Verso una data discovery senza nessun attrito

IBM Fluid Query semplifica il processo

Agosto 2015

White Paper a cura di

Dr. Barry Devlin, 9sight Consulting
barry@9sight.com

Nel mondo di oggi, altamente distribuito, multi-piattaforma, è sempre più probabile che i dati necessari per risolvere qualsiasi esigenza specifica legata ad un processo decisionale risiedano in un'ampia varietà di fonti. Di conseguenza, gli approcci manuali tradizionali, che richiedono prima l'acquisizione, lo storage e l'integrazione di vasti set di dati, nell'ambiente di esplorazione preferito dell'analyst stanno diventando inappropriati. La data virtualization, che offre un accesso trasparente a diverse fonti di dati distribuite, rappresenta un approccio alternativo valido in queste circostanze.

Il presente white paper descrive i problemi associati alla multi-source data discovery dalla prospettiva di un business analyst tipico, mettendo in evidenza le difficoltà incontrate quando si utilizzano prima i metodi manuali tradizionali di integrazione dei dati. L'architettura moderna, di alto livello, presentata - la piattaforma di informazioni integrata - fornisce un inquadramento e spiega la data virtualization. Successivamente, viene esaminato il valore di questa tecnologia nel trarre insights di business dai dati.

Infine, esploreremo una nuova offerta, IBM Fluid Query, e l'insieme di funzionalità di data virtualization che offre attraverso IBM PureData Sistema or Analitico, IBM BigInsights e altri prodotti di data management di IBM. Questa offerta iniziale offre funzioni efficaci e mostra l'orientamento di IBM in questo settore emergente della funzione di data management e della creazione del valore.

Sommario

Sponsorizzato da:
International Business Machines

www.ibm.com

Il Data scientist è il lavoro più affascinante[1]…", borbottò Raymond mentre procedeva faticosamente all'ennesima estrazione dei dati dei clienti aggiornati e anonimi dal data warehouse per il trasferimento nel suo ambiente di sandbox per l'analisi in combinazione con i dati più recenti provenienti dai social media. "Veramente? Ci si sente più come un 'idraulico dei dati'." Cinque minuti più tardi, spuntò sullo schermo il codice di errore SQL relativo al limite di righe superato. Raymond aggiunse imprecando: "O meglio l'ingegnere delle strutture igienico-sanitarie", ricordando la scritta brillante che che aveva letto quella mattina sul camion 'Drain Brain'. Inoltre, copiare i dati da Hadoop era altrettanto difficile; e i messaggi di errore erano ancora più astrusi. "Perché deve essere tutto così difficile?"

Sempre più spesso, non è così. Ma il superamento delle difficoltà incontrate da Raymond, e di molte altre ancora, impone il ripensamento dei vecchi paradigmi. In un mondo in cui i volumi di dati sono enormi, spostare i dati verso la query, come Raymond ha fatto per anni, ha meno senso. Molte volte abbiamo bisogno di spostare la query verso i dati. In un mondo in cui la varietà di dati aumenta sempre di più, la concezione del data warehouse pensato come un unico store fisico di tutte le informazioni di business sta diventando insostenibile... ... al pari di qualsiasi strategia basata sul concetto che una e una sola piattaforma possa memorizzare e gestire tutti i dati. Ora abbiamo bisogno di prevedere un'architettura con più piattaforme (o pilastri, come vedremo più avanti). E dobbiamo disporre dei mezzi per nascondere completamente l'esistenza di queste sedi multiple all'utente o dobbiamo combinare i risultati dei dati provenienti da più fonti, senza soluzione di continuità. Inoltre, abbiamo bisogno di un'infrastruttura che gestisca i dati e crei un contesto di informazioni attraverso queste piattaforme, consentendo la compilazione - temporaneamente o permanentemente - di duplicati di dati parziali in background, come richiesto.

Prenderemo in considerazione questo cambiamento dal punto di vista dell'architettura. Ma, in primo luogo, cerchiamo di capire che cosa sta cercando di fare Raymond alla BigLocal Underwriting Experts, o BLUE in breve.

