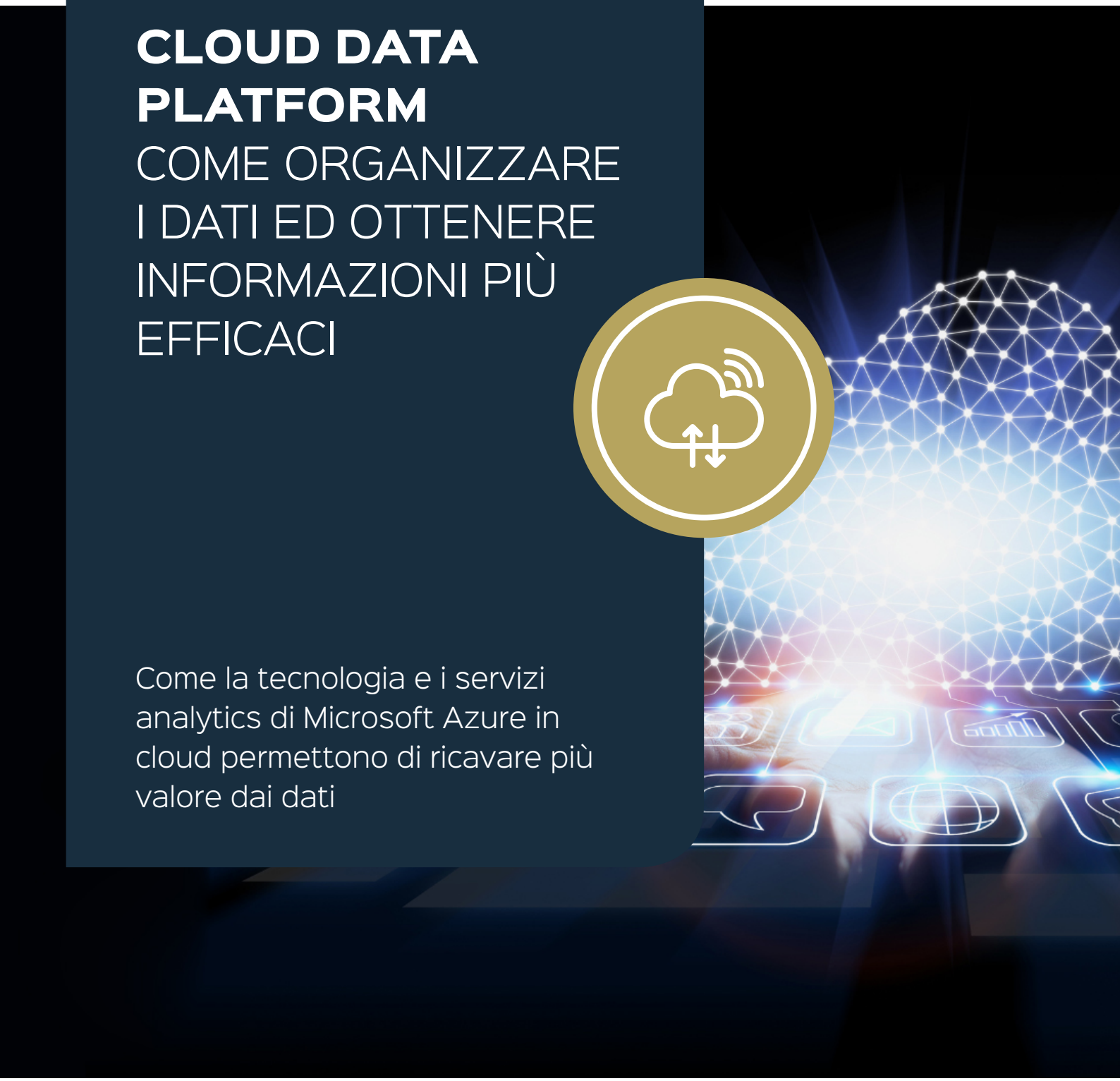
CLOUD DATA PLATFORM

COME ORGANIZZARE I DATI ED OTTENERE INFORMAZIONI PIÙ EFFICACI



Come la tecnologia e i servizi analytics di Microsoft Azure in cloud permettono di ricavare più valore dai dati

