UNIVERSITÀ
DI TORINO

Data lake & Data governance

Tecnologie per abilitare la digital transformation

Ph.D Luca Giraldi

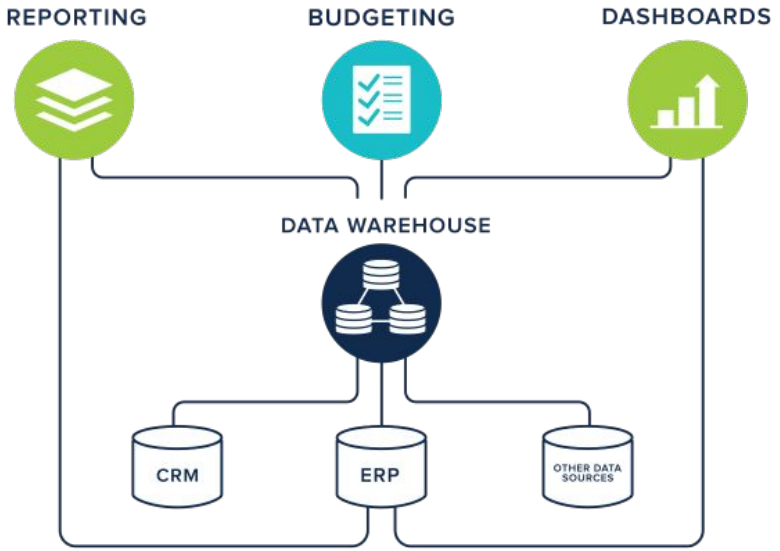
AGENDA

1. Cosa sono e a cosa servono i Data Lake
2. Architettura di un Data Lake
3. Data Governance nell’era dei Big Data
4. Data Lake nel Cloud
5. Case Study: Sophia By SIAE

Data Warehouse o **Data Lake**

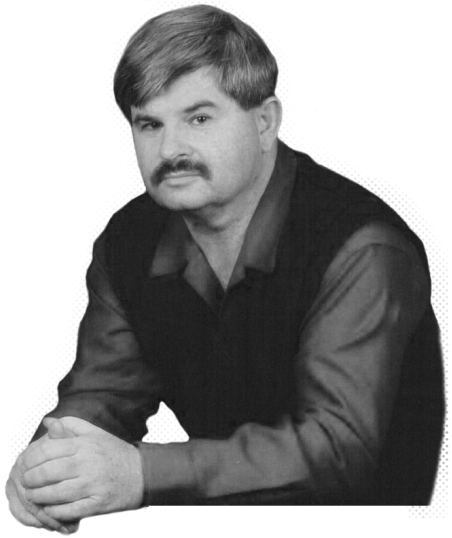
COSA SONO E A COSA SERVONO I DATA LAKE

In informatica un **data warehouse** è un archivio informatico contenente i dati di un'organizzazione, progettato per consentire di produrre facilmente analisi e relazioni utili a fini decisionali - aziendali



Secondo Inmon, il primo a parlare esplicitamente di Data Warehouse, lo definisce come una raccolta di dati:

- **Integrata;**
- **Orientata al tema;**
- **Variabile nel tempo;**
- **Non volatile.**



COSA SONO E A COSA SERVONO I DATA LAKE

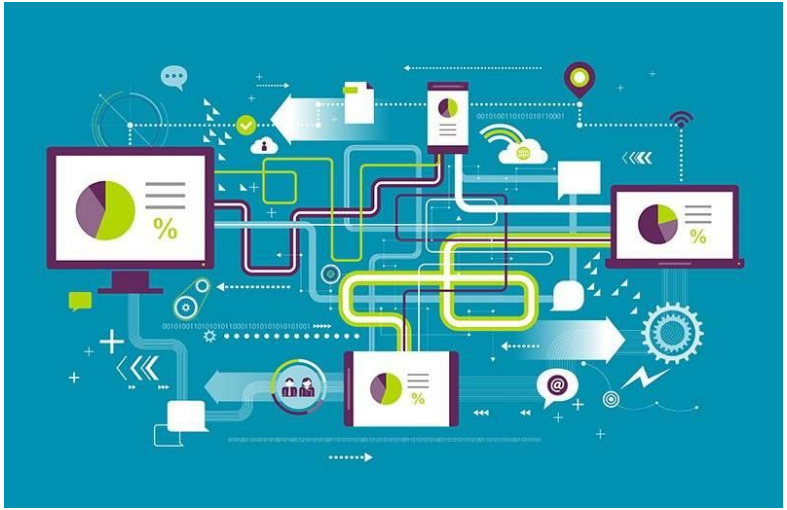
Per **Data Lake** si intende un sistema informatico che può contenere qualsiasi tipo di dato nella sua forma grezza, e che consente di combinarlo con altri dati, elaborarlo e renderlo disponibile nel momento in cui è necessario.

