Lezione 6 modulo 1

In questo modulo parleremo di un'altra tecnologia che si trova a livello applicativo. Riprendiamo un discorso che abbiamo già fatto altre volte che è quello della piramide di Anthony, in cui abbiamo detto che dovremo gestire informazioni sia a livello operativo, sia a livello decisionale, sia dal punto di vista tattico che dal punto di vista strategico e avevamo già detto che per queste fasce potranno esserci dei sistemi di tipo supporto alle decisioni che saranno dei sistemi di tipo specifico per fare delle operazioni che consentiranno di lavorare su dei dati aggregati, esaminando anche l'andamento delle attività di tipo operazionale e avevamo anche detto che potevano esserci due tipi di famiglie di sistemi: i sistemi di supporto alle transazioni, livello operazionale è caratterizzato dal fatto che andiamo a effettuare delle transazioni che devono essere gestite in modo efficiente e abbiamo detto che a questo livello possiamo chiamare i sistemi anche Online Transaction Processing. Un'altra famiglia di sistemi è quella chiamata sistemi OLAP, Online Analytical Processing. Qui siamo interessati soprattutto alla analisi dei dati delle informazioni e abbiamo anche visto che queste informazioni saranno di provenienza, sia interna, quindi ci saranno dei flussi informativi dal livello operativo, ma anche potranno esserci dei dati che proverranno da fonti esterne. Quindi, in questo modulo, andiamo a vedere le caratteristiche di tecnologie che sono pensate specificatamente per questo tipo di sistemi e vedremo, prima le loro caratteristiche, e poi andremo a vedere diverse alternative tecnologiche e la loro architettura. Innanzitutto riprendiamo un concetto che è fondamentale: quando parliamo di analisi dei dati, o anche si parla di Business Intelligence, noi vogliamo avere degli strumenti che ci consentano di fare interrogazioni di tipo complesse. Queste interrogazioni saranno soprattutto caratterizzate dal fatto che vogliamo vedere i nostri dati, non come al livello transazionale, al livello corrente quello che sto facendo in modo specifico, ma in modo tipicamente aggregato, quindi vorremmo aggregare dei dati per estrarre conoscenza dai dati che abbiamo a disposizione e poterli analizzare. Una tipica aggregazione è quella che si fa analizzando i dati, ad esempio, su un periodo di tempo: voglio vedere le vendite che ho fatto in un giorno, oppure il totale delle vendite che ho fatto in un anno. Questi sono esempi di aggregazione: avrò i dati transazionali che registreranno le singole vendite e vorrò andare ad analizzare quelli che sono i risultati su un periodo. Questo ci fa vedere anche un'altra cosa: che i dati che andiamo a trattare non sono solo quelli correnti, ma sono quelli relativi ad un periodo quindi avremo anche dei dati storici. Quindi noi vogliamo fare interrogazioni complesse perché questi dati tipicamente saranno in grandi quantità. Vorrò mettere insieme diverse informazioni riguardo a questi dati, ad esempio, se parlo delle vendite, vorrò anche sapere se sto parlando, ad esempio, di una catena di supermercati, se queste vendite sono state fatte in alcuni luoghi oppure in alcune zone, quindi raccoglierò questa informazione da tante fonti, ad esempio, da tanti punti di vendita e vorrò analizzare le vendite effettuate lungo un periodo, quindi avrò tanti dati e avrò l'integrazione di questi dati e vorrò esaminarli secondo quelle che noi chiameremo successivamente le dimensioni di analisi. Le dimensioni di analisi mi diranno, ad esempio, che voglio occuparmi di zone geografiche, se i punti vendita sono collegati in città e regione oppure di periodi temporali a granularità diversa, il tempo può essere a livello di singolo minuto di vendita, piuttosto che ora, giorno, mese, trimestre oppure anno. Quindi dovrò definire, in generale, quali sono le nostre dimensioni che ci interessano analizzare e anche parleremo di questa granularità e come potremmo definirla. È emerso nel tempo che i tradizionali strumenti di gestione di basi di dati che consentono di gestire molto efficientemente le transazioni, non sono molto adatti per fare queste operazioni. Queste operazioni appunto coinvolgono tanti