

Dove vanno i tuoi dati?

Il database mette in relazione l'utente e i dati di cui dispone, con il vantaggio di accedere a una loro rappresentazione logica. Applicazioni e memorizzazione fisica risultano così indipendenti, garantendo un'elevata flessibilità

MARCO SERGIO



I database devono garantire le performance e l'affidabilità dei sistemi (Fonte: Wonderware)

I database sono chiamati a supportare tutte le attività produttive ed è, quindi, indispensabile che tutte le componenti coinvolte funzionino al meglio e si integrino perfettamente, garantendo agli utenti, performance e affidabilità dei sistemi. Con il termine database si intende, per definizione, una collezione di dati gestita e organizzata da un software specifico: il Dbms (Database Management System, sistema di gestione di database). Si tratta, in pratica, di uno strato software che mette in relazione l'utente e i dati veri e propri. Grazie a questa caratteristica, l'utente o le applicazioni non accedono direttamente ai dati nella forma in cui sono memo-

rizzati nel sistema hardware, ma a una loro rappresentazione logica. Applicazioni e memorizzazione fisica risultano, così, indipendenti, garantendo un'elevata flessibilità e la possibilità di gestire i dati nel modo più opportuno. La maggior parte dei database attualmente in commercio è di tipo relazionale. Si tratta, cioè, di strumenti basati sul modello relazionale teorizzato da Codd nel 1970. In pratica, come indica il nome stesso, fanno riferimento al concetto di relazione, ossia a un insieme di record omogenei defini-

ti sugli stessi campi. A ogni dominio all'interno della relazione viene associato un nome, detto attributo, che ne descrive il ruolo. Inoltre, questa metodologia fornisce sistemi semplici ed efficienti per rappresentare e manipolare i dati, ed è basata su un modello, quello relazionale appunto, con solide basi teoriche.

Tutta colpa del sistema?

Indipendentemente dal tipo di database, a questi strumenti è richiesto, in primo luogo, di consentire l'accesso ai dati attraverso uno schema concettuale, invece che attraverso uno schema fisico. Questi dati, in seguito, devono

essere condivisi e integrati con quelli provenienti da applicazioni differenti. Per questa ragione, un database deve essere in grado di controllare l'accesso concorrente ai dati, ma soprattutto di garantirne la sicurezza, l'integri-



Con il termine database si intende, per definizione, una collezione di dati gestita e organizzata da un software specifico (Fonte: Supersentai)

tà e l'affidabilità. Proprio per assicurare la necessaria affidabilità, soprattutto quando i database sono chiamati a supportare attività produttive, è necessario che tutte le componenti coinvolte funzionino al meglio e si integrino perfettamente tra loro. Ovviamente, il primo elemento da prendere in considerazione è rappresentato dai componenti hardware, la cui affidabilità, nel corso degli anni, è cresciuta progressivamente anche con elementi di tipo standard. Tutti i costruttori, infatti, si sono conformati a una serie di linee guida che garantiscono il livello di affidabilità di questi componenti. Un discorso diverso, invece, è legato all'ambito software. I notevoli progressi, impensabili sino a pochi anni fa, hanno, infatti, permesso di raggiungere valori nell'ordine del 99,99%. Inoltre, con un investimento leggermente superiore, è possibile arrivare anche ai famosi 99,999%, che garantiscono la possibilità di operare con un downtime valutabile in 5-6 minuti all'anno. Il vero punto debole, in molti casi, sembra, invece, rappresentato dal software applicativo, che può avere un impatto sulle attività che gli operatori svolgono sui dati. Una situazione che, secondo alcuni esperti, è dovuta anche alla scelta di alcune aziende di ridurre gli investimenti nella formazione del proprio personale. Inoltre, in molti casi, manca una strategia comune e non vengono

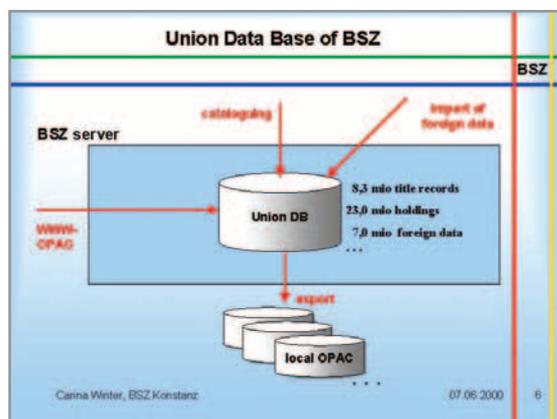
stabilite delle linee guida univoche, lasciando, così, le decisioni all'esperienza e alla sensibilità dei singoli.