Il mondo della multi-source data-discovery

Le informazioni sono state a lungo al centro degli interessi delle compagnie di assicurazione che sono state tra i primi ad adottare il data warehouse per acquisire, gestire e utilizzare i dati di produzione per il reporting, l'analisi dei rischi, la previsione e molto altro ancora. I dati di BLUE provenivano da più fonti e dovevano essere sottoposti al cleansing e consolidati prima dell'uso. A volte la ricezione era un po' lenta. Ma almeno aveva origine dai processi aziendali interni, pertanto si trattava dei cosiddetti dati mediati dai processi o *process-mediated data*[*], come li definisco. Quando ha iniziato ad analizzare i rischi per la prima volta (quando i data scientists non esistevano neanche nei sogni), Raymond lavorava esclusivamente con questo tipo di dati, per lo più nel data warehouse o in data mart specializzati e spesso nei fogli di calcolo. E' diventato un esperto dell'integrazione dei dati *ad hoc* - preparazione, cleansing e copia dei dati necessari nei suoi strumenti

[*] Le descrizioni di questo termine, così come delle informazioni prodotte dall'uomo, *human-sourced information* e dei dati generati dalle macchine, *machine-generated data* sono riportate nel mio white paper del 2012 per IBM *"The Big Data Zoo—Taming the Beasts"*, http://bit.ly/Big_Data_Zoo

preferiti. I suoi stretti rapporti con gli owners e i produttori di molti sistemi operativi di BLUE gli hanno permesso di capire e ottenere l'accesso - relativamente - facilmente ai dati di cui aveva bisogno.

Oggi, le fonti di dati di Raymond sono di gran lunga più ampie e più ricche di quanto ci si aspettasse. Le *human-sourced information* di tipo comportamentale e relazionale provengono in volumi elevati e in continua espansione dai social media e dalle fonti di data aggregator. E' si stanno già accumulando nel nuovo ambiente Hadoop. Naturalmente, l'analisi di questi dati esterni offre, già di per sè, informazioni preziose sul social behavior e sui trend economici. BLUE ha già utilizzato questi dati per rinnovare il processo di valutazione dei rischi per i prospects più giovani. L'azienda ha tratto un alto valore da questo processo, anche se, secondo Raymond, gli strumenti che doveva usare potevano essere un po' più *user-friendly*. Inoltre, emerge già chiaramente che i vantaggi concreti arriveranno quando i dati esterni verranno utilizzati insieme ai *process-mediated data* interni. Raymond ha iniziato i test iniziali per verificare come i dati dei clienti possono essere uniti, in modo sicuro e protetto, alle *human-sourced information* in ingresso. Ha trasferito i dati in entrambe le direzioni tra il warehouse e Hadoop. Ma, già in questa fase, incontra delle difficoltà. I suoi vecchi metodi manuali di gestione delle copie multiple che sta creando non sono sufficienti a gestire la sfida insita nel combinare grandi volumi di dati esterni, ad alta velocità, ma di bassa qualità con i dati interni, ben governati, ma altamente sensibili. Inoltre, per quanto riguarda la comprensione del significato dei nuovi dati ... beh, questa è un'altra storia ancora.

Nell'analytics, il vantaggio reale proviene dall'utilizzo combinato dei process-mediated data esterni e interni.

Per i prossimi anni, Raymond già prevede un'ampia varietà di *machine-generated data* in ingresso, provenienti dall'Internet delle cose, mano a mano che il business delle assicurazioni automobilistiche si reinventa. Quel poco che ha visto della velocità e della portata previste di questi dati sembra estremamente scoraggiante. Ha sentito che potrebbe essere necessaria ancora un'altra piattaforma tecnologica per memorizzare e gestire questi volumi di dati.

Per Raymond - e sicuramente per la maggior parte dei data scientists e degli esperti di dati che operano nel mondo dell'analytics - sta diventando sempre più chiaro che i modi, comprovati e affidabili, per scoprire gli insights da questo insieme in continua espansione di fonti di dati, tipi di dati e dal crescente volume dei dati stessi, sono destinati a divenire inadeguati. La data discovery di questa portata non può essere un prodotto di artigianato. Per cogliere i ricchi frutti promessi dall'analytics su più fonti di dati, si impone la necessità di sviluppare un'architettura di nuova concezione e soluzioni tecnologiche innovative.

Oltre i confini del data warehousing classico - una piattaforma di informazioni integrata