dati, coinvolgono molte operazioni di join e quindi diventano molto onerose dal punto di vista computazionale. Quindi è emersa l'esigenza di avere altri strumenti, strumenti per il supporto all'analisi dei dati. Cominciamo a vedere quali possono essere le tipologie di questi strumenti. La prima tipologia sono i report. Posso raccogliere le informazioni che ho a disposizione, sintetizzarle aggregando i dati secondo gli aspetti che mi interessano in particolare e costruire dei report che possono essere dati ai decisori. Quindi il target per questo tipo di sistemi sono i livelli di tipo manageriale o strategico di un'azienda. Il report tipicamente viene consegnato in un file oppure stampato, ha una caratteristica cioè di essere statico: una volta creato il report, chi lo utilizza lo

potrà leggere ma non potrà elaborare ulteriormente questi dati per fare delle proprie analisi su quanto viene fornito nel report sintetico. Questa esigenza però chiaramente emerge nel momento in cui voglio andare ad esaminare meglio alcune caratteristiche di questi dati, ad esempio, perché trovo degli aspetti che possono essere interessanti, che devono essere approfonditi. Quindi un altro strumento che è stato utilizzato razionalmente per l'analisi dei dati è il foglio di calcolo, ad esempio excel. Anche in questo caso, quello che posso fare è estrarre dei dati anche aggregandoli secondo appunto certe dimensioni di analisi e poi rendo queste informazioni disponibili al decisore che quindi potrà fare delle proprie elaborazioni per esaminare meglio i dati. Quindi ha la caratteristica di essere interattivo. Qual è il problema che può esserci utilizzando un foglio excel? Ho esportato i dati dalle fonti e faccio separatamente quest'analisi e successivamente non ho più dei controlli di consistenza, di correttezza di questi dati, rispetto a quelle che sono le regole di un database da cui, o da più database da cui vengono estratti. Quindi i dati si separano da quello che è il flusso di fornitura dei dati e vengono gestiti separatamente con tutti i problemi che si possono verificare perché si è fatta questa operazione. Quindi è emersa l'esigenza di gestire i dati in un modo più sistematico: un altro aspetto di Excel è che è possibile fare delle elaborazioni ma non è possibile fare quello che si fa tipicamente su una base di dati, cioè delle interrogazioni che vengono fatte attraverso delle interfacce di interrogazione basate su dei linguaggi di interrogazione. Quindi quello che abbiamo come esigenza è di avere delle basi di dati, ma con caratteristiche adatte alle elaborazioni fatte in modo analitico, quindi per sistemi di tipo OLAP. È quindi in grado di contenere grandi quantità di dati e di fare delle interrogazioni che sono, come abbiamo visto, interrogazioni non predefinite ma possono essere fatte dal decisore sulla base di informazioni che vengono visualizzate che quindi il decisore vuole approfondire e analizzare meglio e quindi in modo casuale con delle interrogazioni ad hoc e questo quindi vuol dire che non posso utilizzare le interrogazioni e devo gestirle però in modo che i tempi di risposta siano ragionevoli, quindi con tempi di risposta che, elaborando tante quantità di dati non ci si aspetta necessariamente siano rapidissimi, ma che devono essere ragionevoli per poter avere il risultato e poterlo consultare. Quindi quello che adesso andremo a vedere sono: quali sono queste caratteristiche che vogliamo avere per sistemi di questo tipo, poi andremo a vedere quali sono le tecnologie che sono state sviluppate a livello applicativo per supportare operazioni in questo tipo. Innanzitutto vediamo quali sono gli obiettivi. Vogliamo supportare decisioni, ma, insieme a questo, vogliamo anche identificare i possibili problemi: voglio capire perché un certo punto di vendita vende meno di altri punti di vendita simili un certo prodotto, una certa tipologia di prodotti. Poi abbiamo un'altra caratteristica che l'utilizzatore tipico è a livello manageriale, quindi vogliamo dare degli strumenti che consentono di analizzare i dati, non facendo operazioni solo predefinite, ma anche lasciando la libertà all'utilizzatore di analizzare i dati secondo