Questione di disponibilità

Al problema dell'affidabilità, che spesso potrebbe essere risolto in modo relativamente semplice e a basso costo, si aggiunge quello ancor più complesso della disponibilità. In questo caso, infatti, le barriere sono diverse e, solo in alcuni casi, risolvibili con tecnologie legate al mondo dei database server. I nuovi strumenti, messi a disposizione dai principali vendor, permettono infatti di superare i problemi di guasti fisici, ma anche gli errori applicativi o quelli compiuti dall'utente durante le operazioni di manutenzione. Accanto alle problematiche di tipo 'umano', la disponibilità dei dati può diventare un problema quando c'è concorrenza nell'accesso.

Per esempio, può accadere che un database, per quanto molto robusto e affidabile, venga bloccato dall'accesso di un utente che, in modo del tutto involontario, impedisce a tutti gli altri di utilizzare questa fondamentale risorsa. Un simile problema è molto grave, perché può vanificare tutti gli sforzi e gli investimenti fatti per garantire la necessaria disponibilità.

Il tutto, senza dimenticare che alcuni prodotti creano delle barriere all'accesso, soprattutto nel corso di operazioni di manutenzione o di aggiornamento, impedendo di fatto l'utilizzo dei dati memorizzati. Nella scelta di un database, inoltre, occorre valutare le capacità di scalabilità,



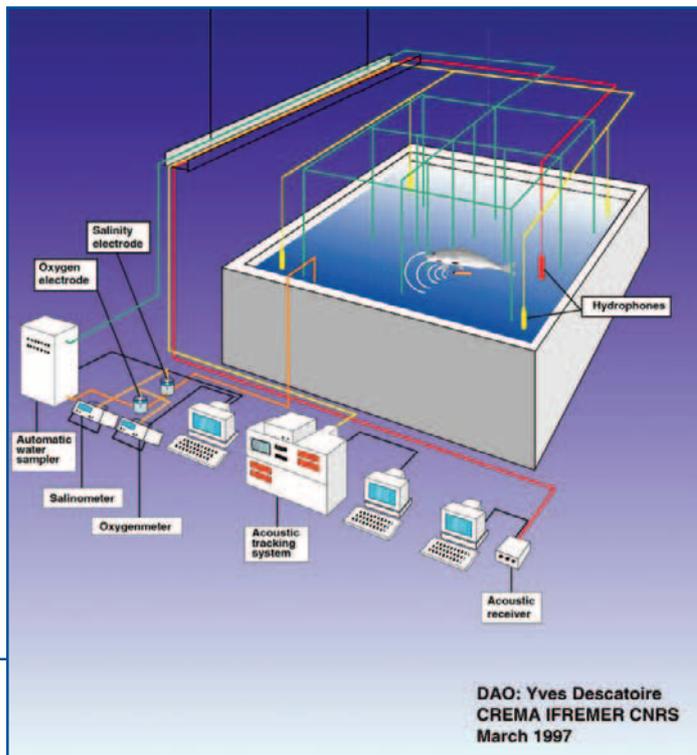
I database di tipo relazionale fanno riferimento al concetto di relazione, ossia a un insieme di record omogenei definiti sugli stessi campi (Fonte: DBZ)

ovvero la capacità di funzionare anche in presenza di un carico di lavoro superiore a quello previsto, con la necessità di aggiungere ulteriori elementi.

E in caso di errore?

