocs.io/en/v0.7.4/`

Ethereum è un altro tipo di blockchain in cui si pone l'attenzione non tanto sulle transazioni quanto sulle **computazioni**. Si ha un modello diverso di blockchain in quanto vengono memorizzate i cosiddetti *ethereum account*. Ogni utente che si registra quindi viene quindi associato ad un account con una certa quantità di criptovaluta, che in questo caso è chiamata **ether (*ETH*)**. Le transazioni avvengono più o meno come in un conto in banca, togliendo *ether* ad uno e dandoli all'altro mentre la validazione delle transazioni avviene quasi come in *bitcoin* anche se si sta cercando di passare alla *proof-of-stake*. La novità di *ethereum* è la possibilità di scrivere/programmare i cosiddetti **smart contract**, ovvero il corrispettivo dei contratti tra persone, dove, per esempio, un utente, per un tot guadagno al mese, concede l'uso di una sua proprietà ad un altro utente. Si usa un linguaggio di programmazione chiamato **Solidity**, simile ad un linguaggio OOP dove al posto degli oggetti si hanno i **contratti** (ma si hanno comunque struct, function etc...). I contratti scritti con Solidity vengono compilati in un bytecode che viene memorizzato sulla blockchain di ethereum (e quindi una copia viene salvata su tutti i nodi della rete P2P). Si possono fare transazioni che trasferiscono soldi oppure posso fare transazioni in cui si invocano funzioni poste all'interno di un certo contratto, in questo caso un miner raccoglierà la richiesta ed eseguirà sulla sua macchina il codice del contratto. Ogni esecuzione di ogni singola operazione del bytecode (che viene eseguito sulla *ethereum virtual machine*) costa una certa quantità di **gas** e quindi bisogna dare una certa quantità di soldi, rappresentati il costo del **gas**, per far eseguire il contratto e, qualora non siano sufficienti, viene sollevata l'eccezione `outOfGas`, il contratto non va a buon fine e i soldi finora spesi vanno persi.
Gli *smart contract* sono codice di cui chiunque può vedere il bytecode e sono una sorta di codice lato server e posso quindi fare dei client che gestiscono le parti "delicate" dei trasferimenti di soldi facendo chiamate ai contratti, che sono elaborati dai miner. La scrittura dei contratti è rischiosa quindi si hanno quindi delle comunità (come *OpenZeppelin*) che controllano i contratti stessi e forniscono delle linee guida e delle librerie da cui attingere, ma nel momento in cui vengono modificati non sono più testati, ovviamente. Un esempio è stato di una startup che ha lasciato una moltiplicazione non protetta da overflow modificando un contratto, preso da *Openzeppelin*, per permettere pagamenti simultanei (banalmente se si doveva dare 10 ETH a due utenti si faceva 10·2 e si controllava che ci fossero effettivamente 20 ETH nel conto prima di effettuare il doppio pagamento). Essendo il bytecode pubblico se

ne sono accorti e qualcuno si è fatto un pagamento a due suoi altri account dando una cifra esorbitante, in modo da scatenare l'overflow, il prodotto per 2 dava 0 a causa dell'overflow e quindi si autorizzava la transazione. Questa cosa non sarebbe stata possibile su *bitcoin* ma la diversa implementazione della blockchain di ethereum rendevano possibile la cosa (e irriconoscibile al miner). Creare quindi *smart contract* è estremamente difficile, dovendo essere estremamente sicuri. Inoltre una volta che il contratto è sulla blockchain ci resta, anche se il creatore può chiedere al contratto stesso di autodistruggersi (nel senso che vengono rese invalide le chiamate alle funzioni di quel contratto).
Ci sono vari strumenti per scrivere contratti, l'ide *Remix* e *MetaMask*, che permette di avere un **wallet** su ethereum e di collegarsi sia alla rete principale che a diverse reti, anche locali sul proprio computer, per testing. Tra le reti di testing principali abbiamo *Rapsten* che cerca di emulare al meglio possibile quella principale (dove gli ETH che trasferisco o pago come **gas** sono finti). In altre reti di test si hanno altri algoritmi di consenso, come per esempio, nel caso della rete *Rinkeby* si ha il **proof-of-autority** (dove chi ha diritto a scrivere una certa informazione lo può fare senza il consenso della rete P2P).
L'iter normale di sviluppo di contratti non prevede in primis l'uso delle reti di test in quanto servirebbe troppo tempo ma si usano strumenti come *Ganache* o *Truffle* che sono strumenti di sviluppo in locale che simulano una rete in locale dove si può testare senza **gas** in modo veloce. Solo dopo si passa alla rete di test e poi alla *main net*, la rete principale.
Esiste anche una libreria chiamata *web3*, scritta in *js*, per permettere alle webapp di connettersi a MetaMask.

4.3 Altre blockchains

Si hanno comunque davvero tante diverse blockchains.
Si ha, ad esempio, *Quorum*, una variante di *Ethereum* in cui si possono usare altri algoritmi di consenso e in cui si possono anche creare dei canali di comunicazione privati tra i nodi della rete P2P.
Si hanno quindi diversi tipi di blockchain e in particolare ci sono le blockchain di tipo **permissioned** e non solo quelle di tipo **permissionless** o **pubbliche**.
per blockchain di tipo *permissioned* si intende che i nodi della rete P2P, che tipicamente sono anche quelli che scrivono informazioni e quindi avviare le transazioni, si conoscono ma eventualmente non si fidano, formando così un **consorzio** in cui i partecipanti sono potenzialmente in conflitto (dove magari

il fallimento di uno è il successo dell'altro). Se le dichiarazione pubbliche siano vere o meno viene deciso con un servizio di *audit* esterno che valuta le dichiarazioni pubbliche. Non c'è quindi ne mining ne *proof-of-work*, magari nemmeno una criptovaluta. La blockchain diventa quindi un punto in cui implementare politiche di trust tra i partecipanti, obbligandosi a vicenda a dichiarare in modo pubblico (e se imbrogli esci dal consorzio perdendone i vantaggi), fattore che viene aiutato dal fatto che è estremamente difficile alterare la blockchain (dovendo rompere una catena e rifarla, imbrogliando sugli hash etc...) rendendo le dichiarazioni praticamente eterne.
Si hanno quindi:

- **permissionless blockchain**, come *bitcoin* o *ethereum*, dove chiunque può scaricarsi il software e diventare miner
- **permissioned blockchain pubblica**, dove possono vedere tutti, come *Quorum*, *EOS* (con una virtual machine basata su *webassembly* e *smart contract* scritti in *C++*), *Hyperledger* (sviluppata da IBM e ospitata dalla *Linux Foundation*, essendo completamente *Open Source*, di tipo modulare, molto complessa ma efficiente, nata per il business con algoritmi di consenso basato sul **problema dei generali bizantini** e che scrive sulla blockchain solo se necessario etc...) etc... Con queste blockchain si perde il discorso della decentralizzazione, non permettendo che partecipi chiunque ma si ha una *certificate authority* che stabiliscono se uno ha il diritto di scrivere o leggere sula blockchain (e si hanno spesso più amministratori che si controllano a vicenda atti a stabilire chi può fare cosa)
- **permissioned blockchain privata**, dove possono vedere solo quelli che scrivono

Grandi aziende, come IBM, Microsoft, SAP etc... che stanno concentrando sullo sviluppo di blockchain ma anche piccole applicazioni, come *CryptoKitties*, basata su Ethereum, che permette di collezionare e crescere "gattini digitali" collezzionabili.
Bisognerebbe inoltre fare un discorso sul rapporto che si ha tra *smart contract* e dispositivi fisici, nonché sull'interpretazione legale degli stessi (e gli avvocati dicono che non sono ne legali ne "smart" per i problemi detti sopra).

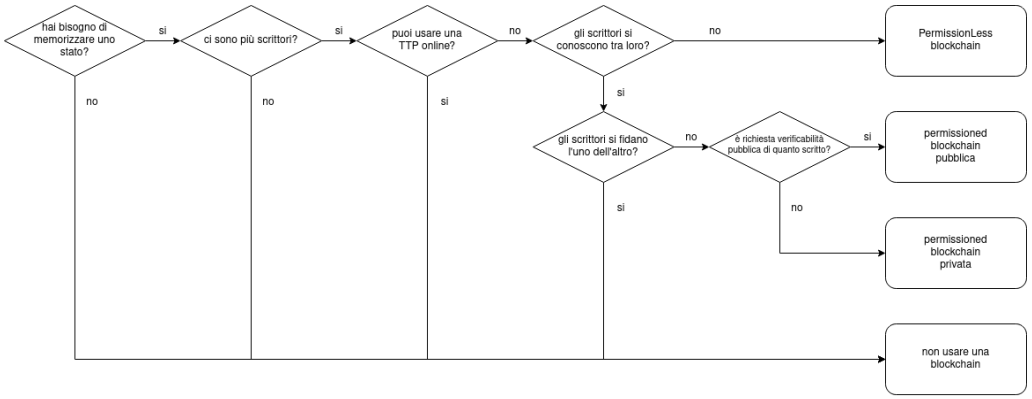


Figura 4.2: Diagramma dei casi di scelta di una blockchain

Vengono qui aggiunte le cose dette in live.
La blockchain potrebbe essere usata per la tracciabilità di prodotti, per evitare *falsi veri* o *veri falsi*. Infatti non è un problema lontano dal **double-spending** (un prodotto viene fatto in numero limitato e quindi non posso avere più di quel numero sul mercato). Per il tracciamento del prodotto associo un hash-code univoco. Metodi come il bar-code non funzionano bene in quanto facilmente duplicabili. Un metodo più moderno prevede l'uso proprio delle blockchain, introducendo anche la tracciabilità dei documenti, permettendo anche di verificare tutta la filiera per arrivare al prodotto finale. Ogni step della filiera è come se avesse un file di log, dove viene segnato tutto quello che viene fatto in quello step. Si deve anche garantire che nessuno step della filiera "bari". Inoltre non si può avere una TTP online (i vari consorzi, gli unici che potrebbero fare da TTP, potrebbero anche loro "barare", non essendo un ente che farebbe solo da TTP) e si ha che gli scrittori "non si fidano" tra loro. Abbiamo quindi tutti i requisiti per una blockchain, come spiegato immagine 4.2. Mantenere una blockchain comunque costa (più di un ipotetico timbro univoco da apporre al prodotto, tramite i concetti di *lotto di produzione*) ma ha più senso in un contesto multi-aziendale.
Ovviamente si ha anche la difficoltà culturale di far interagire piccole imprese con concetti complessi come la *proof-of-work*. Un ragionamento simile si applica sia al mondo del cibo che della moda, che ad altri ambiti.
Il tracciamento dei clienti non è fattibile causa GDPR, in quanto non si può facilmente cancellare i dati (cosa garantita dalla GDPR) nella blockchain, che infatti non è *GDPR-compatible* (e potrei non inserire i dati in chiaro ma cifrati, dimenticando la chiave nel caso il cliente voglia essere dimenticato, in quanto potenzialmente prima o poi sarebbe facile "bucare" i dati).

Capitolo 5

NoSQL

Dopo aver trattato la gestione dei dati distribuita in **ambiente relazionale**, dopo aver trattato le blockchain anche da punto di vista di gestione dati in **ambiente sicuro**, affrontiamo i sistemi **NoSQL**.
Nel 1970 un ricercatore di IBM chiamato Codd pubblica un articolo intitolato *A relational model of data for large shared data banks* in cui presenta un modello di rappresentazione dati indipendente dall'implementazione. Si era passati dal **modello gerarchico** (in cui bisognava usare i puntatori per accedere ai dati) al **modello relazionale**. Questo passaggio è stato fondamentale per lo sviluppo del mondo dell'informatica, un mondo dove l'hardware dedicato allo storage era estremamente costoso, estremamente poco capiente (nel range di pochissimi megabyte) e logisticamente parecchio ingombrante. Il modello relazionale quindi era studiato in primis per questo contesto, proponendo strategie per la rappresentazione compatta dei dati.
Gli aspetti positivi del modello relazionale si possono comunque riassumere in:

- è un modello molto ben definito, con il **principio della minimizzazione**, in merito allo spazio, e il **principio della closed world assumption**, ovvero tutto quello che l'applicazione deve sapere è nel database e a tal fine si ha anche l'uso del `NULL` *value* per:
  - assenza di informazione sul valore
  - assenza di senso di un certo dato
  - assenza di applicabilità di un certo dato

  Quindi il modello relazionale è una sorta di *modello chiuso* che contiene tutto il necessario ed era un modello ragionevole in un

contesto come quello "pre anni 2000", dove un'azienda tipicamente necessitava solo di informazioni interne all'azienda stessa (non c'erano le informazioni dei *social network* etc... infatti, di fatto, non esiste più la *closed world assumption*)

- si hanno circa 35 e più anni di sviluppo su *sicurezza, ottimizzazione* e *standardizzazione*. Da questo punto di vista si ricorda l'uso delle proprietà ACID che comunque rendono ancora valido il modello anche nel 2020. Inoltre ormai si hanno talmente tante informazioni memorizzate tramite il *modello relazionale* che le aziende sono restie a cambiare modello (a causa dei rischi di perdita dei dati durante un porting, anche dati dal fatto che i db preesistenti sono enormi, considerando che i dati persi non sono recuperabili potenzialmente, a meno di repliche)

- è ben conosciuto, è studiato anche nelle scuole e nelle università

- è comunque la scelta migliore per diversi casi d'uso

Si hanno però anche delle limitazioni per il modello relazione, come ad esempio:

- il fatto stesso di essere un **sistema chiuso**, quasi per gli stessi motivi descritti nei vantaggi, può anche essere un aspetto negativo, infatti:

  - il **principio di minimizzazione**, ai giorni nostri, va a discapito, inutilmente, delle performance
  - il **principio della closed world assumption**, come già detto, non è praticamente più valido nell'era dei social network etc...
  - il modello relazionale prevede per un attributo uno e un solo valore e non sono quindi ammissibili array o comunque altre strutture per dati con multipli valori e non si possono fare interrogazioni in modo strutturato all'interno dei valori stessi, dovendo ricorrere a diversi *workaround* e, secondo "il manuale", il metodo standard è quello di fare una tabella di join, detta *tabella di part-of*, ovvero una tabella a parte coi possibili multi valori (questa soluzione porta ad operazioni di join molto dispendiose)

  - non è compatibile con i linguaggi di programmazione moderni. Molti linguaggi moderni sono infatti OOP, con la *object persistency*, paradigma non supportato dal modello relazionale. Si hanno quindi framework (come *Spring*) per far comunicare, tramite l'**object relational mapping** il modello relazionale e gli oggetti della OOP (aggiungendo complicazioni e perdite di performance). Database ad oggetti sono stati creati e implementati nei principali vendors ma non sono usati nelle aziende per varie ragioni, anche di performance e di impatto rispetto ai db relazionali
  - non supportano i **comportamenti ciclici**, comportando limiti modellistici

- i database relazionali sono difficili da modificare dal punto di vista schematico tramite `alter table`. Cambiare gli schemi è complesso e comporta il fermo del database

- per garantire le proprietà ACID nei sistemi transazionali e per poter garantire il 2PC si ha un limite fisico alla **scalabilità**. Dopo un certo punto di crescita i costi di organizzazione rendono impossibile l'esecuzione dei protocolli che garantiscono ACID, in ambiente distribuito. Oggigiorno ci sono varie applicazioni, legate in primis ai social network, per le quali è impossibile mantenere certi protocolli. Il problema della scalabilità è uno degli aspetti più limitanti, specialmente in ottica **big data** (anche se vedremo che le alternative non relazionali coprono casi d'uso anche oltre il solo ambito dei *big data*). I database relazionali si prestano bene alle **scale up (*scalabilità verticale*)** (semplicemente aumentando l'hardware a disposizione, arrivando a costi esorbitanti e fino al punto in cui si raggiunge il limite tecnologico) ma pochissimo allo **scale out (*scalabilità orizzontale*)** (che comporterebbe non l'acquisto di un nuovo hardware intero ma solo di una parte di esso, a prezzo più basso con hardware detto in gergo **comodity**, magari per avere più dischi etc..., ma hardware dedicato storicamente ai db relazionali non prevede questa cosa). Oggigiorno si hanno sempre più spesso le **architetture a microservizi** (che "implementano" bene lo *scale out*)

- il **time in market**, ovvero il tempo in cui un applicativo resta in commercio, è sempre più ridotto (spesso, per esempio, un videogioco ha una vita anche inferiore ad un anno) portando a rendere

obsoleto il processo in cui normalmente si integrano i db relazionali. Anche il *time to market*, ovvero il tempo di preparazione si è ridotto

Oggigiorno lo standard dal punto di vista hardware è drasticamente cambiato con dischi di capienza enorme (nell'ordine dei terabyte per qualche decina di euro). Il rapporto costo-gigabyte è crollato rispetto a pochi anni fa. Il contesto storico è quindi cambiato e in tal senso anche il modello per la rappresentazione dei dati sta cambiando e sta cambiando anche la scelta tra *spazio* e *tempo*, non essendo più la prima una grossa problematica ed essendo la seconda di grande interesse, dovendo rispondere il più in fretta possibile alle interrogazioni.
Sulle slide un elenco dei vari step storici.

5.1 I modelli NoSQL

Il termine **NoSQL** è stato coniato da Carlo Strozzi nel 1998 quando si inventa una sorta di API per Linux per accedere ai dati relazionali senza usare SQL. Il termine è stato ripreso nel 2009 con la "logica" **Not Only SQL** dopo che per circa 9 anni, a partire dal 2000, erano nati nuovi modelli, a grafi, a documenti etc... (da parte di Amazon, Google, Facebook etc...).
Quindi NoSQL è un'insieme di modelli di rappresentazione dati, con relativi software di gestione. Si hanno 3 caratteristiche fondamentali:

1. tutti i modello sono **schema free** o **schemaless**
2. tutti i **DBMS NoSQL** usano il **CAP theorem**
3. si passa da ACID a BASE (da "acido" a "basico") ovvero:
    - *Basic Available*
    - *Soft state*
    - *Eventual consistency*

Schema free

Nel modello relazionale prima veniva definito il modello, ovvero l'insieme degli attributi che descrivevano i dati e le varie relazioni, e poi veniva popolato con i dati veri e propri. Questo iter comportava che se durante lo sviluppo ci si accorgeva che bisognava aggiungere, ad esempio, un attributo bisognava modificare l'intero schema (nonché i dati già caricati).

Con NoSQL si è quindi passati quindi ad un'architettura dove il modello è basato sui dati che vengono inseriti quindi di fatto per ogni "riga" ho potenzialmente uno schema diverso. Questo concetto, detto **schema free** concede una grandissima libertà di sviluppo, al costo di un'alta probabilità di avere inconsistenza tra i dati. Si ha il cosiddetto **open word assumption**, dove i valori `NULL` vengono eventualmente messi solo per motivi applicativi in quanto se un dato non c'è semplicemente non c'è e non bisogna specificare non `NULL` il fatto che non ci sia, garantendo una grande flessibilità (anche se questo può generare problemi). Tale flessibilità ha comunque un costo anche dal punto di vista delle query, in quanto non si può partire dallo schema ma si deve partire da frammenti del modello in studio. Si ha comunque il guadagno che in caso di aggiunta di attributi etc... non bisogna modificare l'intero schema, aggiungo dove serve e basta, anche se ovviamente serve una *governance* del dato più forte (in quanto può anche succedere che non si abbia il tipo di dato).

CAP theorem

Vediamo questo teorema essenziale[1], proposto da Gilber e Lynch nel 2002 dopo che, qualche anno prima, nel 2000, Brewer aveva proposto una sua *congettura sulla consistenza, disponibilità e partizionamento dei webserver*. Il teorema dimostra la congettura di Brewer:

Teorema 1 (CAP theorem)**.** *Preso un* ***sistema distribuito di nodi*** *si considerino tre aspetti:*

- ***Consistency (consistenza)****, ovvero che tutti i nodi abbaino gli stessi i dati nello stesso istante (che è diverso dalla consistenza di ACID)*
- ***Availability (disponibilità)****, ovvero la garanzia che ogni richiesta ricevuta da un sistema riceva una risposta in ogni caso, sia in caso fallimento che successo (non si deve avere quindi un'attesa infinita)*
- ***Partition tolerance (partizionamento)****, ovvero il sistema deve continuare ad operare comunque anche se si hanno perdite di messaggi o fallimenti di parti del sistema*

Il teorema garantisce che tra queste tre caratteristiche, in un sistema distribuito, se ne possano avere contemporaneamente due soddisfatte e mai tre.

[1] così L.T. è contento

Si nota quindi come, prendendo ad esempio i DDBMS, si rinunci alla *partition tolerance*, garantendo le prime due caratteristiche (anche se con un numero limitato di nodi si può arrivare ad una buona resa anche dal punto di vista della *partition tolerance*).
I sistemi NoSQL tipicamente divisi tra:

- sistemi che garantiscono *consistency* e *partition tolerance* (CP), non potendo garantire che il DBMS lavori 24/7. Tra questi troviamo: *BigTable, Hypertable, HBase, MongoDB, Terrastore, Scalaris, Berkeley DB, MemcacheDB, Redis* ...
- sistemi che garantiscono *availability* e *partition tolerance* (AP), non potendo garantire che i dati siano sempre consistenti. Tra questi troviamo: *Dynamo, Voldemort, Tokyo Cabinet, KAI, Cassandra, SimpleDB, CouchDB Riak* ...

Non è detto che due db NoSQL che implementano lo stesso modello, ad esempio documentale, garantiscano le stesse due caratteristiche CAP.
Quindi in fase progettuale la scelta tra due delle caratteristiche CAP è un punto chiave per poi scegliere eventualmente su che DBMS NoSQL appoggiarsi (o anche partire da un DBMS NoSQL che già si conosce ma essendo consci della caratteristica non garantita).
Ci sono contesti applicativi in cui ognuna delle tre caratteristiche può, a seconda, non essere fondamentale. In ogni caso si ha che:

- *consistency* mi garantisce che tutti i client hanno sempre la stessa vista sui dati
- *availability* mi garantisce che ogni client può sempre leggere e/o scrivere
- *partition tolerance* mi garantisce che il sistema continui a funzionare nonostante partizioni fisiche della rete

In fase progettuale quindi ho 3 step:

- scelta del modello
- scelta del DBMS
- scelta delle caratteristiche del CAP theorem

BASE

Analizziamo meglio l'acronimo.

- *Basic Available*, ovvero, comunque vada, la risposta viene sempre data anche nel caso di consistenza parziale
- *Soft state*, ovvero si rinuncia al concetto di consistenza di ACID
- *Eventual consistency*, ovvero, ad un certo punto prima o poi nel futuro (magari anche dopo pochissimi secondi), i dati "convergeranno" ad uno stato di consistenza (in opposizione alla consistenza immediata di ACID)

Vediamo quindi i vari modelli NoSQL:

- **key-value stores** (come *Dynamo, Voldemort, Redis, Riak ldots*) che è il modello più semplice. È infatti il semplice modello *chiave-valore*, semplice da memorizzare, e utile per il caching. Il valore è un *blob* quindi non si ha modo di interrogare internamente il valore me deve essere l'applicazione a manipolarlo. La chiave garantisce tipicamente meccanismi di hash, con la chiave che punta ad un certo valore, massimizzando le performance. Viene garantito l'*hashing distribuito*, ovvero la *distributed hashed table*
- **column family stores**, (come *BigTable, Cassandra, HBase* ...) dove la chiave punta a colonne multiple di valori, ottenendo un formato *pseudo-relazione*. Ogni riga, in questo schema **wide column**, ha una sua struttura interna diversa a quella di altre righe, che sono comunque indipendenti. La chiave ovviamente può essere hashata. Non posso fare *ranged query*
- **document databases** (come *CouchDB, MongoDB* ...) è formato da documenti con coppie chiave-valore o chiave-oggetto, dove l'oggetto può anche essere un'altra chiave-valore, un array di valori, un array di oggetti etc.... Un esempio pratico di documento è il tipico file json, che è ottimo per dire che nel modello documentale si ha un elemento più importante, detto *root* che ha al di sotto tutti gli altri
- **graph databases** (come *Neo4J, FlockDB, GraphBase, InfoGrip* ...) con nodi e relazioni tra gli stessi rappresentate dagli archi. I nodi possono rappresentare entità a cui sono associate varie proprietà (e si chiamano *property graph*). le proprietà sono quindi pertinenti ai nodi. I graph db non scalano bene

- **RDF databases**

(*Spesso l'azienda stessa non è attenta al tipo di modello usato quindi si hanno spesso confusioni*).
A livello di complessità (anche dei dati che rappresentano) possiamo dire che, seguendo l'ordine elencato, si passa dal più semplice al più complesso. Anche dal punto di vista del volume abbiamo lo stesso ordine, dal più voluminoso (quindi il meno complesso è anche il più voluminoso ma più scalabile) a quello meno voluminoso e scalabile.
Il problema fondamentale in tutti questi modelli è come connettere le informazioni e questo diventa un problema fondamentale in termini di prestazioni e analisi da effettuare. Si ha che le relazioni sono implicitamente incluse nei dati. La scelta del modello dipende dall'applicativo e ci sono applicativi che possono obbligare ad avere due o più modelli.
Ovviamente ogni vendor ha il suo linguaggio di query per il suo DBMS NoSQL.