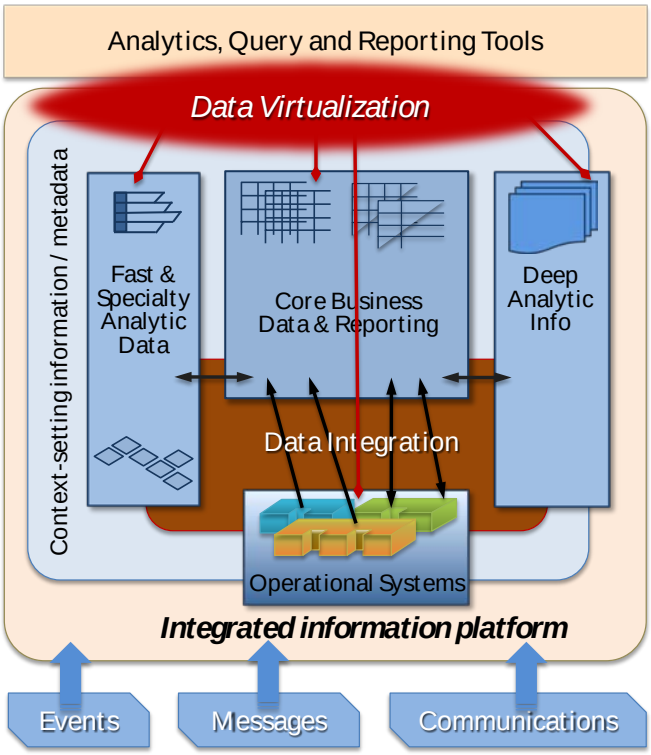
L'Architettura del data warehouse originale[2] risalente alla metà degli anni '80, è stata, in gran parte, determinata dalle esigenze di gestione e controllo degli utenti per una vista coerente e accurata del business, inclusi il reporting delle performance, le prestazioni e l'esame dei problemi attraverso le query e l'analisi. I vincoli tecnologici dell'epoca, che si sono protratti per molti anni, hanno imposto un'architettura in gran parte centralizzata e a più livelli. Tutti i dati richiesti vengono incanalati da più sistemi operativi attraverso un Enterprise Data Warehouse (EDW) per il cleansing e l'integrazione e, nella maggior parte dei casi, vengono consegnati ai data marts ottimizzati per le esigenze di query e reporting dei business users. Questo approccio è rimasto un punto fermo nella comunità di business intelligence (BI) per decenni, nonostante i cambiamenti in corso nel mondo del business orientati verso la tempestività e l'ampiezza delle informazioni piuttosto che verso la coerenza.

Tuttavia, negli ultimi anni si è assistito ad un'enorme esplosione dei cosiddetti big data provenienti dai social media e, ora, dall'Internet delle cose. Questo fenomeno ha indotto cambiamenti radicali nella percezione del business, in merito a come è possibile trarre valore dalle informazioni attraverso la data discovery e gli approcci analitici. Se a tutto questo si sommano alcuni impressionanti progressi compiuti sul fronte delle capabilities tecnologiche - networking, hardware e software - è evidente che i presupposti dell'architettura tradizionale del data warehouse devono essere riesaminati. Negli ultimi cinque anni, o giù di lì, abbiamo assistito a questo trend, in modo evidente, con la promozione di concetti quali i logical data warehouses, i data lakes e i data reservoires. Purtroppo, alcuni di questi termini sono più concetti di marketing che architetture ben definite e non riescono a prendere in considerazione pienamente tutte le esigenze di business emergenti, le possibilità tecnologiche o gli aspetti vitali come la qualità delle informazioni, il contesto, il significato e il pensiero emergente su come vengono effettivamente prese le decisioni dalle persone nelle organizzazioni.

L'avvento dei big data, insieme con i progressi impressionanti compiuti sul fronte della tecnologia, impongono il riesame dell'architettura del data warehouse tradizionale

L'esplorazione completa di tutte queste considerazioni, dalla prospettiva dell'architettura, induce ad una transizione dall'approccio della piattaforma unica per i dati, a più livelli, del data warehouse al nuovo approccio per i dati a pilastri, multi-piattaforma, descritto in modo approfondito nel mio recente libro *"Business unIntelligence"* [3]. Questo cambiamento di prospettiva, apparentemente semplice ma tecnicamente complesso, è illustrato nella Figura 1, che mostra tre colonne di dati con diverse esigenze di gestione ed elaborazione, supportati dalle informazioni che definiscono il contesto, *context-setting information*, (o metadati) che attraversano i pilastri. A seconda delle circostanze aziendali e tecniche, possono esserci più di tre pilastri; mentre è improbabile che ve ne siano di meno.