le esigenze che emergono durante l'analisi. Un altro aspetto che vogliamo considerare è la gestione di dati storici, quindi avremo sempre una serie di dati nel tempo, piuttosto che non lo stato corrente dei dati nel sistema che rappresenta la situazione attuale. Quindi non gestisco la singola vendita, ma sono interessato ad analizzare le vendite che sono state effettuate all'interno di un certo periodo, ad esempio. Poi un altro aspetto da gestire è che vogliamo avere dei dati aggregati. In genere non ci interessa avere i singoli dati che provengono dalle nostre fonti di dati, ma vederli già ad un certo livello di aggregazione su cui possiamo fare ulteriori analisi. Un'altra caratteristica molto importante del sistema di questo tipo è che l'accesso ai dati è tipicamente in lettura. Mentre il tipico sistema transazionale ha l'esigenza di scrivere e di scrivere in modo efficiente, quindi molto velocemente, in questo caso, vogliamo tipicamente leggere dei dati e analizzarli, quindi non andiamo tipicamente a scrivere nuovi dati, quindi genereremo dei report, faremo delle analisi aggregando i dati, ma non andremo a introdurre nuove informazioni di tipo transazionale. Questo caratterizza questi sistemi: noi vorremmo fare quindi delle query molto complesse su dati storici, ma non andremo a aggiornare questi dati, ma vorremmo avere un'efficienza nell'eseguire queste query che sono tipicamente query di lettura. Vediamo adesso quali sono le caratteristiche dei dati. Cosa sono i dati che noi vorremmo andare a gestire all'interno di un modulo applicativo che consente di dare queste funzionalità? Innanzitutto, in genere, il modello relazionale si è visto che può essere non del tutto adatto a gestire dati di questo tipo perché, se vogliamo fare delle operazioni come quelle che abbiamo visto prima, tipicamente

avremo grandi quantità di dati e soprattutto avremo grandi quantità di operazioni di join fra tabelle, che sono tipicamente operazioni lunghe da eseguire e quindi c'è l'esigenza di rendere le query efficienti. Quindi si sono sviluppati i moduli applicativi ad hoc, proprio per risolvere questo problema. Un altro aspetto che abbiamo all'interno di Datawarehouse è che i dati si distinguono in due tipologie di dati principali: avremo i cosiddetti fatti, cioè quello che io posso estrarre dalle fonti, tipicamente abbiamo detto che considereremo più fonti nell'estrazione dei dati, ma l'altro aspetto importante per i moduli di supporto alla Business Intelligence è che abbiamo delle dimensioni di analisi, cioè vogliamo strutturare i nostri dati in modo che sia facile e efficiente fare interrogazioni di tipo complesso, focalizzandoci su alcuni aspetti che vogliamo analizzare, tipo le tipologie di prodotti oppure l'analisi nel tempo, quindi dovremo definire queste dimensioni di analisi. Un altro aspetto che dobbiamo definire per quanto riguarda i dati sono le misure. Dobbiamo definire cosa vogliamo misurare dai fatti, definire le unità di misura e questo sarà alla base poi delle attività di analisi che faremo lungo le varie dimensioni di analisi. Vediamo adesso come questo può essere visualizzato. Ovviamente noi parleremo in generale di n dimensioni, qui per poter visualizzare un esempio, gestiamo tre dimensioni. Parleremo quindi in generale, non di cubi, ma di ipercubi che consentono di darci questa visualizzazione di quello che avviene all'interno di un Datawarehouse. Cosa voglio fare? Vogliamo definire dei fatti e supponiamo di avere come esempio una vendita e queste sono le nostre dimensioni di analisi. Abbiamo detto l'aspetto temporale: nell'aspetto temporale noi dovremmo definire una granularità minima, ad esempio, qua abbiamo deciso di definire le date come giorni in cui io registro le informazioni. Quindi la mia granularità temporale minima è quella del giorno in questo cubo. Un'altra cosa che ci interessa, abbiamo detto, è dove vengono effettuate le vendite, quindi abbiamo, ad esempio, una serie di negozi e supponiamo di avere qui sei negozi rappresentati. L'altro aspetto che vogliamo andare a registrare sono i prodotti venduti, quindi voglio sapere quali prodotti ho venduto e qui sono indicati con dei codici dall'1 