Anche se, come abbiamo sottolineato, l'affidabilità dei componenti hardware è particolarmente elevata, il rischio di un danneggiamento fisico, dovuto anche a cause esterne, non può essere considerato nullo. E il problema diventa ancora più grave se pensiamo che, a fronte di un simi-

le guasto, l'indisponibilità dei dati potrebbe non essere solamente momentanea ma definitiva, con gravi danni per un'azienda. Per prevenire un evento così disastroso, le aziende hanno la facoltà di utilizzare due tecnologie: il clustering o il database mirroring. La scelta, fra le due opportunità, è influenzata sia dal costo, sia dall'obiettivo che un'azienda si pone quando vuole raggiungere un elevato livello di disponibilità. In particolare, il clustering si rivela utile per garantire la disponibilità dei dati in un ambito limitato, come un edificio. Se un nodo dovesse cadere, infatti, gli altri potrebbero assumerne le funzionalità, consentendo, in tal modo, il proseguimento del lavoro. Il database mirroring, al contrario, delega a ogni database il compito di completare il proprio storage, con la possibilità di creare link più o meno forti, e di utilizzare infrastrutture hardware completamente diverse fra loro, sfruttando,



Classificazione dei database

I Dbms vengono generalmente classificati in base al tipo di rappresentazione logica dei dati mostrati. Fra i numerosi modelli esistenti, i più comuni sono:

- database gerarchici: i dati vengono organizzati in insiemi legati fra loro da relazioni di 'possesso'. Questo significa che un insieme di dati può contenere altri insiemi di dati. Si viene, così, a creare una struttura ad albero.
- database reticolari: il modello reticolare è simile a quello gerarchico, in quanto gli insiemi di dati sono legati da relazioni di possesso. La differenza, però, è connessa alla possibilità di appartenere, contemporaneamente, anche a più insiemi, creando, in tal modo, una struttura costituita da un insieme di dati.
- database relazionali: il modello relazionale è basato su una tabella bidimensionale, detta tupla. Per ogni entità rappresentata viene successivamente definita una serie di colonne delle relazioni, dette attributi. Le entità in possesso di caratteristiche comuni fanno parte della stessa relazione.
- database a oggetti (object-oriented): uno schema a oggetti è costituito da un insieme di classi, in grado di definire le caratteristiche e il comportamento degli oggetti presenti nel database. A differenza degli altri modelli, i dati, qui, non sono passivi, ma gli oggetti memorizzati nel database contengono sia i dati sia le operazioni. Per questo, i dati sono in possesso di una loro intelligenza, grazie alla quale possono operare anche senza doversi appoggiare ad applicazioni esterne.

Si viene a creare un accesso ai dati attraverso uno schema concettuale, invece che attraverso uno schema fisico (Fonte: EfreM)

così, al meglio il patrimonio informatico di un'azienda. A questo punto, risulta chiaro come, per molti versi, le due soluzioni possano essere considerate complementari.