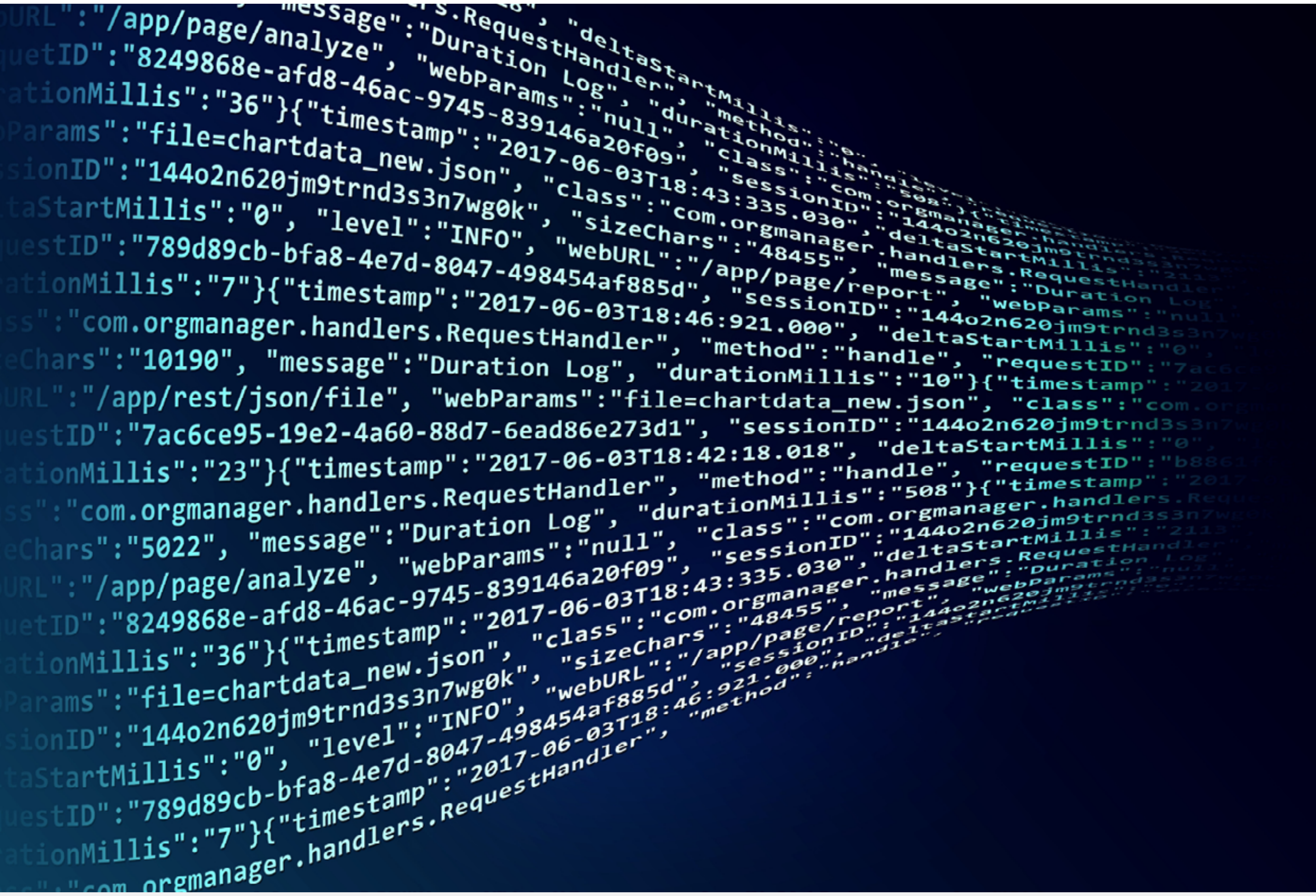
HORSA® GROUP

HORSA® insight

SOMMARIO

GLI ANALYTICS
OGGI E DOMANI

Molte aziende dispongono di data warehouse analitici, spesso on-premises, per supportare il processo decisionale in molti ambiti della loro attività. In alcuni dipartimenti, come vendite e marketing, ci sono utenti che producono report e dashboard partendo da questi sistemi. Impiegano quindi figure dedicate come data analyst per eseguire query e analisi ad hoc sui dati presenti nei data mart, utilizzando strumenti di business intelligence in modalità self-service.