al 6, lotto 1, 1-6. Queste sono le dimensioni di analisi. Ovviamente in generale io potrò avere n dimensioni di analisi. Qual è il fatto che voglio registrare? Abbiamo detto sono le vendite, quindi un cubetto all'interno di questo cubo mi dirà quali sono le vendite in un giorno, in un certo negozio e di un certo prodotto, quindi quel cubetto sarà relativo alle vendite del prodotto 1 nel negozio 6 del primo di marzo. Dobbiamo anche però misurare che cosa vogliamo registrare delle vendite. Ad esempio, qui abbiamo un esempio di misure associate alle vendite. Noi possiamo definire qual è il costo di una certa unità venduta, qual è la quantità venduta e qual è lo sconto che è stato effettuato. Quindi, qui andremo a registrare delle informazioni relative alle vendite che rappresentano queste informazioni per quel giorno, per quel negozio e per quel prodotto. Ovviamente abbiamo detto che questi sono aspetti che dobbiamo a definire, cioè vogliamo parlare di vendite, i nostri fatti sono le vendite, possiamo definire quali sono le dimensioni di analisi, prodotto, punto di vendita e data e possiamo definire quali sono le misure registrate per le vendite per poter poi fare le analisi che quindi potranno essere analisi, ad esempio, relative alla quantità complessivamente venduta in un certo periodo di tempo. Un aspetto che è da sottolineare è che tipicamente noi dovremmo registrare in un'architettura di questo tipo, in un modello in questo tipo, dei dati che sono dei valori distinti. Quindi cosa succede se noi dobbiamo registrare dei dati che tipicamente non hanno questa caratteristica, come potrebbero essere dei dati numerici? Dovremmo definire una discretizzazione di questi dati, ad esempio, all'interno di un certo intervallo di valori, definire degli intervalli di valori in modo da poter ridefinire gli elementi che consentono poi di posizionare un certo dato all'interno del cubo in una certa posizione. Un altro aspetto che è importante, che abbiamo accennato, è che qui vediamo per i dati, per il punto vendite, per i prodotti, quelle che sono le informazioni a granularità più dettagliata, quindi abbiamo deciso che i nostri dati sono, ad esempio, registrati sulla base di giorni. Abbiamo detto però che nella nostra analisi noi vogliamo tipicamente fare analisi per periodo, quindi io devo collegare questa granularità più fine dei giorni a quelle che sono le dimensioni di analisi, ad esempio, abbiamo detto che possiamo parlare di mesi oppure di anni. Quindi non ci interessa solo avere questi singoli dati e il loro valore ma ci interessa anche sapere qual è la gerarchia associata alle dimensioni. Facciamo un esempio: abbiamo detto che registriamo come elementi nel cubo l'ID dei prodotti che sono la granularità più dettagliata dei prodotti. Però io poi potrò analizzare i nostri prodotti, ad esempio, per tipi,

noi abbiamo qua solo degli esempi, ad esempio, un certo prodotto potrà essere di tipo pesce, di tipo carne oppure un prodotto può essere un bagnoschiuma, biscotti e così via e questi tipi potranno essere a loro volta legati, ad esempio, questi sono prodotti alimentari, il bagnoschiuma potrebbe essere collegato alla categoria dei detersivi e così via. Quindi, in generale, e questo sarà definito durante la progettazione, noi definiremo una gerarchia per una certa dimensione di analisi che ci dirà che tipo di aggregazioni potremmo fare. Quindi definire una gerarchia di questo tipo vorrà dire che io potrò fare interrogazioni, chiedendo ad esempio la quantità di bibite vendute in un certo periodo di tempo. Anche per quanto riguarda il tempo, avremo delle gerarchie temporali, ad esempio, io posso avere, se parto dall'unità giorno, gerarchie tipiche potranno essere giorno, mese, trimestre, anno. Definire queste gerarchie, vorrà dire che io potrò fare delle interrogazioni che saranno interrogazioni veloci, ottimizzate, considerando gli elementi che ho definito, quindi quello che io potrò fare facilmente in questo caso è definire tutte le vendite che sono state fatte in un certo mese. È chiaro che se io dovessi fare un'analisi, ad esempio, per settimane questo non è previsto in questo momento, in questa gerarchia, e quindi dovrei