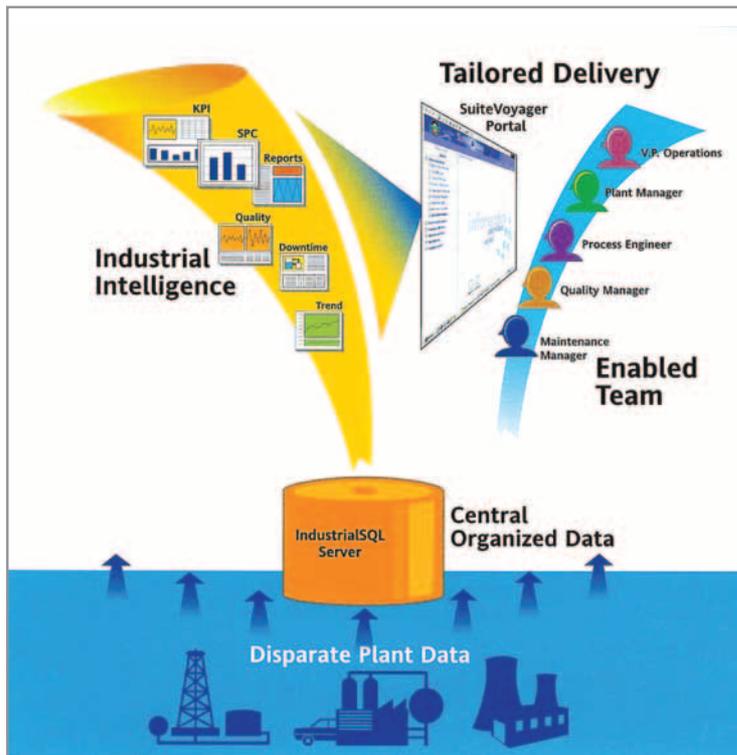
Una replica da... non perdere

La replica dei dati su macchine differenti è fondamentale per consentire di creare soluzioni di standby di un database fra un certo numero di server, anche se questi, in realtà, non sono nati per garantire un'alta disponibilità. La possibilità di operare con soluzioni in standby può essere basata su operazioni automatiche o manuali, anche se in quest'ultimo caso è opportuno che lo switch degli utenti da un server all'altro sia supportato da un'ulteriore serie di funzionalità. Una scelta che, tra l'altro, permette di sgravare il server di produzione da alcuni compiti. Operando in tal modo, però, è necessario prestare una crescente attenzione alla sincronizzazione dei dati. Per questo motivo è opportuno puntare su soluzioni che garantiscano un tempo di latenza molto basso, consentendo così ai server periferici di ricevere rapidamente tutte le informazioni disponibili. In questo ambito negli anni scorsi si è diffusa la tecnologia 'Log shipping', che consente di distribuire il backup dal server di produzione a un certo numero di server in standby, mantenendoli costantemente allineati. Per quanto una simile tecnologia sia molto diffusa, è necessario sottolineare che, stando a quanto hanno

annunciato alcuni produttori, non sono previste innovazioni in questo ambito. La situazione è dettata anche dal fatto che il mirroring, allo stato attuale, offre prestazioni superiori grazie alla sua capacità di mettere a disposizione una serie di funzionalità automatiche in grado di prevenire anche eventuali errori degli utenti. La scelta di utilizzare soluzioni ormai affermate, come il back-up o il restore, è limitata dalla caratteristica di garantire solo la disponibilità dei dati a freddo, senza nessun meccanismo automatico che permetta di distribuire i dati in caso di necessità. Simili tecnologie si rivelano, quindi, utili per le attività di testing e di sviluppo applicativo. Mentre bisogna ricordare che non esiste una logica che garantisca il passaggio automatico dal server di produzione a uno di backup, ma è necessario effettuare lo switch manuale. Decisamente superata anche la logica basata su detach/copy/attach, che impone di scollegare temporaneamente gli utenti. Occorre ricordare che sul mercato esiste un'ulteriore serie di soluzioni ad alta disponibilità basate su tecnologie di replica e failover, con proposte che assicurano un mix di alta disponibilità e scalabilità e prevedono di replicare i dati da un server di produzione e un certo numero di server di sola lettura.

Witness, il testimone

Il database mirroring permette di implementare soluzioni che offrono la possibilità di consultare una sorta di virtual database, per consentire un accesso trasparente ai dati. I

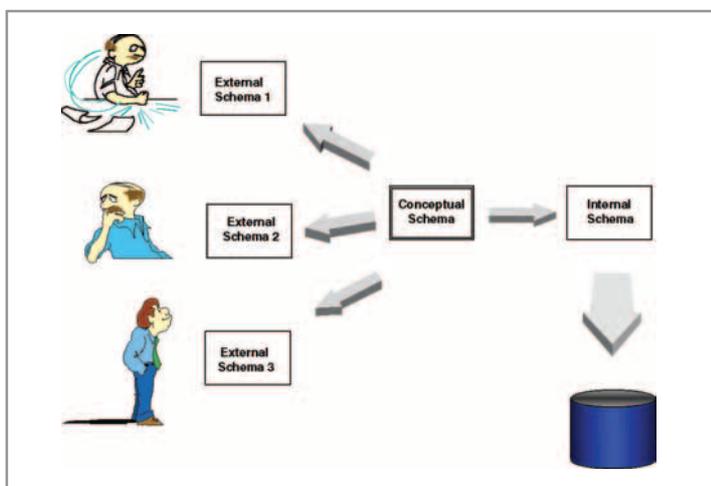


I progressi tecnologici permettono di garantire un'affidabilità pari al 99,999% (Fonte: Wonderware)