Questo approccio è importante perché un’efficace analisi dei dati offre per molte organizzazioni un valore aziendale elevato.

Con l’aumentare della disponibilità di quantità sempre maggiori di dati diversi interni ed esterni all’azienda, l'analisi può fornire ancora più valore.

Per trarre vantaggio da questo cambiamento, le organizzazioni devono abbracciare i nuovi approcci all'analisi dei dati resi possibili dal cloud computing.

Infatti molte organizzazioni stanno **migrando** le loro soluzioni di data warehouse **verso i servizi analytics nel cloud**, per ottenere i diversi vantaggi che offre una **piattaforma di analisi end-to-end** come la disponibilità elevata, la **sicurezza**, la **velocità**, la **scalabilità**, **risparmi** sui costi e **prestazioni** elevate.

I BENEFICI

Esistono diversi benefici per cui si deve prendere in considerazione l'introduzione di servizi di analytics nel cloud, con il fine di avere a disposizione un **ricco ecosistema di dati aggiuntivi e tecnologie analitiche** che possono aiutare a rendere più moderno il data warehouse.

Alcuni dei benefici ricercati dalle organizzazioni che vogliono introdurre questo tipo di soluzione sono da suddividere tra infrastruttura e costi. Di seguiti i punti più rilevanti.

INFRASTRUTTURA

- Un insieme di servizi SaaS che **non richiedono alcuna infrastruttura** da gestire.
- **Scalabilità indipendente** tra lo storage ed il motore computazionale.
- **Adeguamento alle esigenze di caricamento dei dati** (ETL/ELT), all'aumentare del volume dei dati.
- Rispetto degli standard di **sicurezza** richiesti per una piattaforma cloud.
- La piattaforma di analisi deve **integrare nativamente** diversi motori (es. Apache Spark e SQL).
- **Aggiornamenti automatici** costanti ed innovativi.

COSTI

- **Diminuzione dei costi** di implementazione e manutenzione.
- Determinare il **costo in base al consumo effettivo.**
- **Bassi costi di archiviazione** sia per i dati di gestione che di produzione.
- **Sospensione** del consumo dei dati e degli strumenti di analisi se non sono utilizzati.
- **Riduzione dei tempi di sviluppo** del progetto di analisi, e quindi minori costi di gestione.

Tra i **principali provider di soluzioni analytics nel cloud troviamo Azure**, che propone una **soluzione ricca e completa di servizi** utili a rispondere a tutte queste esigenze.

I PROCESSI ED I SERVIZI

Microsoft Azure offre una serie di tecnologie e servizi gestiti in cloud progettati per l'analisi dei dati e per ricavare più valore dai dati.

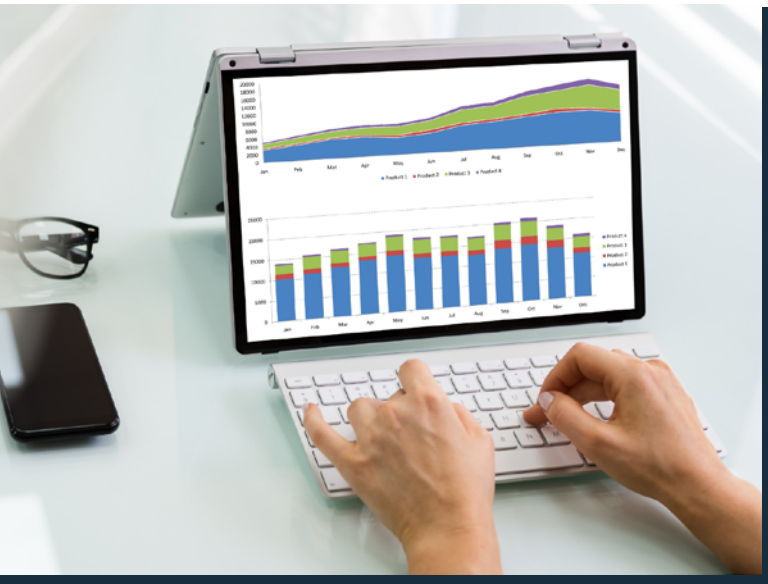
Alcuni di questi servizi sono:

- **Azure Synapse Analytics**: data warehousing relazionale scalabile nel cloud.
- **Azure Blob Storage**: archiviazione cloud a basso costo di dati binari.
- **Azure Data Lake Storage**: file system Hadoop (HDFS) distribuito come servizio cloud.
- **Azure Analysis Services**: servizio basato su tecnologia SQL Server Analysis Services.
- **Azure HDInsight**: supporto per le tecnologie Hadoop, insieme a Spark.
- **Azure Databricks**: una piattaforma di analisi basata su Spark.
- **Azure Machine Learning**: piattaforma per i data scientist utile nella realizzazione e distribuzione di algoritmi di ML.
- **Azure Data Factory**: è un unico servizio cloud per l'integrazione dei dati di diversa natura, siano esse in Azure, in locale o su un altro cloud pubblico. Fornisce un unico set di strumenti e un'esperienza di gestione comune per tutta l'integrazione dei dati.

DATA INTEGRATION NEL CLOUD

È possibile **combinare i servizi in base alle esigenze** per analizzare i dati sia strutturati che non strutturati. L'attivazione ed il coordinamento dei servizi passano da un processo importante, l'integrazione dei dati.

Le attività di integrazione richiedono l'**estrazione dei dati dalle origini**, ed il loro **caricamento nel sistema** in cui devono essere pronti per l'analisi. Il passaggio di questi dati richiede spesso anche la loro trasformazione in alcuni modi. Di solito questi passaggi ha senso automatizzarli.



Azure Data Factory (ADF) è progettato come servizio di integrazione dei dati basato su cloud, e può rispondere a diverse necessità ed essere applicato in diversi ambiti.

Uno degli ambiti di applicazione è quello dei **big data**, dove si ha la necessità di gestire grandi quantità di dati di diversa natura.

ADF offre un modo per **creare ed eseguire pipeline nel cloud**, e può accedere a servizi dati sia locali che cloud e in genere funziona con tecnologie come **Azure Synapse Analytics**, **Azure Data Lake** ed altri servizi di Azure.

Un altro ambito riguarda il data warehousing relazionale, basato su tecnologia SQL Server.

Spesso si utilizzano i SQL Server Integration Services (SSIS) per creare pacchetti SSIS. Per questo tipo di processi, ADF offre la possibilità di eseguire pacchetti SSIS in Azure, consentendo di accedere sia ai servizi di dati locali che cloud.

Caratteristica rilevante di un servizio di integrazione dei dati moderno è quello di **fornire diverse componenti ed eseguire azioni specifiche**. Potrebbe essere necessario copiare i dati da un archivio dati a un altro, o eseguire un processo Spark per elaborare i dati.

In questi casi, ADF fornisce una serie di attività ciascuna focalizzata sullo svolgimento di un tipo specifico di operazioni ed un meccanismo per specificare la logica generale del processo di integrazione dei dati, seguendo le dipendenze e le precedenze indicate nell'intero flusso dei dati.

DATA LAKE E DATA WAREHOUSE NEL CLOUD

Per descrivere una modern data platform partiamo confrontandola con il traditional data warehouse.

Uno dei primi aspetti riguarda la **tipologia dei dati**: invece di essere incentrato principalmente sull'elaborazione dei dati, come facevano i primi data warehouse, la versione "moderna" si basa sull'**archiviazione di molti dati da più fonti**, in vari formati e sull'acquisizione di insight sufficientemente utili da guidare le decisioni aziendali.