andare a identificare quali sono i giorni all'interno di una certa settimana e fare una elaborazione per poter fare la query relativa. Quindi dovremo anche progettare queste gerarchie, dovremo decidere che tipo di elaborazioni tipicamente ci interessano. Vediamo adesso quali sono le proprietà che vorremmo avere da prodotti di questo tipo a livello applicativo. Queste proprietà possono essere riassunte in un acronimo che è FASMI. Adesso andiamo a vedere il suo significato. Allora, innanzitutto, F sta per qualcosa che abbiamo già visto: i fatti. È basato sull'identificazione di quali sono i fatti che vogliamo andare a registrare, le vendite, nell'esempio che abbiamo fatto, ma i fatti ovviamente all'interno di un certo sistema, potranno essere anche di tipo diverso, ad esempio, in un sistema di negozi possiamo parlare delle forniture, è un ambito diverso, quindi un fatto di tipo diverso che voglio analizzare e tipicamente noi avremo ipercubi diversi a seconda dei fatti che vogliamo considerare, possono essere le vendite, possono essere le forniture, possono essere i reclami che ricevo dai clienti, tutti questi sono fatti diversi. Un altro aspetto che abbiamo già visto, passiamo all'inglese per coerenza con la sigla che è stata creata in inglese, A di Analytical. Abbiamo visto quello che è caratteristico di questi sistemi, noi vogliamo fare analisi, quindi questo vuol dire che avremo definito le nostre dimensioni di analisi, le nostre misure che potremmo considerare poi nelle nostre interrogazioni. Poi ci sono altri aspetti che sono importanti. Noi vogliamo tipicamente avere questo sistema, non per il singolo decisore, ma per tutti i decisori all'interno di un'organizzazione, che dovranno esaminare questi dati a seconda delle varie prospettive e condividere eventuali decisioni. Quindi un'altra caratteristica di questi sistemi è che sono condivisi, quindi Shared. Quindi nell'architettura non pensiamo solo al singolo utilizzatore, tipo il singolo utilizzatore che utilizza il proprio foglio Excel, ma avremo un sistema che consente di condividere dati, organizzati in modo appropriato per fare delle analisi, fra più utilizzatori. Poi un altro aspetto che abbiamo già visto è che la M sta per Multidimensional. Dovremo definire più dimensioni di analisi, come abbiamo già detto, e poi la base è che sia Information, vogliamo gestire delle informazioni che sono di interesse per gli utilizzatori specifici di un certo sistema. Inoltre, abbiamo alcune caratteristiche peculiari di questi sistemi: sono orientati agli oggetti. Ad esempio, abbiamo parlato di oggetti da analizzare, le vendite, le forniture, i reclami sono tutti oggetti che possono essere oggetti di analisi. Poi vogliamo che sia integrato. Noi abbiamo più fonti di dati e le fonti saranno sia interne, ma anche esterne. Posso essere interessato ad analizzare dati, nelle mie analisi, che provengono anche da fonti esterne che possono aver condizionato l'andamento della nostra organizzazione. Un altro aspetto che abbiamo è che i dati saranno variabili nel tempo. Ovviamente la mia dimensione storica richiederà di caricare i nuovi dati all'interno del sistema per poter considerare tutto il periodo fino al momento in cui si vuole effettuare l'analisi. Quindi uno degli aspetti è che dovremmo porci il problema di decidere come caricare questi dati in modo che i dati all'interno del sistema di analisi siano disponibili per il periodo di interesse da considerare. Ultimo aspetto è che nel nostro sistema, noi abbiamo detto che vogliamo avere dei dati in lettura, vogliamo che questi dati rimangano gli stessi, non vengano modificati, e quindi una persistenza nel tempo. Vediamo quindi quali sono le caratteristiche della tecnologia che è stata sviluppata per sistemi con queste caratteristiche, parleremo di Datawarehouse come tecnologia separata da quella tipica dei DBMS per la gestione di dati