Il primo pilastro centrale di dati di business fondamentali e di reporting è rappresentato dai dati coerenti, di qualità garantita presenti nell'EDW e nei data marts. In termini di tecnologia, questo pilastro è idealmente basato sulla tecnologia del database relazionale. A seconda delle esigenze di business per una maggiore performance delle query, si possono applicare i database in-memory, Massively Parallel Processing (MPP), colonnari o altre tecnologie specializzate. Gli esempi di sistemi includono IBM DB2 with BLU Acceleration, IBM PureData Sistema or Analitico, dashDB e Spark. Le informazioni analitiche approfondite richiedono un'elaborazione estremamente flessibile, su vasta scala, come la predictive analytics e il text mining spesso eseguiti in ambiente Hadoop o IBM BigInsights. I dati analitici veloci richiedono un'elaborazione analitica rapida che deve essere eseguita "in corso d'opera", come avviene, ad esempio, con IBM InfoSphere Streams. Nel punto di intersezione di velocità e flessibilità, si possono trovare i dati analitici speciali, che utilizzano un'elaborazione specifica come NoSQL, XML, database a grafo e altri database e data store.



Questo approccio a pilastri risponde alle esigenze di business moderne per la tempestività dei dati e per l'accesso a una vasta gamma di tipi di dati e di informazioni. Tuttavia, presenta anche sfide tecnologiche nel garantire la coerenza dei dati tra i pilastri, così come il modo in cui gli utenti possono accedere senza problemi ai dati distribuiti tra i pilastri. Concentriamo ora la nostra attenzione sui componenti di integrazione e virtualizzazione dei dati.

Figura 1: *L'architettura moderna a "pilastri"*

Dai dati, alle informazioni, agli insights

Con i dati, distribuiti ora su diverse piattaforme a seconda del volume, del tipo e della performance dei dati, gli utenti si trovano ad affrontare il problema di come accedere fisicamente ai dati, come combinarli in una varietà di artefatti di informazioni e di come ricavare insights di business utili e preziosi da questi dati. Queste considerazioni sono al centro di data integration e data virtualization. Queste tecnologie, naturalmente, sono esistite in varie forme e livelli di maturità fin dall'esordio del data warehousing. Tuttavia, la piattaforma di informazioni integrata richiede un livello molto maggiore di sofisticazione in entrambe le aree.

Data virtualization — offrire un unico punto di accesso

Il presupposto della data virtualization prevede che i business users non devono mai sapere dove o in quale forma si trovano i dati sottostanti. In poche parole, il compito della data virtualization è nascondere questa complessità. La funzione richiesta può essere caratterizzata da tre livelli di aumento del valore di business:

1. Funzionalità che consente di instradare una query, completamente o in parte, da una piattaforma all'altra, in una forma utilizzabile dalla seconda piattaforma, per ricevere indietro dei risultati, combinabili con i risultati locali, se necessario

2. Possibilità di utilizzare i metadati tecnici su più piattaforme per ottenere e ottimizzare la performance delle query in un ambiente multi-piattaforma attraverso le viste

3. Fornire il collegamento da una visione o modello di informazioni di business logico alla vista tecnica attraverso i metadati che descrivono dove e come risiedono i dati sulle diverse piattaforme

La data virtualization si è affermata come una tecnologia accettata e in fase di maturazione, di vitale importanza nel mondo multi-

La data virtualization è una tecnologia in fase di maturazione, proveniente da due direzioni. In un approccio, i vendor specializzati di data virtualization partono dalla funzionalità del modello di business o tecnico e costruiscono una piattaforma che si estende e poggia su singoli database e data store, massimizzando il numero e la varietà di database e store accessibili. Questo approccio offre una buona generalità di espressione ed evita il lock-in a un vendor di database. Nell'altro approccio, i vendor di database creano la funzione di instradazione ed esecuzione delle query su più piattaforme all'interno del proprio ambiente, di solito concentrato e ottimizzato per un sottoinsieme di piattaforme di interesse. Questo approccio riutilizza gran parte della funzione di database dei vendor ed elimina la necessità di un livello di server aggiuntivo e la relativa struttura di elaborazione delle query. E' particolarmente interessante quando la maggior parte delle query attraversa questa piattaforma.

Data Integration - costruire una piattaforma di informazioni coesa

Nell'ambiente del data warehouse classico, l'obiettivo primario di integrazione dei dati è incentrato sull'acquisizione, preparazione e caricamento dei dati provenienti da diverse fonti nell'EDW e da lì nei data marts. L'ETL (Extract, Transform and Load) e un'ampia gamma di tool specializzati che operano dal real-time al batch, sono considerati il focus primario della data integration. Tuttavia, come nella virtualizzazione, possiamo distinguere tre livelli di aumento del valore di business:

1. Trasferimento dei dati source-target specifici, compresi i diversi source e target, con la gestione di task di cleansing e combinazione localizzate

2. Piattaforme di data integration che offrono un ambiente ETL condiviso, inclusi metadati ETL comuni, interfaccia utente, ecc., spesso in esecuzione su server dedicati o su Hadoop

3. Ambienti di sviluppo e gestione completamente automatizzati che creano un collegamento delle esigenze degli utenti e dei modelli di dati fino alla funzione ETL e all'uso

La data integration è una tecnologia matura per molti aspetti, con diversi vendor che offrono una vasta gamma di soluzioni a tutti e tre i livelli.

