Corso di Laurea Magistrale in Economia

Data Science

A.A. 2018/2019

Lez. 9 – Big Data 2

Fondamenti architetturali

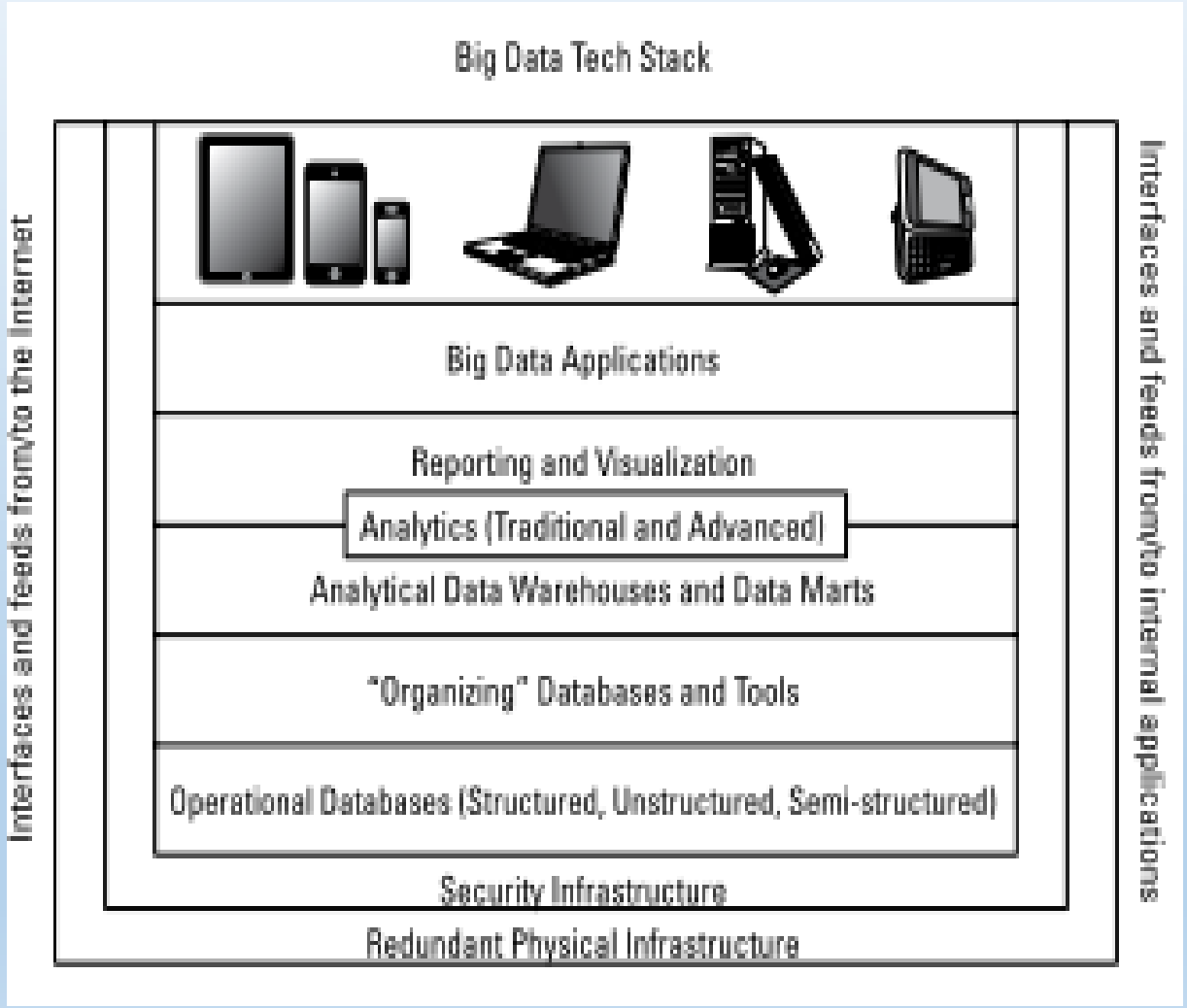
- Oltre ai requisiti funzionali, è importante supportare la performance richiesta
- Bisogni dipenderanno dalla natura delle analisi
- Necessità della giusta quantità di potenza computazionale e di velocità
- Mentre alcune delle analisi saranno effettuate in *real time*, è inevitabile dover memorizzare una certa quantità di dati
- L'architettura dovrà anche avere la giusta quantità di ridondanza in modo tale da essere protetti da possibili latenze e inattività impreviste

Fondamenti architetturali

- Quanti dati dovranno essere gestiti correntemente e in futuro dall'organizzazione?
- Quanto spesso l'organizzazione dovrà gestire dati in *real time* o quasi in *real time*?
- Quanto rischio può affrontare l'organizzazione? Il campo in cui opera è soggetto a vincoli stringenti di sicurezza, affidabilità e *governance*?
- Quanto è importante la velocità per le necessità di gestione dei dati?
- Quanto devono essere certi o precisi i dati?