5.2 Document based system

È uno dei modelli più ricchi nonché più vicini al modello relazionale. Nella sezione approfondiremo anche il più "famoso" DBMS documentale, ovvero **MongoDB**.
Si ha un'evoluzione del modello chiave-valore in quanto si ha una chiave, ovvero un *object identifier*, che può essere indicizzata tramite un *meccanismo di hashing*. I documenti solitamente vengo invece memorizzati in file **json** o **XML** e possono essere cercati a qualsiasi livello.
A differenza del modello relazionale il modello documentale è alla fine un albero e quindi l'accesso ai dati deve comunque partire dal nodo radice, a scendere verso le foglie (anche se si può avere un accesso **a indici** per accedere alle informazioni necessarie). Si ha quindi il concetto di **elemento più importante degli altri**, concetto assente nel modello relazionale, definibile "piatto" da questo punto di vista (a meno delle *data warehouse* dove si ha effettivamente una tabella principale).
La rappresentazione documentale è più vicina al pensiero umano di divisione delle cose in "cartelle" con varie tipologie di informazioni.
Si hanno quindi le seguenti caratteristiche per il modello documentale:

- è multidimensionale, ad albero
- ogni campo può contenere zero valori, un valore, multipli valori o altri documenti

- si possono effettuare query ad ogni campo e in ogni livello
- lo schema è flessibile, infatti due documenti della stessa **collezione** possono avere uno schema diverso di rappresentazione dei dati
- posso aggiungere "in linea" nuove informazioni
- avendo "annegato" uno dentro l'altro le relazioni che sussistono tra gli elementi, sono richiesti molti meno indici per effettuare le query migliorando le performance in quanto le informazioni sono legate tra loro nel documento (e non dovendo, come nel caso relazionale, prendere n tabelle in n file diversi nel *filesystem* facendo poi il join)

Il modello documentale permette due tecniche principali:

- **referencing**, ovvero, come per il modello relazione, si relazionano due documenti tramite un certo attributo, ad esempio:

- **embedding**, ovvero, aggiungere quello che nel *referencing* è un secondo documento come specifica di un certo valore, aggiungendo quindi un livello di profondità (quindi, per esempio, posso avere un documento relativo ad un contatto telefonico contenente anche indirizzo di casa e tale informazione, con i vari attributi, sarà specificata non in un altro documento ma nello stesso), Aumentando lo spazio che però non è più un problema. Ad esempio si ha:

A causa del *time to/in market* ormai sempre più ogni applicazione ha il suo db e questo obbliga la flessibilità dello schema, che garantisce che una modifica allo schema, con aggiunta o modifica degli attributi, non è un problema (mentre con il modello relazionale sarebbe stato un problema). Questa flessibilità però comporta che in fase di query non ho uno schema prefissato tramite il quale interrogare (e non posso nemmeno vedere i primi n documenti in quanto l'$n + 1$ potrebbe essere comunque diverso). Questa problematica è relativa all'uso di database costruiti da terzi.
Si hanno diversi linguaggi di query, tra cui *jsoniq, jmespath, jaql, JLINQ, etc...*, non avendo più quindi un unico linguaggio come era SQL per i modelli relazionali, ma avendone uno per vendor.
Dal punto di vista delle terminologia si hanno le seguenti "relazioni" con il modello relazione:

SQL	NoSQL
database	database
tabella	collezione
riga	documento
colonna	chiave

5.2.1 MongoDB

MongoDB è un DBMS non relazionale documentale con i dati memorizzati in un modello chiamato **binary json (*Bson*)**, che consente maggior efficienza tramite una rappresentazione binaria. Tramite l'embedding dei dati consente di non eseguire i *join* e permette anche l'accesso tramite indici. Supporta comunque il referencing anche se con cali di performances.
Dal punto di vista architetturale si hanno una serie di *engine* diversi fra loro (attualmente si usa **WiredTiger (*WT*)**, dopo acquisto di una società che si occupava di db relazionali distribuiti con transazioni, mentre prima si usava *MMAP V1*). Grazie a WT si è introdotto, nella v4, il supporto alle transazioni. Si ha poi il *data model*, il *MongoDB query language (MQL)* (proprietario), i componenti di controllo degli accessi, il meccanismo di gestione etc.... Le repliche sono organizzate in **replicaSet**.
Nel caso di MongoDB si ha:

RDBMS	MongoDB
database	database
tabella/view	collezione
riga	documento (Bson)
colonna	campo
indice	indice
join	documento con embedding
chiave esterna	referencing
partizione	shard

Vediamo quindi un esempio di schema in Mongodb:

Listing 2 Esempio di aggiunta di un documento p nella collezione *posts*

Analizzando il codice abbiamo:

- la definizione di p come un oggetto json
- tra gli elementi abbiamo *tags* che contiene un array di valori

- tra gli elementi abbiamo *comments* che è un array di documenti
- l'ultima riga è tra le chiavi di successo di Mongodb in quanto ha un linguaggio simile a quelli di programmazione. Nel dettaglio si ha il caricamento del documento p nella collezione *posts*. Si ha un meccanismo simile alla OOP e l'integrazione coi linguaggi di programmazione è facilitata dall'ormai molto diffuso supporto ai json (come esistono già meccanismi per il passaggio da oggetti a json e viceversa)

Si possono inoltre fare indici, tramite `ensureindex` (che prende come argomento un json), a ogni campo del documento, per esempio:

Si nota che posso creare indici a livello che voglio, a documenti o array, o anche, come nell'ultimo esempio, usare indici geo-spaziali.
MongoDB, rispetto al modello relazionale, si avvicina tanto alla programmazione.

Repliche in MongoDB

Innanzitutto si ha che MongoDB rientra nella categoria CP rispetto al *CAP theorem*, non garantendo disponibilità in caso di guasto.
La replica viene effettuata in modalità master-slave anche se, nel gergo di MongoDB si ha **primary-secondary**, avendo scrittura sul primary e replica sui secondary. Si ha in realtà che uno dei nodi replica viene definito a posteriori *primary*. La replica avviene tramite il file di log del *primary*. Si ha un sistema che procede per repliche parziali incrementali, all'inizio il nodo fa un sync, prendendo i dati dal primary e poi prende dal file di log del primary solo gli ultimi n movimenti.
La garanzia della tolleranza si ha tramite il meccanismo di elezione dei nodi, nel caso il nodo primary vada down.
Si ha che un nodo può assumere uno dei seguenti stati:

- STARTUP: un nodo che non è membro effettivo di nessun replica set
- PRIMARY: l'unico nodo nel replica set che accetterà le operazioni di lettura
- SECONDARY: un nodo che di effettuare la sola replicazione dei dati (può essere promosso a PRIMARY in fase di elezione)

- RECOVERING: un nodo che sta effettuando una qualche operazione di recupero dei dati, magari dopo un *rollback* (può essere promosso a PRIMARY in fase di elezione)
- STARTUP2: un nodo che è appena entrato nel set (può essere eletto a PRIMARY)
- ARBITER: un nodo non atto alla replicazione dei dati con lo scopo di prendere parte alle elezioni per promuovere i nodi a PRIMARY (può essere eletto a PRIMARY in fase di elezione)
- DOWN/OFFLINE: un nodo irraggiungibile
- ROLLBACK: un nodo che sta svolgendo un rollbackper cui sarà inutilizzabile per le letture (può essere promosso a PRIMARY in fase di elezione)

L'elezione elegge il nodo con la versione più aggiornata dei dati e in questa fase le scritture sono interrotte (motivo per cui non è garantita la A di CAP). L'elezione dura comunque pochissimi.
Durante il processo di elezione ogni *replicaSet* manda, con timeout per assumere che un nodo sia down, un "hearbeat" (ovvero un bit ogni due secondi) agli altri nodi. Se il nodo primario risulta down si procede con l'elezione, che in termini umani è di tipo *maggioritario con liste bloccate* nel replicaSet. Ogni nodo ha un identificativo con un punteggio rappresentante la priorità nelle elezioni (più alto più probabile che venga eletto). Quindi appena il primary va down il secondario con valore più alto chiamerà le elezioni. Si ha che un replica set può avere al massimo 50 membri di cui solo 7 votanti, che esprimono un solo voto. I nodi SECONDARY non votanti hanno priorità 0 e accettano solo letture ma sono comunque in grado di votare. Nel caso di partizionamento to della rete si rischia un problema di riconciliazione tra vecchio primary e nuovo, dovendo quindi bloccare la scrittura.

Scrittura

La scrittura avviene sul nodo primario e si hanno diverse politiche in base tramite l'opzione *writeConcern* che dice cosa fare per le replicaSet. Si hanno tre parametri:

1. w, indicante il numero di nodi in cui il dato deve essere replicato prima di essere considerata conclusa l'operazione. Ovvero è il numero di nodi sui quali deve essere confermata la scrittura per essere confermata a livello globale

2. j, ovvero l'intenzione di scrivere sul log di *Wiredtiger*, tramite la logica *write ahead logging* (ovvero prima scrivo sul file di log e poi sul disco), prima che l'operazione sia stata eseguita. Questo per garantire la persistenza del dato, avendolo comunque nel file di log. Senza l'opzione lo scriverà in un secondo momento

3. *wtimeout*, ovvero il tempo limite, espresso in ms, da aspettare nel caso in cui $w > 0$ (se 0 non devo aspettare risposta da nessuno)

Aumentare w comporta aumentare il tempo di esecuzione. Nel dettaglio:

- $w = 0$ non da alcuna certezza di inserimento, utile per operazioni veloci
- $w = 1$ implica la scrittura sul nodo primary
- $w = n$ implica la scrittura su n nodi
- $w = majority$ dove almeno la metà dei nodi più 1 deve aver scritto prima della conferma dell'operazione (nella realtà si tenta di scrivere su più nodi ma appena si ha un ack dalla metà più 1 si conferma la scrittura anche se magari si sta ancora scrivendo su altri nodi)

Qualora si debba caricare una gran quantità di dati, di cardinalità conosciuta, per motivi di efficienza non conviene aspettare che il nodo dia l'ack ma si caricano tutti subito sul primary, controllando poi con `count` se tutto è stato caricato.

Frammentazione

Parliamo ora di frammentazione e scalabilità.
Si hanno 3 tipi di sharding per la scalabilità orizzontale:

- *hash-based*, quindi sulla chiave
- *range-based*, quindi in base ad un certo range di elementi
- *tag-aware*, quindi in base a certi attributi

Si ha un meccanismo dinamico basato su *pay as you go*, ovvero nel momento in cui ho lo sharding sarà MongoDB ad adattarlo nel momento in cui cambia il volume di dati, facendo un *bilanciamento automatico*. Si ha uno "spazio" della mia chiave di shard che posso dividere in n chunks per $n + 1$ chiavi di shard. Il numero massimo di chunk che è possibile definire su una sharded key può definire il numero massimo di shard in un sistema.
Si hanno 3 ruoli fondamentali nella frammentazione:

1. **mongos**, che è il punto di accesso software per le query, è l'entry point del cluster, tutte le query verranno eseguite su questo processo. Appena lanciato mongos va a leggere il config server
2. **config server**, ovvero un file che conosce la struttura degli shard, conoscendo frammentazioni e repliche. Dalla versione 4.0 ogni nodo distribuito è anche replicato
3. i vari **shard**, ovvero istanze di MongoDB

Quindi il router quando riceve una query capisce dove sono i frammenti, grazie al config server, e quindi procede con la query di tipo:

- **target query** se deve interrogare solo un nodo
- **broadcast query** se deve interrogare più nodi

Sia gli shard che il config server sono replicati su replicaSet per evitare che diventino *single point of failure* (anche per questo dalla 4.0 ogni volta che frammento replico).
Si ha un **primary shard**, che contiene l'intero db (ovvero tutte le collezioni) e i **secondary shard** che contengono frammenti e repliche di alcune collezioni.
Solitamente si ha quindi una configurazione a 3 nodi:

1. uno primary
2. due nodi secondari con frammenti diversi e le repliche del primary

Letture

Ho una politica di *readConcern* (che deve essere consistente come del caso del *writeConcern*, facendo scegliere se è più dispendioso scrivere o leggere):

- *local*, che è il valore di default t per leggere i dati dai nodi in un replicaSet. La query è fatta secondo la località spaziale e quindi se legge un nodo non primary rischio di leggere dati non ancora replicati
- *available*, di default per lettura su nodi secondari, che fa vedere che il dato c'è ed è consistente
- *majority*, quando almeno metà più 1 repliche ha ricevuto un ack in scrittura sul dato allora posso leggere

Vediamo quindi un esempio di interrogazione, tramite il metodo `find`:

db.user.find(age:$gt:18,name:1,address:1).limits(5)

che equivale a:

$gt significa *greater than* e si hanno vari identificativi per i confronti (per non avere i simboli $>$, $<$ etc... che darebbero fastidio nelle stringhe).
Vediamo anche una *insert*:

con `insertOne` che specifica che voglio inserire un solo documento mentre avrei `insertAll` per tutti i documenti.
Un *update* avviene con la stessa logica con `updateAll`:

$upsert sarebbe un *insert if not exist*, per permettere l'inserimento e non la modifica in caso di assenza. Tale cosa è assente in SQL ma servirebbe un doppio comando.
Si hanno varie utility per importare i dati, da specificare come opzioni al comando `mongoimport`, ad esempio:

mongoimport –d database –c collection –type csv -ignore blanks –file PATH

per importare un csv al path PATH ignorando gli spazi bianchi.

Transazioni

Dalla versione 4.0 MongoDB ha introdotto nel modello documentale le transazione tramite l'engine *Wiredtiger*. Introduce quindi il concetto di *log* che indica che tutto è nel nodo primary. Si opera sempre sul nodo facendo un *lock* che può essere di tipo:

- **sharded (*S*)**, per le letture
- **exclusive (*X*)**, per le scritture
- **intent sharded (*IS*)**, per esprimere l'intenzione di bloccare in modo condiviso uno dei nodi che discende dal nodo corrente
- **intent exclusive (*IX*)**, per esprimere l'intenzione di bloccare in modo esclusivo uno dei nodi che discende da quello corrente

Il lock garantisce transazioni a livello di singola collezione mentre prima l'atomicità era garantita a livello della singola operazione. Usando l'embedding, anche prima della v4.0, potevo gestire un sacco di modiche della singola collezioni. Nel caso del referencing invece serviva una gestione più approfondita delle transazioni. Essendo WT creato originariamente per un sistema distribuito funziona solo per sistemi distribuiti, infatti avendo un solo nodo si presuppone che si abbia una sola applicazione che accede, cosa che renderebbe a priori più semplice la gestione.

Modello contro efficienza

esempio a partire dalla slide 37
Spesso il cambiamento del modello può comportare un miglioramento delle prestazioni non indifferente, più del cambiamento dell'hardware (**verrà approfondito in live**).
Avere più piccoli documenti separati migliora il tempo, rendendolo lineare (mentre un grosso documento embedding aveva tempi esponenziali nell'inserimento delle righe).
Con questa soluzione poi devo sistemare le query.
Quindi l'importanza del modello è decisiva dal punto di vista delle prestazioni.
Nelle slide brevissima parte su CouchDB inutile.

Vengono qui aggiunte le cose dette in live.
Nella storia di NoSQL compaiono sempre grandi aziende, come Google, Amazon o Facebook che creano il software, fanno il paper e rendono il software open. Il rilascio dei codice in modo open source viene fatto in quanto in primis la community permette continui miglioramenti "a costo zero", inoltre le aziende in questione guadagno in primis su pubblicità (Google e Facebook in primis) e cloud (vedisi Amazon che fa il 90% degli utili dal cloud). Nessuno di questi vende direttamente tecnologia (anche perché, esempio, Android ha un valore economico basso) ma la usano. Si ha quindi esigenza applicativa, non affidandosi a società esterne ma facendosi le cose internamente, rilasciando poi le cose open source per far abituare gli utenti alle proprie tecnologie, facendo così in modo che la gente voglia usare quelle tecnologie in ambito cloud. In tutto ciò si contribuisce a creare pubblicità mirate etc... in quanto il vero "bene" sono i dati non l'hardware o le tecnologie (e quelli non sono open). Si ha la **data-driven economy**, i dati sono il nuovo petrolio. IBM, Oracle etc... che vendevano invece la tecnologia ora sono più in crisi e hanno dovuto cambiare *business*.
Per progettare una review di prodotti servono informazioni in merito a chi fa la review (utente verificato, numero di vecchie review etc...). In ogni elemento del documento salvo queste informazioni. Non è poi possibile avere tutte le review nello stesso documento, quindi le primissime review (che vengono visualizzate subito) le lascio nel documento principale e per le altre lascio una reference ad un altro documento. Le review possono anche essere usate per valutare l'affidabilità dell'utente ed eventuali scelte alternative dei prodotti. Si punta comunque all'efficienze e all'efficacia.
I db documentali scalano tramite **distributed hashed table (*DHT*)**. Defi-

nendo una funzione di hash definisco implicitamente il dominio della funzione, posso quindi distribuire in modo efficiente sui nodi i vari valori di hash. Posso quindi avere diversi server con un nodo che salva le coppie chiave-valore con tutte le chiavi di sua responsabilità, è il **chord ring**. Ogni nodo quindi sa che dati deve gestire tramite la funzione di hash, avendo un algoritmo di scelta del nodo tramite *finger table*, con complessità logaritmica. L'aggiunta di un nodo può portare a buchi di sicurezza ma permette una scalabilità orizzontale incredibile.
sistemare questa parte!

5.3 GraphDB

È il modello più ricco e complesso tra i NoSQL.
Un grafo, da un certo punto di vista, ha proprietà simili a quelle di un modello ER. È infatti un insieme di nodi e di relazioni tra loro, dove i concetti più importanti (quelle che in ER sarebbero le *entità*) sono modellate come nodi. Bisogna quindi studiare come connettere i nodi, infatti il modo con cui le entità sono semanticamente associate si modellano come relazioni, tramite archi.
Si hanno campi di applicazione dei db a grafo:

- social network
- sistemi di raccomandazione, che suggeriscono oggetti simili a utenti simili
- ambito geografico
- reti logistiche (per spedizioni etc...)
- transazioni finanziarie per la *fraud detenction* (riconoscendo certi pattern sospetti, anche a livello geografico)
- master data management, rappresentazione unica della realtà
- bioinformatica
- controllo di autorizzazioni e accessi

Esempio 5. *Per esempio su Twitter avrei i nodi con gli utenti e gli archi rappresentanti la relazione follow.*

Si usa una **descrizione estensionale dei dati**. Si può avere infatti:

- **descrizione estensionale dei dati** quando si descrivono le istanze via-via, descrivendo anche istanze di tipo diverso (magari utenti e messaggi) usando la stessa sintassi del nodo (ugualmente per gli archi). La semantica viene data dall'etichetta
- **descrizione intensionale dei dati**, che bisognerebbe inferirla direttamente dai dati e quindi non è applicabile

Nei RDBMS si aveva il *join* basato su valori, facendo uno *scheme understanding* sulle varie tabelle, ottenendo quindi il risultato tramite anche più operazioni di *join*. Si può avere però il problema applicativo, avendo magari **relazioni ad anello** tra le varie tabelle ad esempio una tabella "persona" che tramite una tabella "amici" ritorna su se stessa, avendo la tabella "amici" che esplode in numerosità (magari cercando amici di amici etc... complicando le cose di livello in livello). Avendo potenzialmente livelli infiniti di *join* si va incontro al **join bombing**, che fanno soffrire i DBMS. Si può considerare comunque il **six separation degree**, uno studio sociologico, che dice che con sette connessioni si può raggiungere chiunque (ora coi social ne bastano meno). Simile è il **Kevin Bacon number**, che dice il numero medio di film fatti tra artisti per ottenerne uno fatto con Kevin Bacon, anche in questo caso si arriva allo stesso risultato visto con il **six separation degree**.
Nei sistemi a grafo abbiamo quindi gli archi mentre nei documentali avevamo identificatori che puntavano ad altri documenti (con un modello basato sui valori come se fossero *foreign keys* dei RDBMS), comportando, nel caso dell'embedding, task di modellazione complessi, scarsa efficienza e complessità nel gestire dati fortemente connessi, mentre, con il referencing starei simulando una sorta di modello a grafo.
I db a grafo hanno forte potenza espressiva permettendo di caratterizzare i nodi con un certo *tipo* e definire le varie relazioni tramite archi. Coi grafi si possono facilmente rappresentare relazioni umane, sociali, affettive, lavorative etc... usando dei semplici archi. Si possono connettere elementi tramite un'infinità di tipi di relazioni.
Nei nodi posso rappresentare anche entità complesse, con vari attributi, ovvero varie **proprietà** (ovviamente ogni nodo può avere un set di attributi diversi, avendo fortissima libertà di schema).
È un modello molto più immediato rispetto ad un RDBMS. Inoltre, nel caso si abbiano potenzialmente più relazioni tra due nodi basta contare quanti cammini orientati (che possono quindi passare per nodi intermedi diversi) esistono tra quei nodi per ottenere la cardinalità di tali relazioni.
Un grafo con le proprietà nei nodi è detto **property graph** È un modello

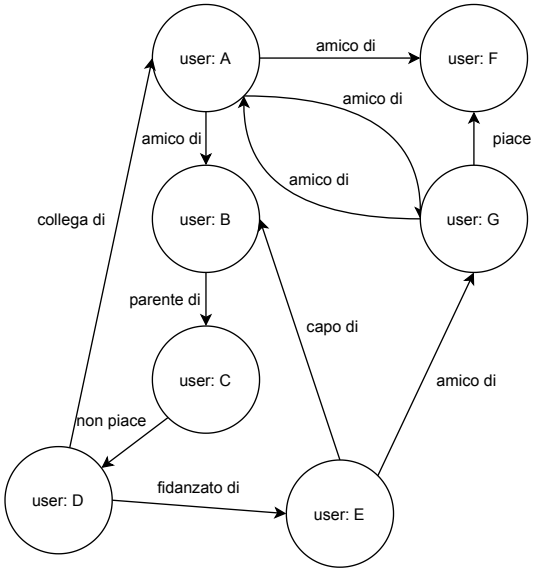


Figura 5.2: Esempio di db a grafo

Figura 5.3: Esempio di db con rappresentato da un **property graph**, con *user* come chiave e un array di valori

simile a quello che si fa su una lavagna, è una modellazione simile a quella umana, basti pensare alle mappe concettuali.
Si hanno comunque vari tipi di *property graph*;

- quando sia nodi che relazioni possono avere proprietà espresse da chiave-valore, accettando quindi che i nodi abbiano un tipo
- altri ammettono array di valori

Si nota che invece in altri modelli NoSQL come RDF i nodi possono avere un tipo ma non proprietà e le relazioni non hanno proprietà, dovendo sempre rappresentare il tutto tramite triple "soggetto-predicato-oggetto".
Il problema di modellazione dei grafi si ritrova quindi nella scelta di cosa mettere come nodo e cosa come relazioni (ed eventualmente le proprietà), avendo varie scelte modellistiche con vari e pro e contro.
Ovviamente anche qui ogni vendor ha il suo linguaggio, anche se ne esiste uno *standard*, chiamato **gremlin**, che più che essere un linguaggio di interrogazione è un *linguaggio di attraversamento dei grafi* (dato che cercare sottografi e attraversare il grafo è la principale tecnica di studio degli stessi che si ha a disposizione). Questo tipo di linguaggio viene detto **vertex-based** che si usano per studiare la validità di predicati tramite l'attraversamento dei nodi.
Si hanno vari vendor quindi e si hanno due tipi di approccio:

1. **graph processing**, dove ci si concentra sul processamento dei dati, processamento che viene eseguito tramite i grafi
2. **graph storage**, dove i dati vengono memorizzati in forma di grafo

In entrambi i casi la cosa può essere fatta in modo:

- **nativo**:
  - *native graph storage* per indicare che sono ottimizzati e progettati per le gestione stessa dei grafi. In questo caso comunque i dati vengono memorizzati in modo intelligente, memorizzando in modo vicino i nodi e il loro archi, facendo *index free adjacency*, in modo che seguire gli archi uscenti da un nodo è molto più immediato. Tra gli esempi di db abbiamo **Neo4j**, che in memoria salva, per ogni nodo, tutti i suoi riferimenti, tramite pointers. In scrittura comunque si perdono performance, a vantaggio di una lettura molto rapida (al costo comunque di molto uso di RAM). Si evita il *join bombing* seguendo i riferimenti vicini ad ogni nodo, facendo l'attraversamento

- **non nativo**:
  - *non native graph storage* per indicare che i dati vengono trasformati da grafo ad un formato diverso tra cui relazionale, object oriented db etc... Si rischia comunque il *join bombing*

Potrei avere un db con una rappresentazione a tabelle (e non a grafo) studiata come se fosse un grafo (traducendo poi le query in SQL), questo è l'approccio di **FlockDB**.
Un'alternativa è avere, come nel caso di **Titan**, dove posso comunque usare *Gremlin* ma con dati memorizzati in modo colonnare.
Si hanno quindi due approcci principali per gestire un grafo:

- **graph database**, ovvero un DBMS che gestisce in maniera persistente un grafo che viene usato per le transazioni. Si ha in questo caso la dicitura OLTP
- **graph compute engine**, ovvero tecnologie per l'analisi *off line* dei grafi. Si ha in questo caso la dicitura OLAP

I graph database supportano le operazioni CRUD (come tutti i modelli NoSQL):

- Create
- Read
- Update
- Delete

Come abbiamo visto si studiano gli attraversamenti del grafo, che sono equiparati ai *join* del modello relazionale, per fare le query in modo performante non facendo join ma attraversando e basta. Sono inoltre costruiti per essere usati nei sistemi transazionali (OLTP).
Si ha un problema grave coi grafi, motivo per cui si hanno i *non native graph storage*: **i grafi non scalano bene e non possono essere facilmente frammentati**.
Partizionare un grafo raramente è possibile, in quanto comporta un taglio di relazioni. Se quindi si vuole scalare bisogna passare, alla fine, ad un altro modello, aggiungendo però costo di comunicazione tra il modello a grafo e il nuovo modello scelto, aggiungendo un layer software.

5.3.1 Neo4j

Abbiamo già introdotto precedentemente Neo4j. Sappiamo che è nativo sia per il *graph processing* che per il *graph storage* (sovvrendo quindi la scalabilità).
Approfondiamo quindi il **linguaggio di interrogazione di Neo4j**, chiamato **Cypher** (*OpenCypher* nella versione open).
Cypher è definito come un **pattern-matching query language**. È un linguaggio simil-umano, espressivo, dichiarativo (si dice il cosa e non il come), che consente operazioni di aggregazione, sorting e di limit (trovare ad esempio le n top key etc...) e permette di aggiornare il grafo.
Un formato in modo testuale viene rappresentato con le parentesi tonde per i nodi e le quadre per gli archi. Le relazioni orientate vengono indicate tramite "->" o "<-".