COSA SONO E A COSA SERVONO I DATA LAKE

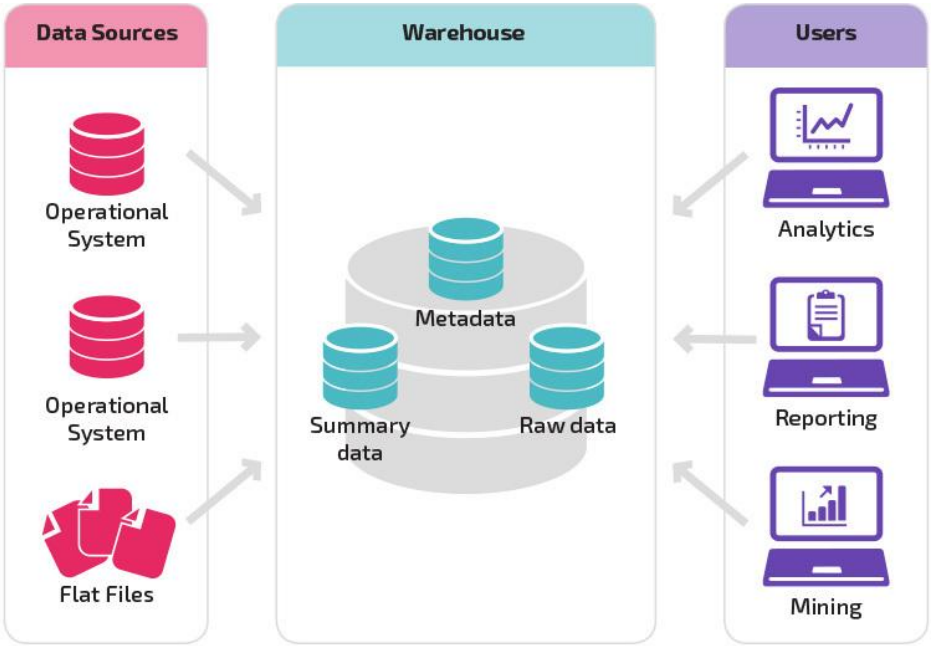
La produzione di Big Data provenienti dalle fonti più disparate (social media, sensori, dispositivi, etc), combinata con la nascita di nuove tipologie di dato, ha dato luogo a problematiche inedite con le quali le organizzazioni hanno dovuto confrontarsi.

Contestualmente all'esponenziale moltiplicazione del volume, varietà e velocità di dati prodotti, si è assistito ad un fiorire di tecnologie e tecniche orientate al trattamento ed alla gestione di questi Big Data.

Queste tecnologie e i nuovi approcci di governance dei dati si combinano e concretizzano nei data lake.