Fondamenti architetturali

- Per comprendere i big data può essere di aiuto considerare i componenti dell'architettura. Un'architettura di gestione dei big data deve includere una varietà di servizi che consentono alle imprese di utilizzare una miriade di sorgenti di dati in modo veloce ed efficace
- Per avere consapevolezza di questo, è possibile porre i componenti in un diagramma che aiuti a vedere dove essi si trovano e che relazioni ci sono tra essi



Interfacce e alimentazione

- È importante notare che in ciascun lato del diagramma ci sono indicazioni sulle interfacce e sull'alimentazione sia per i dati gestiti internamente che per quelli provenienti da sorgenti esterne
- Per comprendere come operano i big data nel mondo reale, è importante capire questa necessità. Infatti, ciò che rende i big data «big» è il fatto che essi si basano su enormi quantità di dati provenienti da moltissime sorgenti
- Pertanto, delle API aperte rappresenteranno l'elemento chiave per ogni architettura di big data
- Le interfacce, comunque, esisteranno a ciascun livello e tra ciascuna coppia di strati dello stack
- Senza servizi di integrazione, i big data non possono funzionare

Architettura fisica ridondante

- L'infrastruttura fisica di supporto è fondamentale per l'operatività e la scalabilità di un'architettura di big data
- Per supportare un volume sconosciuto e non predicibile di dati, un'infrastruttura fisica per i big data deve essere diversa da quella tradizionale. Essa, infatti, deve essere basata su un modello di calcolo distribuito.
- Questo vuol dire che i dati possono essere memorizzati fisicamente in molte locazioni differenti e possono essere collegati tra loro attraverso le reti, un *file system* distribuito, vari *tool* di analisi e varie applicazioni
- La ridondanza è importante perché si sta operando con molti dati provenienti da sorgenti differenti

Ridondanza

- La ridondanza si può avere in molti modi:
  - Se una compagnia ha realizzato un *cloud* privato, è possibile avere la ridondanza gestita nell'ambiente privato che potrà, quindi, scalare per supportare carichi mutevoli
  - Se una compagnia vuole contenere la crescita delle infrastrutture interne, può usare dei servizi di *cloud* esterni per incrementare le sue risorse interne
  - In alcuni casi, questa ridondanza può essere nella forma di offerte di Software as a Service (SaaS) che consente alle compagnie di fare sofisticate analisi di dati sotto forma di un servizio. L'approccio SaaS offre dei costi più bassi, uno startup più veloce e una evoluzione della tecnologia sottostante senza soluzione di continuità

Infrastruttura di sicurezza

- Più è importante l'analisi dei big data per le compagnie e più importante sarà mettere al sicuro i dati
  - Una compagnia che opera nel contesto sanitario vorrà probabilmente usare applicazioni sui big data per determinare i cambiamenti demografici o i cambiamenti nei bisogni dei pazienti
- I dati su cui lavorare devono essere protetti sia per soddisfare le richieste di conformità sia per proteggere la privacy dei pazienti
- È necessario stabilire chi può vedere i dati e in quali circostanze lo può fare
- Potrà essere necessario verificare l'identità degli utenti nonché proteggere l'identità dei pazienti
- Queste tipologie di richieste di sicurezza devono essere parte della definizione dei big data sin dall'inizio e non in un secondo momento

Sorgenti di dati operazionali

- Necessario includere tutte le sorgenti di dati che daranno una visione completa dell'azienda e vedere come i dati impattano il modo con cui quest'ultima porta avanti il suo business
- Tradizionalmente, una sorgente di dati operazionale consisteva di dati altamente strutturati gestiti mediante un DBMS relazionale
- I dati operazionali attualmente possono comprendere un insieme più esteso di sorgenti di dati dove sono incluse le sorgenti non strutturate quali i dati sui clienti e i social media in tutte le loro forme
- Nel mondo dei big data si troveranno nuovi approcci emergenti alla gestione dei dati, comprese le architetture per la gestione dei documenti, dei grafi, dei database a colonne e di quelli geospaziali. Nel complesso, tutti questi dati sono conosciuti come DBMS NoSQL, o «not only SQL»

Sorgenti di dati operazionali

- Necessità di mappare le architetture dei dati alle tipologie di transazioni.
- Possono essere necessarie delle architetture dei dati che supportino dei complessi contenuti non strutturati
- Per sfruttare al meglio i big data è necessario includere sia database relazionali che database non relazionali, come pure sorgenti di dati non strutturati
- Tutte queste sorgenti di dati operazionali hanno alcune caratteristiche in comune:
  - Esse rappresentano sistemi di registrazione che tengono traccia dei dati critici richiesti per attività in *real time* e giornaliere
  - Esse sono aggiornati continuamente sulla base delle transazioni che avvengono all'interno delle unità di business e sul web
  - Affinché queste sorgenti possano fornire una rappresentazione accurata dell'azienda, esse devono mischiare dati strutturati e non strutturati
  - Questi sistemi devono anche essere scalabili per supportare migliaia di utenti mantenendo la consistenza. Essi possono includere sistemi transazionali di e-commerce, sistemi di gestione delle relazioni con i clienti o applicazioni per call center