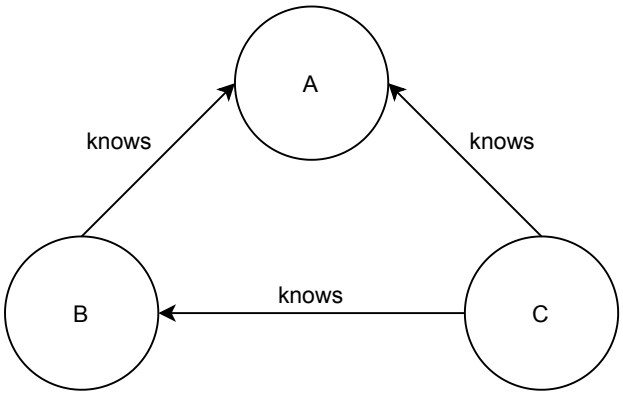
Esempio 6. *Tramite:*

(c)-[:KNOWS]->(b)-[:KNOWS]->(a), (c)-[:KNOWS]->(a)

con la "," che significa $\wedge$.
Equivalentemente potevamo avere:

/c)-[:KNOWS]->(b)-[:KNOWS]->(a)<-(c)-[:KNOWS]

per rappresentare:



Vediamo quindi un'interrogazione classica in Cypher, dove si cercano tutti gli utenti tali per cui conosce qualcuno a che conosce qualcuno b e che sono etichettati come "Michael", ritornando a e b:

Il risultato sarà una ***tabella*** *che contiene tutte le coppie a e b. Avrà dei dati ridondati a causa della dualità intrinseca di avere a che conosce b, cosa che probabilmente porta ad avere che b conosce a.*

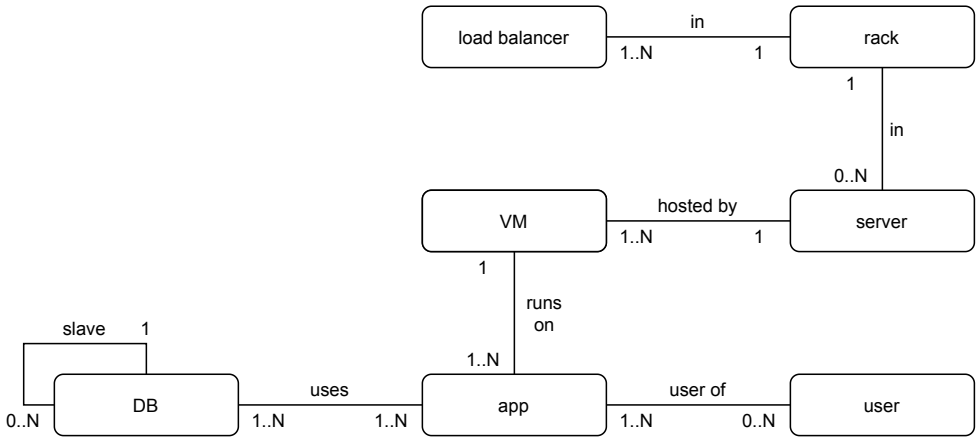
Quindi il risultato **non sempre** è un grafo (mentre nel documentale era sempre un documento e nel relazionale una tabella) ma può anche essere altro, ad esempio una tabella. Non è quindi sempre possibile concatenare query in quanto potrei non avere un grafo in uscita.
Si hanno comunque clausole (coprendo le operazioni CRUD):

- *where* per imporre criteri di filtraggio tramite pattern-matching dei risultati
- *create/create unique*, per creare nodi e relazioni
- *delete*, per cancellare nodi, relazioni e proprietà
- *set*, per impostare dei valori alle proprietà
- *foreach*, per fare un update per elementi del grafo posti in una lista (per ogni nodo, per ogni arco etc...)
- *union*, per unire i risultati più query
- *with*, per concatenare query propagando i risultati in pipe

5.3.2 Grafi e modello relazionale

Compariamo ora il modello a grafi con il modello relazionale.
Per farlo sfruttiamo un esempio:

Esempio 7. *Vogliamo costruire un modello relazionale per la gestione di una server farm. Si ha che l'utente accede ad una applicazione che gira su una VM ed accede ad un db (primario o secondario). I server che ospitano le VM sono installati su sistemi rack controllati da un sistema di load balancing.*
Una tipica attività è quella di controllare se ogni elemento del sistema funzioni e, in caso di guasto, capire cosa non stia funzionando. Per farlo si hanno dei sensori, sia software che hardware, che monitorano le varie componenti.
Modello uno schema del mio sistema collegando in una mappa concettuale le varie componenti che comunicano tra loro o che sono in relazioni (anche solo per dire che un server è in un certo rack). Si modellano così le varie istanze.
Tale modellazione relazionale può avere il seguente ER:



Ogni elemento ha l'attributo on per indicare se i sensori dicono che la componente funzioni o meno, attributo che verrà usato per le query.
Qualora quindi un'utente mi dica che una certa applicazione non funziona devo risalire step by step al server, eventualmente al rack, al db etc. . ., avendo una quantità di join non indifferente (6 join diversi nel caso peggiore).
Una rappresentazione a grafo sarebbe molto più semplice, avendo dei semplici cammini anziché avere tutti quei join. Si avrebbe quindi un nodo per singola componente reale con le relazioni tra le varie singole componenti.
Pensiamo quindi, per esempio, alla seguente query:

Trovare gli asset (server, app, machine virtuali) in stato down per l'utente user 3

Parto quindi dal nodo che come nome "user 3", trova tutti quegli asset con distanza da 1 a 5 rispetto al nodo utente, arrivando quindi fino a 5 livelli di profondità. Se tale asset è in down viene restituito una sola volta.

5.3.3 Ancora su Neo4j

Tutta la teoria e la letteratura sui grafi viene applicata ai db a grafo, dai cammini minimi alle clique. Il vero problema resta la modellazione e la scelta del DBMS. Approfondiamo quindi ancora di più Neo4j.
Abbiamo già visto che non ha meccanismi nativi di frammentazione, avendo scelto di sfruttare la *index free adjacency* e la modellazione a grafo, non permettendo scalabilità. In compenso garantisce varie caratteristiche funzionali:

- permette le transazioni con tutte le proprietà ACID, con meccanismi di locking

- garantisce la disponibilità
- garantisce la recoverability

In merito alla scalabilità, nonostante le limitazioni sopra esposte, garantisce l'**alta disponibilità**. L'idea è che è possibile andare a definire un cluster di nodi Neo4j, alcuni chiamati **core server**, ovvero nodi per cui tutte le scritture sono fatte in contemporanea, essendo quindi sincronizzati. Si hanno poi gli altri nodi che sono chiamati **replica server**, che sono di sola lettura, che vengono creati tramite meccanismi di replica simili a quelli già trattati. I *replica server* vengono anche usati per reporting analysis etc.... Fatta una query essa comunque non viene distribuita ma eseguita sul nodo più vicino. Tutto questo sistema è detto **casual clustering**.
Si ha una bassa **latenza** nelle query in quanto si usa un indice per risalire al primo nodo e poi si procede coi cammini, avendo quindi che le performances non dipendono dalla grandezza del dataset ma solo da quali sono i dati richiesti dalla query.

5.4 Modelli poliglotta

In situazioni reali, ad esempio un e-commerce, posso avere i dati salvati in vari DBMS, sia relazionali che NoSQL, a seconda dei dati che contengono (magari avrei le raccomandazioni in un db a grafo con Neo4j, il catalogo prodotti in un documentale come MongoDB, le sessioni utente in un db key-value come Redis, dati finanziari e reporting in un RDBSM, per i log delle attività un modello NoSQL colonnare etc...). Dover gestire tutti questi modelli diversi è molto completo, ma sensato in base al tipo di dati. Ho quindi una sorta di **mondo poliglotta**.

5.4.1 ArangoDB

In questo mondo poliglotta complesso ci arriva in soccorso un particolare DBMS: **ArangoDB**.
Il DBA dovrebbe conoscere tutti i modelli in uso e tutti i linguaggi associati, essendo anche in grado di gestire i vari conflitti. ArangoDB introduce il concetto di **db poliglotta**, supportando i seguenti modelli:

- key-value
- documentali
- a grafo

- a ricerca full-text (come fa anche **Elastic Search**)

Rispetto ai modelli precedentemente descritti a grafo con storage non nativo, tipo **Titan**, che appunto accettano solo un linguaggio di query a grafo, il linguaggio di query di ArangoDB, chiamato **AQL (*Arango Query language*)**, garantisce la possibilità di interrogare dati in formato json (quindi documentale), a grafo, in formato key-value o per il full text search.
Ad alto livello Arango organizza i dati in database e collezioni (per fare un parallelo schemi e tabelle), dove le collezioni memorizzano documenti o similari. I documenti, come già accennato, sono documenti json di profondità arbitraria (sub-object, sub-arrays etc...), non avendo alcun limite da questo punto di vista. Nel caso semplice tali documenti sono omogenei, avendo stessi attributi e tipi, e effettuare una query AQL è molto semplice:

Avendo un linguaggio simile a uno di programmazione.
Abbiamo quindi, nella prima query, che selezioniamo tutti i *c* in '*categories* (che è una collezione) trovando solo le chiavi che sono o "book" o "kitchen" ritornando *c*. Nella seconda selezioniamo selezioniamo tutti i *c* in '*categories*, ordiniamo i risultati in base al titolo e ritorniamo chiave e titolo.
Posso comunque avere documenti eterogenei, con attributi diversi. Possiamo comunque interrogare usando il modello a grafo e tali query mostrano quali documenti sono collegati (direttamente o indirettamente) ad altri documenti, specificando il cammino per ottenere il collegamento. Queste connessioni tra documenti sono chiamate **edges** che sono salvati in "edge collections". Gli *edges* hanno sempre due attributi che si riferisocno ai nodi entranti e uscenti:

- *_from*
- *_to*

Avendo sempre archi direzionati del tipo:

$$\_from \rightarrow \_to$$

anche se si possono effettuare query anche nell'ordine opposto.

Esempio 8. *Preso il documento json:*
*posso avere un **edge**, per una certa relazione, del tipo:*
(Potrei comunque continuare a interrogare il primo json come se fosse un documentale)

Rispetto a quanto visto per i modelli a grafo cambia drasticamente la scalabilità.
Si hanno vari metodi di scalabilità, avendo le modalità di deploy:

- a singola istanza
- master/slave, scritture su master e repliche su slave
- active failover
- cluster
- multiple datacenters

Approfondiamo soprattutto gli ultimi tre.

Active failover

Nel modello master/slave quando il master fallisce lo slave non può essere sostituito e quindi si è pensato al modello **active failover**, dove esistono delle istanze *single-server*, chiamate **leader**, che sono in lettura e scrittura e delle istanze, sempre *single-server*, chiamate **followers**, che non sono altro i vecchi slave, passivi e non scrivibili. Il leader invia le operazioni di scrittura ai followers che replicano in modo asincrono i dati tramite la cosiddetta **Write-Ahead Log (*WAL*)**, ovvero il leader prima salva su un file di log e poi sul disco e tale file di log, tramite replica incrementale viene aggiornato, tramite un meccanismo di replica chiamato **apache mesos**. Esiste poi almeno un **agency** che controlla le configurazioni di leader e followers e controlla che siano tutti disponibili. Qualora l'agency si accorgesse che un server leader non è più disponibile promuove, con un meccanismo elettivo, un follower a leader. Resta quindi una sorta di master/slave ma con gli agency che procedono con il controllo e la rielezione.

Cluster

Il cluster è un'evoluzione, avendo più **coordinator** che si connettono con i client. I coordinator effettuando le query e gli **agency** non si occupano di altro se non l'elezione degli stessi e della sincronizzazione dei servizi per l'intero cluster. Si hanno poi i **DB server**, di tipo slave, che, in caso di fallimento, non comporta problemi in quanto l'agency attua le solite politiche di aumento repliche, se esse sono previste, bilanciando nuovamente i dati in caso di sharding.
Lo sharding avviene in partizioni da almeno 40gb, con un meccanismo simile a quello di MongoDB: ogni server avrà associato uno shard, e ogni server, avendo ciascuno circa 3 repliche, avrà le repliche di altri server (oltre alla sua stessa replica, di cui è **leading shard**) (??????). Questo sistema permette

di interrogare quasi direttamente tutto sullo stesso server (avendo molte repliche) e permette, in caso di server non funzionante, di poter stabilire che un altro server diventa il **leading shard** per un certo frammento (tramite meccanismi soliti di elezione tramite gli agency), replicando per avere comune le tre repliche tra i vari server e potendo continuare a fare interrogazioni. Si hanno due soluzioni per la scrittura:

1. il protocollo **read once write all (ROWA)** che obbliga il coordinator a scrivere su tutte le repliche
2. il coordinatore scrive solo dove si ha la copia primaria e poi si hanno le repliche, è la soluzione **oneshard** (che era chiamato **write concern** in MongoDB)

Arango quindi scala molto bene.

Cluster to cluster

L'ultimo livello è il **cluster to cluster**, una soluzione asincrona di tipo **one way** in cui prendo un cluster e lo copio interamente in un altro, usando una coda come **kafka**. La replica è one way per cui scarico dal primary alla replica o viceversa in tempi diversi, non contemporaneamente. Non è pensato per fare una replica del single-server (userei il master/slave).

5.5 Modello key-value

È un modello abbastanza semplice dove si hanno appunto rapporti chiave-valore. I valori sono considerati come dei blob, non essendo accessibili/interpretabili direttamente dal db(??????). Si hanno poche operazioni:

- inserire un valore data la chiave
- trovare un valore data la chiave
- modificare un valore data la chiave
- cancellare una chiave e il suo valore

È inoltre possibile associare, in alcune soluzioni come **Redis**, un tipo ad un valore, per esempio *string, integer, list etc...*.
Si ha accesso veloce ai dati tramite strutture di hash e posso avere scalabilità orizzontale. I modelli key-value funzionano solitamente *in memory*, anche

nel caso della condivisione dati.
Dal punto di vista della nomenclatura si ha:

relazionale	key-value
tabella	bucket
riga	key-value
id-riga	key

Tra i vari DBMS abbiamo:

- Redis
- Memcached
- Riak KV
- Hazelcast
- Ehcache

5.5.1 Redis

nato del 2009 da Salvatore Sanfilippo mentre stava lavorando su una soluzione di realtime web analytic con mysql. Nasce a causa dei limiti del modello relazionale, non riuscendo a fare le interrogazioni in realtime. Assunto da VMware continua a lavorare su Redis fino a mettesi in proprio: *Redislab* (con la parte di ricerca e sviluppo in Sicilia e la sede principale negli USA).
Redis può gestire chiavi espresse in ASCII. I valori primitivi sono le *string* ma si hanno vari container di stringhe:

- hashes
- lists
- sets
- sorted sets

Non si può sempre fare una ricerca per valore all'interno di questi elementi.
Sono stati poi aggiunti i **moduli** per fari formati di file, per memorizzare file *json* ma anche per *video in streaming*. Offre quindi blob più strutturati per memorizzare i dati, alcuni a pagamento.
Si ha una versione distribuita enterprise, a pagamento. Si possono fare repliche e avere una soluzione in cluster con un certo numero di shard.

Studiamo quindi come avviene lo sharding.
Bisogna stabilire una *policy* per capire come dividere i dati e si è scelto di dividere le chiavi in base al valore di hash. Si può, volendo, anche dividere le chiavi in base a delle RegEx.
Inoltre bisogna capire dove distribuire i dati e si hanno due soluzioni:

1. quella di default, detta **dense**, dove il meccanismo di sharding è gestito tramite un meccanismo incrementale, ovvero quando si riempie la memoria della macchina si crea un'altro shard su un'altra e così via

2. una detta **sparse**, dove, avendo già definito il numero di nodi, vado a popolarli fin da subito

Si ha l'alta disponibilità in base a due soluzioni (???):

1. una soluzione, detta **Redis replica of setup** in cui si hanno dei nodi che formano il cluster su un datacenter su cui si effettuando le operazioni e altri nodi che formano altri cluster su altri datacenter verso cui sincronizzo. È quindi un master/slave, replicando per disaster-recovery

2. una soluzione, chiamata **active-active setup**, in cui posso scrivere e leggere sui cluster presenti in due datacenters diversi (ma magari nello stesso rack, parlando di **Redis cluster with rack zone awareness**) avendo in entrambi i datacenter sia i db principali che le repliche. Magari avendo protocolli ROWA etc...

Di recente si è aggiunta la possibilità di rendere persistenti dati anche se il loro punto di forza resta la gestione *in memory* (piuttosto che andare a costruire dei sistemi di *cache* etc...).

5.6 Wide column

Studiamo ora il modello **wide column store**.
Anche se storicamente non è così il modello *wide column* può essere visto come un'evoluzione del modello *key-value* (è nato prima il modello *wide column*). L'idea è che accanto ha la chiave non ho un valore/blob ma una riga con degli attributi mono-valore. Si ha quindi una via di mezzo tra il modello *key-value*, essendo possibile fare interrogazioni su queste righe e quindi all'interno delle colonne, e il modello documentale, non riuscendo a garantire la stessa potenza descrittiva (i documenti sono comunque più espressivi essendo più

strutturati, permettendo sub-array e sub-document).
Tra le implementazioni troviamo:

- **BigTable**, il primo modello *wide column* introdotto da Google
- **Hbase**
- **Cassandra**, che però viene ritenuto spesso *key-value* e anche loro stessi si definiscono *key-value*, basando la loro distribuzione sulle **distributed hashed table (*DHT*)**

5.6.1 BigTable

L'idea dietro **BigTable** è quella di una mappa multidimensionale ordinata, persistente e sparsa, ovvero si ha una mappa indicizzata tramite tre informazioni:

1. row key, di tipo *string*
2. column key, di tipo *string*
3. timestamp che caratterizzano ogni singolo valore, compreso la column key (da immagine sembra così perlomeno). Sono di tipi *int64*

All'interno si hanno poi varie **column family** con i vari dati, memorizzati come *string*.
BigTable è molto usato in ambito web grazie alla sua capacità di memorizzare contenuti HTML etc...
Grazie ai timestamp si garantisce un controllo di concorrenza basato sul multiversioning. Possiamo dire che i sistemi che implementano *wide column* con timestamp supportano transazioni ACID compliant tramite **MultiVersion Concurrency Control (*MCC o MVCC*)**.
Un insieme di righe viene definito **tablet**. Abbiamo visto che ogni riga può avere diversi **column family** e si ha che ciascun column family può avere diversi **qualifier** e per ogni qualifier più valori diversi. La libertà dello schema deriva che ogni column family può avere un numero diverso di qualifier. Si ha quindi l'approccio *key-value* per l'accesso tramite chiave e si ha un sistema di indicizzazione per migliorare le prestazioni. Ogni column family ha un nome, una *string*, è può contenere altre colonne, che a loro volta appartengono ad una column family, specificata tramite `familyName:columnName`. Un esempio nella figura 5.4. Non è semplice facile le range query a causa del meccanismo di hashing.
Il timestamp (anche se può essere usato altro) garantisce un **version number** univoco.

row key	timestamp	column "column1"	column "column2"	
"key 1"	t_{12}	"<html>..."		
	t_{11}	"<html>..."		
	t_{10}		"column2:cnnsi.com"	"CNN"
"key 2"	t_{15}		"column2:apache.com"	"APACHE"
	t_{13}		"column2:my.look.ca"	"CNN.com"
	t_6	"<html>..."		
	t_5	"<html>..."		
	t_3	"<html>..."		

Figura 5.4: Esempio di modello *wide column* implementato da BigTable

5.6.2 Hbase

La versione distribuita di BigTable venne rinominata **Hbase** (dove *H* sta per **Hadoop**).
In questo caso i dati sono divisi in tables e ogni table è composta di colonne che sono a loro volta raggruppate in column family (che possono avere un numero variabile di colonne, che erano i qualifier in BigTable). Ovviamente supporta il multiversioning.
Un'istanza di Hbase è formata da varie righe, identificate tramite una chiave, a cui sono associate, per ciascuna riga, varie column family, una per ogni tipo di dato. La column family viene rappresentata tramite un insieme di coppie attributo-valore e i valori possono avere timestamp diversi, specificati tramite `@timestamp` (`@ts=value`).
Si possono avere le stesse column family, dal punto di vista del nome, in più righe ma queste possono avere un insieme diverso di valori, garantendo che la table è **sparse**, rivelandosi utile per il mapping "uno a tanti". I dati sono tutti *bytes*. Un esempio nella figura 5.5. Il modello colonnare non presta attenzione ai *join*, che in questo caso implicherebbero scorrere l'intera tabella, ma si concentra sul **denormalizzare** ovvero a replicare le informazioni nel momento in cui si crea una column family nuova (???).
Quindi lo schema è formato dalle tables e dalle column family, **le colonne non fanno parte dello schema**.
Hbase implementa quindi le **dynamic Columns** in quanto ogni riga ha

row key	data
A	info: {'altezza': `1.8m', 'stato':'italia} ruolo: {'fiat': 'direttore', 'alfa': 'co-fondatore'}
B	info: {'altezza': `1.75m', 'stato':'italia} ruolo: {'fiat': 'direttore'@ts=2010, 'fiat': 'amministratore delegato@ts=2011, 'lancia': 'consulente'}

Figura 5.5: Esempio di istanza di Htable, dove si nota quanto detto, avendo le stesse column family ("info" e "ruolo") ma con all'interno anche attributi diversi (esempio "consulente")

differenti colonne, i quali nomi sono codificati nella riga, non facendo parte dello schema.
Il versioning è essenziale per garantire lo storico delle informazioni.
SU Hbase ho alcune operazioni fondamentali:

1. aggiungere valori
2. ricercare per valore esatto (e non in modo range-based)

Vediamo quindi come si distribuisce un cluster Hbase. Si hanno tre componenti principali (ricordando che si basa su Hadoop), avendo un'architettura master/slave:

1. un master, detto **HbaseMaster**
2. tanti region server, detti **HRegionServer**, che sono i vari slave
3. il client, detto **Hbase client**, che dialoga con il singolo master

Nel dettaglio una *region* è un sottoinsieme di righe della tabella, che vengono partizionate orizzontalmente tramite regole standard di frammentazione (partizionamento che viene fatto in automatico). Quindi un *HRegionServer* gestisce un insieme di righe di una o più tabelle e la sincronizzazione (lettura e scrittura) avviene tramite file di log. Il master coordina gli slave, assegna le region, riconosce fallimenti degli slave etc. . ..
Il client quindi fa la richiesta al master il quale cerca il dato e sviluppa l'operazione all'interno della singola region. Nel momento in cui il region server scrive i dati, attraverso il sistema distribuito chiamato **Hadoop Distributed File System (*HDFS*)**, viene fatta la replica e avviene la sincronizzazione dei dati. Potrei avere più master coordinati tra loro. Il coordinamento

tra i master e i region server avviene tramite il meccanismo di **ZooKeeper**, ovvero un orchestratore software. HDFS permette di avere la memorizzazione orizzontale dei dati e tramite il **write ahead log (*WAL*)** prima scrivo sul log e poi sul disco, garantendo la ricostruibilità dei dati in caso di guasto e garantendo che prima o poi i dati saranno allineati. ZooKeeper quindi coordina master e region server per lettura/scrittura mentre HDFS consente l'allineamento e la memorizzazione su un file system distribuito.
Hbase ha un linguaggio di query poco ricco, potendo fare solo una *select* di dati con un certo valore. Si può comunque fare la ricerca sia sulle righe che su alcuni attributi (definiti precedente come "chiave"). Il sistema di ricerca è estremamente performante.

5.6.3 Cassandra

Come abbiamo detto **Cassandra** lo studiamo come modello *wide column* anche se loro stessi si definiscono *key-value*, essendo sostanzialmente tale ma usando concetti dei modelli *wide column*.
Cassandra è l'evoluzione open source del db NoSQL utilizzato da Facebook e risulta interessante perché implementa un sistema di distribuzione completamente diverso da quello master/slave di Hbase.
In cassandra esiste il concetto **column family**, definito all'interno di un **key space**, come insieme di coppie key-value e quindi una column family altro non è che una tabella con le coppie key-value come righe. Il concetto di column family quindi varia rispetto a quello di Hbase. Una riga è una collezione di colonne con un nome (sono quindi come le righe di una tabella relazionale). La chiave corrisponde al nome della colonna e ogni riga (che è appunto usa serie di coppie key-value definite come colonne) contiene almeno una colonna. Un esempio di modellazione è visibile alla figura 5.6. Si possono aggiungere attributi a piacere e modificarli (mentre in Hbase si creavano/modificavano le column family). Se i valori corrispondenti a tutte le column name sono uguali a "-" si specifica che essi non sono usati. Posso anche specificare di avere un singolo valore non in uso sempre con "-".

	name	id	time
key1	A	1	t_1

	name	id	time
key2	-	2	t_2

Figura 5.6: Esempio di due righe della stessa column family di Cassandra con due chiavi e tre colonne (con i tre column name “nome”, “id”, “time”)

Il linguaggio di query di cassandra è chiamato **Cassandra Query language (CQL)** molto simile ad SQL. Vediamo un esempio di creazione di tabella (dove per ogni attributo si specifica il tipo), con la specifica di una primary key:

(si specifica che i campi di tipo timestamp sono specificati in millisecondi).
Per intenerimento dati si procede in modo simile, anche qui, ad SQL:

Inutile dire che anche per le query si ha un modello simile ad SQL dove le *select* leggono uno o più records dalle column family di Cassandra ritornando un insieme di righe:

La particolarità di Cassandra è l’architettura.
Se BigTable usava:

- Column families
- Memtables
- SSTables

e altri DBMS come **Amazon Dynamo** dove si ha (???):

- hashing consistente per mantenere la replica
- partizionamento e replica
- routing *one-hop*

in Cassandra i ha un’architettura distribuita *Peer-to-Peer* che si basa sostanzialmente su una modalità simile a **distributed hash tables (*DHT*)**.