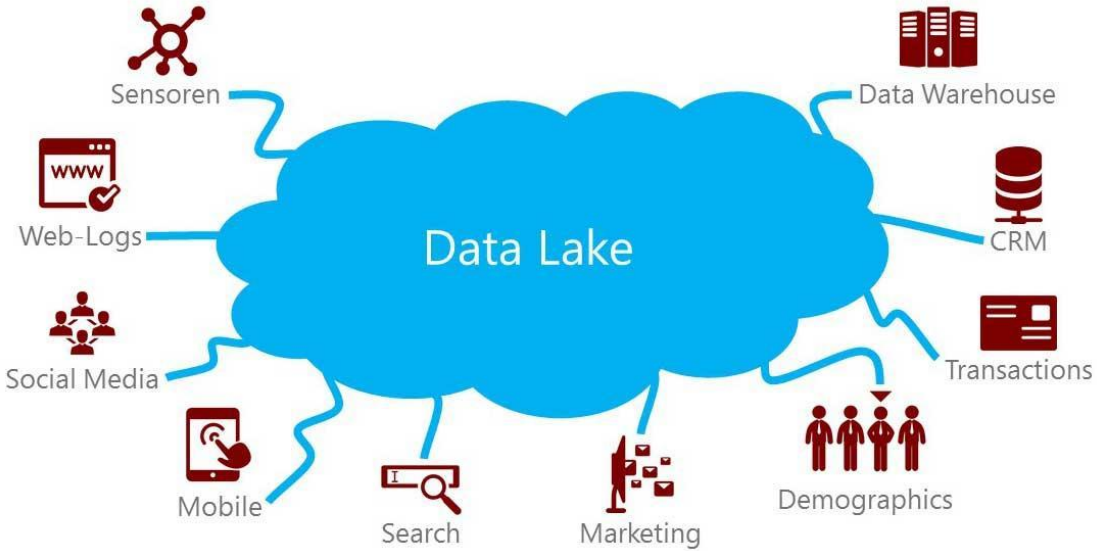


COSA SONO E A COSA SERVONO I DATA LAKE



Traditional Data Warehouse

COSA SONO E A COSA SERVONO I DATA LAKE



Data Lake

Data Warehouse	vs	Data Lake
Strutturati, Elaborati	Dati	Strutturati, semi-strutturati, non strutturati, grezzi
Schema on write	Processamento	Schema on read
Costosi per elevati volumi di dati	Storage	Progettati per lo storage a basso costo
Scarsamente agili, configurazione fissa	Agilità	Altamente agili, facilmente riconfigurabile
Tecnologie mature	Sicurezza	Tecnologie in fase di maturazione
Business analyst	Utenti principali	Data scientist ed altri

Data Warehouse Vs Data Lake

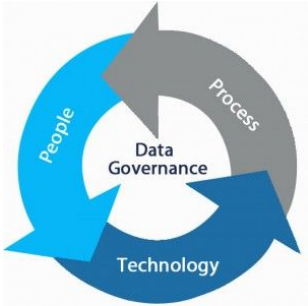
Data Warehouse	vs	Data Lake
Centralizzato	Storage	Distribuito
No	Gestisce stream alta velocità	Sì
Elevata	Scalabilità	Big Data
Tradizionali (es. Regressioni lineari)	Algoritmi	Cutting Edge (deep learning, NLP, machine learning)
Report e drill down	Visualizzazioni tipiche	Tag cloud, heat maps, esplorazione interattiva

Data Warehouse Vs Data Lake

Cosa è la data governance

Per Data Governance si intende, tradizionalmente, l’insieme di processi formali all’interno di un’azienda che hanno come obiettivo che **i dati:**

- Siano acquisiti da fonti affidabili
- Siano conformi a standard qualitativi definiti
- Siano conformi a specifiche regole aziendali
- Siano creati e modificati dalle persone giuste
- Siano adatti alle analisi e processamenti successivi
- Seguano un processo di modifica documentato in dettaglio
- Siano in linea con la strategia organizzativa aziendale
- Siano sempre validi e affidabili in qualsiasi stadio di trasformazione



La Data Governance per un’azienda **è fondamentale per:**

- Minimizzare i rischi nell’uso dei dati (es. errate analisi basate sui dati, data breach, violazioni privacy)
- Massimizzare il valore delle informazioni estratte dai dati mediante la loro analisi.