Ruolo della performance

- L'architettura dei dati non può non tenere conto dell'infrastruttura di supporto di un'organizzazione
  - Per esempio, un'organizzazione potrebbe essere interessata ad eseguire modelli per determinare se è sicuro trivellare per cercare il petrolio in un'area offshore, una volta forniti dei dati in *real time* sulla temperatura, sulla salinità, i sedimenti e altre proprietà biologiche, chimiche e fisiche dell'acqua.
  - Potrebbero essere necessari giorni per eseguire questo modello utilizzando una configurazione di server tradizionali. Invece, utilizzando un modello di calcolo distribuito, ciò che prima richiedeva giorni può richiedere minuti.
- La performance può anche determinare il tipo di database da utilizzare
  - Per esempio, in alcune situazioni si può voler comprendere come due elementi di dati molto diversi sono correlati. Qual è la relazione tra il «brusio» in una social network e la crescita in alcune vendite?
  - Questa non è la tipica *query* che si può fare in un database strutturato relazionale. Un database a grafo potrebbe essere la scelta migliore, dal momento che esso è progettato specificatamente per separare i «nodi» dalle loro proprietà, nonché gli «archi» dalle corrispettive proprietà
- Altri importanti approcci di database operazionali includono i database a colonne, che memorizzano efficientemente le informazioni in colonne anziché in righe. Questo approccio porta a delle performance più veloci perché l'input/output è estremamente veloce
- Quando è necessaria la memorizzazione di dati geografici, è il caso di utilizzare un database spaziale, capace di memorizzare e gestire dati e interrogazioni sulla base di come gli oggetti sono correlati nello spazio

Organizzare i servizi di dati e i tool

- Non tutti i dati utilizzati dalle organizzazioni sono operazionali. Una quantità crescente di dati proviene da una varietà di sorgenti che non sono ben organizzati o immediati; si pensi ai dati che provengono da macchine o sensori nonché da sorgenti massive pubbliche e private
- Nel passato gran parte delle compagnie non erano capaci di catturare o memorizzare questa vasta quantità di dati: era semplicemente troppo costoso o troppo travolgente
- Anche se le compagnie erano capaci di catturare i dati, esse non avevano i tool per fare niente con essi. Pochissimi *tool* potevano dare un senso a questa enorme quantità di dati
- I *tool* esistenti erano complessi da usare e non producevano risultati in un lasso di tempo ragionevole
- Alla fine, coloro che veramente volevano gestire enormi quantità di dati erano costretti a lavorare con dei campioni degli stessi. Ciò aveva il rischio indesiderabile di perdere degli eventi importanti se questi non erano catturati nel campione

Tool

- Con l'evoluzione delle tecnologie di calcolo è ora possibile gestire immensi volumi di dati che prima potevano essere gestiti soltanto da supercomputer con grandi costi
- I prezzi dei sistemi sono crollati e le nuove tecnologie per i sistemi distribuiti sono abbordabili
- Il vero passo in avanti nei big data si ebbe quando compagnie come Yahoo!, Google e Facebook realizzarono di avere la necessità di guadagnare dalla enorme quantità di dati che i loro prodotti stavano realizzando
- Queste compagnie dovevano trovare nuove tecnologie per memorizzare, accedere e analizzare enormi quantità di dati quasi in *real time* in modo tale da ricavare benefici economici dal possesso di tanti dati sui partecipanti alle loro reti
- Le loro soluzioni a questo problema stanno trasformando il mercato della gestione dei dati. In particolare, le innovazioni MapReduce, Hadoop e BigTable hanno rappresentato la scintilla per una nuova generazione di strumenti per la gestione dei dati
- Queste tecnologie risolvono uno dei problemi principali, ovvero la capacità di elaborare efficientemente, efficacemente e velocemente enormi quantità di dati

Analisi tradizionali e avanzate

- Cosa fa, ora, un'organizzazione con tutti i suoi dati in tutte le sue forme per cercare di dare un significato ad essi?
- Alcune analisi utilizzeranno un Data Warehouse tradizionale, mentre altre analisi si avvarranno di tool predittivi avanzati
- Gestire globalmente dei big data richiede molti approcci diversi per supportare l'azienda a pianificare con successo il proprio futuro
- Questi Data Warehouse e Data Mart forniscono la compressione, il partizionamento multi-livello e un'architettura di elaborazione massivamente parallela

Big data analytics

- La capacità di gestire e analizzare petabyte di dati consente alle organizzazioni di trattare cluster informativi che potrebbero avere un impatto sul business
- Ciò richiede dei motori di analisi che possano gestire dati altamente distribuiti e fornire risultati che possono essere ottimizzati per risolvere un problema di business
- Le analisi possono diventare piuttosto complesse con i big data. Per esempio, alcune organizzazioni stanno usando modelli predittivi che integrano dati strutturati e non strutturati per predire le frodi
- Analisi sui social media, analisi sui testi e nuovi tipi di analisi vengono utilizzati dalle organizzazioni per poter «guardare dentro» ai big data.

Introduzione al *cloud*

- Il potere del *cloud* sta nel fatto che gli utenti possono accedere alle risorse di calcolo e di memorizzazione necessarie con poco o nessun supporto IT e senza la necessità di comprare hardware e software
- Una delle caratteristiche chiave del *cloud* è la grande scalabilità: gli utenti possono aggiungere o eliminare risorse quasi in *real time* sulla base delle loro necessità
- Il *cloud* gioca un ruolo importante all'interno del mondo dei big data. Possono accadere enormi cambiamenti quando questa infrastruttura si combina con le innovazioni nella gestione dei dati
- Il *cloud computing* è un metodo per fornire un insieme di risorse di calcolo condivise che includono applicazioni, piattaforme per il calcolo, la memorizzazione, il networking, lo sviluppo e il *deployment*, come pure processi di business