Unire virtualizzazione e integrazione negli insights - il contesto è la chiave

Facendo ricorso al pensiero laterale, si nota che la virtualizzazione e l'integrazione dei dati sono due facce della stessa medaglia. Entrambe mirano a fornire i dati provenienti da fonti remote e spesso multiple all'utente che ha bisogno di eseguire delle query o comunque di utilizzarli. Nell'integrazione dei dati, secondo la modalità tradizionale, tutti i dati che possono essere richiesti vengono copiati in una destinazione comune, in previsione e in anticipo rispetto alle esigenze degli utenti. Originariamente veniva eseguita tramite trasferimenti in bulk, mentre attualmente il trickle feeding sta diventando sempre più comune. La data virtualization, invece, attende che l'utente richiede dei dati e quindi tenta di trasferire la quantità minima necessaria nel corso della query. La virtualizzazione, pertanto, è più limitata dalle esigenze immediate degli utenti di tempestività dei risultati, anche se l'integrazione è sempre più soggetta agli stessi vincoli, mano a mano che la maggior parte delle imprese sta passando alle operazioni in tempo reale. E' possibile riscontrare una simmetria funzionale. Ciò implica la necessità di un'infrastruttura comune e condivisa al fine di garantire che gli utenti ricevano risultati compatibili a prescindere da quale tecnologia viene utilizzata per una particolare esigenza di business, in un ambiente nel quale coesistono entrambe le tecnologie.

Un aspetto di particolare importanza è che i metadati sottostanti sono condivisi da entrambi i set di tool. Questi metadati riguardano entrambi gli aspetti di business e tecnici dei dati memorizzati e le trasformazioni subite. Alla loro massima estensione, descrivono l'intero contesto dei dati, inclusa la descrizione, il significato, il lineage, la qualità, i valori validi e gli utilizzi, ecc., che ne consentono l'uso come vere informazioni di business. In realtà, i metadati sono troppo limitati, in termini di percezione, per coprire tutte queste esigenze. Questo materiale costituisce, in realtà, una componente chiave delle informazioni di business stesse e, di conseguenza, preferisco utilizzare l'espressione *context-setting information (CSI)*. Solo questo contesto completo consente agli utenti di ottenere insights di business reali. E solo quando le CSI vengono pienamente condivise tra l'integrazione e la virtualizzazione, questi insights possono essere garantiti in termini di validità e coerenza.

Le context-setting information (CSI) forniscono le basi per un ambiente di virtualizzazione/integrazione consolidato che offre ai business users insights reali.

La piattaforma di informazioni integrata illustrata nella Figura 1 richiede un insieme di tool sofisticati di virtualization, integration e context-setting information, così come livelli significativi di integrazione tra questi aspetti funzionali. Possiamo già sentire Raymond rispondere, nel suo stile inimitabile: "Un'architettura bellissima, certo, ma io ho bisogno di un aiuto concreto! E ne ho bisogno ora ... "

IBM Fluid Query — virtualizzazione e integrazione moderne

A cominciare dall'annuncio di IBM Fluid Query, nel marzo 2015 e con un aggiornamento a luglio con la versione 1.5[†], IBM ha iniziato a rispondere alle richiesta di aiuto di Raymond. Fluid Query offre due punti di vista distinti dai quali partire per affrontare il mondo multi-piattaforma. Il primo punto di vista è quello dei power users classici, come Raymond, il cui background e competenze li portano ad iniziare il lavoro sui dati dall'ambiente relazionale - in questo caso, IBM PureData Sistema or Analitico - e a collegarlo ad altre fonti di dati. Il secondo punto di vista è quello dei più recenti data scientists che lavorano principalmente dall'ambiente Hadoop - in questo caso IBM BigInsights - per poi analizzare il data warehouse relazionale e altre piattaforme. Esaminiamo questi due approcci separatamente, in quanto si rivolgono a diversi casi d'uso e offrono funzionalità leggermente diverse. Ma prima, diamo un'occhiata ad alcuni casi d'uso tipici.