che tipicamente sono DBMS relazionali e quindi basati su tabelle. L'architettura che possiamo considerare è quella che vediamo rappresentata in questa figura. Innanzitutto consideriamo che partiamo da dati che vengono estratti da sorgenti. Abbiamo detto che le sorgenti possono essere di vario tipo e possono essere interne ed esterne. Tipica sorgente interna sono i dati operazionali, che saranno contenuti in uno o più basi di dati e registreranno le transazioni che sono state effettuate all'interno dell'organizzazione. Posso avere anche dati non strutturati, quindi ad esempio dei report che sono disponibili e che anch'essi possono essere utilizzati poi per esaminare le caratteristiche delle attività. Un altro tipo di dati che posso avere, indicati qui come i Big Data, sono i dati che arrivano da strumenti di misura tipo sensori che rilevano dati ad esempio relativi al passaggio delle persone in un negozio oppure relativi a temperature all'interno di un certo punto di vendita, nei vari punti del punto di vendita e così via. Quindi tanti dati che tipicamente vengono registrati a una granularità temporale abbastanza fine e che vorrò analizzare per vedere se ci sono, ad esempio, correlazioni con degli eventi che sono stati registrati, ad esempio, per identificare dei problemi. Poi potrò avere diverse sorgenti esterne, ad esempio, sorgenti esterne potranno essere altri report, altri basi di dati oppure potranno essere informazioni che io posso avere dal servizio, ad esempio di tipo meteorologico, quindi vengono dall'esterno e però possono di nuovo essere utili per capire l'andamento di una certa attività nell'organizzazione, quindi per analizzare i fatti. Ovviamente questi dati avranno tutte caratteristiche diverse, avranno schermi diversi, avranno formati diversi e così via. Cosa dovrò fare? Dovrò fare un'operazione che ci porterà a caricare questi dati nel nostro Data Warehouse che vediamo come una base di dati di dati da analizzare, quindi organizzati secondo fatti, dimensioni e misure e che idealmente vediamo come centralizzata, quindi raccoglierò tutti i dati che saranno oggetto dell'analisi. Questa operazione viene fatta tramite operazioni di tipo ETL. ETL sta per estrazione - Extraction e Transformation. Abbiamo detto che i formati, ad esempio, possono essere diversi, la struttura dei dati può essere diversa nelle varie fonti. Dovremmo fare in modo di ricondurre tutti questi dati ad un formato unico per poterli registrare all'interno del Data Warehouse. Quando i dati saranno pronti, potremmo fare l'operazione di caricamento, Loading. Nel fare questo, a volte viene utilizzata anche un'area di memorizzazione intermedia che viene chiamata Staging Area, che ci consentirà di caricare i dati e lavorarci su per fare queste operazioni. Noi vogliamo vedere più in dettaglio queste operazioni, ma prima di passare a questo, vediamo anche la parte destra di questa figura. Questi dati informazionali vengono caricati in un Data Warehouse, quindi da questo possiamo fare direttamente delle operazioni di analisi. Ovviamente noi possiamo generare le analisi anche con strumenti di analisi di vario tipo, poi abbiamo la possibilità di generare report con strumenti di report, oppure possiamo avere un motore che ci consente di fare le interrogazioni, secondo la nostra filosofia di utilizzare un ipercubo, quindi fare operazioni di interrogazione oppure di correlazione fra i dati, estrazioni caratteristiche particolari dei dati. Questo lo possiamo fare direttamente dalle Data Warehouse, quindi prendendo i dati dalle Data Warehouse, focalizzandoci su un ipercubo e analizzando questi dati. Spesso però i decisori lavorano solo su una parte dei dati, quindi si focalizzano su un aspetto, ad esempio, se abbiamo il gestore del negozio potrà essere interessato ad analizzare ad un certo punto le vendite, le caratteristiche delle vendite, per decidere come, ad esempio, rifornire il negozio in modo da avere sempre i prodotti disponibili. Questo ci porta spesso a creare degli altri strumenti di supporto che vengono rappresentati qui con il nome di Data Mart che sostanzialmente danno agli utilizzatori un sottoinsieme delle Data Warehouse, focalizzato solo su alcune tipologie di fatti. Quindi io potrò, per utilizzatori diversi, creare Data Mart diversi che ovviamente avranno una dimensione più contenuta e quindi potranno essere anche utilizzati, ad esempio, su strumenti che hanno direttamente i manager a loro disposizione per poter