Cloud e Big Data

- Nel *cloud computing* ogni cosa (potenza di calcolo, infrastrutture di elaborazione, applicazioni, business process, data e analisi) può essere rilasciata come un servizio
- Per essere operazionale nel mondo reale, il *cloud* deve essere implementato con processi e automatismi comuni standardizzati
- Molte aziende gestiscono servizi cloud per ogni cosa dal backup alle opzioni Software as a Service (SaaS)
- Con la crescita del *mobile computing*, più consumatori, professionisti e aziende stanno creando e accedendo dati con servizi basati sul *cloud*

Modelli di *cloud deployment*

- Due modelli *cloud* chiave sono importanti nella discussione sui big data, ovvero i *cloud* pubblici e quelli privati
- Gran parte di quelle organizzazioni che adottano modelli di *deployment* e rilascio a *cloud* useranno una combinazione di risorse di elaborazione private (data center e *cloud* privati) e servizi pubblici (operati da una compagnia esterna per l'uso condiviso con una varietà di clienti che pagano una quota per l'utilizzo)
- Come queste compagnie bilanciano fornitori pubblici e privati dipende da un certo numero di fattori che includono la privacy, la latenza e gli obiettivi
- In questo modo è possibile determinare se si vuole utilizzare un *cloud* IaaS pubblico – per esempio per i progetti – o se si vuole continuare a mantenere i propri dati in locale
- Ovviamente, è anche possibile utilizzare una combinazione delle due soluzioni

Il *cloud* pubblico

- Il *cloud* pubblico è un insieme di infrastrutture hardware, di rete, di memorizzazione, di servizi, di applicazioni e di interfacce possedute e operate da una terza parte per l'uso da parte di altre compagnie e individui
- Questi fornitori commerciali creano un data center altamente scalabile che nasconde i dettagli dell'infrastruttura sottostante al consumatore
- I *cloud* pubblici sono praticabili perché gestiscono dei carichi di lavoro relativamente semplici e ripetitivi
  - Per esempio, la posta elettronica è un'operazione molto semplice. Perciò, un fornitore *cloud* può ottimizzare l'ambiente in modo tale che sia perfettamente adatto per supportare un gran numero di clienti, anche se essi generano molti messaggi
  - Analogamente, *cloud* provider pubblici che offrono servizi di memorizzazione e di elaborazione ottimizzano il loro hardware e software per supportare questi specifiche attività
- Al contrario, il tipico data center supporta tante applicazioni e carichi di lavoro differenti che lo rendono difficile da ottimizzare

Il *cloud* pubblico

- Un *public cloud* può essere molto efficace quando un'organizzazione sta eseguendo un progetto di *data analysis* complesso e necessita di più potenza di calcolo per gestire le proprie attività
- In aggiunta, le compagnie possono scegliere di memorizzare i dati in un *cloud* pubblico dove i costi per gigabyte sono relativamente bassi se confrontati con i costi di acquisto di dispositivi di memorizzazione
- I principali problemi con i *public cloud* per i big data sono le richieste di sicurezza e la quantità di latenza accettabile
- Non tutti i *cloud* pubblici sono uguali. Alcuni sono scalabili con un alto livello di sicurezza e di gestione dei servizi. Altri sono meno robusti e sicuri, ma anche molto meno costosi.
- La scelta dipende dalla natura dei big data e dalla quantità di rischio che ci si vuole assumere

Il *cloud* privato

- Un *cloud* privato è un insieme di risorse hardware, di rete, di memorizzazione, di servizi, di applicazioni e di interfacce possedute e gestite da un'organizzazione per l'utilizzo da parte dei propri impiegati, dei propri partner e dei propri clienti
- Un *cloud* privato può essere creato e gestito da una terza parte per l'uso esclusivo di un'organizzazione
- Il *cloud* privato è un ambiente altamente controllato, non aperto all'uso da parte del pubblico.
- Un *cloud* privato è altamente automatizzato con un focus sulla governance, sulla sicurezza e sull'affidabilità
- L'automazione rimpiazza dei processi di gestione dei servizi IT più manuali per supportare i clienti. In questo modo le regole e i processi di business possono essere implementati all'interno del software in modo tale che l'ambiente diventa più prevedibile e maneggevole
- Se le organizzazioni stanno gestendo un progetto di big data che richiede l'elaborazione di enormi quantità di dati, il *cloud* privato può essere la scelta migliore in termini di latenza e di sicurezza
- Un *cloud* ibrido è una combinazione di *cloud* privato e di utilizzo di servizi di *cloud* pubblico con uno o diversi punti di contatto tra i due ambienti
- Lo scopo è quello di creare degli ambienti *cloud* ben gestiti che possono combinare servizi e dati da una grande varietà di modelli di *cloud* per creare un'ambiente di calcolo unificato, automatizzato e ben gestito

Infrastructure as a service - IaaS

- Infrastructure as a Service (IaaS) è uno dei più diretti servizi di *cloud computing*. Esso consiste nel rilascio di servizi di computing che includono l'hardware, le reti, la memorizzazione e gli spazi di data center basandosi su un modello di affitto
- Il consumatore del servizio acquisisce una risorsa e paga per essa sulla base della quantità usata e della durata dell'utilizzo
- E' possibile trovare sia versioni pubbliche che private di IaaS
- Nell'IaaS pubblico l'utente paga per acquisire queste risorse; quando l'utente smette di pagare, la risorsa scompare
- In un servizio IaaS privato, è generalmente l'organizzazione IT o un integratore che crea l'infrastruttura progettata per fornire le risorse su richiesta agli utenti interni e, qualche volta, ai partner