due server, in alternativa, svolgono il ruolo di server di applicazione e di mirror con quest'ultimo che, in caso di necessità, assume immediatamente il compito di distribuire le informazioni senza tempi morti. In alcune applicazioni è possibile aggiungere un witness (un terzo server) che garantisca il failover automatico fra i due server, in caso di necessità. Il tutto, in modo completamente trasparente per il cliente, senza imporgli di collegarsi a un nuovo server. Un ulteriore vantaggio di questa tecnologia è legato alla capacità del witness di tenere sotto controllo il mirroring di più server contemporaneamente, indipendentemente dalla loro posizione geografica, garantendo sempre un failover molto breve.

Sicurezza o performance?

Il mirroring fra due server garantisce, in genere, un buon livello di sicurezza per ciò che riguarda la disponibilità dei dati, ma quali sono i suoi impatti sulle performance del sistema? Nel caso in cui il database mirroring sia impostato su full, il commit delle informazioni avviene solo se il dato è stato confermato su entrambe le macchine. In questo modo, il failover automatico può avvenire e si ha la sicurezza che tutto è



Ogni utente accede ai dati in funzione delle proprie esigenze (Fonte: Essential Strategies)

andato a buon fine. Al contrario, se la funzionalità è impostata su off, il commit avviene anche se il dato risulta confermato solo sul server principale e non necessariamente su entrambi. Questo significa che, in alcuni casi, potrebbe essere necessario ripetere le operazioni di backup.

Scegliendo quest'ultima modalità è possibile migliorare le performance, in termini di velocità pura, ma non è garantito il buon esito di ogni operazione. In ogni caso, avendo predisposto un opportuno witness, è possibile segnalare immediatamente il suo mancato funzionamento e intervenire per porvi rapidamente rimedio. Il failover clustering e il database mirroring hanno in comune la rilevazione dei failover hardware e il passaggio in automatico da un server all'altro.

Entrambi, inoltre, possono supportare lo switch in automatico, con la sicurezza di non perdere nessuna informazione nel corso del passaggio da uno all'altro e la totale trasparenza per tutti gli utenti che vi stanno accedendo. E', però, importante sottolineare che il failover clustering avviene a livello di sistema e, pur risultando relativamente rapido, non è velocissimo e non permette di accedere a un'istanza. Il mirroring, invece, avviene, a livello di ogni singolo server, in modo pressoché istantaneo. Inoltre, è possibile svolgere liberamente una serie di attività sul singolo server di standby. In entrambi i casi, comunque, si gode di un'alta disponibilità, una peculiarità completata da soluzioni di disaster recovery anche in località distanti dal punto di vista geografico.

Pronti per la foto?

I database administrator, come qualunque persona, possono commettere errori. Per questa ragione, è necessario predisporre test con verifiche e simulazioni, per verificare l'effettiva capacità di effettuare un restore, soprattutto quando i database sono molto grandi. Un'attività che, in realtà, non sempre viene svolta in modo corretto e con la necessaria frequenza. Le nuove tecnologie di snapshot permettono, invece, di fare una sorta di 'fotografia' del database, con dati che vengono fotografati e salvati in modo perenne e non modificabile. Gli snapshot si rivelano utili anche perché, oltre a operare in modo istantaneo, non influiscono con il funzionamento della macchina. Inoltre, grazie a queste soluzioni, è possibile sia accedere ai dati di un certo istante, sia recuperarli, riportando la vita di un database a un tempo precedente.

Questa tecnologia è particolare in quanto non viene fatta ogni volta una copia di tutti i dati, ma si provvede a salvare solo i dati che sono stati modificati. Il numero di snapshot eseguiti dipende, quindi, dalle modifiche apportate. Perciò, alla creazione di un nuovo database, o quando vengono modificate alcune pagine, quest'ultime sono marcate per consentire agli utenti di realizzarne una copia. Combinando lo snapshot e il mirroring si può, quindi, replicare a caldo il contenuto di un database su un altro server, con la possibilità di permettere agli utenti di accedere alle informazioni contenute in determinati snapshot. ■