fare le proprie attività di analisi. Quindi vediamo che il nostro obiettivo è caricare i nostri dati su Data Warehouse e poi da questi saranno associati i vari strumenti, noi ci concentreremo soprattutto su questo, sul fatto che vogliamo caricare i dati, quindi su queste operazioni ETL e poi vorremmo andare ad analizzare che cosa vorrà dire dal punto di vista dell'ottimizzazione dei dati, dare un supporto a query di tipo multidimensionale e quindi quali saranno le proposte tecnologiche disponibili per poter fare questo tipo di interrogazioni. Vediamo un po' più in dettaglio il discorso dell'ETL. Abbiamo detto Extraction. Il problema principale che qui avremo è quando fare l'estrazione. Normalmente l'estrazione

viene fatta in certi momenti, attingo alle fonti, ovviamente posso anche pensare di alimentare in modo continuativo il nostro Data Warehouse, ma una cosa che andrà decisa è se voglio fare un'estrazione di tipo statico, cioè prendere i dati dalle fonti in un certo momento oppure di tipo incrementale. Quindi alimento, ad esempio periodicamente, il nostro Data Warehouse oppure posso anche considerare un'alimentazione in streaming che però è meno tipica per questo tipo di sistemi, perché si suppone che l'utilizzatore stia analizzando i dati su un certo periodo, quindi quello che interessa in genere non è un periodo, un' analisi continua, ma è focalizzato sull'analisi di periodi e quindi, quando si parla di alimentazione incrementale, di solito si intende incrementale, però basata su dei periodi di aggiornamento che vengono considerati man mano. Passiamo all'altra fase che è quella di trasformazione. Abbiamo detto, i dati arrivano da diverse fonti strutturate in modo diverso con diverse caratteristiche. Ci potranno essere vari problemi che vengono dal fatto che io vado a prendere i dati da fonti diverse che richiederanno di lavorare i dati, prima di poter ricaricare nelle Data Warehouse. Ad esempio, ci potrebbero essere problemi relativi alla qualità dei dati. Quindi un tipo di operazione che viene fatta è quella di pulizia dei dati o cleaning per ottenere dati di qualità. Ad esempio, per vari motivi, le fonti potrebbero contenere dei dati non corretti, evidentemente non corretti, perché, ad esempio, avrò definito l'età delle persone che potrà andare, all'interno delle Data Warehouse, supponiamo dai 100 ai 120 anni. Questo può essere un'informazione utile, perché, se io trovo un valore inserito in modo non corretto, ad esempio, un valore negativo oppure un valore mille, chiaramente sono in presenza di un dato che non è accettabile per il nostro sistema. Quindi supponiamo che ci siano dei possibili errori all'interno dei dati oppure un altro tipico problema, non stiamo considerando dati storici, che si verifica in questi casi è il dato mancante, ad esempio, sto registrando i dati per una serie di giorni, però per un certo giorno per qualche motivo non ho registrato quell'informazione, quindi dovrò decidere cosa fare per gestire correttamente e poi successivamente le query su quei dati. Quindi dovrò andare ad esempio a completare una serie temporale con dei dati mancanti o cercare altre fonti, ovviamente le operazioni che si dovranno fare dipenderanno dal problema riscontrato. Quello che voglio fare inizialmente è esaminare i dati per esaminare la loro qualità e scartare quelli che sono, o correggere, quelli che sono evidentemente sbagliati. Un altro problema che si può verificare è che da diverse fonti mi arrivino dati con informazioni diverse. Ad esempio l'indirizzo di una persona: se io trovo due indirizzi associati alla stessa persona, dovrò capire qual è l'indirizzo corretto e assumere come dato corretto solo un indirizzo oppure un'altra possibilità in questo caso è verificare che effettivamente una persona può avere associati due indirizzi con caratteristiche diverse. Quindi quello che si deve fare è quella che viene chiamata la riconciliazione. Ovviamente, all'interno di questo problema ci sono anche problemi legati al fatto che voglio riconoscere che effettivamente una certa persona 'Mario Rossi' sia effettivamente quel 'Mario Rossi' e non un altro 'Mario Rossi', quindi ci sono problemi di