Se un traditional data warehouse può essere pensato come un data store, la versione moderna di oggi assomiglia più da vicino a un **grande data distribution center**. Nel tempo per rispondere alle necessità di analisi sempre più evolute a fianco dei traditional data warehouse, sono stati introdotti i Data Lake, un repository che accetta dati da più origini e può archiviarli in qualsiasi formato, legati principalmente sui casi d'uso della data science.

Rimanevano però due ambienti di analisi separati, e c'era la necessità di trovare un modo per unificarli. Da qui nasce il concetto di **modern data platform** dove si ha la possibilità di **unire in un unico ambiente tutti i needs di analisi** e si possono gestire adeguatamente i dati multi-strutturati in un'unica piattaforma.

La modern data platform è sia una **piattaforma di analisi completa che un "warehouse"** organizzato dei dati, e a differenza di un Data Lake esalta i concetti di governance e certificazione del dato.

Azure Synapse Analytics copre tutti questi aspetti ma con una infrastruttura al passo con i tempi. Tutti coloro che utilizzano i dati possono avere a disposizione un'**esperienza unificata** per la **preparazione**, la **gestione dei**

dati, il **data warehousing**, i **big data** e le attività di **intelligenza artificiale**.

Con due modi per analizzare i dati tramite **carichi di lavoro provisioned** o tramite il **modello di consumo serverless**, le organizzazioni possono scegliere l'opzione più conveniente per ogni caso d'uso. Inoltre, quando si tratta di dati, la **sicurezza** e la **privacy** sono della massima importanza, e sono integrate nell'infrastruttura di Azure Synapse.

Per un controllo degli accessi dettagliato, le aziende possono contribuire a **garantire che i dati rimangano al sicuro e privati** utilizzando la sicurezza a livello di colonna e la sicurezza nativa a livello di riga, nonché il mascheramento dinamico dei dati per proteggere automaticamente i dati sensibili in tempo reale.

L'introduzione di Azure Synapse Analytics permette di ottenere dei benefici.

Una delle caratteristiche principali sono le **Performance**, garantite dalle le migliori prestazioni del database relazionale che utilizza tecniche come l'elaborazione massiva in parallelo (MPP) e la memorizzazione automatica della cache.

Non di secondaria importanza la **Velocità**. Il data warehousing per sua natura è ad alta intensità di processo, e questo implica l'acquisizione, la trasformazione, la pulizia, l'aggregazione, l'integrazione dei dati e la produzione di visualizzazioni e report sui dati raccolti.

I numerosi processi coinvolti nello spostamento dei dati dalle origini a un data warehouse che sono spesso complessi e interdipendenti, vengono gestiti in modo semplice e si evita che un singolo collo di bottiglia possa rallentare l'intero processo di caricamento e creare un picco imprevisto nel volume dei dati che richiederebbe la necessità di aumentare le capacità di velocità.

I sistemi cloud devono attenersi a livelli elevati di Sicurezza e compliance. **Azure** è una piattaforma cloud **sicura**, altamente **scalabile** e **disponibile a livello globale** ed **Azure Synapse Analytics**, che risiede all'interno dell'ecosistema di Azure, ne eredita tutti i benefici.

L'intera piattaforma è una infrastruttura gestita, si tratta di **servizi cloud di tipo SaaS**. Non ci si deve preoccupare di gestire data center e delle relative operazioni per il data warehouse, e questo consente alle aziende di riallocare risorse preziose dove viene prodotto valore e di concentrarsi sull'utilizzo del data warehouse per fornire

le migliori informazioni. Questo riduce il TCO e consente un migliore controllo dei costi sulle spese operative.

Una infrastruttura di questo tipo permette di avere una Scalabilità di alto livello. Il volume di dati aumenta con il passare del tempo e con la raccolta del dato storico, così Azure Synapse Analytics può essere ridimensionato per adattarsi a questa crescita aggiungendo risorse in modo incrementale all'aumentare dei dati e dei relativi carichi di lavoro. **È una infrastruttura che racchiude in una sola soluzione sia il Data Lake che il Data warehouse**.

Azure Synapse può eseguire **query su dati strutturati o semistrutturati** con risorse di data warehousing ed eseguire rapidamente una query serverless su dati non strutturati direttamente sul data lake, permettendo ai professionisti dei dati di creare soluzioni di analisi senza dover unire una moltitudine di servizi.