Platform as a service - PaaS

- Platform as a Service (PaaS) è un meccanismo per combinare l'IaaS con un insieme astratto di servizi *middleware*, *tool* di sviluppo del software e di *deployment* che consentono all'organizzazione di avere un modo consistente per creare e mettere in esercizio applicazioni su un *cloud* o nei propri locali
- Un PaaS offre un insieme consistente di servizi di programmazione o *middleware* che assicura che gli sviluppatori abbiano un modo ben testato e ben integrato per creare applicazioni in un ambiente *cloud*
- Un ambiente PaaS mette insieme lo sviluppo e il *deployment* per creare un modo più maneggevole di costruire, rilasciare e scalare le applicazioni
- Un PaaS richiede un IaaS

Software as a service - SaaS

- Software as a Service (SaaS) è un'applicazione di business creata e ospitata da un provider in un ambiente condiviso.
- La condivisione si riferisce alla situazione in cui una singola istanza dell'applicazione gira in un ambiente *cloud* servendo più clienti che mantengono comunque separati tutti i loro dati
- I clienti pagano per il servizio con un modello di contratto basato sull'utente oppure mensilmente o annualmente
- Il modello SaaS prevede che al di sotto ci siano il PaaS e l'IaaS

Data as a service - DaaS

- Data as a Service (DaaS) è strettamente correlato al SaaS
- Il DaaS è un servizio indipendente dalla piattaforma che consente di connettersi al *cloud* per memorizzare e recuperare i propri dati
- In aggiunta, è possibile trovare un insieme di servizi specializzati per i dati che sono di grande beneficio per un ambiente di big data
- Per esempio, Google offre un servizio che può processare una *query* con 5 terabyte di dati in soli 15 secondi. Questo tipo di *query* richiederebbe tipicamente 10 volte questo tempo in un tipico data center

Cloud imperativo per i Big Data

- Svariate caratteristiche del *cloud* rendono quest'ultimo un componente importante dell'ecosistema dei big data

1) Scalabilità:
   - La scalabilità rispetto all'hardware si riferisce alla capacità di passare da potenze di calcolo piccole a potenze di calcolo grandi con la stessa architettura
   - Riguardo al software, essa si riferisce alla consistenza delle performance per unità di potenza quando le risorse hardware crescono
   - Il *cloud* può scalare a grandi volumi di dati. Il calcolo distribuito, che è parte integrante del modello *cloud*, lavora realmente secondo un modello «divide et impera»
   - Pertanto, se si hanno grossi volumi di dati, essi possono essere partizionati su più server *cloud*
   - Una caratteristica importante dell'IaaS è che essa può scalare dinamicamente. Questo vuol dire che se si scopre di avere bisogno di più risorse di quanto si era immaginato, queste si possono ottenere. Ciò è strettamente correlato con il concetto di elasticità

Cloud imperativo per i Big Data

2) Elasticità:
   - L'elasticità si riferisce alla capacità di espandere o ridurre la richiesta di risorse computazionali in tempo reale basandosi sulla domanda
   - Uno dei benefici del *cloud* è che i client hanno il potenziale di accedere a quanti servizi necessitano e quando ne hanno necessità
   - Questo può essere utile per progetti di big data dove si potrebbe dover espandere la quantità di risorse computazionali necessarie per far fronte al volume e alla velocità dei dati
   - Naturalmente, questa importante caratteristica del *cloud* che lo rende attrattivo agli utenti finali implica che il provider del servizio deve progettare un'architettura della piattaforma ottimizzata per questo tipo di servizio
3) Raggruppamento di risorse:
   - Le architetture *cloud* consentono la creazione efficiente di gruppi di risorse condivise che rendono il *cloud* economicamente fattibile

Cloud imperativo per i Big Data

4) Self-service:
   - Con il self-service l'utente di una risorsa *cloud* può usare un browser o l'interfaccia ad un portale per acquisire le risorse necessarie, ad esempio per eseguire un modello predittivo enorme
   - Ciò è drammaticamente differente da come si possono acquisire risorse da un data center, dove sarebbe necessario richiedere risorse al reparto IT
5) Costi iniziali spesso bassi:
   - Se si utilizza un *cloud* provider, i costi di up-front possono essere spesso ridotti perché non si stanno comprando enormi quantità di risorse hardware e non si sta affittando nuovo spazio per gestire i propri big data
   - Grazie all'adozione delle economie di scala associate all'ambiente cloud, quest'ultimo può risultare attrattivo
   - Ovviamente, l'organizzazione deve fare le proprie valutazioni per verificare se è interessata a un *cloud* pubblico, a uno privato, a uno ibrido, oppure a nessuna forma di *cloud*

Cloud imperativo per i Big Data

6) Pagamento in base all'utilizzo:
   - Una tipica opzione di pagamento per un provider *cloud* è la «Pay as You Go» (PAYG), che vuol dire che si pagano le risorse che effettivamente si utilizzano
   - Questo può essere utile se non si è sicuri di quali risorse sono necessarie per il proprio progetto di big data
7) Tolleranza ai guasti:
   - I provider di servizi *cloud* dovrebbero avere la tolleranza ai guasti intrinseca nella loro architettura, fornendo dei servizi senza soluzione di continuità anche in presenza di guasti in uno o più componenti del sistema