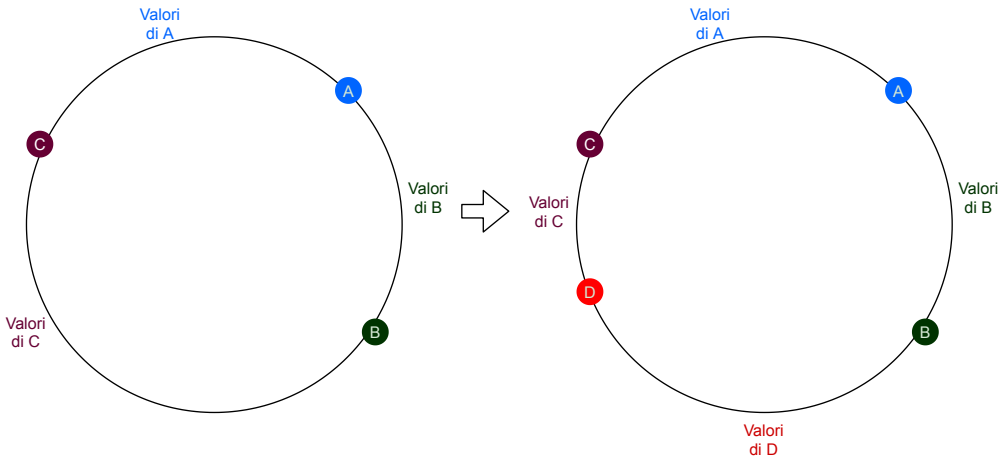


Figura 5.7: Esempio di inserimento di un nodo ("D") nell'architettura ad anello che viene posto, tramite lo studio sull'hash della chiave, tra il nodo "B" e il nodo "C". Con "valori di X" si intende che quei dati si trovano nel nodo "X"

Cassandra implementa un meccanismo in cui si garantisce l'**elasticità** in modo **trasparente**, aggiungendo i vari nodi in modo molto efficace e performante. Per farlo si prende lo spazio di indirizzamento delle chiavi e si assegna l'organizzazione prendendo lo spazio di indirizzamento, dividendolo nei nodi disponibili. Tramite ZooKeeper si garantisce anche in questo caso la ridondanza, supportando anche la **rack awareness** (dati possono essere replicati tra diversi rack per "proteggersi" da guasti di macchine/rack). Inoltre non si ha un **single point of failure**.
Analizzando nel dettaglio il partizionamento si ha che i nodi sono divisi in modo logico in una forma ad anello, detta **ring topology**. Si usano quindi i valori hash delle chiavi, associate alle partizioni dati, per assegnare la chiave e i dati collegati ad un nodo dell'anello. L'hashing viene limitato ad un certo valore massimo per permettere la struttura ad anello (banalmente quello che si potrebbe fare anche con una semplice operazione *mod*) e si ha che i nodi con meno carico si "spostano" sull'anello per alleggerire quelli con più carico (inserendo un nodo in modo da "spezzare" i dati in carico ad un certo nodo). L'aggiunta di un nodo influisce solo sul nodo precedente e il nodo successivo. Spesso i nodi vengono replicati almeno 3 volte e viene fatto nei primi 3 nodi successivi al nodo di cui si sta facendo la replica (nell'esempio sopra, ipotizzando voler per semplicità fare solo due repliche di "A", essere sarebbero su "B" e "d"). Le repliche di un nodo vengono fatte ad ogni scrittura dello stesso. Sapendo che ogni nodo monitora il successivo, qualora questo sia guasto, si può chiedere a quello successivo ancora, garantendo che i dati non vengono

persi, il quale a sua volta replicherà su un nodo successivo all'ultimo sui cui aveva repliche per avere sempre le 3 repliche.
Per monitorare i guasti si usano i **protocolli gossip**. L'idea è appunto quella che un nodo chiede al successivo e al successivo ancora se sia vada tutto bene, ottenendo, di passaggio in passaggio, una conoscenza completa e robusta dello stato del cluster (ricordando che essendo una rete *Peer-to-Peer* non si ha alcun master). In caso di failure si procede alla riscrittura dei dati sul nodo successivo tramite il file di log.
Le operazioni di write sono quindi **write ahead log**, prima scrivo quindi sul **commit log** e poi sulla tabella persistente.
Parliamo quindi di consistenza dei dati.
Cassandra nel suo linguaggio offre la **read/write consistency** avendo delle politiche per cui si hanno letture/scritture consistenti:

- read consistency: la lettura è data per certa se un nodo risponde, politica **ONE to ALL** o, altrimenti, se un certo numero di nodi è d'accordo prima di ritornare il risultato della lettura

- write consistency: la scrittura è data per certa se un certo numero di nodi viene aggiornato prima di considerare la scrittura valida, politica **ANY to ALL**, oppure se almeno un nodo mi risponde o anche se tutti i nodi rispondono

Si ha inoltre una politica di **quorum** per il calcolo della "maggioranza" (per la quale anche possono avvenire read/write consistency) che si basa sul **replication_factor**. La richiesta di consistenza viene fatta nelle query:

Dove per esempio si specifica, con `USING CONSISTENCY ONE` che basta la conferma di un nodo per validare l'operazione. Potrei usare poi `USING CONSISTENCY QUORUM`, che si basa sul valore k del *replication_factor*, che stabilisce che servono almeno k nodi che validino l'operazione. Con `USING CONSISTENCY ALL` specifico che tutti i nodi devono confermare l'operazione. Il problema è che se chiedo un solo nodo che verifichi l'operazione rischio di avere inconsistenza, avendo i nodi precedenti che non sono stati magari in grado di comunicare a tale nodo di aver ottenuto i dati.
Posso avere poi `USING CONSISTENCY ANY`, specificando anche il booleano *hinted_handoff_enable*, che autorizza la scrittura anche se si ha il nodo offline in quanto un'altro nodo prende la richiesta in carico e, non appena il nodo originale torna accessibile, ne trasferisce i dati, per procedere dopo la validazione del primo nodo che risponde.
In merito alle operazioni di *delete* si ha che esse semplicemente rendono il dato non disponibile, essendo più veloce cambiare un flag piuttosto che cancellare. A causa di ciò comunque si creano nelle tabelle fisiche dei problemi di

spazio, quindi periodicamente si fanno dei **merge**, ovvero ogni singolo nodo procede alla "compattazione" dei propri dati, sovrascrivendo i valori non più disponibili.
Per assicurare la sincronizzazione dei nodi e per evitare perdita di consistenza si utilizzano dei *checksum* per comparare i dati di un nodo con quelli dei successivi. Per effettuare questo controllo, nel dettaglio, vengono usati degli hash-tree detti **Merkle tree** nel seguente modo:

- vengono mandati degli snapshot dei dati ai nodi successivi
- creati e trasmetti ad ogni "compattazione" principale (non lo dice ma c'è scritto questo nella slide ????)
- se due nodi prendono uno snapshot con un intervallo pari a `TREE_STORE_TIMEOUT` allora i due snapshot sono comparati e in caso di successo i dati vengono sincronizzati

Vediamo nel dettaglio le operazioni di lettura.
Le letture inconsistenti vengono fatte usando `ANY`, posso poi specificare che mi risponda un nodo con `ONE` etc. . ., come per le scritture.
Quindi in lettura, i nodi vengono interrogati fino a quando il numero di nodi che rispondono con il valore più recente non raggiunge un livello di coerenza specificato da **ONE to ALL**. Si ha quindi che:

- se il livello di consistenza richiesto non viene raggiunto i nodi vengono aggiornati con il valore più recente che viene quindi restituito
- se viene raggiunto il livello di coerenza, il valore viene restituito e tutti i nodi che riportavano valori vecchi vengono aggiornati

Si hanno quindi, per esempio:

5.7 Considerazioni finali

Si è notato come bene o male la scelta di algoritmi di gestione siano alla fine sempre gli stessi, variano invece al più le architetture (master/slave *Peer-to-Peer*) ma i vari modelli NoSQL restano comunque da un certo punto di vista (in merito ad eventuali repliche, distribuzione etc. . .) tutti simili tra loro. La differenza di modello invece consente la scelta di un certo sistema NoSQL in base al tipo di dato che bisogna inserire nel db. Un'altra differenza è il linguaggio usato e la facilità d'uso dello stesso (e delle altre tecnologie legate al DBMS).
La scalabilità è un'altro punto da considerare durante la scelta.

Vengono qui aggiunte le cose dette in live.
DynamoDB, essendo *in memory* e quindi non persistente, viene usato per associare su Amazon ogni persona al suo carrello, che deve essere sempre disponibile mentre navigo nell'e-commerce. Sono altri sistemi che poi manipolano il carrello stesso ma la sola associazione è gestita, in modo estremamente performante, tramite DynamoDB, *in memory*.
Parliamo di **knowledge graph**. Studiati dai creatori di Google nel 1996. Nel 2000 poi Tim Berners-Lee introduce il *semantic web*. Si ha, in un motore di ricerca:

- crawling
- page ranking
- word indexing
- ricerca di istanze prossime a quanto richiesto
- vista dei risultati tramite il ranking

Si è poi iniziato a cercare parole e non fatti, con la conoscenza rappresentata tramite un grafo. La ricerca così più approfondita, riconoscendo cose legate alla ricerca ed eventuali altri risultati simili o di interesse. Si hanno quindi risultati più specifici, con risultati di ricerca arricchiti con studi semantici. Un knowledge graph è solitamente in formato RDF ed è quindi un supporto alla ricerca online. Un nuovo campo di ricerca è l'estrazione di conoscenza anche dal linguaggio naturale, con concetti espressi in formato testuale. Per le query si usano SparQL, Gremlin etc...

Capitolo 6

Architetture di integrazione

L'80% del lavoro nel caso di gestione dati si ritrova nella preparazione degli stessi e nella progettazione. Nella preparazione dei dati si parla anche di **data integration**.
Si hanno diverse possibilità a seconda di carico di lavoro, tecnologie, qualità dei dati etc...
Posso avere (approfonditi più avanti):

- consolidamento dei dati
- gestione distribuita dei dati, senza consolidare ma facendo un'integrazione virtuale, tramite il modello wrapper-moderator

Dal punto di vista organizzativo si hanno vari problemi:

- autonomia, ovvero il grado di indipendenza tra i diversi DBA nelle scelte progettuali
- frequenza delle query
- valore economico dell'integrazione, ovvero la rilevanza per il processo operativo aziendale e decisionale di avere informazioni integrate in input in modo da produrre output efficaci
- volatilità delle fonti nel tempo, ovvero la frequenza di aggiunta o eliminazione di fonti e frequenza di modifica degli schemi delle fonti
- complessità gestionale, ovvero lo sforzo da dedicare alle attività di gestione relative a database e infrastrutture sia hardware che software, a causa della corrispondente complessità delle organizzazioni che utilizzano i db

- costi di eterogeneità, ovvero i costi nascosti ed espliciti relativi ai processi aziendali dovuti all'utilizzo di dati eterogenei

Si hanno anche problemi tecnologici:

- rilevanza dei dati storici e conseguente necessità di memorizzare periodicamente nuovi dati senza cancellare quelli vecchi
- complessità della query, in termini di numero di dati e tabelle visitate e numero di operatori su di esse, e conseguente complessità temporale nell'esecuzione della query
- rilevanza delle query rispetto alle transazioni, ovvero l'importanza relativa e frequenza delle query rispetto alle modifiche dei dati

In base a tutti questi criteri si hanno varie soluzioni (nelle slide tabella).
Con l'introduzione a NoSQL si è visto come sia possibile memorizzare i dati in tanti modi diversi. Ci si chiede quindi come "unire" diversi (per sorgente, formato natura etc...) dataset anche di tipologia diversa, procedendo all'**integrazione** dei dati. Questo processo può avvenire anche su dataset di volume ridotto avendo complessità anche su dati piccoli.
Si parla solitamente di:

- **data integration**, dove si hanno diversi set di dati (anche non disgiunti) con attributi diversi. Si costruisce un unico insieme con tutti gli elementi dei due insiemi, con gli elementi presenti nell'eventuale intersezione con tutti gli attributi dei due insiemi di partenza (mentre gli altri tengono solo gli attributi iniziali). In pratica si fa un *full outer join* (ovvero prendo tutti i valori della tabella di sinistra e li matcho, se posso, con quelli della tabella di destra, prendendo però poi anche tutti gli elementi che non joinano). Il risultato finale potrebbe non avere a priori dell'informazione in più sui dati
- **data enrichment**, dove voglio arricchire i dati di un insieme con quelli di altri insiemi, qualora possibile. In altre parole è una specie di *left outer join*, prendendo tutti gli elementi della tabella di sinistra e aggiungo le informazioni della tabella di destra qualora ci sia un match con la tabella di destra, altrimenti non aggiungo nulla. Si ha un match diverso dalla semplice uguaglianza tra il set che voglio arricchire e quello che contiene i potenziali arricchimenti. Potrei avere quindi uguaglianza, similitudine, appartenenza etc... perché posso rilassare i vincoli di match quando voglio arricchire l'informazione (sfruttando altre relazioni più

leggere, magari basate su informazioni terze, per esempio relazioni geografiche etc...). Il dataset arricchito contiene quindi solo gli elementi del dataset iniziale che volevo arricchire ma questi potrebbero (potrei non essere riuscito ad arricchire tutti i dati) contenere informazioni in più

In entrambi i casi per rappresentare l'assenza di un attributo uso un NULL nella tabella se si parla di un modello relazionale. In caso dei vari modelli NoSQL seguo gli standard dei vari modelli e dei vari DBMS per rappresentare l'assenza di valore per un certo attributo.
Solitamente si cerca di aggiungere, per integrazione o arricchimento:

- informazioni spaziali, con informazioni geografiche dei dati che mancano e quindi vanno aggiunte
- informazioni temporali, con informazioni "storiche" dei dati da integrare/arricchire. Si ha però un problema di capire cosa venga orima e cosa dopo, comportando ambiguità nel rapporto causa-effetto

ma potrebbero esserci infinite altre casistiche.
Si parlerà solo di data integration ma ogni tanto verranno aggiunti incisi in merito all'enrichment.

6.1 Data integration

Immaginiamo di avere due dataset con due schemi, espliciti o meno a seconda che siano basati sul modello relazionale o siano modelli NoSQL (e in quanto caso bisogno inferirli in modo abbastanza preciso vedendo magari i primi documenti etc...), da voler integrare.
Ipotizziamo che i due schemi rappresentino la stessa informazione ma da due punti di vista. Si hanno, per esempio, uno schema *Libro*, con titolo, ISBN e autori (specificati da nome e data di nascita), e uno schema *Autore*, con nome, data di nascita e libri scritti (specificati da titolo e ISBN). Ho quindi due schemi ugualmente informativi ma con diversa rappresentazione (ma stesso modello), rendendo quindi difficile l'integrazione che prevede diverse operazioni. Ovviamente la situazione si complica ancora di più se ho modelli eterogenei, dal punto di vista dei modelli, avendo modelli sia relazionali che non. Bisogna anche scegliere il modello per lo schema integrato finale e non esiste una soluzione ottima a priori ma deve essere scelta in base allo scopo dello schema finale. Tale scelta influenzerà ovviamente le tecniche di costruzione dello schema finale. Per procedere quindi scelgo in primis il modello

dello schema finale e come prima cosa, dopo averlo scelto, converto il modello di partenza di modello diverso in quel modello, ottenendo quindi omogeneità, dal punto di vista dei modelli, dei modelli di partenza.
Formalizziamo un poco quanto detto. Si hanno vari livelli di eterogeneità/-conflitti:

- **eterogeneità di nome**, legato a come sono state etichettate le eterogeneità. Ho quindi concetti di primo livello (che a seconda del modello sono entità, documenti, attributi, nodi etc...) che rappresentano le stesse informazioni ma con label diverse
- **eterogeneità di tipo**, ovvero a come modellizzo effettivamente un particolare "pezzo" della realtà

Oltre a questi si ha appunto l'**eterogeneità di modello**, che abbiamo visto sopra come risolvere fin dal principio, che "copre" ogni altro tipo di eterogeneo.

6.1.1 Eterogeneità di nome

Partiamo con questa prima categoria.
Si hanno situazioni di:

- **sinonimia**, ovvero rappresento lo stesso concetto con nomi diversi, che sono sinonimi. Sintassi diversa e semantica uguale
- **omonimia**, ovvero rappresento concetti diversi con gli stessi nomi. Questo caso è più complicato da riconoscere (ad esempio potrei avere *Città* in due schemi ma in uno rappresenta quella di nascita e in un altro quella di residenza) e rischia di comportare diversi errori, magari portando a dover decidere quale valore associare senza sapere come decidere, integrando in modo errato e perdendo informazione. Raramente si ha accesso alla documentazione dei dataset e quindi bisogna sperare che i vari attributi siano il meno ambigui possibile. Sintassi uguale e semantica diversa
- **iperonimia**, ovvero tra i due concetti uno è più di alto livello rispetto all'altra, secondo magari una relazione *IS-A* (esempio il concetto *Persona* rispetto a quelli *Uomo* e *Donna*, con il primo che include i secondi)

In fase di progettazione dello schema integrato devo quindi riconoscere e gestire i casi appena elencati prima di procedere con l'integrazione.

6.1.2 Eterogeneità di tipo

In questo caso si ha che lo stesso concetto viene rappresentato in modo strutturalmente diverso in due schemi, potendo quindi avere, ad esempio:

- domini di definizione differenti per lo stesso attributo in due schemi (ad esempio da una parte il dominio delle città italiane e dall'altro quelle islandesi)
- un attributo in uno schema e avere un valore derivato in un altro schema, avendo quindi, nel primo caso, un attributo indipendente mentre nel secondo si ha una dipendenza funzionale con un altro attributo
- un attributo in uno schema e un entità in un altro schema
- un attributo in uno schema e una gerarchia di generalizzazione in un altro schema
- un'entità in uno schema e una relazione in un altro schema o magari, nel modello a grafo, avendo in uno schema un nodo e un arco nel secondo
- diversi livelli di astrazione per lo stesso concetto in due schemi (ad esempio due entità con nomi omonimi legati da una gerarchia IS-A in due schemi)
- diverse granularità nei domini di definizione
- diverse cardinalità nelle stesse relazioni
- key conflicts (la chiave in uno schema è diverso da quella dell'altro)

Si hanno quindi tantissimi casi in cui di presentano eterogeneità.

6.1.3 Data integration system

Succede quindi che si ha una **trasformazione di schema**. Si procede con uno **schema matching** per capire come unificare i vari schemi, capendo, per esempio, quali attributi coincidono (anche se magari con sintassi diverse). Si procede poi con l'**integrazione** vera e propria, ottenendo quindi lo schema integrato. Il software che procede con queste fasi è detto **data integration system** che procede prendendo e mettendo insieme delle sorgenti dati e cerca di ottenere uno schema integrato. Vengono presi e uniti tutti gli attributi

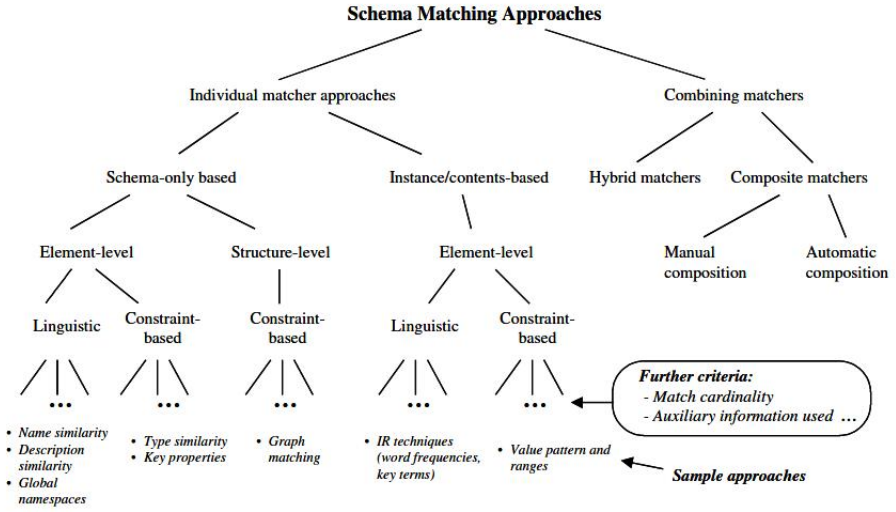


Figura 6.1: Esempio di albero per lo schema matching

degli schemi.
In modo analogo funzionano i datawarehouse dove ho schemi diversi, solitamente relazionali ma non sempre, che vengono messi insieme e quindi manipolati.
Si hanno tantissimi modi per fare **schema matching** e si fa ancora molta ricerca in merito. Si hanno match singoli, match combinati, match basati sullo schema (vedendo label etc...), match basati sui contenuti (vedendo i valori etc...) etc... (per un albero più dettagliato [1] vedere figura 6.1) Spesso bisogna fare schema matching *on the fly* dovendo integrare nuovi schemi in modo assai rapido. Si hanno anche sistemi su grande scala con sistemi *pay as you go*, dove serve altrettanta velocità.
Vediamo quindi meglio le fasi:

1. **schema transformation o pre-integration**, è una fase quindi propedeutica che presi in input n schemi me li restituisce omogenei tramite tecniche di model trasformation o reverse engineering

2. **Correspondences investigation**, dove si studiano le corrispondenze tra gli schemi tramite tecniche specifiche

3. **schemas integration e mapping generation**, dove si definiscono lo schema integrato e le regole di mapping necessarie

[1] Rahm, Erhard & Bernstein, Philip. (2001). A Survey of Approaches to Automatic Schema Matching.. VLDB J.. 10. 334-350. 10.1007/s007780100057.

Avendo da integrare/arricchire decine di schemi si hanno diverse strategie. Si hanno due scelte fondamentali:

1. integro a coppie. Grazie a questa scelta binaria ho due possibilità:
   (a) prendere uno schema, integrarlo con un altro e integrare il risultato con un altro ancora etc.... Ottengo una soluzione a scala con un albero binario non bilanciato
   (b) uso un approccio bilanciato aggregando schemi "vicini" e aggregando poi i vari risultati delle integrazioni. Ottengo quindi un albero binario bilanciato
2. integro con più schemi contemporaneamente. Ho due approcci:
   (a) integrare tutti gli schemi insieme (**questo bisogna evitarlo** anche solo per $n > 3$)
   (b) integrare in modo iterativo, in modo analogo a quanto fatto con il bilanciamento binario ma ovviamente non con un numero fisso di schemi da integrare ogni volta (magari ne integro due o magari tre, raramente di più per ovvi motivi)

Si ha appunto anche il problema che posso dire le stesse cose con strutture completamente diverse. Si cercano quindi **corrispondenze semantiche** sfruttando la conoscenza pregressa sul dominio applicativo. Si possono sfruttare sia le **equivalenze** che le **generalità** (*IS-A*) per cercare le eventuali corrispondenze. Si rischia di avere anche **disgiunzioni** tra attributi completamente diversi magari di schemi uguali rischiando di avere problemi.
In uno schema non relazionale identificare i mapping è quindi molto arduo ma bisogna anche generare i nuovi mapping e le nuove regole di associazione. Classificando i matching si trovano anche nuove tipologie di conflitto che devono essere risolti con trasformazioni opportune. Dobbiamo quindi meglio definire in primis i conflitti. Si hanno diversi tipi di conflitti:

- **conflitti di classificazione** quando elementi omologhi descrivono insiemi diversi di oggetti del mondo reale (ad esempio docenti e relatori di tesi, che sono comunque docenti ma non tutti i docenti sono relatori). Si risolve tramite generalizzazione o specifica della gerarchia (i relatori sono una parte dei docenti)
- **conflitti descrittivi** quando tipi omologhi hanno proprietà diverse o le stesse proprietà sono descritte diversamente. Si hanno:

  - conflitti di nome, ovvero nomi diverse per le stesse cose
  - conflitti di composizione, ovvero diversi attributi e/o metodi (magari avendone di più o di meno o anche diversi)

  La soluzione dipende dai casi e può, per esempio, prevedere l'inclusione di tutti gli attributi (usando l'uguaglianza) o la scelta di alcuni (usando l'ereditarietà) (???)

- **conflitti strutturali** dove si hanno gli stessi elementi rappresentati tramite strutture diverse (magari da una parte un elemento è un'intera entità e dall'altra solo un attributo). Per risolvere si sceglie, per massimizzare l'informazione ottenuta, è prendere la struttura meno vincolata (quindi tra entità e attributo scelgo entità). Scelgo quindi la struttura "più importante" secondo la logica del "più cose ho meglio è" (mentre nel caso di datawarehouse si sceglie di non salvare determinate informazioni)

- **conflitti di frammentazione** dove magari lo stesso elemento è in un singolo oggetto in uno schema mentre è frammentato nell'altro. Si risolve tramite l'aggregazione dell'elemento frammentato in quello unico (???). Si integrano i dati quindi anche nell'ottica di "rimetterli insieme" (questo tipo di tracciamento è comodo anche nella lotta evasione fiscale)

Parliamo quindi delle **regole di mapping** definite tra gli schemi "sorgente" e quelli ottenuti dopo l'integrazione e possono essere trovate per lo schema sorgente guardando alle trasformazioni effettuate per la risoluzione di omologie e conflitti avute durante il processo di integrazione effettuato sullo schema sorgente. Le regole mi dicono che un elemento chiamato in un modo nello schema integrato ha una precisa corrispondenza con un certo elemento di uno schema sorgente.
Se si hanno conflitti a livello direttamente di istanze, avendo che li stessi dati (lo stesso oggetto del mondo ideale) in sorgenti diversi hanno valori diversi, i problemi aumentano. Questo è un problema fondamentale nel data integration, anche nel caso dell'arricchimento. Posso avere:

- un "semplice" **conflitto di attributi** avendo due entità che rappresentano lo stesso oggetto reale con attributi diversi (anche solo uno) per la stessa proprietà. Le chiavi in questo caso sono coerenti

- un **conflitto di chiave** (detto anche **conflitto di entità** o **conflitto di tupla**) se ho chiavi primarie diverse in due entità diverse che però rappresentano lo stesso oggetto (banalmente entità

che rappresenta lo studente nei db della Bicocca usa la matricola mentre in un altro il codice fiscale) mentre gli altri attributi coincidono

Una scelta di soluzione, detta **currency**, prevede, in caso di conflittualità di valore di attributo, di scegliere il valore inserito temporalmente per ultimo, nella speranza che sia il più aggiornato (ma spesso non è esattamente così). Le tecniche che trattano i conflitti a livello di istanza possono essere applicate in due diverse fasi del ciclo di vita di un sistema di integrazione dati, vale a dire:

- a **design time**
- a **query time**

I problemi principali si hanno con i conflitti di chiave non sapendo bene se gli stessi oggetti sono presenti nei vari dataset e questi problemi saltano fuori a query time. Tuttavia, durante il design time decide la strategia da seguire per risolvere i conflitti prima che le query vengano elaborate, ovvero nella fase di progettazione del sistema di integrazione dei dati e le tecniche a query time incorporano le specifiche della strategia da seguire all'interno della formulazione della query. Si hanno quindi le **funzioni di risoluzione dei conflitti** che cercano, dati due valori conflittuali, di restituire quello più probabile. Un criterio di misura si basa sull'affidabilità delle sorgenti stesse da cui proviene il conflitto. In altri casi si ragiona in termini di funzioni matematiche di minimo, massimo, media, varianza etc... in presenza di valori numerici (introducendo probabilmente errori, di sottostima, sovrastima etc... ma l'importante è esserne consapevoli per stimare l'impatto dell'integrazione sui dati). Per valori non numerici si hanno altre funzioni, ad esempio la concatenazione nel caso di stringhe o la scelta delle più lunghe/corte. Ovviamente non si può sapere la risposta giusta ma si può imporre una logica di stima. È comunque buona pratica tenere traccia tali conflitti e in caso siano troppi preoccuparsi dell'effettiva validità dei dati. Il data integration assume quindi due forme:

1. **deduplicazione**, dove si scopre se ci sono concetti del mondo reale replicati nella tabella, avendo che l'integrazione viene fatta solo su quella tabella (non avendo schema matching), internamente. Posso avere quindi valori discordanti nella stessa tabella, magari per errori inserimento avendo stringhe con distanza sintattica bassa (qualche carattere, secondo la **distanza di edit**). Si usano tecniche empiriche. Questo check va fatto di valore in valore ma non si può avere certezza avendo che distanze brevi si

hanno anche tra parole potenzialmente corrette (basti pensare a Carlo e Carla, entrambi potenzialmente corretti per un ipotetico Carl, che potenzialmente è errato in lingua italiana). Si procede quindi, in assenza di chiave, ad una pulizia di eventuali duplicati cercando di "fonderli" del risultato corretto più probabile. La distanza di edit va bene solo per stringhe piccole, per quelle grosse si hanno altre funzioni (come la Jaccard index function)

2. integrazione in due tabelle diverse, avendo schema matching etc. . .. Si supponga quindi di voler fondere due tabelle senza chiave primaria. Si cerca quindi una pseudo-chiave primaria (usando l'insieme più grande di attributi comuni alle tabelle che siano significativi, ad esempio nome, cognome e luogo di nascita) matchando le tabelle su di essa. Ma potrei anche qui non matchare per errori di scrittura. Faccio quindi un matching approssimato e concatenando le due tabelle e procedendo poi con la deduplica della tabella risultante. Potrei comunque sempre ragionare con la distanza di edit etc... restando con le tabelle separate, procedendo con una combinazione lineare degli score delle distanze (riscalate su $\{0, 1\}$, con 1 match perfetto, facendo una somma pesata e dividendo per il numero di score), decidendo una soglia oltre la quale si ha il match. Si usano tecniche probabilistiche. Si procede quindi la fusione delle tabelle e al raffinamento del risultato con l'eliminazione dei duplicati

Oltre a tecniche empiriche e probabilistiche si hanno anche tecniche miste e/o basate su conoscenza a priori.
Si sono finora trascurati quei casi in cui si sta palesemente rappresentando la stessa cosa ma con termini diversi (magari con nomi contratti, esempio Leonardo che diventa Leo.).
Si ha infatti il **record linkage**, ovvero la tecnica di risoluzione dei conflitti a livello di istanza, dove, date le tabelle in input, posso avere in output:

- **matching tuples**, ovvero i valori che sicuramente matchano (anche se non perfetti ma con score oltre una certa soglia)
- **non matching tuples**, ovvero i valori che sicuramente non matchano
- **possible matches**, ovvero una sorta di "area grigia" dove non si sa come dare risposta (come appunto l'esempio dl nome abbreviato). La scelta della soglia per il match e di quella per il non match

modifica la cardinalità di questi casi, ad esempio una coppia di soglie troppo cattiva produrrà come conseguenza molti casi dubbi, essendo quest'area quella compresa tra le due soglie

In mezzo subentra l'umano che "a mano" sistema i valori, non avendo altro modo. Da qualche hanno si è anche introdotto il machine learning, tramite **Deepr** (basato su reti neurali, trainato tramite specifici dati di training), per sostituire l'uomo nell'analisi di questa zona di mezzo. Le soluzioni di machine learning sono fortemente legate al dominio. Un esempio classico di problema è quello relativo all'indirizzo di casa nelle anagrafiche. In tal caso prima si cerca di normalizzare l'indirizzo (espandendo abbreviazioni di vie, piazze etc..., esempio v. D. Alighieri diventa Via dante Alighieri) facendo poi il confronto sulle stringhe. altra tecnica è usare la geocodifica (tramite maps etc) usando poi latitudine e longitudine (con comunque un margine di differenza, dato dall'errore anche del GPS in uso civile e dagli standard statali di definizione degli indirizzi). Si nota quindi come il dominio degli indirizzi ha una risoluzione particolare (oltre al fatto che lo standard degli stessi varia di stato in stato) e molti domini specifici hanno tecniche specifiche (difficilmente apprese tutte da una rete neurale (???)).
Si è dato per scontato un altro problema: le dimensione delle tabelle. La il check dei valori tra le tabelle esplode nel numero di confronti. Bisogna quindi ridurre lo **spazio di ricerca**, non potendo fare il prodotto cartesiano tra le tabelle ($A \times B$), anche perché nichilizzerebbe la scalabilità. Si hanno tantissimi **blocking method** per ridurre lo spazio di ricerca. Una volta ottenuto un sottoinsieme dello spazio di ricerca posso partire con vaie tecniche empiriche (ad esempio ordinare sui cognomi, e sperare che siano corretti o con al più pochi errori, per effettuare poi il check). Si applica poi un **modello di decisione** per il check, tramite deep learning, semantica, funzione di risoluzione delle stringhe etc..., ovvero un qualsiasi meccanismo per ottenere match, unmatch e possible match. Per farlo si necessita anche di un benchmark che sono anch'essi fortemente dipendenti dal dominio e quindi non sempre disponibili. Si hanno benchmark sugli indirizzi etc....
Vediamo, riassumendo, nel **record linkage probabilistico** si ha:

1. **preprocessing**, normalizzazione dei formati (espansione abbreviazioni, lowecase, eliminazione spazi e caratteri speciali etc... o passando ad altre codifiche, come la geocodifica) secondo uno standard

2. **blocking**, riduzione dello spazio di ricerca. Uno dei metodi è il **sorted neighbour** che implementa il blocking scegliendo una chiave, ordinando i record in base alla chiave univoca e spostando

una finestra sui file ordinati, confrontando a destra e a sinistra nelle due tabelle gli elementi in questa finestra. Le finestre sono dinamiche in quanto potrei avere match sul bordo che andrebbero persi

3. **compare**, scegliendo una funzione di distanza e cercando un sample con coppie conosciute di match e unmatch. Bisogna anche valutare per ogni valore di distanza la frequenza di matching e unmatching

4. **decide**, valutando la distanza e le soglie

Per confrontare se due date sono uguali prima le normalizzo al modello o europeo o americano, le normalizzo in base al carattere separatore e poi confronto.
Per la fusione posso:

- ignorare un conflitto, mettendo tutti i valori disponibili
- cercare di evitare i conflitti basandomi sui metadati, sul sorgente o sulle istanze
- gestire i conflitti tramite le funzioni di risoluzione dei conflitti e alle altre cose discusse sopra

Si hanno elenchi delle varie strategie possibili (ignorare, scegliere il più recente, scegliere il valore medio e molte altre), ognuna delle quali, come anche il record linkage, può introdurre errori. La scelta delle soglie inoltre può portare ad essere meno conservativi, introducendo falsi positivi, o troppo conservativi, riducendo drasticamente i match. Le varie scelte dipendono comunque dai singoli casi.

6.2 Architetture per data integration

L'ordine di questa sezione è dubbio, forse andava ad inizio capitolo o forse no.
Si parte come abbiamo detto da una serie di schemi separati in vari db, con modelli diversi. Posso avere anche file csv etc... Bisogna quindi raccogliere le varie informazioni e integrarle in un'unico dataset. Si hanno più tecnciche:

- la tecnica del **consolidamento** dove si raccolgono vari dataset dalle varie sorgenti informative, li integro come visto precedentemente e infine salvo il risultato in un nuovo db, che sarà la mia nuova architettura dati

- la tecnica che prevede di rappresentare i vari schemi separati in modo che essi siano mantenuti inalterati, non memorizzando il risultato dell'integrazione, ma dando possibilità all'utente di scrivere query sui dati integrati. Si produce quindi un nuovo schema globale che comunica tramite un mediatore software per prendere le query. Questo è la vera e propria tecnica di **virtual data integration**

Nel capitolo si parla soprattutto di virtual data integration anche se spesso non specificato.
Il data integration è un è un'importante area di ricerca e business che ha lo scopo principale di consentire ad un utente un accesso uniforme a più fonti di dati autonome ed eterogenee, attraverso la presentazione di una visione unificata di questi dati. Trovare questa uniformità tra elementi eterogenei è complesso perché bisogna trovare differenze e somiglianze in ogni schema per potersi conformare. Bisogna quindi riconciliare tutte le eterogeneità di schema e di istanza.
Il vantaggio di architetture d'integrazione rispetto alle architetture federate è la capacità di gestire meglio le eterogeneità, soprattutto se complesse, a livello di schema. Inoltre i sistemi federati non possono gestire eterogeneità a livello di istanze a causa della qualità (accuratezza, incompletezza inconsistenza etc...) dei dati. D'altro canto nelle architetture federate si ha un livello di autonomia in poco più basso nel senso che nei modelli integrati si assume che si hanno sorgenti provenienti da chissà dove mentre nei sistemi federati i nodi sono scelti in modo più consapevole.
Si hanno due approcci principali alla data integration:

1. **integrated read-only views (*Mediation*)**, dove si ha integrazione solo per lettura. I dati letti nei vari db restano quindi invariati. Questa soluzione a livello aziendale viene scartata in favore delle datawarehouse, avendo persistenza dei dati e profondità storica maggiore (anche se perdo flessibilità nel momento in cui si rimuove una sorgente dati)

2. **integrated read-write views (*Mediation with update*)**, dove si hanno anche gli update. È estensione dell'architettura di Mediation per supportare gli aggiornamenti in una vista integrata (dovendo poter accedere ai vari db anche in scrittura). Questa cosa è difficile e anche poco studiata in letteratura e quindi non la tratteremo. In questo caso si preferiscono i modelli federati

L'integrazione dei dati è il problema di combinare i dati che risiedono in diverse fonti e fornire all'utente una visione unificata di questi dati può quindi

essere letteralmente definita come **global virtual schema (*GS*)**.
Parto dai vari db coi vari modelli e schemi e creo un GS che ha un certo modello. Si hanno mapping tra i vari schemi e il GS e quindi l'utente può effettuare delle query passando per il GS e a runtime il GS capisce dove sono le informazioni richieste, converte il linguaggio di query in quello necessario per il db dove si trova l'informazione ed effettua la query (che viene eseguita localmente sul db). Sempre on the fly si risolvono i conflitti e viene restituito il risultato della query. Potrei avere anche integrazione pay as you go.
Si hanno quindi tre elementi fondamentali in un'architettura di integration:

1. il GS
2. le varie sorgenti
3. il mapping tra le sorgenti e il GS

Si hanno quindi due componenti fondamentali:

1. un **mediator**, che data una query la frammenta e la riscrive per poter lavorare con i vari schemi locali. Inoltre il mediator mette insieme i vari risultati, risolve i conflitti e risponde alla query
2. vari **wrapper**, che agganciano ogni schema locale delle sorgenti al GS, rappresentando la sorgente in uno schema compatibile con quello del GS. Essi ricevono query nel linguaggio del GS e rispondono di conseguenza

Le architetture **wrapper-mediator** sono lo standard nel mondo del virtual data integration, nel dettaglio della **lazy integration** (*capire che dice al minuto 12.05*) consentendo maggior flessibilità (**dal minuto 12 al minuto 12.30 si capisce davvero poco, controllare**).
Bisogna studiare GS e mapping.
Si hanno due tipiche architetture del sistema, a seconda della direzione dei mapping:

1. una in cui si parte dagli schemi locali per arrivare al GS
2. una in cui si parte dal GS e si arriva gli schemi locali

In ogni caso lo scopo è interrogare l'informazioni come se fosse unica.
Nelle slide qualche esempio.
Il mediator costruisce uno schema unificato di diverse fonti di informazioni (eterogenee) e consente a un utente di formulare una query su di esso. La query dell'utente viene quindi trasformata in una serie di sottoquery, una

per ciascuna fonte di dati coinvolta nella query e il risultato viene sempre raccolto e mergiato dal mediator che lo restituisce all'utente.
Il mediator quindi deve riconciliare le varie istanze mettendo insieme le varie risposte dei vari db. Si hanno varie azioni, da svolgere in breve tempo (essendo tutto on the fly), fatte dal mediator:

- raggruppare informazioni sulla stessa entità del mondo reale
- rimuovere la replicazione dei tra le varie fonti di dati
- risolvere le inconsistenze tra le varie fonti di dati
- ottenere accuratezza, completezza, etc... tra i dati provenienti da diverse fonti di dati

L'interazione con il mediator è quindi divisa in due fasi:

1. la creazione della rappresentazione unificata (**Publishing phase at design time**), creando GS e mapping
2. la formulazione e l'esecuzione di una query nella rappresentazione unificata (**Querying Phase**). Questa fase è a runtime

Nella fase di **publishing** i problemi sono quelli di modellare e definire linguaggi per lo schema unificato, costruendo anche i mapping. Eventualmente devo gestire gli update. Devo anche estrarre lo schema dei vari db per capire come procedere.
Nella fase di **querying** bisogna stabilire il linguaggio, fare query unflodin-g/rewriting grazie ai mapping. Devo inoltre cleaning e fusion dei risultati. Bisogna inoltre convertire i vari linguaggi da GS a schemi locali.
Il mapping è centrale in entrambi. Possiamo vedere il sistema di virtual data integration come una tripla:

$$(G, S, M)$$

dove:

- G indica lo schema globale
- S l'insieme degli schemi sorgente
- M il mapping

Dobbiamo appunto costruire uno schema globale virtuale. Il problema è capire quali dati reali nelle sorgenti di dati corrispondono a quei dati virtuali rappresentati per scelta dallo schema globale. Per il mapping si hanno vari approcci, a seconda della relazione tra lo schema globale e quelli locali, rispetto al concetto di vista (che in SQL ricordiamo costruire una sorta di tabella virtuale):

- **GAV (Global As View)**, in cui lo schema globale viene creato sulla base dell'osservazione degli schemi sorgente, attraverso un processo di integrazione intenzionale degli schemi sorgenti (si pensi anche al processo di consolidamento, o ad una situazione in che vogliamo rappresentare in modo integrato tutto il contenuto informativo dell'architettura dati di un'organizzazione). L'approccio si può riassumere dicendo che lo schema globale è una vista di quelli locali (tutte, contemporaneamente). Ogni tabella dello schema globale è una vista degli schemi locali. In questo caso il mapping va dagli schemi locali a quello globale. Il vantaggio è che si prendono tutte le informazioni disponibili, procedo tramite arricchimento dei dati ma potrebbe capitare che non si ottiene lo schema voluto. GAV si presta poco agli update in caso di rimozione (in caso di aggiunta semplicemente rifaccio l'integrazione e aggiungo elementi allo schema globale, anche se questo ha un costo) in quanto tutti i concetti che prima avevano a che fare con la sorgente tolta vanno cambiati, avendo poi problemi anche a livello di istanze. Questo approccio ci dice come recuperare i dati da istanze di origine locale (reale) per ciascuna tabella dello schema globale (virtuale). Con il GAV il mediatore procede tramite unfloding delle query, cioè la sostituzione nella query dei riferimenti allo schema globale con l'espressione della view globale in termini delle view locali. Bisogna però riportare i *join* operazione che abbiamo già visto rischia di essere complessa e poco efficiente. Nonostante tutto gli approcci GAV sono solitamente quelli più usati

- **LAV (Local As View)**, dove lo schema globale viene progettato a priori indipendentemente dagli schemi sorgente, da cui prenderò solo i pezzi di informazione di interesse. Il mapping tra i sorgenti e lo schema globale si ottiene definendo ciascun sorgente sullo schema globale. Può capitare che concetti locali non siano nello schema globale ma ogni concetto globale avrà una qualche rappresentazione locale. Questo approccio rende più semplice l'evoluzione temporale dello schema globale in quanto in caso di update aggiungendo uno schema locale non cambia lo schema globale (che è fisso) ma si ha solo un eventuale arricchimento del mapping con un nuovo query mapping. Il mapping va quindi dallo schema globale a quelli sorgente/locali avendo che il contenuto di ogni schema locale è descritto in termini di vista sullo schema globale. Se una colonna non è presente nel globale ma solo nel

locale rispondo NULL. Questo approccio ci dice come le istanze sorgenti (reali). Il query management è più complesso del GAV ma è più semplicemente estendibile del GAV contribuiscono/sono associati alle tabelle dello schema globale (virtuale)

- **GLAV (Global and Local As View)** dove il mapping tra sorgenti e schema globale si ottiene definendo un insieme di viste, alcune sullo schema globale e altre sulle origini dati. È un approccio ibrido

Esempio 9. *Un mapping GAV è del tipo:*
(prendo tutto insieme. facendo una union)
Con la seguente query per sapere i nomi dei professori che hanno più di 50 anni:
Avendo poi un'operazione di unfolding infatti nella query il mediator sostituire i riferimenti a GProf con la specifica della vista in termini degli schemi locali

Esempio 10. *Vediamo la stessa query in approccio LAV.*
Si hanno nello schema globale delle create view di quello locale:
Con la seguente query per sapere i nomi dei professori che hanno più di 50 anni:
Il mediator quindi parte dalle viste per cercare i nomi dei richiesti in entrambe le viste mettendo poi insieme i risultati.
Questo esempio è semplificato in quanto ho schemi locali dello stesso tipo.

In generale i sistemi di integrazione hanno ancora scarse prestazioni.

Vengono qui aggiunte le cose dette in live.
Guardare slide incontro, c'è tutto.

Capitolo 7

Data quality

Il tema del **data quality** è particolarmente importante perché, dopo aver approfondito come gestire i dati e le loro architetture è bene studiare anche i dati in se.
un esempio di scarsa data quality è quella di avere lo stesso dato con valori diversi nello stesso posto, anche semplicemente una pagina web.
I dati sono una rappresentazione della realtà e questo porta al fatto che la realtà viene da noi modellata tramite alcuni dati specifici, tramite un pezzo della realtà e tale rappresentazione potrebbe non essere a priori oggettiva, anche a causa di strutture linguistiche e/o basate sui sensi. Si definiscono quindi:

- **utilità** come precisione della rappresentazione interna del dato rispetto a il compito svolto (e quindi, ad esempio, un'immagine ritoccata potrebbe essere più utile per determinati scopi)
- **fedeltà** come l'aderenza del dato alla realtà

Si è capito che la qualità è un concetto complicato.

Definizione 9. *Diamo una serie di definizioni:*

- ***qualità**: caratteristiche di un artefatto che influiscono sulla sua capacità di soddisfare le esigenze e le aspettative dell'utente, dichiarate o implicite. Un dato è di qualità se è adatto all'uso che se ne deve fare. Si ha il concetto di **fitness for use**. "La qualità dei dati è negli occhi di chi li usa e non nelle mani di chi li produce"*
- ***dimensione**: una caratteristica specifica che descrive la qualità delle informazioni, solitamente non misurabile ma potrebbero esserlo*

- ***sottodimensioni**: sotto-caratteristiche che spesso classificano una certa dimensione*

*Le dimensioni misurabili lo sono tramite **metriche**, che, secondo le definizioni dello standard ISO 9126-1 e secondo il framework ISM3 sono un'insieme di procedure che comprende:*

- *una procedura (o metodo) di misurazione, cioè un algoritmo che prende l'elemento da misurare e lo associa a una misura (sia esso un valore ordinale o un intervallo)*
- *una corretta unità di misura (o scala), cioè il dominio dei valori restituiti dalla procedura di misurazione*

Si possono avere metriche diverse per la stessa dimensione

La qualità dei dati è un concetto che può essere espresso attraverso molteplici dimensioni, ad esempio la accuratezza (magari anche solo per errori di typo nei dati), la comprensibilità, la completezza (avendo magari valori a NULL), l'inconsistenza (avendo contraddizioni nei dati) etc...
Su slide tabella con le metodologie.
Un altro aspetto da tenere in considerazione è che si hanno metriche:

- **oggettive**, modi formali e precisi per misurare le metriche per una dimensione di qualità in termini di valori di un dominio, indipendentemente dalla percezione/valutazione umana
- **soggettive**, modi per misurare le metriche per una dimensione di qualità che dipendono dalla percezione (di solito valutazione esplicita e ordinale) delle persone coinvolte nel processo di misurazione, dipendente dalla percezione/valutazione umana

Vediamo qualche aspetto nel dettaglio.
Partiamo dall'**accuratezza**.

Definizione 10. *La precisione di un valore v è definita come la vicinanza tra v e un valore v' considerato come la corretta rappresentazione del fenomeno del mondo reale che v intende rappresentare.*
Impatta valori alfanumerici e si può applicare alle tuple e alle relazioni.

Un esempio di inaccuratezza è un typo.
Si hanno due sottodimensioni di questa qualità:

1. **accuratezza sintattica**, magari controllando se la stringa è presente in un insieme di valori di riferimento per il dominio trattato (magari un vocabolario, una lista di città etc...). Ci sono dati su cui non è applicabile. Generalmente si fa:
    - in base alle stringhe
    - in base ai singoli token di una stringa
2. **accuratezza semantica**, molto difficile da verificare (magari tramite un'altra sorgente dati per cross validare). Per piccolissimi db si fa a mano ma è generalmente impensabile

Conoscere lo schema aiuta a valutare la accuratezza dei valori e si hanno diverse metriche. In primis posso dire se un dato è vicino a quello che mi aspetto in un certo dominio tramite una funzione di distanza che confronta il valore che ho e il dominio di riferimento. Dopo questo posso pensare a correggere eventualmente il dato. Si cerca il valore più vicino al dominio di riferimento. Potrei anche fare una stima statistica per capire quale sia il valore corretto. Per il calcolo delle distanze su stringhe uso la solita **distanza di edit (non normalizzata)** ma non è la sola. In fase di assesment considero solo valori con una massima distanza di edit stabilita e nella fase di improvment posso stimare la giusta correzione per ottenere la stringa magari corretta (magari in corrispondenza di typo). La distanza sintattica potrebbe però essere poco informativa in merito alla semantica (si pensi a Domenico spesso abbreviato in Mimmo). Possiamo introdurre la **distanza di edit normalizzata** come:

$$1 - \frac{edit(a, b)}{n}$$

con n numero massimo di simboli. Si ottiene un valore tra 0 e 1, con 1 che indica che i valori che sono identici, accuratezza pari a uno implica distanza nulla (misuro il complemento a 1 della distanza). Tali valori tra 0 e 1 sono comodi in quanto spesso le metriche danno risultati che sono in questo range e posso quindi effettuare una somma pesata tra le varie metriche.
Si hanno vari tipi di funzione di edit (ma limitate a stringhe corte):

- Hamming distance
- Levenshtein distance
- Jaro-Winkler

o di distanza di token (per stringhe composte da più stringhe, esempio "Sesto San Giovanni"):

- Jaccard index
- Sorensen-Dice
- N-gram

Si possono avere problemi nel trovare la reference table di un certo dominio. Passiamo alla completezza.

Definizione 11. *Definiamo la* ***completezza*** *di un insieme di dati come la copertura con la quale il fenomeno osservato è rappresentato nell'insieme di dati. Avere ad esempio dei NULL impedisce la completezza.*
Impatta valori alfanumerici e numerici e si può applicare alle tuple, agli attributi, agli oggetti e alle relazioni.
Si ha che, normalizzando sempre con $1-$ *per avere massima completezza a 1:*

- *la completezza di tupla è basata sul numero di NULL di una certa riga rispetto al numero di attributi presenti*
- *la completezza di attributo è basata sul numero di NULL di una certa colonna rappresentante un attributo*
- *la completezza di tabella è basata sul numero di NULL dell'intera tabella*
- *la completezza di oggetto è il numero di valori non nulli in tutti gli oggetti rappresentati nelle tuple*

Nelle precedenti definizioni, abbiamo assunto una ipotesi di **mondo chiuso**, ovvero tutto ciò che è rappresentato nella db è vero, tutto il resto è falso. Questa è la tipica ipotesi che si fa nei DBMS. Posso ipotizzare un **mondo aperto**, come nei DBMS NoSQL, dove si tace di ciò che non si sa, avendo **object completness**, tenendo conto del fatto che gli oggetti rappresentabili sono più delle tuple della tabella, e di tale cardinalità serve una stima indiretta. Quindi nel mondo n on relazionale è difficile stimare la completezza, non avendo certezza del numero di oggetti.
Parliamo ora di **proprietà temporali**, di livello di aggiornamento, ovvero di **currency**. Si parla anche di tempestività. Trattando il tempo è difficile dare una definizione formale e specificare la semantica trattata. Si ha che la **currency** misura con quale rapidità i dati sono aggiornati rispetto al corrispondente fenomeno del mondo reale. Una prima misura della currency è il ritardo temporale tra il tempo $t1$ dell'evento del mondo reale che ha provocato la variazione del dato, e l'istante $t2$ della sua registrazione nel sistema informativo ma questa misura è costosa in quando generalmente l'evento non

è ben noto, non sapendo quando si hanno cambiamenti etc.... Un'altra metrica di currency è vederla come come differenza tra tempo di arrivo alla organizzazione e tempo in cui è effettuato l'aggiornamento (cosa misurabile in presenza di log, ad esempio, sui tempi di update). Un altra metrica è la differenza rispetto al metadato "ultimo aggiornamento effettuato" che, per valori con periodicita' di aggiornamento nota, currency calcolabile in maniera approssimata ma poco costosa (ma non si hanno informazioni in merito alle modifiche del dato).
La **tempestività** misura quanto i dati sono aggiornati rispetto a un particolare processo (o ai processi) che li utilizza. Quindi è dipendente dal processo, a differenza della currency, ed è associabile al momento temporale in cui deve essere disponibile per il processo che utilizza il dato. Posso avere dati obsoleti per il processo di chi li usa ma con alta currency.
Trattiamo infine la **consistenza**.

Definizione 12. *Si definisce la* ***consistenza*** *in due modi:*

1. *come consistenza dei dati con i vincoli di integrità o dipendenze funzionali definiti sullo schema (ad esempio i vincoli di integrità nel modello relazionale). Si hanno anche dei vincoli di consistenza, detti business rule, che possono riguardare uno o più attributi, relazioni etc... e possono essere espressi in termini di probabilità. Si usano anche vincoli di integrità in logica e si hanno dei data edit nelle indagini statistiche*

2. *come consistenza delle diverse rappresentazioni di uno stesso oggetto della realtà presenti nella base di dati (come visto nella parte di data integration)*

Invece con **accessibilità** si esprime la capacità di un utente di accedere ai dati a partire dalla propria cultura, stato fisico e psichico e dalla tecnologie disponibilità, magari dal punto di vista culturale, dello stato fisico, delle tecnologie etc...
Un altro aspetto importante che va sempre considerato è il **tradeoff tra varie qualità**, ad esempio consistenza e completezza nel modello relazionale possono essere non conciliabili quando si voglia rispettare l'integrità referenziale. Nel dominio statistico si ha tradeoff tra tempestività e completezza/accuratezza, che vengono però privilegiate (bisogna quindi studiare quanto è "sporco" un certo dato). Nel web si preferisce la tempestività rispetto ad accuratezza/completezza.