Caso d'uso di business per la virtualizzazione e l'integrazione moderne

1. Il business alla velocità della luce: Il business analyst, nell'ambiente di data warehouse, sta esaminando la tendenza al ribasso delle vendite. I dati presenti nel warehouse (in questo caso) sono un'istantanea delle attività a fine giornata precedente, che corrisponde alla vista che la maggior parte degli utenti preferisce o che forse è dettata dal tempo necessario per consolidare ed eseguire il cleansing dei dati. L'analisi (utilizzando i dati del giorno precedente) porta ad una determinazione del problema e vengono adottati provvedimenti correttivi. Idealmente, l'analyst vorrebbe, quindi, vedere immediatamente i risultati dell'azione. Interrogando i dati in tempo reale dall'ambiente operazionale e combinando il risultato con i dati del data warehouse standard di fine giornata, l'analyst può vedere il trend delle vendite aggiornato al minuto in caso di necessità, senza richiedere l'aggiunta di capabilities di carico in tempo reale con un aumento dei costi per il data warehouse.

2. Collegare le isole di dati relazionali: La maggior parte delle imprese esegue più database relazionali, a volte a causa dei requisiti applicativi, altre volte come conseguenza degli sviluppi storici. Ma questo non implica che i dati contenuti in un sistema siano inaccessibili da un altro. La data virtualization offre la possibilità di collegare un altro sistema di database relazionale per accedere e combinare i dati richiesti. Inoltre, inserendo dati più remoti nell'ambiente PDA (BigInsights o DB2), è possibile collegare più RDBMS nella stessa query, a condizione che i possibili impatti sulle performance siano compresi.

3. Creazione di un contesto approfondito: I numeri tratti dal data warehouse, da soli, non sempre raccontano la storia completa. Consideriamo, ad esempio, la seguente query: "Quali sono i nostri prodotti più venduti che ottengono recensioni buone o le migliori recensioni?" Questa query sembra ragionevole per un business user, ma contiene sia *process-mediated data* che *human-sourced information*: i dati di *top selling* derivano dal sorting dei volumi di vendita nel data warehouse o mart, mentre le *recensioni buone o le migliori recensioni* richiedono l'analisi dei social media e di altri dati contenuti in Hadoop. In molti casi, dal customer relationship management al supporto del call center, le informazioni testuali (e talvolta in forma di immagini) creano un contesto molto più approfondito intorno ai numeri del business. Questo tipo di analisi permette alle query SQL, che hanno avuto origine nell'ambiente relazionale, di accedere ai big data contenuti negli stores di Hadoop, spingendo

[†] Le funzioni descritte in questo white paper sono disponibili in IBM Fluid Query versione 1.5.

l'elaborazione verso il basso, in MapReduce e sfruttando la potenza delle applicazioni che vi risiedono, come il pattern recognition, la predictive analytics, ecc.

4. **Attività agili:** Chi ha tempo per realizzare un nuovo data mart ogni volta che le esigenze di business si espandono o cambiano? A volte basta un dettaglio extra o due su un data mart esistente. O forse si vuole semplicemente verificare se l'aggiunta di una nuova colonna può essere utile. La capacità di unire i dati da un mart ai dati che risiedono in un altro in una semplice query può fornire risposte rapide, sia per scoprire se il risultato è utile che per soddisfare rapidamente una nuova esigenza, senza richiedere la creazione di un nuovo mart da parte di un progetto IT. L'agilità per reagire alle nuove esigenze o ai cambiamenti del mercato è di vitale importanza nel mondo del business di oggi.

Le esigenze di business, quali la tempestività, la contestualizzazione, l'agilità e il supporto storico sono alla base dell'adozione delle funzionalità di

5. **Supporto storico:** Nell'era dei big data, spostare periodicamente i vecchi dati, utilizzati meno di frequente in un ambiente più economico è conveniente. Ma il solo fatto che vengono utilizzati meno comunemente, non significa che l'azienda sarà felice se ci vogliono ore per recuperare questi dati, se è necessario. Lo spostamento regolare dei dati più vecchi in Hadoop, pur consentendo l'accesso attraverso una query virtualizzata dal data warehouse, offre il meglio dei due mondi - costi di storage inferiori con accesso ai dati da parte degli utenti, senza soluzione di continuità, quando necessario.

Punto di vista 1: operatività e supervisione

Non esiste praticamente nessuna azienda al mondo che non utilizzi un data warehouse per supervisionare le operations con tool di reporting e query. In realtà, molte organizzazioni ne utilizzano più di uno - nonostante il prezioso consiglio che io stesso ho espresso insieme ad altri esperti. Ne deriva che molte esigenze di carattere decisionale richiedono dati provenienti da più fonti. Inoltre, molti data warehouse tradizionali contengono dati che costituiscono un'istantanea di un punto nel tempo, corrispondente, di norma, alla fine della giornata precedente. Che cosa succede, dunque, quando un utente di un data mart ha bisogno di dati che risiedono solo in un altro warehouse o mart, o quando sono necessari dati operazionali più tempestivi per rispondere a una richiesta di business?