- In alcune situazioni un service provider non può anticipare i bisogni di un cliente. Pertanto, è una situazione comune aggiungere delle capacità addizionali da un provider di terze parti. Tipicamente, il consumatore non si dovrebbe accorgere che sta utilizzando un service provider *cloud* addizionale

Utilizzo del *cloud* per i Big Data

- Esistono vari modi con cui il *cloud* può essere utilizzato per i big data. Eccone alcuni:

1) IaaS in un *cloud* pubblico:
   - In questo scenario un'organizzazione utilizza una infrastruttura di public *cloud* per i propri servizi di big data perché non vuole utilizzare risorse interne
   - L'IaaS può consentire la creazione di macchine virtuali con capacità di memorizzazione ed elaborazione quasi illimitate
   - L'organizzazione può scegliere il sistema operativo che preferisce ed ha la flessibilità di scalare dinamicamente per soddisfare i propri bisogni
   - Un esempio potrebbe essere quello di utilizzare il servizio Amazon Elastic Compute Cloud (Amazon EC2) per eseguire un modello predittivo in *real time* che richiede che i dati siano processati utilizzando un'elaborazione massicciamente parallela
   - Potrebbe essere un servizio che elabora dati di vendita di una catena relativi a miliardi di pezzi per classificare i consumatori in *real time*

Utilizzo del cloud per i Big Data

2) PaaS in un *cloud* privato:
   - PaaS è un'intera infrastruttura personalizzata che può essere usata per progettare, implementare e rilasciare applicazioni e servizi in un ambiente *cloud* pubblico o privato
   - PaaS consente ad un'organizzazione di utilizzare dei servizi middleware fondamentali senza dover far fronte alla complessità di dover gestire elementi hardware e software proprietari
   - I venditori PaaS stanno iniziando ad includere tecnologie dei big data, quali Hadoop e MapReduce, all'interno delle loro offerte PaaS
   - Per esempio, si potrebbe voler costruire un'applicazione specializzata per analizzare grandi quantità di dati medici. L'applicazione farebbe uso di dati sia real-time che non real-time. Essa richiederebbe Hadoop e MapReduce per le attività di memorizzazione ed elaborazione
   - La cosa importante di PaaS in questo scenario è quanto velocemente l'applicazione può essere rilasciata. Non è necessario aspettare i team di IT interni per aggiornarsi sulle nuove tecnologie ed è possibile sperimentare più liberamente
   - Una volta che è stata identificata una soluzione solida, questa può essere portata *in house* quando il proprio reparto IT è pronto a supportarla

Utilizzo del cloud per i Big Data

3) SaaS in un *cloud* ibrido:
   - In questo caso si potrebbe desiderare di analizzare «la voce dei clienti» proveniente da più canali
   - Molte compagnie hanno compreso che una delle sorgenti di dati più importanti consiste in ciò che i clienti pensano e dicono sulla loro compagnia, sui loro prodotti e i loro servizi. Avere accesso a dati di questo tipo può fornire un supporto enorme nei comportamenti e nelle azioni della compagnia
   - In uno scenario di questo tipo, il venditore SaaS fornisce la piattaforma per l'analisi, come pure i dati dei social media
   - In aggiunta, la compagnia potrebbe utilizzare il proprio CRM aziendale nel proprio ambiente cloud privato per contribuire all'analisi
- Alcuni attori del settore *cloud* stanno usando il termine «big data application» per descrivere applicazioni che operano nel *cloud* e usano i big data. Esempi di questi includono Amazon.com e LinkedIn
- Si potrebbe obbiettare che queste sono in realtà applicazioni SaaS che risolvono un particolare problema di business. È comunque solo un problema di denominazione in un contesto emergente

Criticità nell'utilizzo del cloud per i Big Data

- I servizi basati sul *cloud* possono fornire una soluzione economica ai bisogni dei big data. Tuttavia, il *cloud* ha i suoi problemi ed è importante conoscerli prima di spostare i propri big data su di esso. Alcune problematiche da considerare sono le seguenti:
- Integrità dei dati:
  - Bisogna assicurarsi che il fornitore abbia i giusti controlli capaci di assicurare il mantenimento dell'integrità dei dati
- Conformità:
  - Bisogna assicurarsi che il provider possa far fronte a qualunque problematica di conformità relativa alla propria azienda o al proprio settore industriale
- Costi:
  - I costi bassi possono trarre in inganno. Bisogna stare attenti a leggere tutti i dettagli di un contratto e bisogna assicurarsi di sapere cosa si vuole fare veramente nel *cloud*