7.1 Quality improvment

Avendo quindi già introdotto la fase di **quality assesment** passiamo a quella di **quality improvment**, ovvero al miglioramento dei dati stessi.
L'obiettivo della fase di miglioramento, una volta misurata la qualità dei dati a aver scoperto che magari sono di bassa qualità rispetto alle esigenza, è quello di migliorare i dati. La qualità dei dati è un problema di tipo multidimensionale, potendo decidere di migliorare anche solo alcuni aspetti relativi alla qualità dei dati (come la completezza, la consistenza, l'aggiornamento temporale etc...), sapendo che ci sono dei tradeoff quindi il miglioramento di alcuni aspetti va a discapito della qualità di altri. Si hanno in generale due strategie in fase di miglioramento:

1. **data-driven**, migliorando il dato stesso in quanto tale. Si punta a migliorare la qualità dei dati modificando direttamente il valore dei dati attraverso il confronto con altri dati ritenuti di buona qualità. Ad esempio, i valori dei dati obsoleti vengono aggiornati aggiornando un database caratterizzato da una valuta più alta. Si migliora il dataset stesso. Questo è una soluzione molto usata in data science. Questa strategia si applica se la raccolta dati è continuativa.
   Si hanno varie procedure:
   - **acquisizione di nuovi dati**, che migliora i dati acquisendo dati di qualità superiore per sostituire i valori che sollevano problemi di qualità. Questa strategia può essere utilizzata per tutte le dimensioni esaminate nella fase di valutazione
   - **record linkage**, detta anche **identificazione degli oggetti**, che confronta i dataset con valori "sporchi" con una fonte certificata o di qualità superiore, identificando le tuple/record nei due dataset che potrebbero fare riferimento allo stesso oggetto del mondo reale e pulendo i "valori" sporchi con corrispondenti valori di qualità superiore
   - **affidabilità della fonte**, dove si selezionano le fonti di dati sulla base della qualità dei loro dati
2. **process-driven**, migliorando il processo di acquisizione dei dati (datti errati possono portare ad errori sistematici), ottenendo che il dataset viene alimentato con dati corretti. Si punta quindi a migliorare la qualità ridisegnando i processi che creano o

modificano i dati. Ad esempio, un processo può essere riprogettato includendo un'attività che controlla il formato dei dati prima dell'archiviazione. Tale strategia fa riferimento all'ambito detto **Business Process Reengineering (BPR)**, ovvero avendo la possibilità di riprogettare i processi, magari scoprendo che una sorgente era di qualità troppo bassa.
Si hanno due tecniche:

(a) **process control**, dove si aggiungono elementi e/o procedure di controllo nel processo di produzione dei dati quando:

- vengono creati nuovi dati
- vengono aggiornati i set di dati
- il processo accede a nuovi set di dati

Verificando errori e la qualità dei dati stessi.
In questo modo, viene applicata una strategia "reattiva" agli eventi di modifica dei dati, evitando così la degradazione dei dati e la propagazione degli errori

(b) **process redesign**, dove si ridisegnano i processi per rimuovere le cause della scarsa qualità e introduce nuove attività che producono dati di qualità superiore. Se la riprogettazione del processo è radicale, questa tecnica viene definita *reingegnerizzazione dei processi aziendali*. Queste tecniche sono puntuali e non sono facilmente integrabili in un sistema di raccolta continua dei dati

Nel lungo termine le tecniche process-driven sono più efficaci, eliminando il problema alla radice, ma sono estremamente costose nel breve termine. Le tecniche data-driven sono più economiche ma costose nel lungo e quindi sono adatte per un'applicazione "one-shot" e, quindi, sono consigliati per i dati statici.
Approfondiamo il data-driven.
Si hanno miglioramenti specifici per la dimensione:

- **accuratezza** tramite confronto dei valori con un dominio di riferimento. Questo è il tipo di tecnica che abbiamo considerato per l'accuratezza sintattica nella fase di quality assesment

- **completezza** tramite completamento di dati incompleti con tecniche specifiche che sfruttano la conoscenza sui dati

- **consistenza** tramite l'identificazione dei dati corretti sfruttando vincoli di integrità, dipendenze funzionali o dati derivati etc..., correggendo poi tramite sorgenti esterni o acquisendo nuovamente i dati

Come tecnica si usa anche la **error localization and correction** che identifica ed elimina gli errori di qualità dei dati rilevando i valori che non soddisfano un dato insieme di regole, dette **edits** nelle metodologie per i le analisi statistiche (queste tecniche sono infatti studiate principalmente nel dominio statistico). Rispetto ai dati elementari, i dati statistici aggregati (come media, somma, massimo etc...), sono meno sensibili alla localizzazione probabilistica e alla correzione dei valori probabilmente errate. Sono state proposte tecniche per la localizzazione e correzione degli errori per incongruenze, dati incompleti e valori anomali, ovvero valori significativamente diversi da tutti gli altri valori in un set di dati per dati elementari e dati statistici.
Un altra tecnica di quality improvment è la **deduplica** che corrisponde al raggruppamento di record di dataset che si riferiscono alla stessa entità nel mondo reale (raggruppo i doppioni). Praticamente la deduplica è un'operazione di record linkage fatta su un'unica tabella. Si hanno tre gruppi in output alla deduplica:

1. coppie/gruppi di tuple che matchano, corrispondenti allo stesso oggetto reale

2. coppie di tuple che non matchano che corrispondono a entità distinte nel mondo reale

3. possibili coppie/gruppi di tuple, per le quali non è stato possibile giungere una conclusione definitiva, e sono necessarie ulteriori indagini (e costi) per assegnarle al gruppo match o mismatch

Un'altra attività classica, propedeutica ad altre attività (è una sorta di fase di preprocessamento), è la fase di **normalizzazione**, trasformando le stringhe, scritte nello standard del dominio di riferimento, effettuando, per ogni dimensione, il riconoscendo della stessa e il relativo miglioramento. Per la normalizzazione di domini specifici, tipo i nomi delle vie etc..., esistono tool specifici.
Si ha quindi il miglioramento di singoli valori o di intere tuple, migliorando accuratezza sintattica, completezza, currency e consistenza, in base a certe metriche.
Ogni valore che non matcha con la tabella di riferimento (spesso disponibili online a seconda del dominio ma non sempre) deve essere corretto. Il miglioramento dell'accuratezza sintattica viene effettuato sostituendo il valore

errato con quello a distanza inferiore nella tabella di riferimento, dove la distanza può essere valutata dopo una precedente verifica dei valori e delle loro caratteristiche (si nota come le attività data-driven sono costose dal punto di vista temporale).
Mentre per migliorare l'accuratezza possiamo adottare procedure standard di confronto con le tabelle di riferimento, per la completezza è molto più complesso, non avendo un elemento da cui partire, avendo a disposizione solo un valore NULL, non riuscendo sempre a ricostruire il valore corretto per sostituirlo al NULL. Spesso comunque si ha un contesto per aiutare a rimuovere i NULL, quindi, la procedura più intuitiva è eseguire una nuova acquisizione di dati incompleti riempiendo di volta in volta più valori NULL possibili. Poiché di solito si tratta di un'attività molto costosa, possiamo adottare diverse euristiche che di solito dipendono dal contesto.
Per la completezza si usano anche i **dati derivati**, definiti come dati per i quali esiste una formula matematica per la quale questi dati possono essere ottenuti a partire da un altro set di dati già conosciuto (esempio banale un prezzo con o senza IVA).
Bisogna migliorare la consistenza vedendo se la dipendenza funzionale sui dati derivati e altri vincoli di integrità intra-relazionale possono essere sfruttati, oltre che per la completezza completezza, anche per la consistenza.
Esempio case study su slide.
Dopo tutte queste operazioni si ottiene un db già abbastanza pulito ma si hanno, solitamente, ancora problemi con spazio di miglioramento. Si userà quindi il **record linkage**.

7.1.1 Record Linkage

Il **record linkage** viene usato anche in altri contesti oltre il data quality con anche altri nomi (come object recognition, object matching, deduplicazione, se si tratta di un solo dataset, etc...).
Prima della fase di data integration si ha lo *scheme integration*, con approcci architetturali. Dal punto di vista di integrazione si hanno, si ricorda, gli approcci GAV o LAV, ma in entrambi i casi si hanno elementi rappresentanti lo stesso oggetto reale in db diversi che potrebbero non essere facilmente integrati a causa, magari, dell'assenza di una chiave primaria univoca. Bisogna quindi studiare come collegare i record presenti nei vari db tramite **record linkage** e mettere poi insieme i risultati con un'operazione di **fusion**, cercando di studiare eventuali ridondanze o ambiguità.
Il record linkage consiste quindi nell'identificare le stesse osservazioni in diversi file/db. In generale si parla di **entity resolution task/record linkage** avendo varie "varianti":

- **deduplicazione**, che considerando una sola tabella è usata principalmente con il modello ER. Si procede raggruppando record che corrispondono allo stesso oggetto reale normalizzando lo schema e riducendo il numero di record nel dataset. Come variante si hanno i cluster di calcolo

- **record linkage**, dove appunto si matchano da un archivio dati deduplicato a un altro (bipartito). Si ha anche la versione k-partita lavorando con più data store. Generalmente questo metodo proposto nei datastore relazionali, ma più frequentemente applicato a record non strutturati da varie fonti

- **canonicalizzazione**, che generalmente fornisce il record più completo, attribuisce i valori tramite la fusione, costruendo dei **single version of truth**, usando metodi probabilistici. Questo metodo è tipico nel *master data management*, dove si raccolgono tutti i dati per poi integrarli

- **referencing**, detto anche **entity disambiguation**, dove si matchano record "sporchi" con uno "pulito", con una tabella di riferimento deduplicata che è già stata canonizzata. Questo metodo generalmente utilizzato per atomizzare più record sulla stessa chiave primaria e donare informazioni aggiuntive al record

Si hanno vari tool per la entity resolution:

- *NTLK*, un natural language toolkit
- *Dedupe**, per lo structured deduplication
- *Distance*, un'implementazione in C per distance metrics
- *Scikit-Learn*, per machine learning models
- *Fuzzywuzzy*, per fuzzy string matching
- *PyBloom*, per probabilistic set matching

In input al record linkage si ha quindi una serie di tabelle e in output un insieme di tuple che matchano e un insieme di tuple che non matchano, nel caso relazionale. Si hanno anche tuple per cui non si sa dire nulla o per le quali non si ha certezza su match/mismatch. Lo stesso ragionamento si applica anche a modelli non relazionali (dati nei nodi per un modello a grafo, valori dei documenti nel documentale etc...).
Si hanno varie tecniche per la comparazione delle coppie:

- quella **empirica**, basata sulla distanza di simboli nei valori delle tuple
- quella **probabilistica**, dove presumo di conoscere un campione di frequenze di "distanze" tra coppie di tuple, note come corrispondenti o non corrispondenti e dove potrei decidere quelli corrispondenti e quelli non corrispondenti nell'universo proiettando tale conoscenza sul campione nell'universo della coppia. Si usano soglie per poter discernere match e mismatch
- quelle **knowledge based**, dove si decide in base a regole per il match delle tuple
- **mista**, sia probabilistica che knowledge based
- le recenti tecniche di **machine learning**

Ovviamente grosse tabelle portano, confrontando tutti gli elementi ogni volta, a complessità assurde. Si hanno quindi vari step per il record linkage:

- costruzione dello spazio di ricerca iniziale come prodotto cartesiano di tutti i dataset in input. Tale spazio è immenso. Questa è una fase di preprocessamento e in questa fase si cerca di ottenere un formato unico per tutti i sorgenti, sia dal punto di vista del modello che da quello dei dati (normalizzando i vari valori di uno stesso dominio, ad esempio avendo un solo modo per definire via, senza abbreviazioni, rimuovendo spazi etc...)
- si usano tecniche di blocking per ridurre lo spazio di ricerca, ottenendo uno spazio di ricerca ridotto. Si hanno varie tecniche di blocking in letteratura e vari algoritmi. Tra tutti si ha il *sorted neighbour* dove si prende un attributo e cerco di ordinare la tabella sugli insiemi degli attributi in modo tale che i gruppi degli attributi che confronto siano vicini tra loro, tutti con tutti (esempio ordino per cognome in modo da mettere insieme record con cognomi vicini per vedere magari errori, ovviamente funziona entro certi limiti). Ci si muove tra gli elementi tramite sliding window per specificare meglio i controlli, riducendo i confronti (senza la finestra ho n^2 confronti, con n righe, con la finestra di lunghezza m righe ho $m^2\frac{n}{m}$, quindi $m \cdot n$ confronti, avendo $m \ll n$). Questa fase è facilmente distribuibile, facendo finestre di controllo in parallelo i più nodi

- sullo spazio ridotto si usa un metodo di comparazione, a cui è associato un metodo di decisione, per definire i match, i mismatch e quelli di cui non si ha certezza, che chiamiamo possible match (non sempre si hanno, dipende dalle soglie usate, una soglia di certo non permette di creare i possible match due invece si, nel mezzo tra le due si hanno i possible match). Si hanno metodi probabilistici, metodi non supervisionati come il clustering e anche supervisionati come:
  - svm
  - logistic
  - random forest
  - boosting gradient

  Confrontando quindi le varie coppie di dati

Questo è lo schema di base diciamo e in ogni fase si ha un processo di quality assesment.
Il primo problema è che i dataset potrebbero essere di formato diverso, relazionale o meno, e quindi, come i data integration, si cerca di trasformare un modello nell'altro. Potrei dover unire in un unico formato anche immagini e mappi, oltre a documenti json o csv etc...
Esempio del processo su slide.
Approfondiamo quindi l'uso di tecniche **probabilistiche** per la fase di comparazione, avendo un **probabilistic record linkage**, dove comunque le prime fasi restano normali.
Quindi si cerca la probabilità di match tra le coppie usando dei vettori di attributi:

$$P_{match} = \sum_{score \in A} w \cdot score$$

Avendo:

- w pesi tra 0 e 1, avendo peso maggiore per feature meglio predittibili
- A vettore degli attributo ciascuno con un punteggio *score*

Se il punteggio così pesato supera una soglia ottengo un match.
Si hanno anche meccanismi basate su regole (avendo magari che lo score di un attributo deve essere maggiore di una soglia e quello di un altro di un'altra soglia). Sebbene la formulazione di regole sia difficile, è possibile applicare regole specifiche del dominio rendendo più specifico l'approccio.

La soluzione proposta da **Fellegi Sunter**, quella probabilistica, è ottimale quando gli attributi di confronto erano condizionatamente indipendenti, permettendo comunque di avere un'area di ambiguità.
Bisogna quindi per ogni coppia definire una funzione di distanza, imponendo un insieme finito di attributi da usare per valutare tale distanza, potrebbero esserci attributi non rilevanti. Un modo "ottimale" è scegliere attributi dello stesso significato e con la maggior accuratezza possibile per evitare errori. Si hanno anche attributi discriminatori per i quali i domini contengono un alto numero possibile di valori (ad esempio la lista di nomi, anche se in questo caso la distribuzione è eccessiva, o dei cognomi, già più significativa). Meno attributi si prendono in considerazione più rischio falsi positivi, aumentando i matching, anche se aumento le prestazioni. Scegliere troppi attributi può portare a falsi negativi, avendo tantissimi mismatch. È quindi un discorso che va approfondito di caso in caso.
L'algoritmo di Fellegi Sunter, per la fase decisionale, suggerisce di campionare la frequenza di match e non match, prendendo una coppia che si sa già che matcha e valutando per ogni distanza la frequenza di match/mismatch (???). Per ogni distanza avrò alla fine associata una percentuale di match. Si fa un ragionamento in stile SVM anche se senza iperpiani ma con un separatore più complesso.
Esempio su slide.
All'aumentare della distanza ovviamente aumentano i match. Alla fine del ragionamento posso produrre le soglie (o la soglia) da usare per lo studio dell'intero dataset.
Si rischiano comunque falsi positivi e falsi negativi che possono essere ridotti aumentando i possible match, allargando le soglie.
Il record linkage, nel calcolo della distanza, è comunque molto legato al dominio.
Approfondiamo ora le tecniche di comparazione basate sul machine learning. Si hanno approcci supervisionati, SVM, random forest, conditional random fields (nel caso di formati non relazionali) etc...
Ci sono anche tecniche non supervisionate, tramite clustering, con l'idea di usare K-means per raggruppare elementi simili. Posso fare anche clustering gerarchici p usando la expectation maximization. Si hanno anche tecniche di apprendimento attivo come il commitee of classifiers.
Matchare testi descrittivi lunghi è difficile ma con tecniche di distanza non euclidea e di embedding, in un contesto comunque di machine learning, si riesce ad ottenere dei risultati.
In generale comunque il record linkage deve essere veloce, anche a discapito di un minimo di qualità.
Bisogna quindi valutare la qualità in termini di:

- true positive, TP
- true negatives, TN
- false positive, FP
- false negative, FN

avendo:

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

Avendo un'elevata conoscenza di dominio posso usare tali conoscenze per la comparazione e la cosa è difficilmente generalizzabile.

7.1.2 Data fusion

Passiamo quindi alla fusione dei record che rappresentano lo steso oggetto nel mondo reale.
Bisogna gestire i conflitti. Si hanno varie tecniche *(traduzione a spanne da slide, il prof non ha detto quasi nulla*):

- **ignorare il conflitto**, non è una buona soluzione. Queste strategie non consentono di prendere la decisione in base a ciò che è in conflitto con i dati e a volte non vengono rilevati i conflitti di dati. Esempio di strategia di ignoramento è **Pass It On**, che presenta tutti i valori e quindi rimuove il conflitto con la soluzione all'utente
- **evitare il conflitto**, con strategie basate su istanze e metadati, scegliendo di prendere istanze senza conflitti, prendendo silo valori consistenti senza dire niente degli altri. Queste strategie riconoscono l'esistenza di possibili conflitti generali, ma non rilevano e risolvono i singoli conflitti esistenti. Invece, gestiscono dati in conflitto applicando una decisione unica allo stesso modo a tutti i dati, ad esempio preferendo un dato da una fonte speciale con la strategia **Trust your friend**
- **risolvere il conflitto**, prendendo una sorgente più sicura e facendo una scelta a caso su vari aspetti, non arrivo ad una soluzione ottima (???). Queste strategie considerano i dati e i metadati prima di risolvere il conflitto. Si hanno due strategie:

1. decidere quando scegliere un valore tra tutti i valori già presenti (ad esempio il dato più recente vedendo i log)
2. strategie di mediazione, quando scelgono un valore che non necessariamente esiste tra i valori inconsistenti

La scelta di fusione deve essere comunicata a chi poi usa i dati altrimenti si rischiano molti problemi (magari si è scelto di scegliere lo stipendio minimo in caso di inconsistenza e un addetto della contabilità deve saperlo). È il tema della **provenance dei dati**, il tenere traccia delle decisioni prese. Con al data fusion si ottiene potenzialmente qualcosa con pochi NULL.
Su slide esempio con intero caso di studio.
Alcune note con conclusive.
La **data preparation** è un buon modo per ottenere ottimi risultati, tramite normalizzazione di dati e schemi e *imputation*. Si rendono più facili i riscontri tramite distanze sintattiche. Nel data normalizazion si cerca di:

- convertire tutto in lowercase e rimuovere gli spazi
- usare uno spell checker per rimuovere typo conosciuti
- espandere abbreviazione
- convertire nickname
- fare lookup lessicali
- applicare tecniche di stemming, tokenizzazione lemmatizzazione delle parole (bho ???)

Per lo schema normalizazion si ha il tentativo di:

- matchare nomi di attributi che specificano la stessa cosa
- spezzare attributi dove necessario
- gestire valori nulli
- rappresentare i dati in modo strutturato

Si ha poi l'imputation, la fase di decisione, dove si cerca di:

- capire come gestire/comparare i NULL
- capire come gestire/comparare, in modelli non relazionali, l'assenza di valori, magari mettendo nei none o valori medi con una certa distribuzione, a seconda del caso

Queste scelte impattano molto sul record linkage.

Data conflict

Studiamo quindi i vari **data conflict** che possono occorrere parlando di data fusion.
Si hanno vari tipi di conflitti, tra cui quelli di formato, di battitura, con typo, similarità di stringhe con semantica diversa etc... Un altra fonte di conflitti è data dall'uso di software diversi per diverse sorgenti per rappresentare però le stesse informazioni. Un altro caso e avere tanti dati storicizzati messi in db diversi che devono essere integrati ma che sicuramente per i discorsi appena fatti avranno conflitti.
Si hanno conflitti :

- **intra-source**, avendo la stessa informazione nello stesso sorgente rappresentata in modo diverso. Questo può accadere non avendo applicato controlli su integrità, consistenza, errori di typo, calcoli scorretti, dati ridondanti, valori obsoleti, varianti degli stessi valori etc...
- **inter-source**, avendo la stessa informazione in diversi sorgenti rappresentata in modo diverso (non avendo consistenza globale), anche se consistenti localmente. Questo può accadere a causa di diversi tipi di dato, avendo varianze sulla lingua, nello "spelling", nelle convezioni linguistiche o negli standard (si pensi a rappresentazione di indirizzi, numeri telefonici etc...)

Per risolvere si fa in primis data fusion delle tabelle, facendo un lavoro di data management per fondere i dati di diverse sorgenti dati. Il secondo step è combinare tali informazioni, integrando tanti dataset, che rappresentano le stesse entità (tramite *join* sulle primary key), anche da dataset diversi (che non hanno le stesse chiavi primarie, spesso). Bisogna quindi eliminare i conflitti e un primo problema è quello di correggere tali errori. Si hanno varie tecniche, spesso in sequenza:

- usare una **tabella di riferimento**, magari per città, prodotti o cose simili, applicando misure di similarità tra la nostra tabella e quella di riferimento
- standardizzare e trasformare i dati a seconda di un unico standard (vie nello stesso formato etc...)
- sfruttare la conoscenza del dominio, conoscendo le varie convenzioni, ontologie, dizionari (anche di sinonimi e contrari), tesauri

(ovvero il lessico dei termini relativi a un ambito generale o specifico di conoscenze, collegati tra loro in una rete gerarchica e relazionale) etc...

- fare *outlier detenction* ed eliminare i dati riconosciuti come errori

Si passa solo ora al vero e proprio *data fusion* nel processo di integrazione. Riprendendo più in generale i vari step di data integration (anche solo tra due sorgenti) si ricordano quindi (**esempio si slide**):

- **schema mapping**, creando i mapping tra gli attributi dei due sorgenti, scegliendo, in caso di riconosciuta similarità (con differenza riconosciuta per un typo o per una traduzione), quale nome di attributo mantenere. Si fa quindi una sorta di **schema integration**
- **data trasformation**, ovvero le istanze due due sorgenti vengono trasformate e normalizzate per la nuova sorgente con lo schema già integrato, avendo nuove entità per questo nuovo schema
- **duplicate detenction**, ovvero vengono riconosciute duplicazioni che rappresentano la stessa informazione (sempre con un discorso di similarità)
- **data fusion**, dove le duplicazioni appena riconosciute vengono effettivamente fuse, secondo varie strategie

Si effettua **schema matching**, considerando tre dimensioni di qualità:

1. completeness, completezza
2. conciseness, concisione
3. correctness, correttezza

Si cercano attributi semanticamente uguali, integrando su quegli attributi, ma questo può generare record duplicati. Si fa quindi duplicate detection risolvendo con data fusion incertezze e contraddizioni. Si hanno quindi (*io non ho capito*):

- **extensional completeness**, aggiungendo a livello di record informazioni esterne
- **extensional conciseness**, per l'unione di dati a livello di record

- **intensional completeness**, aggiungendo a livello di schema informazioni esterne
- **intensional conciseness**, per l'unione dei dati a livello di attributi

Parliamo meglio di *duplicate detenction*. Lo scopo è trovare oggetti che rappresentano la stessa entità del mondo reale ma questi duplicati non sono identici ma si hanno misure di distanza per la similarità, tra cui:

- distanza di Lavenshtein
- distanza di Soundex
- distanza di JAccard
- ...

Avendo grandi quantità di dati non posso poi pensare di comparare ogni singola coppia di dati provenienti dai vari dataset. Si hanno quindi stategie di *partitioning* come:

- sorted neighbourhood
- blocking
- ...