Fino ad ora sono state disponibili due soluzioni: (1) scaricare i dati da molteplici fonti e procedere in Excel, o (2) rivolgersi all'IT e chiedere loro di costruire un nuovo mart. Come molti business analysts, Raymond ha preferito il primo approccio, avendo sperimentato qualche dolorosa vicenda a causa dei ritardi di consegna dell'IT. Con i nuovi dati che appaiono nei sistemi Hadoop e NoSQL, la sfida non fa che aumentare. IBM Fluid Query offre una terza opzione: l'accesso ai dati remoti richiesti direttamente come parte di una query SQL.

Le business operations e la supervisione avvengono in un ambiente relazionale tradizionale che sta diventando sempre più diversificato, richiedendo soluzioni di

Con questa opzione, la visione del mondo dell'utente è centrata sui dati che risiedono nei sistemi di database relazionali di IBM come IBM PureData Sistema or Analitico (PDA), DB2, ecc. Utilizzando la sintassi SQL standard, il campo di applicazione della query può essere esteso per includere dati su altre piattaforme, sia relazionali che non relazionali. Nel mondo relazionale, questo include altre implementazioni PureData Sistema or Analitico, IBM DB2, dashDB, PureData Sistema or Operational Analitico e Oracle. Gli obiettivi non relazionali includono IBM BigInsights, Hortonworks, Cloudera e Spark.

Le piattaforme relazionali e IBM BigInsights possono restituire i risultati delle query inviate e questi risultati vengono uniti nel database PDA originario. Un'ulteriore estensione potrebbe anche consentire la memorizzazione nella cache (o lo snapshotting) dei dati comunemente utilizzati a livello locale per evitare il trasferimento dei dati e di ricreare più volte i data set richiesti. Hortonworks, Cloudera e Spark, d'altro

canto, possono trasferire i set di dati non qualificati e potenzialmente di dimensioni superiori se non sono in grado di elaborare la sintassi SQL appropriata localmente. In questo modo, il sistema PDA centrale esegue la funzione di qualificazione prima del join. Ciò può determinare volumi di trasferimento dei dati superiori attraverso la rete e un carico aggiuntivo sulla macchina di origine. Tuttavia, mano a mano che la funzionalità SQL dei sistemi Hadoop migliora, volumi crescenti di lavoro possono essere spinti in tali sistemi, migliorandone le prestazioni globali.

Punto di vista 2: ottimizzazione e previsioni

Se si ascolta la stampa IT e il coro degli analysts, verrebbe da pensare che l'intero settore è concentrato esclusivamente sulla predictive analytics e su altre forme di analytics per prevedere gli andamenti e i comportamenti futuri dei mercati e per ottimizzare i sistemi operativi di conseguenza. Si potrebbe essere indotti a immaginare che l'intero universo stia passando automaticamente ad Hadoop e alle piattaforme collegate. Questo, naturalmente, è falso. In realtà, non sarà mai vero - i costi associati alla migrazione sarebbero proibitivi e, in ogni caso, tra qualche anno, apparirà probabilmente un altro messia del software. Tuttavia, fonti di dati nuove e significative, principalmente esterne, sono in corso di attuazione su Hadoop. E alcune aziende emergenti, prive di sistemi legacy, stanno sperimentando la conduzione del business in questo ambiente. Innegabilmente, la piattaforma Hadoop costituisce oggi il terreno privilegiato dei data scientists (e degli unicorns). In questi casi, l'analisi e il reporting hanno inizio qui, nell'ambiente Hadoop, e i tool per eseguire tutto questo stanno migliorando e si stanno estendendo ad un ritmo sostenuto.

La predictive analytics e altre forme di analytics sono eseguite al meglio sui dati tradizionali e sui big data combinati, richiedendo soluzioni di virtualizzazione e

La stragrande maggioranza delle imprese continua ad operare e a controllare le proprie operations in un ambiente relazionale tradizionale, mentre i data scientists concentrano il proprio lavoro nell'ambiente Hadoop. Tuttavia, anche loro hanno sicuramente bisogno di combinare i dati provenienti dai sistemi relazionali esistenti con i risultati delle query nell'ambiente dei big data. Così, un data scientist che lavora in IBM BigInsights può accedere ai dati in tabelle contenute nel sistema PureData Sistema or Analitico (o DB2, PDOA, dashDB, ecc.) direttamente dalla analisi eseguita in Hadoop. Con BigInsights, Fluid Query offre la possibilità di restituire sottoinsiemi di dati qualificati dalla query. Per un data scientist che lavora su Hortonworks, Cloudera e Spark, può essere necessaria un'elaborazione intermedia per qualificare ulteriormente i risultati della query. In entrambi i casi, il valore deriva dalla possibilità di collegare informazioni di produzione, quali numeri di clienti o informazioni convalidate del data warehouse, per esempio, i dati esterni, dati di qualità inferiore, come i nomi utente di Twitter.