Criticità nell'utilizzo del cloud per i Big Data

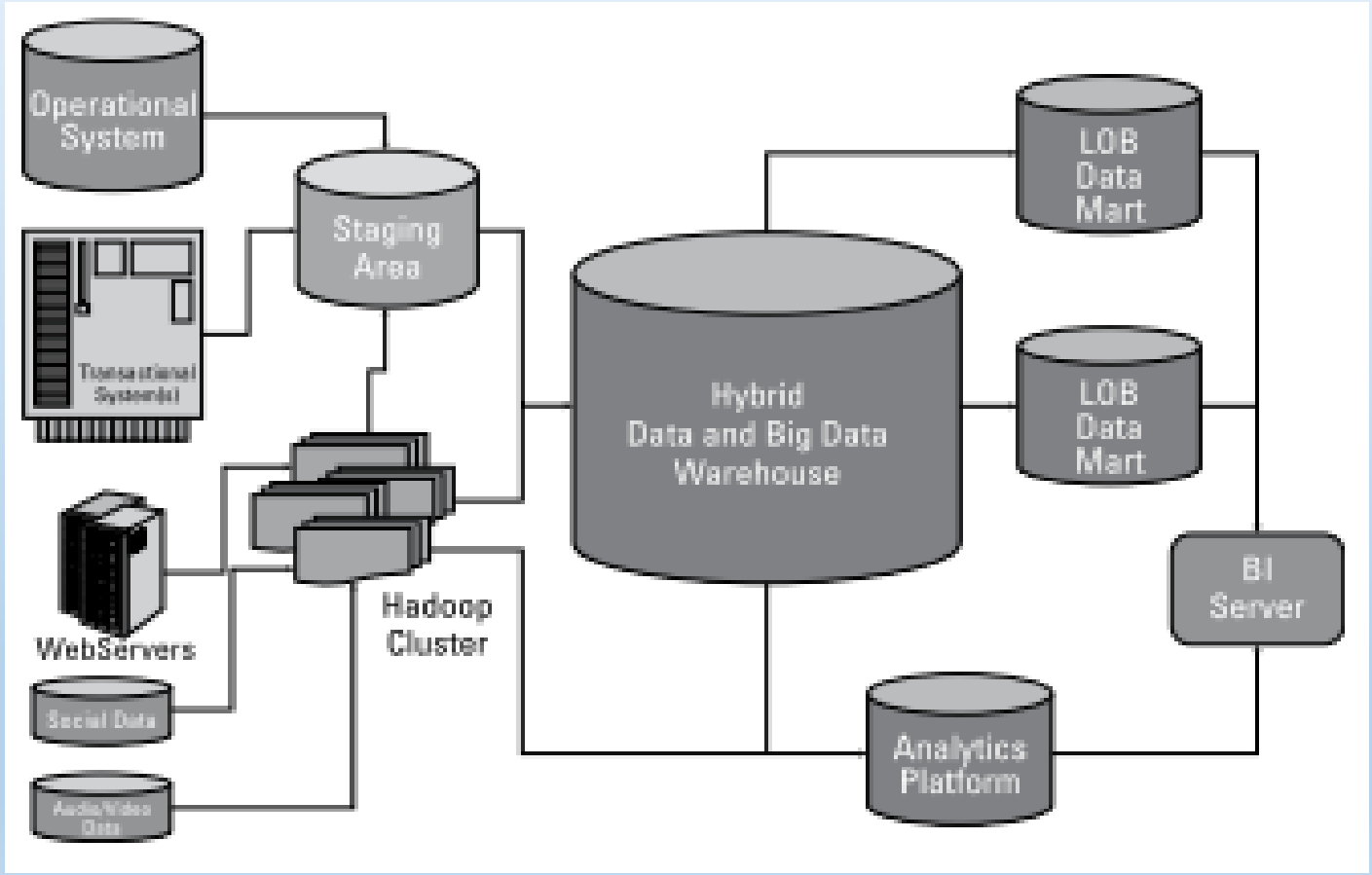
- Trasporto di dati:
  - Bisogna cercare di capire come portare su *cloud* i dati la prima volta. Per esempio, alcuni provider consentono di inviarli su dei dispositivi tramite posta tradizionale. Altri insistono nel fare il loro upload attraverso la rete (molto più costoso)
- Performance:
  - Poiché si è interessati ad ottenere determinate performance dal proprio service provider, bisogna assicurarsi che esistano delle esplicite definizioni di Service Level Agreement per ciò che concerne la disponibilità, il supporto e la performance
  - Per esempio, il provider può dire che è in grado di garantire l'accesso ai dati il 99.999% delle volte; è opportuno, però, leggere il contratto. Questo valore, ad esempio, include la manutenzione schedulata?
- Accesso ai dati:
  - Quali controlli vengono fatti per assicurarsi che il cliente e solo il cliente può accedere ai propri dati? In altre parole, che forme di accesso sicuro vengono adottate?
  - Questo potrebbe includere la gestione delle identità, dove l'obiettivo principale è quello di proteggere le informazioni sulle identità personali in modo tale che sia opportunamente controllato l'accesso alle risorse del computer, alle applicazioni, ai dati e ai servizi
- Localizzazione:
  - Dove sono localizzati i propri dati? In alcune compagnie e in un alcune nazioni, ci sono delle disposizioni che impediscono che i dati siano memorizzati o elaborate su macchine di nazioni differenti

Data Warehousing e Big Data

- Nel futuro è presumibile che sia necessario costruire ambienti ibridi dove i Big Data possano operare fianco a fianco con i Data Warehouse classici
- Innanzitutto è importante riconoscere che il Data Warehouse così come è progettato oggigiorno non cambierà nel breve termine. Pertanto, è più pragmatico usarlo per ciò per cui è stato pensato, ovvero fornire una versione ben controllata della realtà riguardo ad un argomento che l'organizzazione vuole analizzare
- Il Data Warehouse potrebbe fornire informazioni su una particolare linea di prodotti dell'azienda, sui suoi clienti, sui suoi fornitori e sui dettagli delle transazioni
- L'informazione gestita nel Data Warehouse o nel Data Mart dipartimentale è stata costruita attentamente in modo tale che i metadati fossero accurati
- Con la crescita di nuove informazioni basate su web, è pratico e spesso necessario analizzare questa enorme massa di nuove informazioni insieme ai dati già presenti sul Data Warehouse