Il problema di *data fusion* consiste in creare un singolo oggetto a partire dai duplicati, risolvendo i conflitti a livello di valore. Si hanno però valori a NULL da capire come risolvere, si hanno contraddizioni, si ha incertezza sulla verità dei valori (dovendo quindi "modellare" l'incertezza), si hanno metadati (come preferenze, aggiornamenti dei dati, correttezza), si ha che bisogna cercare di conservare il valore originale e l'origine dei dati (si parla di *provenance*) e infine dell'implementazione di tutto questo dei DBMS. In figura 7.1 possiamo notare vari campi del data fusion di cui approfondiamo alcuni aspetti. Parliamo ora di **resolution stategies** e di **resolution functions**.
Uno dei problemi del data fusion è appunto quello di modellare incertezze e contraddizioni. L'incertezza è creata dal confronto tra valori e valori NULL per lo stesso oggetto reale mentre la contraddizione è data da valori non NULL diversi per lo stesso oggetto reale. I NULL infatti hanno diversi significati dal punto di vista della semantica:

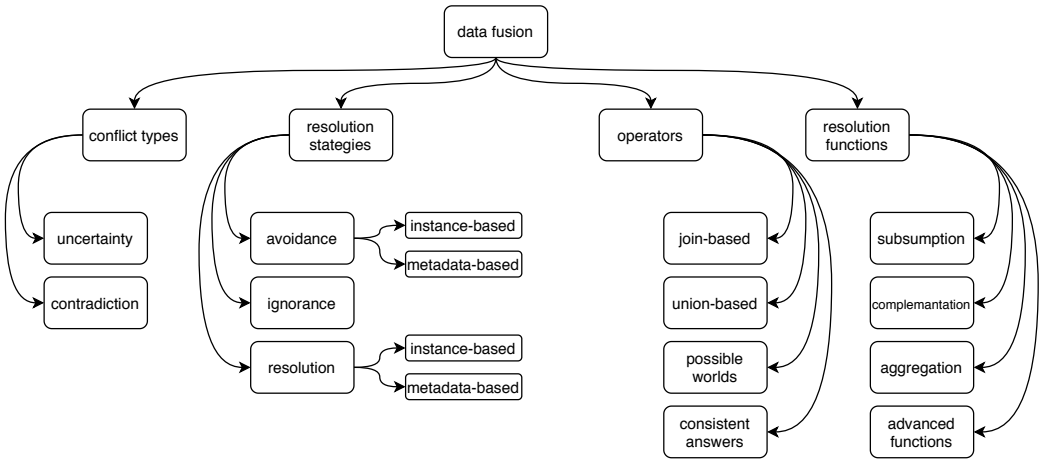


Figura 7.1: Albero con i vari campi del data fusion

- dovrebbe esserci un valore ma non lo si conosce (esempio so che una persona ha una data di nascita ma non la so)
- non è applicabile un valore ad un certo attributo in un certo contesto (esempio una persona single ha valore NULL per sposato)
- c'è un valore nascosto che non sono autorizzato a vedere (esempio il numero di telefono in alcuni casi)

Si hanno varie strategie quindi. Per il **conflict avoidance** si hanno:

- **instance-based**: **take the information, non gossiping**
- **metadata-based**: *trust your friends*

Per il **conflict resolution** si hanno:

- **instance-based**:
  - **deciding**: *cry with the wolves*, *roll the dice*
  - **mediating**: *meet in the middle*
- **metadata-based**:
  - **deciding**: *nothing is olther than the news of yesterday*

Come funzioni che risolvono i conflitti abbiamo (da usare a seconda del contesto):

- *min, max*, ovvero funzioni standard di aggregazione usabili per quantità misurabili (*conflict resolution, instance-based, deciding*)
- *sum, avg*, ovvero funzioni standard di aggregazione usabili per quantità misurabili (*conflict resolution, instance-based, mediating*)
- *count*, una funzione standard di aggregazione usabili per quantità misurabili (*categoria non specificata*)
- *random*, ovvero prendere valori random (*conflict resolution, instance-based, deciding*)
- *longest, shortest*, prendendo valori più lunghi/corti (*categoria non specificata*)
- *choose(source)*, scegliendo valori da particolari sorgenti (*conflict avoidance, metadata-based, deciding*)
- *chooseDepending(value, column)*, scegliendo valori dipendenti da valori in una certa colonna (*conflict avoidance, instance-based, deciding*)
- *vote*, con una votazione maggioritaria, ovvero se più sorgenti dicono che un si ha un certo valore allora probabilmente è quello giusto (*conflict resolution, instance-based, deciding*)
- *coalesce*, ovvero una funzione che prende il primo valore non NULL (*conflict avoidance, instance-based, deciding*)
- *group, concat*, per raggruppare o concatenare i valori (*conflict avoidance, instance-based, deciding*)
- *mostRecent*, per scegliere il valore temporaneamente più recente (*conflict resolution, metadata-based, deciding*)
- *mostAbstract, mostSpecific*, per scegliere il valore più astratto/-specifico (*conflict resolution, metadata-based, deciding*)
- *commoAncestor* per prendere il common ancestor (*conflict resolution, metadata-based, mediating*)
- *escalate*, per esportare i valori in conflitto (*conflict ignorance*)

Dal punto di vista degli **operator** si ha che, dati due sorgenti A e B (di tre colonne coincidenti solo nelle prime due, quindi A ha X, Y, Z e B ha X, Y, W):

- due tuple sono identiche se ho una cosa del tipo, a partire dai due sorgenti (con $-$ si indica che i valori dalle due NULL):

  $$a, b, -$$

  $$a, b, -$$

  Quindi sia Z che W non hanno valori mentre le colonne coincidenti hanno lo stesso valore. Quindi in fase di fusione avrei una cosa del tipo:

  $$a, b, -, -$$

  $$a, b, -, -$$

  e quindi:

  $$a, b, -, -$$

- due tuple sono *subsumed* ovvero una contiene l'altra se ho (sottoinsieme di tuple):

  $$a, b, c$$

  $$a, b, -$$

  Quindi in fase di fusione avrei una cosa del tipo:

  $$a, b, c, -$$

  $$a, b, -, -$$

  e quindi:

  $$a, b, c, -$$

- due tuple si completano se ho:

  $$a, b, c$$

  $$a, b, d$$

  Quindi in fase di fusione avrei una cosa del tipo:

  $$a, b, c, -$$

  $$a, b, -, d$$

  e quindi:

  $$a, b, c, d$$

- conflitto di tuple se ho:

  $$a, b, c$$

  $$a, e, d$$

  Quindi in fase di fusione avrei una cosa del tipo:

  $$a, b, c, -$$

  $$a, e, -, d$$

  e quindi:

  $$a, f(b, e), c, d$$

  indicando con $f(b, e)$ che devo usare una funzione per risolvere il conflitto

Quindi si hanno gli operatori per:

- tuple identiche: *union, outer union*, dove con la *union*
- tuple *subsumed* (con quindi incertezza): *minimum union*
- tuple complementari (con quindi incertezza): *complement union, merge*
- tuple in conflitto:
  - approcci relazionali: *match, group, fuse*
  - altri approcci: *possible worlds, probabilistic answers, consistent answers*

I vari operatori risolvono i problemi di duplicazione associati alle tuple.
Su slide tabella degli operatori ed esempi... anche perché non ha detto nulla la prof.
Riprendendo le tecniche di **truth discovery**, ovvero capire quale sorgente dice il vero su un certo record si hanno strategie come *trust your friend* e *cry with the wolves*.

Capitolo 8

Big data

Si studia ora la tematica dei **big data** ovvero come definire piattaforme per la gestione di un grosso volume di dati.
Non si ha una definizione formale di big data ma nel 2012 Gartner disse:

Big data is high volume, high velocity, and/or high variety information assets.

Anche se si resta molto nel generico. Si identificano comunque le prime **tre "V" dei big data**:

- **volume**, parlando di *data at rest*
- **variety**, parlando di *data in many form*
- **velocity**, parlando di *data in motion*

Successivamente si è anche aggiunta la **quarta "V" dei big data**:

- **veracity**, parlando di *data in doubt*

indicando che non si ha effettiva fiducia nei dati prodotti. Si hanno poi altre due "V" (spesso legate al concetto di **open data**, al come si usano i dati):

- **visibility**, parlando di *data in the open*
- **value**, parlando di *data of many value*

Analizzare i dati su singola macchina è lento, costoso e difficile. La prima idea è quindi quella di distribuire i dati, distribuendo il loro storage ma anche il calcolo, in sistemi paralleli. Bisogna quindi studiare sincronizzazione, deadlock, bandwith, coordinazione, fallimenti etc. . ., tipici dei sistemi paralleli. Negli ultimi anni, prima grazie ai grossi player e poi grazie alle soluzioni

cloud, si hanno ottime soluzioni dal punto di vista del parallelismo soluzioni complesse "sotto la scocca" ma semplici da usare ad alto livello.
L'elemento fondamentale è quindi il passaggio da solo storage all'aggiunta dell'elaborazione di grandi quantità di dati, possibilmente con scalabilità lineare (all'aumentare dei nodi scalo linearmente). Le elaborazioni vengono quindi spostate dove sono i dati e non viceversa infatti normalmente ci si connetteva al db, si effettuava la query, si scaricavano in locale i risultati e si elaboravano. Ma questo non è efficiente con grandi quantità di dati (anche nell'ordine di terabyte) quindi conviene spostare l'elaborazione (che pesa molto meno in termini di codice sorgente). Ovviamente bisogna approfondire il tema della gestione dei fallimenti (i nodi possono cadere), soprattutto girando su hardware economico (commodity). Si hanno anche problemi di estensibilità che devono essere gestiti.
I dati ormai sono il "nuovo petrolio" in qaunto bisogna *trovarli*, *estrarli*, *rifinirli*, *distribuirli* e usarli per guidare l'economia.
Si hanno:

- **grandi giacimenti di dati**, come i social, il web e gli opendata
- **piccoli giacimenti di dati**, come quelli locali di sensori, IoT, CRM (di piccoli volume)

Per estrarre i dati si hanno query, app dedicate, wrapper, log, stream etc...
Si hanno sorgenti nativamente relazionali (come CMR), alcuni nativamente non relazionali, alcuni sono fermi altri in movimenti, alcuni da sensori etc... comportando varie tecniche di estrazione.
Dal punto di vista dello storage si hanno due soluzioni:

- **on-premises** (in casa)
- **on-cloud**

La scelta è fondamentale in ottica **ecosistema digitale**, ovvero un insieme di tecnologie che si usano bene insieme. Sceglitore dove mettere i dati comporta scelte specifiche sull'ecosistema digitale. Al momento i grandi produttori di cloud (Amazon, Microsoft, Google, SAP etc...) hanno un ecosistema digitale per cui diventa anche difficile scegliere il migliore in base ai propri utilizzi. Spesso si hanno comunque accordi commerciali con alcuni produttori. Le soluzioni cloud comunque "vincolano" ad un certo ecosistema, rendendo scomodo lì'uso di tecnologie di altri produttori.
I dati fanno comunque raffinati e puliti, si hanno procedure di estrazione, trasformazione e caricamento, studi di privacy, integrazione e analisi dati con data mining e analytics.

Si ha anche un problema di trasporto, garantendo interoperabilità per i dati grezzi, mentre le piattaforme di big data costruiscono dei **wallet garden** ovvero dei "giardini murati", dove si ha integrazione solo all'interno di quel dato ecosistema digitale. Si creano dei lock-in per cambiare provider molto importanti (dovendo in caso riscrivere molto codice per "trasportare i dati da un giardino all'altro").
In una architettura di big data si ha quindi, oltre al sistema di storage, anche un sistema di analisi dati. I darti vengono caricati, gestiti e messi in un'area di staging. I dati vengono quindi processati e poi gestiti dal punto di vista della sicurezza, del monitoraggio e della pulizia. Bisogna ovviamente anche gestire l'accesso ai dati. In generale si ha un hardware fisico, un framework per lo storage etc...
Le nuove piattaforme di analisi dati non nascono in ambiente relazionale e i risultati di tale analisi, insieme ai dati relazionali, finiscono nel data warehouse. Un'alternativa posso usare, se ho solo dati relazionali, un datawarehouse. Se si aggiungono dati non relazionali posso anche pensare di aggiungere semplicemente un **data lake**, un'architettura di big data analytics che fa analisi anche con machine learning (usando ad esempio R) e poi contribuisce a popolare il datawarehouse.
Non sempre avere più fati comunque è la soluzione migliore, in quanto da la percezione di avere analisi per forza migliori (ma non sempre è vero).
Il rischio di overfitting comunque è sempre dietro l'angolo, non sempre i dati sono la risposta e non sempre l'analisi dati porta a correlazioni sensate (spesso da modelli predittivi, basati comunque su dati biased, si passa a modelli prescrittivi).

8.1 HDFS

Concentriamoci sullo storage di grandi quantità di dati.
I dati devono poter essere distribuiti per essere poi elaborati. Il sistema software che permette di memorizzare dati che potenzialmente possono essere di qualunque volume, con qualunque tasso di velocità di arrivo e di qualunque tipo (per le prime tre "V", volume, velocity, variety, avendo *any size, any rate, any type*). La tecnologia più efficace non è nulla che abbiamo già visto (pensare ad un NoSQL obbliga ad un certo formato di dato) ma dobbiamo ragionare in ottica di file system. Abbiamo quindi file system apposta per dati che devono essere distribuiti, tra cui **Hadoop Distributed File System (*HDFS*)**, derivato open di **Google Distributed File System (*GDFS*)**. HDFS è molto performante anche sulle commodity, consente replica tra nodi anche con fallimento di nodi, ha un meccanismo nativo per cui ogni frammen-

to di dato è replicato su tre nodi diversi e ha un approccio **write once read many** (d'aiuto nella tipica situazione di raccolta dati dai social e successiva analisi, sentiment analysis etc...). I dati non si spostano verso i "workers" ma è l'opposto. Prima distribuisco i dati sui vari nodi di elaborazione ed eseguo lì i vari task. Questo permette di gestire il "collo di bottiglia" dei dischi, che hanno accesso lento ma velocità effettiva accettabile, e di gestire meglio la RAM.
Si fa quindi *scale out* e non *scale up*, avendo hardware economico, si ragiona la gestione di pochi file grandi da elaborare di volta in volta e si immagina un approccio di tipo *write once read many*.
L'idea di HDFS consiste nel prendere un file lo si spezza in blocchi da 64MB (o multipli, spesso 128MB), ogni pezzo, detto *chunk*, viene distribuito in almeno tre nodi. Si ha poi un'architettura master-slave che coordina la distribuzione e l'accesso. Si ha il nodo **namenode** che ha tutte le informazioni dell'infrastruttura HDFS (il master) e una serie di **datanode** (gli slave) su macchine poco costose dove si salvano i chunk e che mandano periodicamente degli heartbit al namenode (in modo che il namenode percepisca i guasti e provveda con nuove repliche). Le richieste si fanno al **namenode** che provvede poi a reindirizzare verso i giusti **datanode**. Il namenode è il single point of failure ma si ha sempre un **secondary namenode** verso il quale periodicamente si fa una transazione dei log del namenode. I datanode non dialogano tra loro ma solo con il namenode avendo un concetto di **cluster membership** per permettere al namenode di coordinarli.
Il write once read many contribuisce alla coerenza dei dati (tante richieste e poche scritture, assumendo un carico di lavoro di tipo **OnLine Analytical Processing (*OLAP*)**) e tramite il namenode si permette al client di accedere direttamente ai giusti datanode coi chunk desiderati.
Tipicamente si ha una replica nello stesso rack, una in un rack vicino e una terza molto lontano (ulteriori repliche i nodi arbitrari).
Il namenode ha quindi tutte le informazioni in memoria, l'elenco dei file, l'elenco dei blocchi, i soliti attributi/metadati di un file system (nome dei file, nome delle directory, owner, dimensione etc... non il tipo, che non è dato dall'estensione) e i log.
Il namenode quindi coordina i vari elementi e le operazioni tra file. Il namenode si occupa anche di curare il "benessere" del sistema facendo repliche e ribilanciamento, eliminando nodi morti.
I datanode invece ricevono i dati e verificano che siano corretti. Inoltre segnalano il loro stato al namenode. La correttezza del dato viene verificata tramite checksum, usando CRC32.

8.2 Map reduce

Si passa ora all'elaborazione dei dati.
Map reduce è un modello di programmazione/engine, ormai in disuso, utile per capire nuove tecnologie. Questo engine permette di scrivere programmi che sono realizzati in stile funzionale e eseguiti in parallelo su grandi cluster di macchine poco costose. In un programma di map reduce si hanno due fasi:

- **map**, ovvero una funzione che processa coppie chiave/valore per generare coppie intermedie chiave/valore. Indicando con k key e v value:
$$(k, v) \rightarrow [(k', v')]$$
- **reduce**, ovvero una funzione che mergia tutti i valori intermedi associati alla stessa chiave intermedia:
$$(k', [v']) \rightarrow [(k', v')]$$

Si hanno poi:

- **partition**, ovvero il numero di partizioni di k'. Spesso poi si usa un hash della chiave:
$$hash(k')\ mod\ n$$
dividendo lo spazio delle chiavi per l'esecuzione parallela della reduce
- **combine**, ovvero:
$$(k', [v']) \rightarrow [(k', v'')]$$
dove si hanno dei mini-reducers in memoria dopo il map usati come task di ottimizzazione per ridurre il traffico di rete

In pratica si prende l'insieme di attività e lo si divide, nella *partition*, in vari "worker" che producono risultati che poi vengono messi insieme, nella *combine*. Ovviamente bisogna studiare ogni step di divisione lavori, gestione degli stessi e unione dei risultati:

- come assegnare lavori alle unità di lavoro
- cosa succede se ci sono più work che workers
- cosa succede se i workers hanno bisogo di condividere risultati parziali

- come aggregare risultati parziali
- come capire che tutti i workers hanno terminato le loro attività
- cosa fare se un workers muore

L'idea base consiste, nella map, nell'iterare su un gran numero di record, possibilmente eseguiti localmente, ed estrarre qualcosa di interessante da ognuno. Si mettono poi insieme i risultati ordinando i risultati intermedi. Infine, nella reduce, si aggregano i risultati e si genera il file di output. Tra map e reduce su ha una fase di **shuffle & sort** . Il framework si occupa di tutto il resto, il developer si occupa solo della componente funzionale (map, reduce, partition e combine) .
Quindi map reduce si occupa, su un filesystem distribuito, di:

- gestione scheduling, rassegando i task di map e reduce ai worker
- gestione di data distribution, spostando i processi verso i dati
- gestione della sincronizzazione, ottenendo, ordinando e unendo i dati intermedi
- gestione di errori e fallimenti hardware, effettuando l'eventuale ripristino

La comunicazione tra i nodi avviene in HDFS. Si istanzia un job di map reduce e il client chiede al **job tracker**, che tiene conto dei vari job, di far partire il lavoro. Il job tracker cerca tra i nodi liberi e assegna i job da far eseguire, i cosiddetti **task tracker**. Il task tracker viene eseguito localmente e riceve i parametri dal client e lancia su una vm locale o la map o la reduce. Il job tracker periodicamente cheide se tutto funziona tramite heartbit.
(**esempi su slide per map reduce**)
Soluzioni moderne si trovano con **Spark**, che lavora in memory e non su file system etc...

8.3 Hadoop

Uniamo HDFS e map reduce e otteniamo **Hadoop**.
Hadoop quindi è un framework dove si ha un sistema di storage distribuito e un sistema di calcolo parallelo. Su alcuni nodi si ha il job tracker e il namenode (magari anche su macchine diverse). Sui nodi meno costosi si hanno poi datanode e task tracker. Di fatto è sempre master/slave con single point of failure con gestione repliche etc...
Nell'ecosistema di Hadoop si hanno anche:

- **pig** come linguaggio di scripting che viene tradotto in map-reduce. Esegue problemi di data analysis come un workflow ETL
- **hive** che fornisce un'interfaccia SQL-like per i dati in HDFS, in pratica si tratta il tutto alla datawarehouse (**Apache Hadoop based data warehouse**). Infatti map reduce aveva limiti (non riusabilità, complessità etc...) e quindi si è pensato di aggiungere questo strato per l'interazione con SQL. Hive è intuitivo, genera automaticamente piani di esecuzione per le query, consente analisi alla datawarehouse e consente di tradurre query tramite esecuzioni di piani Hadoop Map Reduce. **Esempio su slide**. Architetturalmente si ha un'interfaccia web, si scrive la query, si ha la fase di parser, optimizer e planning, si eseguono script di map reduce user-defined o job già presenti di default (???). I dati vengono quindi salvati in HDFS. Si ha un software quindi che serializza/deserializza i dati traducendo l'SQL (???). LA cosa viene usata in machine learning, data/text mining etc...

8.3.1 YARN

Nella versione 1.0 di Hadoop map reduce doveva fare troppe cose quindi alla versione 2.0 si è deciso di disaccoppiate le funzioni di gestione del cluster e quelle di gestione dei singoli job, usando gli slave per la gestione dei job. Si ha quindi **Yet Another Resource Negotiator (YARN)**. Ora quindi hanno un **resource manager (RM)** globale, che gestisce l'intero cluster, e un **application manager (AM)** per ogni applicazione. Per creare un nuovo job si chiede all'RM le risorse e poi l'AM le gestisce come in map reduce.
In Hadoop 2.0 si ha quindi HDFS, YARN come *data operating system* e vari motori di elaborazione dati (Map Reduce, spark etc...). Su questi poi si hanno, per esempio:

- hive
- programmi java
- Accumulo, key/value
- Hbase, colonnale
- Storm
- Spark, un processing engine in memory

Un esempio di architettura è **Cloudera** dove si hanno, oltre a HDFS, Hbase, YARN:

- Impala, un engine SQL
- Mahout, per il machine learning
- Sqopp, per il passaggio tra SQL e HDFS e viceversa
- Flume, un servizio distribuito, affidabile e disponibile per raccogliere, aggregare e spostare in modo efficiente grandi quantità di log di dati

Si ha quindi HDFS per lo storage (insieme a soluzioni specificatamente non relazionali come Hbase) e poi si ha un ecosistema digitale con vari servizi.
Le piattaforme cloud offrono soluzioni simili con varie data platform (con sistemi di storage ed elaborazione, con un data lake dove si buttano i dati che poi vanno analizzati e processati):

- AWS
- AZURE
- Google Cloud Platform (un filo meno evoluto e meno strutturato)

Bene o male sono tutti molto simili.
Su slide immagini delle varie data platform e breve spiegazione dei vari strumenti

8.4 Big Data integration

Queste esercitazioni sono tra le peggiori mai sentite nella mia vita. La prof parlava in modo completamente sconnesso, ho cercato invano un senso logico, non l'ho trovato. Non ho messo particolari "(???)" per segnalare dubbi perché ne servirebbe uno ogni 5 parole. Auguri!
Con Big Data integration si intende l'unione di big data e data integration. Con i big data ci si concentra su volume dei dati, velocità di collezione e analisi, priorità, valore dei dati, varierà dei dati, veridicità dei dati etc...
Si hanno due soluzioni principali:

1. **data warehouse** che consiste nel creare un unico archivio (tramite materialized view) di dati da diverse fonti offline. Questo è un business milionario. Essendo i dati materializzatati (tramite liste

materializzate) insieme è veloce fare query ma si hanno problemi di consistenza

2. **virtual integration** che consiste nel supportare la query su uno schema mediato (sistema wrapper/mediator (???)) applicando la riformulazione della query online. Si ha il link e la fusione delle varie risposte ricevute

Bisogna in ogni caso accedere a tante sorgenti (e non tanto di dati) che cambiano molto nel tempo. Tutto quanto appena detto sui big data in merito ai dati diventa in merito ai sorgenti: volume dei sorgenti, velocità di collezione e analisi, priorità, valore dei sorgenti, varierà dei sorgenti, veridicità dei sorgenti etc...
Fare big data integration è importante per costruire grafi di conoscenza di grandi dimensioni, detti anche **knowledge bases** (si pensi ad esempio a quella di Google, **google knowledge graph**). Una knowledge base è un'insieme di entità con relazioni tra loro dove le classi delle entità e le relazioni stesse vengono definite in un vocabolario. Si ottiene quindi una rete di connessioni. L'altro fattore per cui è importante è perché dai knowledge graph possiamo ottenere dati geo spaziali, utili in vari scopi della società. Un altro fine è per fare analisi scientifica, magari di geni, malattie etc...
Fin'ora si è studiato il "small" data integration, che era molto più semplice, dove comunque ogni "pezzo" arrivava da un certo sorgente, ma in ongi caso in termini "piccoli". Si avevano quindi, nello "small" data integration:

- **scheme alignment**, che vuole risolvere il problema della struttura. Se si hanno due sorgenti con schemi diversi si "allineano" ad uno schema unico. Spesso si ha eterogeneità nei modi di esprimere classi, relazioni e attributi (anche in sorgenti con lo stesso schema). Ci si concentra quindi sullo schema

- **record linkage**, dove si individuano delle istanze che sono le stesse nonostante l'ambiguità di rappresentazione nei vari sorgenti. Sposto quindi l'attenzione sul contenuto, avendo spesso eterogeneità nel rappresentare le stesse entità

- **data fusion**, dove si uniscono e riconciliano i contenuti che sono in mismatch nei vari sorgenti. Si cerca di risolvere, secondo varie tecniche i conflitti di valore

risolvendo le varie ambiguità, di semantica, rappresentazione (nonché inconsistenze) etc... tramite pipeline con appunto queste tre fasi.
Nel caso dei big data si hanno però moltissimi dati e conseguenti inconsistenze, incompletezze, duplicazioni etc... Si hanno quindi varie tecniche.

8.4.1 Schema alignment

Analizziamo un attimo la prima fase.
Partiamo nel caso, passando al big data, in cui si hanno dati molto specifici in un certo dominio. In primis ci si deve chiedere come sono distribuiti tali dati nel web, studiando eventuali provider di tali dati, che li raccolgono da diverse sorgenti etc... Si studiano eventuali incongruenze. Analizziamo meglio la distribuzione dei dati. Bisogna capire come queste informazioni possono essere, tramite tecniche di information extraction, raccolte da diversi sorgenti. Bisogna anche capire il volume e la copertura di tali informazioni, che sono sullo stesso dominio ma presi da sorgenti diverse. Vediamo uno studio. Con DMP12 si sono studiati 9 domini con accesso a grandi db di dominio specifico, con entità che hanno attributi che possono valere come identificatori. In questo caso la metodologia dello studio comprendeva:

- utilizzare tutta la cache nell'engine di ricerca
- le pagine web hanno entità con un identificatore
- si aggregano insiemi di tutte le entità da ogni sorgente

Con lo studio si è ottenuto che fin da subito ha valori molto alti di copertura delle informazioni **1-coverage** (**grafici su slide**) e che resta stabile aumentando i sorgenti. Già con 100 sorgenti su 100000 si aveva copertura massima. La statistica **k-coverage** mi raccoglie la ridondanza delle informazioni presenti, avendo lo stessa frazione di dati in almeno k sorgenti diversi. Per una **5-coverage** ho copertura al 90% con 5000 sorgenti e al 95% con 100000 sorgenti. I dati erano in merito ai numeri telefonici dei ristoranti. Al variare dei dati cambiano le statistiche: una **1-coverage** sui siti dei ristoranti si ha l'80% con 100 sorgenti e il 95% con 10000 (copertura che calava molto aumentando il k, grafici su slide).
In merito a quanto bene erano connessi i sorgenti in un dato dominio si può ragionare in merito all'engine di ricerca. Con DMP12 si è considerato il grafo entità-sorgente per vari domini. Si ha un grafo bipartito con entità e sorgenti come nodi, collegati da archi se un'entità è in un certo sorgente. Si è quindi studiato il grafo, studiando la sua connettività si studia la robustezza dell'algoritmo di estrazione delle informazioni al variare delle scelte iniziali di architettura. Si parte da un insieme di entità, dette **seed**, e a partire da queste se ne cercano altre su altri sorgenti, contando le iterazioni necessarie per arrivare ad una *convergenza*, alla quale ci si ferma. Si studia anche il *diametro* del grafo (esempio su slide). Si studiano anche ridondanza dei dati e eventuali overlap.
Facciamo quindi delle considerazioni fin ali dello studio.