L'architettura centralizzata, **evita la creazione dei silos di dati** in quanto, con l'implementazione della funzionalità HTAP (Transational Analytical Processing) di tipo ibrido e nativo per il cloud è possibile rimuove le barriere tra i servizi di database di Azure ed Azure Synapse Analytics, consentendo alle organizzaizioni di ottenere informazioni dettagliate dai propri dati transazionali in tempo reale direttamente nei database operativi, senza gestire lo spostamento dei dati o gravare sui propri sistemi operativi.

È una architettura che permette in molti casi la **massimizzazione delle skill**: sono presenti i **motori Apache Spark** e **SQL integrati in Azure Synapse**, e questo permette ai professionisti dei dati che preferiscono SQL di collaborare senza problemi con coloro che preferiscono Spark e viceversa.

Significa che chi ha familiarità o preferisce SQL può eseguire query sulle tabelle Spark utilizzando il linguaggio T-SQL, ed i data engineer o i data scientist che preferiscono linguaggi come Python, Scala, SparkSQL o C# possono trasformare i dati nello stesso servizio che ospita pipeline di dati, data lake e data warehouse.

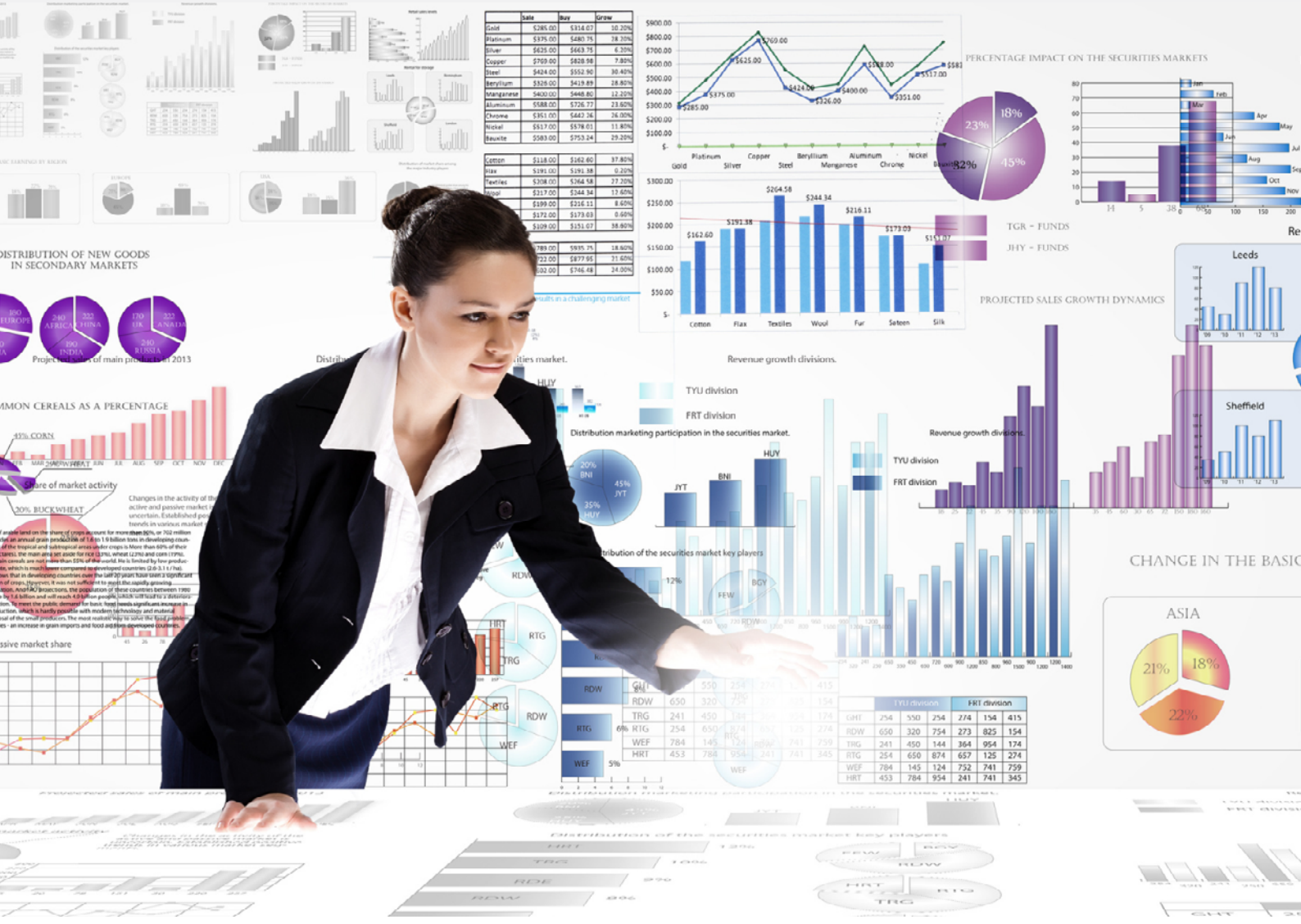
Per quanto riguarda l'ambito dei **costi**, nel cloud ha molta importanza il concetto di **Cost saving**. L'implementazione e la gestione di un data center on-premise è costosa, per tutta una serie di fattori come i costi di server e hardware, spazio fisico occupato e personale dedicato.