Il Data Lake per la governance dei Big Data

I **sistemi tradizionali** sono stati concepiti per trattare **dati strutturati** con **bassi tassi di acquisizione.**

Gli approcci e strumenti tradizionale per la Data Governance hanno difficoltà a tenere il passo con dati non strutturati, acquisiti in tempo reale e in grandi quantità finendo per **concentrarsi su un piccolo sottoinsieme di tutti i dati disponibili.**

I Data Lake forniscono strumenti per la Data Governance su diversi aspetti:

- Nuovi approcci al Lifecycle Management perché grandi quantità di dati possono coesistere in un unico ambiente per tempi più lunghi
- Applicazione di tecniche big data alla classificazione e verifica della qualità dei dati
- Applicazione ai big data di policy per valutare l’attendibilità di dati esterni
- Applicazione di politiche di sicurezza e protezione privacy ai flussi big data

Professioni legate alla data governance

Chief Data Officer

Un ruolo dirigenziale e non tecnico, per una persona con doti di gestione del cambiamento e leadership, in grado di comprendere appieno il business, orientata alla collaborazione e in grado di evangelizzare le unità aziendali.
Ha competenze tecniche e conoscenza dei dati e del loro potenziale valore come risorsa. Si occupa della data governance, del design dei dati e dell’infrastruttura per la loro gestione. Conosce metodologie e strumenti per la data governance.

Data steward

Mentre la data governance si focalizza su politiche e procedure di alto livello la data stewardship si concentra sul coordinamento pratico e sull'implementazione delle direttive di alto livello.
Il data steward è responsabile di realizzare le politiche di sicurezza e utilizzo dei date identificate dalle iniziative di data governance aziendale, agendo come collegamento tra il dipartimento IT e quelli business.

Funzioni di base per la data governance nel Data Lake

Security & privacy
Proteggere i dati

Definire e applicare ai big data le politiche di accesso e protezione dei dati rispetto al loro valore per il business e per gli interessati alla data privacy.

Metadata
Conoscere i dati

Classificare, profilare, valutare in termini qualitativi, tracciare, versionare, correlare, identificare i dati in tutti i loro passaggi all’interno del data lake.

Information Lifecycle
Gestire la vita dei dati

Gestire il tempo di vita dei dati, dalla loro creazione alla loro utilizzazione ed, infine, alla loro distruzione in uno scenario big data: velocità, variabilità, volume.

Case study: Sophia by SIAE

Sophia è un progetto di innovazione nato dalla relazione tra SIAE, il cliente e owner del progetto e del sistema, e Alkemy società a cui è affidata la Digital Transformation di Siae.

Grazie all'iniziativa di Alkemy Lab (stream di innovazione della stessa Alkemy) ed al coinvolgimento della società di ricerca e sviluppo Linkalab s.r.l. (specializzata su Big Data e Data science) all'interno di un processo di partnership ed aggregazione, è stato possibile creare valore per il cliente, oltretutto in un valore nobile quale il diritto autoriale.

SIAE | DALLA PARTE DI CHI CREA

Alkemy
digital_enabler

Problema

gestire e **monetizzare** il diritto d'autore che deriva dall'utilizzo multimediale online.

Dal 2009 al 2016 c'è stato un **aumento del 2067% (!)** nella fruizione di opere digitali online.

I Digital Service Provide (DSP)r trasmettono i report sulle utilizzazioni di opere (DSR) a SIAE che li gestisce con Sophia:
Youtube: 10,2 miliardi di utilizzazioni all'anno
Spotif: 7,2 miliardi di utilizzazioni all'anno
Deezer 1,2 miliardi di utilizzazioni all'anno
iTunes 1,2 miliardi di utilizzazioni all'anno
Apple Music 2,7 miliardi di utilizzazioni all'anno
GooglePlay 600 milioni di utilizzazioni all'anno

Grazie all’applicazione delle tecnologie **Big Data Open Source** e allo sviluppo di architetture basate su piattaforme **Cloud (AWS)**, è stato possibile in breve tempo (10 mesi) la messa in produzione del nuovo Data Lake di SIAE.

La realizzazione del progetto Sophia, è risultata strategica per la gestione dell'identificazione delle opere musicali di SIAE.

Grazie a esso SIAE è ora in grado di **processare in modo ottimizzato i flussi big data,** inviati dalle piattaforme digitali quali Spotify, iTunes o YouTube, che contengono i report di utilizzo delle opere in standard **Digital Sales Report Message Suite (DSR).**


Grazie