In entrambi i casi, IBM Fluid Query è in grado di trasferire data sets completi da una piattaforma all'altra, consentendo la creazione e la gestione di data store storici su Hadoop o di data caches sulla piattaforma PureData Sistema or Analitico. E, dato il progresso registrato sul fronte di data sources/targets e delle funzionalità nel breve periodo di tempo intercorso tra la release iniziale e la versione 1.5, è ragionevole aspettarsi che IBM continuerà ad espandere entrambi gli aspetti in futuro.

Conclusioni

Anche se si trova ancora nelle prime fasi di evoluzione, IBM Fluid Query sta determinando la direzione che sta già iniziando a soddisfare le esigenze di Raymond per un modo migliore di lavorare con i dati provenienti da più fonti, sia i dati interni tradizionali che la crescente ricchezza di dati esterni. Fluid Query è, in questo senso, sia un programma di funzionalità necessarie per supportare una piattaforma di informazioni integrata che un insieme di prodotti che forniscono tale funzione in fasi.

Risponde non solo alle esigenze dei business analysts più tradizionali come Raymond, ma anche della categoria emergente di data scientists che esegue gran parte del proprio lavoro sulle piattaforme Hadoop.

Il supporto al processo decisionale, oggi, è in fase di transizione da un ambiente in cui tutti i dati venivano immessi e utilizzati in un data warehouse relazionale ad un mondo approccio, multi-piattaforma, a volte chiamato data warehouse logico. In questo mondo moderno, i volumi di dati e i punti di forza/debolezza della tecnologia indicano che la funzione deve essere sempre di più spostata verso i dati, ovunque si trovino, e non viceversa. Tale virtualizzazione e integrazione dei dati assumerà un'importanza e una popolarità sempre maggiore, al passo con l'aumento della complessità e delle dimensioni dell'ambiente dei dati. IBM Fluid Query rappresenta un punto di partenza in questo percorso e un punto fermo verso una piattaforma di informazioni integrata che orienta gli insights approfonditi e preziosi provenienti da tutti i dati gestiti e acquisiti dal business.

Mano a mano che ci spostiamo verso un mondo ricco di dati, multi-piattaforma, IBM Fluid Query offre un buon punto di partenza e la direzione utile nell'adozione della virtualizzazione e

Il Dr. Barry Devlin è annoverato tra le massime autorità nel business insight ed è uno dei fondatori del data warehousing, con la pubblicazione del primo architectural paper sul tema nel 1988. Con oltre 30 anni di esperienza nel settore IT, tra cui 20 anni con IBM come Distinguished Engineer, è un prestigioso analyst, consulente, docente e autore dell'autorevole libro,"Data Warehouse—from Architecture to Implementation" e di numerosi white paper. Il suo nuovo libro, ***"Business unIntelligence—Insight and Innovation Beyond Analitico and Big Data"*** *(http://bit.ly/BunI-Technics) è stato pubblicato nell'ottobre 2013.*

Barry è fondatore e titolare di 9sight Consulting. E' specializzato nelle implicazioni legate a risorse umane, aziendali e IT delle soluzioni di business insight approfondite che combinano ambienti operazionali, informativi e di collaborazione. Un tweeter regolare, @BarryDevlin, scrittore, blogger e collaboratore allo sviluppo del settore delle informazioni, Barry risiede a Città del Capo, Sud Africa e opera in tutto il mondo.

[1] Davenport, T.H. and Patil D.J., *"Data Scientist: The Sexiest Job of the 21st Century"*, Harvard Business Review, (Ottobre 2012), https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

[2] Devlin, B. A. and Murphy, P. T., *"An architecture or a business and information system"*, IBM Systems Journal, Volume 27, Numero 1, Pagina 60 (1988), http://bit.ly/EBIS88

[3] Devlin, Barry, *"Business unIntelligence—Insight and Intuition Beyond Analitico and Big Data"*, Technics Publications, New Jersey, (2013), http://bit.ly/BunI-TP1