Per quanto riguarda la distribuzione anche per i domini con forti aggregatori, dobbiamo studiare a fondo i sorgenti per costruire un database ragionevolmente completo, sopratutto per un k alto di coverage (per aumentare la confidenza). In merito alla connettività si ha che le orgenti in un dominio sono ben collegate, con un alto grado di ridondanza e overlap dei contenuti e questo resta vero anche se alcuni aggregatori principali vengono rimossi.
Un secondo studio (grafici sempre su slide), LDL+13, è stato fatto su due domini, sulla qualità del web. Bisogna sperare in questo caso di avere dati "puliti", sapendo che una bassa qualità comporta un grande impatto.
Si è quindi studiata la consistenza dei dati, scoprendo una forte inconsistenza, a causa di ambiguità semantica in primis. Un altro problema sono gli errori di unità (precisione diversa) o errori veri e propri (con valori incompatibili). Spesso le sorgenti si copiano a vicenda senza controllare la validità dei dati copiati, portando a problemi di stima (facendo sembrare statisticamente corretto un valore che non lo è).
Un altro problema è quello del volume dei dati, avendo milioni di sorgenti con dati specifici per certi domini. Nel web, secondo CHW+08, si hanno 154 milioni di tabelle, 10 milioni di sorgenti ad alta qualità, per MKK+08, 10 milioni di tabelle relazionali, per EMH09. Per i big data si hanno evidenti difficoltà quindi di schema alignment. Fare data warehouse sarebbe costosissimo ed è complicato fare virtual integration.
Continuando coi problemi si ha la velocità:

- da 43.000 a 96.000 fonti Deep Web (con form HTML), secondo B01
- 450.000 database, 1,25 milioni di interfacce di query sul Web, secondo CHZ05
- 10 milioni di fonti Deep Web di alta qualità, secondo MKK+08
- Molte fonti forniscono dati in rapida evoluzione, ad esempio, i prezzi delle azioni

È quindi difficile capire l'evoluzione della semantica, è quasi impossibile catturare dati che cambiano rapidamente (nonché pensare a dei data warehouse a causa dei costi).
Un altro problema è la varietà dei sorgenti e la loro veridicità (da LDL+13 si ha che scarsa qualità dei dati dei sorgenti del deep web). Tutti questi problemi sono delle challenge moderne da risolvere.
Vediamo quindi alcune tecniche per lo schema alignment.
Mediamente si hanno 3 step, per abilitare il linkage e ottenere risultati semanticamente significativi (per esempi guardare slide):

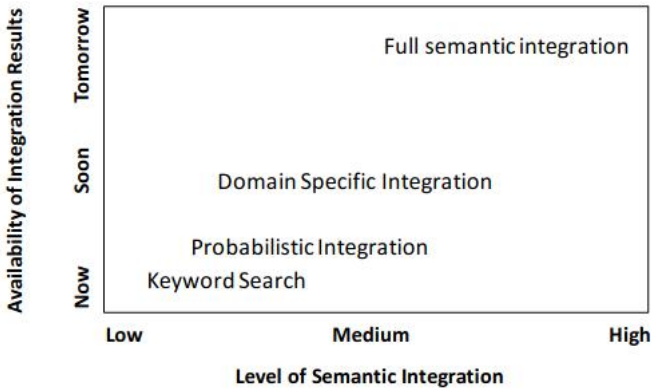


Figura 8.1: Schema dello spazio delle strategie di schema alignment

- **mediated schema**, dove si specifica la modellazione delle informazioni con uno schema che mi permetta di unificare e di allineare tutti i sorgenti, ottenendo una sorta di unico schema finale "mediatore". Si crea una vista unica con tutte le sorgenti e si catturano le informazioni principali del dominio studiato. Solitamente questa fase è manuale e lo schema mediato contiene più informazioni degli schemi dei singoli sorgenti ma potrebbero essrci casi in cui si hanno informazioni mancanti
- **attribute matching**, dove si cercano corrispondenze tra gli attributi dello schema mediato e quelli dei singoli sorgenti. Si "connettono" gli attributi dei sorgenti con quelli del mediatori, di solito si ha una corrispondenza "1:1". Un attributo del mediatore può essere una combinazioni di attributi del sorgente (o viceversa)
- **schema mapping**, dove si costruisce il mapping tra ogni sorgente e il mediatore. Tale mapping specifica la relazione di semantica tra il contenuto dei sorgenti e quello del mediatore e può essere usato per riformulare una query indirizzata al mediatore (che verrà poi suddivisa in ogni sorgente). Ci sono tre tipi di mapping, già visti (esempi su slide):
  - GAV
  - LAV
  - GLAV

Si hanno diverse tecniche che considerano volume e varietà:

- integrazione di interfacce di query di deep web in GMG+04 e CHZ05
- scansione, indicizzazione di dati deep web in MKK+08
- estrazione di dati strutturati da tabelle web in CHW+08, LSC10, PS12 e DFG+12 ed elenchi web in GS09 e EMH09. Lo studio delle tabelle è difficoltoso per il loro numero (154 milioni di tabelle relazionali, con 5.5 milioni di attributi e 2.6 milioni di schemi). Devo usare la ricerca per keyword. SI hanno problemi di ambiguità per le tabelle, si ha che non tutte le tabelle sono rilevanti per le query e spesso hanno anche specifiche funzionalità. Si usano delle featureRank per stimare la qualità delle tabelle basandosi su:
  - query independent features (numero di righe/colonne, intestazioni, numero di NULL etc...)
  - query dependent features
  - regressione lineare
  - feature pesate

  Si ottiene quindi uno score delle tabelle.
  Si ha anche lo schemaRank che specifica anche la coerenza dello schema. Data $p(x)$ la frazione dello schema univoco che contiene l'attributo x definisco il **point-wise mutual information** tra due attributi a e b come:

$$pmi(a,b) = \log_2 \left(\frac{p(a,b)}{p(a) \cdot p(b)} \right)$$

  e la coerenza è la media di tutti i $p(a,b)$ per ogni coppia di attributi nello schema.
  Per l'annotazione delle tabelle in LSC10 ci si chiede, data una tabella Web, quali entità si verificano in quali celle, quali sono i tipi di colonna e le relazioni tra le colonne. Il testo nelle celle descrive entità ma può essere ambiguo e le intestazioni non usano vocabolari controllati. Risolvere questi problemi mi permette di fare query sui metadati relazionali ed estrarre conoscenza dalle tabelle. Per annotare le tavelle uso cataloghi che consistono in una gerarchia di tipi, entità che sono istanze di tipi (possibilmente multipli) e relazioni binarie (rappresentati come alberi, esempi su slide). I cataloghi si usano per stimare la qualità dei contenuti delle celle della tabella, studiando la similarità. Annotare mi permette di

estrarre la semantica e di arricchire le tabelle. Si hanno diversi approcci per etichettare i varie elementi (righe, colonne, relazioni etc...).
Con DFG+12 si cercano invece tabelle simili in un "corpus" di altre tabelle. Cerco tabelle simili ad una tabella data. Le tabelle connesse sono ad esempio:

- tabelle candidate ad una *union* con l'aggiunta di nuove entità. Si usano tre segnali per stabilire se sono candidate alla *union*:
  - il rumore di entità e tipi
  - la cura di entità e tipi
  - il conto delle co-occorrenze di entità e tipi

  Alcuni approcci usano:
  - la similarità di Jaccard pesata sui set di etichette per la coerenza delle entità corrispondenza del peso massimo bipartito per la coerenza dello schema
- tabelle candidate per una *join* con l'aggiunta di nuovi attributi. Le tecniche sono analoghe alla *union*

Si cercano le migliori tabelle tramite un "punteggio di relazione" con quella in analisi. La metodologia è a tre step:

1. si verifica la consistenza delle entità tra la tabella data e quella di cui si sta studiando la relazione
2. si verifica se si possono espandere le entità della tabella data con l'altra
3. si verifica che i due schemi delle due tabelle siano consistenti, avendo schemi simili

Un approccio naive calcola il punteggio per ogni coppia di tabelle. Si usano quindi filtri per ottenere meno comparazioni (usando i filtri come criteri di blocco per classificare le tabelle ed esegui solo confronti di correlazione anll'interno dei sottoinsiemi creati dai blocchi) e più veloci (applicando in sequenza vari filtri, basandosi ulla selettività e l'efficienza computazionale dei filtri). Come filtri utili si hanno che le tabelle condividano colonne o entità.

- sistemi di dataspace in FHM05, HFM06 e DHY07. È un approccio quasi impossibile per volume, varietà e velocità, a causa di un approccio pay-as-you go (inizia con poco e poi "esplode")
- integrazione dei dati basata sulla ricerca per parole chiave in TJM+08 e con dati dinamici in TIP10

Una big data integration automatica è possibile in DDH08 ma si ha necessità di modellare l'incertezza sulla semantica degli attributi nei sorgenti. Si crea automaticamente uno schema mediato da un insieme di origini tramite schemi probabilistici (tramite mapping da sorgenti a mediatori pesati) che offrono il beneficio di una modellazione dell'incertezza. Si ha una sorta di clustering degli attributi, anche se si ha incertezza sulla sua accuratezza.

8.4.2 Record linkage

Passiamo quindi alle seconda fase, il **record linkage**, che si basa sul contenuto.
Si hanno tre step principali per studiare scalabilità, semantica e similarità:

1. **blocking**, usata per rendere gli step successivi più efficiente. Si creano blocchi per creare sottoinsiemi più specifici per gli step successivi. Si ragiona quindi in termini di scalabilità producendo blocchi più piccoli con record simili tra loro. Si usano funzioni di blocking su valori di uno o più attributi. La creazione dei blocchi deve essere il più efficiente possibile
2. **pairwise matching**, ogni record viene comparato con un altro per stabilire le coppie che matchano. Si computa la similarità di record nello stesso blocco. Si hanno decisioni locali all'interno dei blocchi che possono non essere consistenti a livello globale
3. **clustering**, dove l'obiettivo è assicurare la consistenza globale. Si studia come partizionare l'insieme di tutti i record in modo che ogni partizione si riferisca in modo distinto ad una entità, garantendo la semantica nelle partizioni.

Anche in questo caso si hanno tecniche legate al volume, dovendo gestire milioni di record:

- map-reduce, una tecnica per rendere più efficacie la parallelizzazione in presenza di molti dati, in VCL10 e KTR12

- blocchi adattivi in DNS+12, MKB12 e VN12
- blocchi in data space eterogenei, in PIO+12 e PKP+13

Per la velocità si ha la tecnica del record linkage incrementale in WGM10 e WGM13. Il numero dei sorgenti esplode velocemente e rende i confronti estremamente costoso. Le tecniche tradizionali di record linkage non sono infatti applicabili per i big data ma appunto conviene una strategia incrementale, aggiornando i link esistenti quando si hanno aggiornamenti dei dati.
In merito alla prioritizzazione si ragiona su un ordine di priorità degli schemi. Si cerca di matchare dati strutturati e testi/dati non strutturati, in KGA+11 e KTT+12, per quanto riguarda la varietà (esempio su slide). Si cerca inoltre di matchare tabelle e cataloghi, con LSC10.
Le sfide quindi nel record linkage sono quelle di matchare specifiche strutturate con altre non strutturate (e non strutturate con altre non strutturate). Si procede parsando il testo, semanticamente, offline o online. L'operazione ha tre step (**non si capisce nemmeno quali siano, sta delirando, aiuto!**):

1. **string tagging** per identificare tramite tag parse plausibili per ottenere il parser ottimo dei un certo testo. Si crea un indice invertito sui record strutturati tale per cui l'indice invertito ritorna un insieme di nomi di attributi associati alla stringa nei record strutturati. È importante che ci sia un parse plausibile in modo che ogni attributo abbia un parse distinto. Si studiano quindi le ambiguità
2. si usa una funzione di matching che stabilisce uno score tra il record strutturato e il testo. Si usa un vettore di feature similarity per la similarità tra gli attributi
3. Bho

Matchare dati strutturati non è così banale, a causa di limiti nei dati, all'eterogeneità etc... Spesso si usano gli identificatori se disponibili.
In merito alla veridicità, con LDM+11, si provvede ad un link temporale dei record, studiando quando sono veri.
Per esercitazione 5 guardare le slide e pregare (se avete qualcuno da pregare).

8.5 Big data & data fusion

Parliamo ora di **data fusion** nel caso di big data.
L'idea è sempre quella di riconciliare il contenuto che è in conflitto e si hanno tre passaggi importanti per risolvere le inconsistenze tra diversi sorgenti:

1. **voting**, avendo che se un valore è presente in più sorgenti allora l'algoritmo lo deve preferire

2. **source quality**, avendo che ogni sorgente potrebbe aver associato un punteggio di qualità che lo porta ad essere preferito o meno rispetto ad altri, dando un "peso" ai sorgenti

3. **copy detenction**, avendo l'identificazione di sorgenti che sono copiate da altre sorgenti (cosa molto comune), fattore che potrebbe interferire con la qualità dei passaggi sopra citati. Si riduce quindi il "peso" si sorgenti che hanno copiato

Si hanno tecniche per la **veracity** tra cui:

- usando l'affidabilità del sorgente, in YHY08, GAM+10, PR11, YT11, GSH11 e PR13
- combinando accuratezza dei sorgenti e *copy detection*, in DBS09a, QAH+13
- usando multipli valori di verità, in ZRG+12
- usando dati numerici errati (???), in ZH12
- usando comparazioni sperimentali sui dati del web, in LDL+13

Per le altre "V" si hanno altre tecniche:

- per **volume** la data fusion online, in LDO+11
- per **velcoity** la scoperta di valori veri per dati dinamici (dove i dati evolvono molto velocemente), in DBS09b, PRM+12
- per **variery** la combinazione di record linkage e data fusion, in GDS+10

La soluzione base è detta per il *voting* è detta **naive voting** e supporta differenze sulle opinioni e cerca di risolvere i conflitti. Funziona bene per sorgenti indipendenti con accuratezza simile (se sono con diversa accuratezza bisogna dare pesi maggiori ai voti sulle sorgenti più "famose"). In caso di sorgenti che hanno copiato si procede con la riduzione del "peso" del voto di tali sorgenti.
È importante misurare l'accuratezza delle sorgenti (in YHY08 e DBS09a) e per farlo bisogna calcolare una media di tutti i valori di un certo sorgente

S, ovvero quanti valori sono veri date tutte le osservazioni. Si ha quindi il calcolo dell'accuratezza di S, ovvero $A(S)$:

$$A(S) = \underset{v_i(D) \in S}{\text{avg}} \, Pr(v_i(D)true \mid \Phi)$$

Avendo:

- $v_i(D) \in S$ per dire che S ha il valore v_i per l'elemento D
- Φ osservazioni di tutti gli elementi per il sorgente S
- $Pr(v_i(D)true \mid \Phi)$ probabilità di $v_i(D)$ di essere vero (è una probabilità condizionata)

In input si hanno quindi i vari elementi D, con ognuno associato i vari valori per gli attributi:

$$val(D) = \{v_0, v_1, \ldots, v_n\}$$

(quindi $val(D)$ è in input). In input si hanno anche le osservazioni Φ.
In output si ha $Pr(v_i(D)true \mid \Phi)$, $\forall\, i = 0, \ldots, n$, sapendo che la somma di tali valori è pari a 1. Ci si basa quindi sulla formula di Bayes e quindi bisogna calcolare:

$$Pr(\Phi \mid v_i(D)true)$$

questo:

- sotto le condizioni di indipendenza, dovendo avere che $Pr(\Phi_D(S \mid v_i(D)true)$
- se S fornisce v_i allora: $Pr(\Phi_D(S \mid v_i(D)true) = A(S)$
- se S non fornisce v_i allora: $Pr(\Phi_D(S \mid v_i(D)true) = \frac{1-A(S)}{n}$

Si calcola l'accuratezza fino a che non converge, avendo in ciclo i seguenti conti:

- calcolo dell'accuratezza del sorgente:
  $A(S) = \text{avg}_{v(D) \in S} \, Pr(v(D) \mid \Phi)$
- calcolo del conteggio dei voti del sorgente: $A'(S) = \ln \frac{nA(S)}{1-A(S)}$
- calcolo del conteggio dei valori del voto:
  $C(v(D)) = \sum_{s \in S(v(D))} A'(S)$.
  Si usa un conto di similarità per il conto dei voti, avendo
  $C^*(v) = C(v) + \rho \sum_{v' \neq v} C\,(v') \cdot \text{sim}\,(v, v')$

- calcolo della probabilità: $\Pr(v(D) \mid \Phi) = \frac{e^{C(v(D))}}{\sum_{v_0 \in \text{val}(D)} e^{C(v_0(D))}}$

Ci si pone la domanda di voler integrare **tutti** i dati. Alcuni di essi possono essere ridondanti o di bassa qualità. Si è scoperto che i dati ridondanti non portano guadagni (all'aumentare il numero di sorgenti raggiungo un asintoto orizzontale quasi subito, avendo sulla y i risultati voluti e sulla x il numero di sorgenti, **esempio su slide** dove con 35 sorgenti si hanno praticamente gli stessi risultati di 894 sorgenti). Integrare tutti i sorgenti quindi non sempre porta guadagni.
Bisogna misurare l'accuratezza dei sorgenti, avendo che dati errati rovinano la qualità. Per misurare l'accuratezza spesso si usa una raccolta selezionata e verificata manualmente di sorgenti, definendo un *gold standard*. Si ha l'accuratezza delle sorgenti in input (come percentuali dei valori correttamente forniti rispetto al *gold standard*) e si assegna voto maggiore in base all'accuratezza (**esempio su slide, qui non si è capito una mazza**).
Bisogna quindi studiare come scegliere i sorgenti prima di fare integrazione per trovare il giusto bilanciamento tra guadagno e costo, massimizzando la qualità sotto un certo budget (**grafico su slide**) o di minimizzare il costo entro un certo range di qualità (**grafico su slide**). Si ha la cosiddetta **teoria del marginalismo**, presa dalla *teoria economica*, ovvero andare a integrare ma sapendo che l'integrazione ha un costo e quindi il costo deve essere essere bilanciato con il guadagno. Si studiano il margine di guadagno e il margine di costo per ottenere il "massimo profitto" (*inutile dire che si continua a capire poco di questa lezione*).
Il problema è quindi, spesso quello di comprendere il guadagno e il costo del processo, di integrazione dovendo affrontare varie difficoltà:

- ambiguità e record linkage
- decidere quali dati integrare con il data fusion
- i costi del processo di integrazione

Su slide (*che allo stato attuale non sono online, in caso auguri con la rec*) descrizione di un progetto che aveva come fine l'approfondimento delle tematiche legate all"integrazione e l'uso di algoritmi per ottimizzare il processo di selezione, dimostrando quanto integrare molte sorgenti senza uno studio precedente non porti risultati interessanti. Si segnala che ci sono le rec anche su altri due progetti: *Temporal scoping of fact* e *MantisTable*.

Capitolo 9

Data management for machine learning

Vediamo un argomento che unisce architetture dati e machine learning.
Normalmente in machine learning si parte da dati di training già puliti, si fa il training e si ottiene il modello che a sua volta, con una serie di dati reali, produce dei risultati. In produzione tutto questo non è sufficiente e non si ha tanta importanza nella fase di *train* e *serve* ma nei dati in se. Bisogna prepare bene i dati per il ML, avendo il rischio di **garbage in, garbage out** in quanto i computer elaborano in modo acritico anche un insieme di dati in entrata palesemente insensati (garbage in) producendo, a loro volta, un risultato insensato (garbage out). La fase di *prepare* dei dati è quindi fondamentale, mediamente:

- 80% del tempo è *data preparation*
- 5% del tempo è *identificazione del modello*
- 5% del tempo è *training*
- 5% del tempo è *deployment* e altri lavori minori

Bisogna poi, per i risultati, fare *data validation*, *data monitoring* ed eventuali correzioni.
I dati di input vengono mediamente divisi in training (grandi) e serving (piccoli) come già sappiamo.
Nella fase di *prepare* dei dati di training bisogna valutare quali sono le feature rilevanti, fare data exploration dei valori delle feature, quali sono le migliori pratiche per estrarre dai valori delle feature i migliori valori per il modello di machine learning etc... Si parla di **feature selection** e **feature engineering**. Bisogna poi valutare il modello, capendo se è abbastanza buono, se

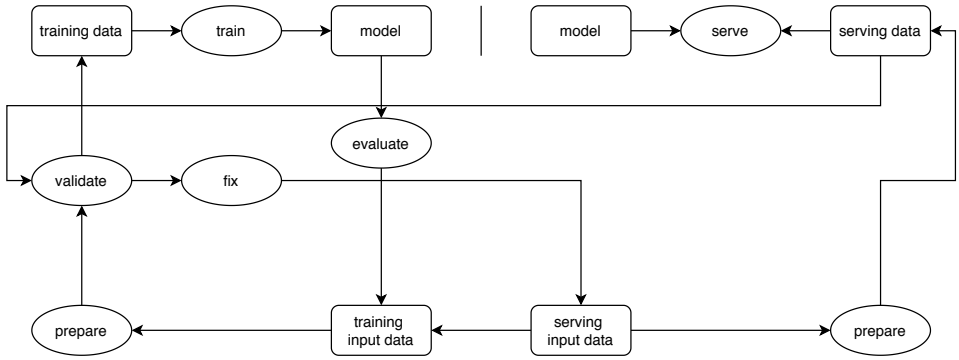


Figura 9.1: Esempio di pipeline completa di machine learning in produzione.

deve essere modificato o se servono più dati o più feature.
In produzione però bisogna andare oltre, basti pensare che magari una fonte dei dati subisce un refactoring e tutto il modello si "rompe", avendo **data failure**, con performance che crollano o peggio, se si parla di reinforcment learning, si ha il modello che viene allenato con dati errati (e riaddestrare è costoso). Si vede quindi che il problema non è il modello in se ma i dati.
Bisogna quindi validare i dati in input al modello prima di darli in pasto al modello per capire se è tutto corretto. Bisogna tenere conto anche delle "deviazioni" dei dati che dipendono dalle situazioni a contorno dell'environment (ad esempio un modello per il traffico addestrato l'anno scorso non vale ora causa pandemia). I problemi vanno quindi fixati, per poter poi ricreare i dati di testing serving e riprendere l'intero ciclo di vita.
Dal punto di vista del personale si hanno almeno tre figure:

- **l'esperto di ML**, che studia il modello
- **il software/web engineering (*SWE*)**, che effettivamente sviluppa il modello
- **il site reliability engineering (*SRE*)**, che si occupa appunto di fixare i vari problemi (sia dei dati che di interazione con il cliente)

In figura 9.1 può essere studiata la **data understanding** nell'ambito del machine learning, avendo un *sanity check* prima di trainare il primo modello e avendo altre analisi a ciclo dopo. Tra i check classici nel sanity check abbiamo:

- il controllo di valore massimo, minimo e più frequente di ogni feature, studiando l'accuratezza sintattica come dimensione di qualità

- l'istogramma dei valori continui e categorici per vedere la distribuzione dei valori
- se una feature è presente in un numero sufficiente di esempi per lo studio in analisi, studiando quindi la completezza come dimensione di qualità
- se una feature ha il giusto numero di valori, dando un vincolo di consistenza ai dati

Non sempre i dataset reali sono "puliti".
Se sapessimo a priori cosa cercare basterebbero una seria di query per la *data exploration*. La realtà però è che spesso non si sa come è fatto un dataset e quindi si sfrutta la *data visualization*, ovvero strumenti di visualizzazione per poter guardare graficamente i vari elementi per procedere poi al sanity check. Mediamente per la visualizzazione si usano metriche basate sulla deviazione e, in ogni caso, si hanno informazioni da visualizzare più interessanti/utili di altre.
Questo è all'inizio del processo mentre durante la fase di messa in produzione e delle varie iterazioni si hanno altri aspetti per la data understanding. In ordine di importanza:

- **analisi feature-based**, studiando la sensitività di una feature rispetto ai dati e studiando le conseguenze di eventuali variazioni degli stessi. Si usano approcci visuali o alla datawarehouse (???)
- **analisi del ciclo di vita dei dati**, studiando dipendenze delle feature nascoste (fattore necessario in alcuni modelli, come quello bayesano che assume indipendenza delle feature). Può essere successo che un dato sia stato spezzato in due dati tra loro dipendenti
- **altre questioni aperte**, tra cui la fairness del modello di machine learning, ovvero studiando il "pregiudizio" di alcuni modelli rispetto a certe classi (magari perché addestrati solo su alcune classi o avendo pochi dati per alcune classi, ad esempio predire il nuovo presidente degli USA e ottenere solo uomini, non avendo donne nello storico dei dati), avendo quindi modelli biased perché i dati sono biased (il modello è corretto ma i dati no). Questo orimo aspetto è molto importante nel mondo del lavoro. Un'altra questione è quello di cercare di "fregare" i modelli di ML tramite dati di tipo *adversarial*. Si parla in questo caso di **reti GAN** e **deep fake**, ovvero sistemi in cui si usa un secondo sistema con

cui si cerca di ingannare il primo classificatore con dati simili ma "sintetici". In pratica si cerca di imparare dal passato per ricreare qualcosa di simile. Questa seconda questione ha più valenza nel mondo della ricerca

Per la fase di **data validation** si studiano eventuali cambiamenti nei dati in input al modello (anche solo un uppercase che diventa lowercase con le feature in uppercase che diventano *rare*), fattore che può rovinare i risultati (è una situazione tipica in produzione).
Bisogna capire come fare le correzioni e si deve riconoscere quando accadono queste cose. Si hanno delle *best practice*:

- **alert**, facendo un *continuous data cleaning* posso avere alert per errori tecnici ma anche per cambi di dati (magari per variazioni dei tempi a cui si riferisocno i dati ad esempio i dati presi in pandemia sono diversi da quelli di due anni fa)
- **playbook**, che contengono le classi di problemi e vanno consultati manualmente (per ora)

Si hanno errori che hanno anche impatto basso sulle prestazioni. Bisogna considerare anche che eventuali correzioni possono avere conseguenze su altre fasi della pipeline.
Passiamo alla **data preparation**. Bisogna fare **feature engineering** per capire come migliorare le performance, aggiungendo/rimuovendo dati/feature/attributi (e studiandone la qualità prima dell'aggiunta). Si hanno tecniche per passare da un dato grezzo in una rappresentazione (magari booleana, tramite *One Hot Encoding*) più adatta ai modelli di ML (queste tecniche si fanno a Data Analytics). Tra le altre tecniche standard abbiamo:

- Normalization
- Bucketization
- Winsorizing
- Feature crosses
- usare un modello pre-addestrato o l'embedding per estrarre feature

Anche i risultati di reti neurali possono essere usati per alimentare i dati, tramite *rapresentational learning* (???).
La feature engineering richiede molta conoscenza di dominio e avere buoni

dati richiede anche una cardinalità minore degli stessi per ottenere ottime prestazioni. Conviene quindi migliorare la qualità dei dati più che aumentarli a dismisura, nella maggioranza dei casi di classificazione. In altri campi, come i sistemi generativi di linguaggi (naturali), quindi questo trade-off sembra essere non così ben definito in ogni caso. Si hanno tecniche di data management per migliorare la qualità delle label etc...
Aggiungere feature in produzione è comunque complesso, anche dal punto di vista computazionale, oltre al fatto che aggiungere dati è costoso.