Queste spese possono essere ridotte con l'introduzione di una piattaforma come Azure Synapse Analytics. Infatti, offre la versa flessibilità del cloud, con

pagamento in base al consumo senza la necessità di complesse riconfigurazioni all'aumentare dei dati o dei carichi di lavoro crescono.

L'efficienza dei costi, in un data warehouse, subisce la variabilità nel tempo delle richieste di elaborazione e dei carichi di lavoro, con variazioni spesso non lineari che hanno dei picchi in particolari momenti.

Si pensi ad esempio, ad aziende che hanno un prodotto stagionale, queste aumenteranno drasticamente i carichi di lavoro del data warehouse solo in un certo periodo dell'anno. Con l'elasticità di Azure Synapse si può facilmente e rapidamente aumentare o diminuire la propria capacità in base alle richieste senza alcun impatto sulla disponibilità, la stabilità e la sicurezza, pagando solo per l'effettivo utilizzo.



CONCLUSIONI

Adottare una soluzione di cloud data platform permette alle organizzazioni di avere una **piattaforma di analisi a 360°**, dove tutti gli utilizzatori possono avere sempre a disposizione tutti i dati e le informazioni utili per generare gli insight di business.

Infine, una nota sulla piattaforma da adottare. Un **ruolo importante spetta al cloud**, che offre maggiore scalabilità delle risorse e flessibilità nella gestione dei costi operativi, mantenendo un alto livello di servizio sia in termini di disponibilità che di performance.

Perché Azure Synapse Analytics?

È un servizio di analisi che **permette di avere in un unico ambiente l'integrazione dei dati, funzionalità di data warehousing e di Big Data**.

Permette di **eseguire query sui dati in base alle esigenze di analisi**, usando sia opzioni serverless o dedicate, con un'esperienza unificata per tutto il processo di gestione del dato che va dall'inserimento alla distribuzione.

In conclusione, contiene tutti gli elementi che permettono di avere a disposizione una **soluzione efficace per rispondere a tutte le esigenze di analisi delle aziende**.

L'AUTORE

Simone Bocchi

È "Data Platform Practice Director" in Horsa S.p.a.
Più di 15 anni di esperienza nel settore data & analytics, con rilevanti competenze in sviluppo business e project management. Il suo lavoro è nell'ambito della gestione di complesse architetture per avviare e migliorare il processo di digitalizzazione della gestione dei dati, dei processi e delle informazioni in aziende strutturate operanti in diversi settori.

Horsa® insight

Horsa Insight è la Business Unit
di Horsa Group che si occupa di
Advanced & Predictive Analytics.
Un consulente di fiducia che
accompagna e guida le aziende nel
proprio percorso di data strategy.

A cura di Marketing
& Communication Dept.
Horsa Group

Horsa® GROUP

www.horsa.com