identificazione degli oggetti che sto analizzando che possono essere anche piuttosto complessi. Altro problema che abbiamo detto è guardare proprio come i dati sono scritti nelle basi di dati di provenienza. Noi vogliamo ricondurre questi dati a un formato che è quello della base di dati del Data Warehouse, quindi faremo delle operazioni di standardizzazione e queste potranno essere di vario tipo. Vediamo le tipologie principali: noi abbiamo un certo dato, può essere rappresentato in una base di dati con più campi, ad esempio supponiamo di partire con tabelle relazionali, abbiamo, supponiamo, un indirizzo, l'indirizzo può essere rappresentato come via, numero civico, città, in altri casi io potrei trovare un'unica stringa all'interno della base di dati che contiene un campo indirizzo. Standardizzazione vuol dire che io deciderò quale sarà il formato dell'indirizzo e quindi inserirò nel mio Data Warehouse gli indirizzi secondo un formato che è quello deciso per i Data Warehouse. Quindi abbiamo un discorso di struttura dei dati nel Data Warehouse. Un altro discorso riguarda i valori. Tipicamente cercherò di identificare i miei dati all'interno del sistema, attraverso dei codici, quindi i codici che standardizzano questo dato. Ad esempio, possiamo considerare dei codici per i prodotti oppure dei codici per delle categorie merceologiche che sono state definite in modo standard, qualche entità di standardizzazione, o definire dei propri codici interni. Altro aspetto riguarda il formato dei dati. Esempi tipici sono il fatto che i numeri decimali, ad esempio, in italiano si rappresentano con la virgola, nei Paesi anglosassoni col punto, il formato delle date, le date possono essere scritte in tanti modi, in formato

giorno-mese-anno oppure, sempre nello stesso formato l'anno potrebbe avere quattro cifre, oppure in una notazione anglosassone prima il mese poi il giorno e poi successivamente l'anno. Devo decidere un formato delle date all'interno del mio sistema perché ovviamente la data sarà sempre un elemento importante nell'analisi, ma dovrò riconoscere le date e rappresentare tutte nello stesso modo. Altro aspetto da considerare nella trasformazione è la ricerca dei duplicati. L'obiettivo è quello di eliminarli, ad esempio, ritornando al caso di 'Mario Rossi', se io ho informazioni su 'Mario Rossi' e scopro che è la stessa persona e non voglio caricare due volte una persona 'Mario Rossi', ma voglio caricarlo una volta sola. Quindi questo vuol dire che devo riuscire a identificare gli oggetti e poter eliminare quelli che vengono riconosciuti come pubblicati per non avere due volte informazione su una stessa persona il che ovviamente causerebbe dei problemi nel momento in cui vado ad analizzare questi dati. Una volta fatta questa operazione di estrazione prima e di trasformazione, i dati sono pronti per essere caricati nel Data Warehouse perché avranno una struttura, una qualità adatte poi per supportare operazioni di tipo di analisi. Nel parlare di operazioni di trasformazione, abbiamo accennato più volte al fatto che facciamo riferimento a informazioni che abbiamo sui dati stessi. Ad esempio qua nel cleaning, avevamo detto che possiamo avere un range predefinito per l'età oppure abbiamo detto che il formato dei dati sarà rappresentato in un certo modo, ad esempio, utilizzo la notazione con il punto per le cifre. Quindi avremo definito come rappresentare i nostri dati. Tutto questo in un Data Warehouse viene rappresentato all'interno di altri dati che sono dati sui dati e vengono chiamati quindi metadati. I metadati ci diranno quale è la struttura del database, ci diranno quali sono state le operazioni di trasformazione, quindi se ho modificato il dato in una delle operazioni precedenti, ne terrò memoria per poter eventualmente andare ad analizzare aspetti che si possono verificare di tipo critico, poi abbiamo detto che avrò delle regole di trasformazione, ad esempio, le date: se trovo una data in un altro formato, dirò come trasformarlo nel formato che è stato deciso. Ultimo aspetto, terrò traccia delle operazioni di analisi, quindi operazioni sui database, per poter fare le statistiche sull'utilizzo del Data Warehouse.