Data Warehousing e Big Data

- L'architettura di un sistema ibrido viene mostrata nella seguente figura:



Data Warehousing e Big Data

- Certi aspetti dell'interazione del Data Warehouse con i Big Data possono essere relativamente facili da gestire.
  - Per esempio, molte delle sorgenti di Big Data includono i propri metadati ben progettati. I siti di e-commerce complessi includono elementi di dati ben definiti (clienti, prezzi, etc.).
  - Perciò, quando si conducono analisi tra il Data Warehouse e la sorgente di Big Data per gestire le informazioni si interagisce con due insieme di dati con modelli di metadati attentamente progettati che sono stati razionalizzati
- Naturalmente, in alcune situazioni, le sorgenti informative mancano di metadati espliciti. Prima che un analista possa combinare dati di un Data Warehouse con Big Data meno strutturati è necessario effettuare integrazioni
- Tipicamente, l'analisi iniziale di petabyte di dati rileverà dei pattern interessanti che possono aiutare a predire dei sottili cambiamenti dell'azienda o potenziali soluzioni a un problema

Data Warehousing e Big Data

- L'analisi iniziale può essere completata utilizzando tool quali Hadoop MapReduce
- A questo punto, si può iniziare a comprendere se si è capaci di supportare la valutazione del problema che si sta affrontando
- Nel processo di analisi è tanto importante eliminare dati non necessari quanto identificare dati rilevanti per il contesto di business
- Quando questa fase è completa, i dati rimanenti devono essere trasformati in modo tale che le definizioni dei metadati siano precise. In questo modo, quando i Big Data saranno combinati con i dati del Data Warehouse i risultati che si otterranno saranno accurati e significativi
- Ovviamente, nel far comunicare i Big Data con il Data Warehouse diventa fondamentale un processo e un tool di integrazione
- Il processo di ETL è fondamentale anche per la componente Big Data

Data Warehousing e Big Data

- Il loading dei Big Data è differente rispetto a quanto avviene per i Data Warehouse.
- Nel Data Warehouse, dopo che i dati sono stati codificati, non cambiano mai
- La struttura distribuita dei Big Data porterà spesso le organizzazioni a caricare i dati in una serie di nodi e solo dopo ad effettuare l'estrazione e la trasformazione
- Quando si crea un ambiente ibrido, la natura distribuita dell'ambiente di Big Data può modificare drammaticamente la capacità delle organizzazioni di gestire enormi volumi di dati nel contesto di riferimento

Cambiamento del ruolo del Data Warehouse

- È utile pensare alle similarità e alle differenze tra il modo con cui i dati vengono gestiti nel Data Warehouse tradizionale e il modo in cui lo sono quando il Data Warehouse è combinato con i Big Data
- Le similarità tra i due metodi di gestione dei dati comprendono:
  - Requisiti per le definizioni di dati comuni
  - Requisiti per estrarre e trasformare le sorgenti di dati più importanti
  - La necessità di conformarsi ai *business process* e alle regole richieste
- Le differenze tra il Data Warehouse tradizionale e i Big Data possono essere così riassunte:
  - Il modello di calcolo distribuito dei Big Data sarà essenziale per consentire al modello ibrido di funzionare
  - L'analisi dei Big Data sarà il principale obiettivo degli sforzi, mentre il Data Warehouse tradizionale sarà usato principalmente per aggiungere un contesto di business storico e transazionale

Il futuro dei Data Warehouse

- Con l'avvento dei big data il mercato dei Data Warehouse è iniziato a cambiare e ad evolversi
- Nel passato, era semplicemente non economico per le compagnie memorizzare enormi quantità di dati da un gran numero di sistemi di registrazione. La mancanza di efficacia nei costi e di architetture pratiche per il calcolo distribuito implicava che un Data Warehouse doveva essere progettato in modo da ottimizzarlo per operare su un singolo sistema unificato
- Pertanto, i Data Warehouse erano costruiti attorno ad un obiettivo
- In aggiunta, il Data Warehouse doveva essere attentamente controllato in modo tale che i dati venivano definiti e gestiti in modo preciso
- Questo approccio rendeva i Data Warehouse accurati e utili alle organizzazioni per poter interrogare le corrispettive sorgenti di dati

Il futuro dei Data Warehouse

- Tuttavia, questo stesso livello di controllo e precisione ha reso difficile fornire all'azienda un ambiente capace di gestire le molto più dinamiche sorgenti di Big Data. Il Data Warehouse evolverà troppo lentamente
- I Data Warehouse e i Data Mart continueranno ad essere ottimizzati per l'analisi aziendale
- Tuttavia, una nuova generazione di offerte combinerà i depositi di dati storici e altamente strutturati con diversi tipi di sorgenti di Big Data
- Prima di tutto, le sorgenti di Big Data forniranno la capacità di analizzare enormi volumi di dati quasi in *real time*
- Secondo, una sorgente di Big Data prenderà i risultati di un'analisi e fornirà un meccanismo per fare match tra i metadati dell'analisi dei Big Data e i requisiti del Data Warehouse