BENVENUTI!

who

why

what

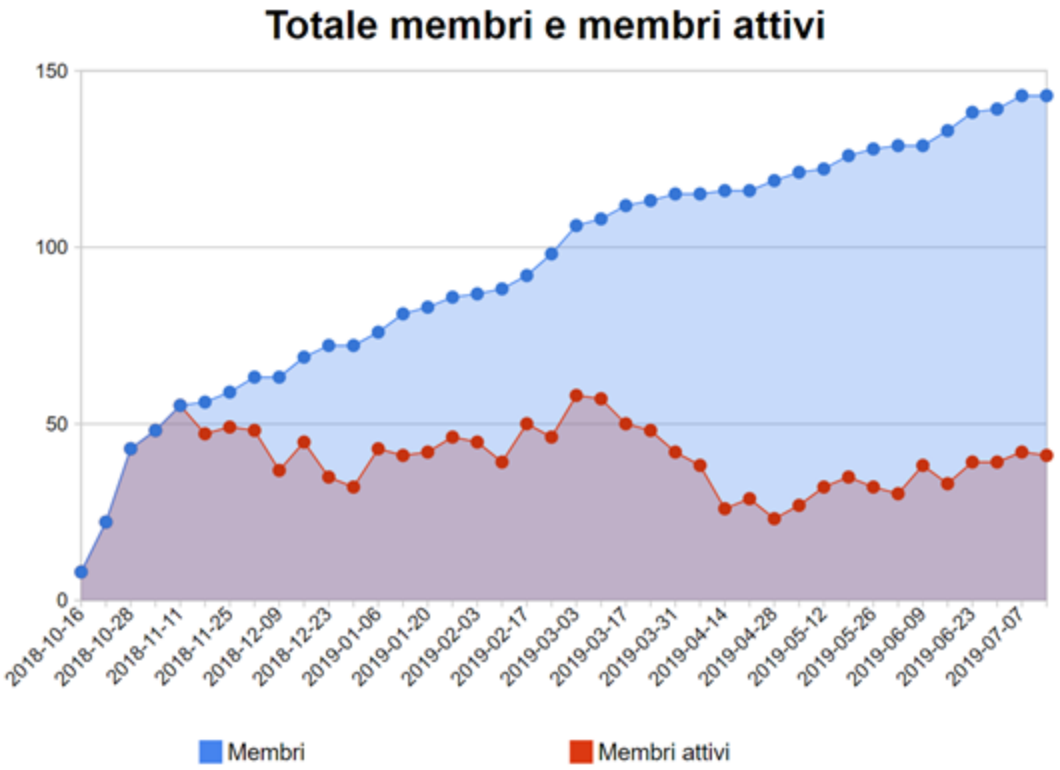
when

where

who

community

who


Totale membri e membri attivi
150
100
50
0
2018-10-16
2018-10-28
2018-11-11
2018-11-25
2018-12-09
2018-12-23
2019-01-06
2019-01-20
2019-02-03
2019-02-17
2019-03-03
2019-03-17
2019-03-31
2019-04-14
2019-04-28
2019-05-12
2019-05-26
2019-06-09
2019-06-23
2019-07-07
Membri
Membri attivi

what

```
for (var oneDay of month) {
  if (you.Code 💻 || you.Beer 🍺 || you.Tech 📱 ) {
    CremaWeb.meetup.join()
  }
}
```

what

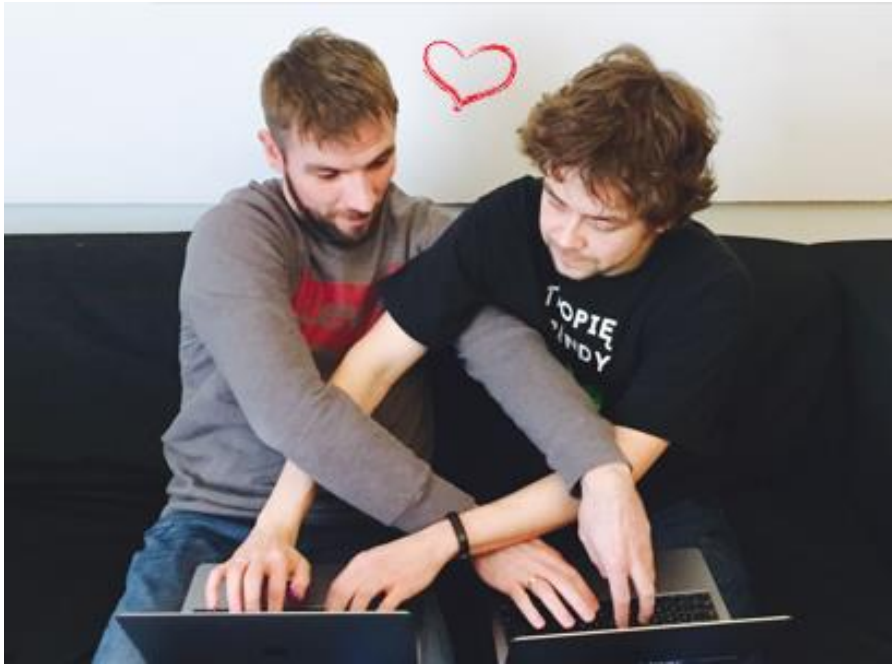
Organizzazione eventi su temi Tecnologici

Area di discussione comune su Web (Discord)

Coinvolgimento diretto o indiretto dei dipendenti delle aziende della zona

Riferimento propositivo per discussioni e eventi dedicato ad appassionati del campo IT

why



when
where

meetup.com/CremaWeb

REI
Reindustria Innovazione

per ora grazie a

REI
Reindustria Innovazione

GSLabs

esc ENTERPRISE SOFTWARE CONSULTING

iDigital3 HUMAN CENTERED

...aspettando i contributi di chi si vorrà unire...

6 Febbraio 2025

Databricks: I processi ETL al giorno

databricks

APACHE Spark™

Parquet

DELTA LAKE

kafka

Mirko Porcu

Presentiamoci …. :)

Indice

1. Introduzione

Nell’era digitale le aziende raccolgono quotidianamente una quantità enorme di dati da **diverse fonti eterogenee**, che sarebbero impossibili da gestire se non attraverso processi specifici di gestione del dato.

La gestione del dato, o **Data Management**, include processi come la **data integration**, la **data quality**, la **data governance** e la **data security**, che permettono alle aziende di ottenere una visione completa e affidabile del proprio business, ad esempio, comprendendo meglio il comportamento dei clienti o le performance di vendita.

Attraverso una corretta gestione dei dati, le aziende possono, ad esempio, individuare nuove opportunità di mercato, ottimizzare le campagne di marketing e migliorare l'efficienza operativa, trasformando i propri dati in una risorsa strategica.

2. Il data management e i processi ETL

il Data Management è l'insieme di processi e tecniche che permettono alle aziende di:

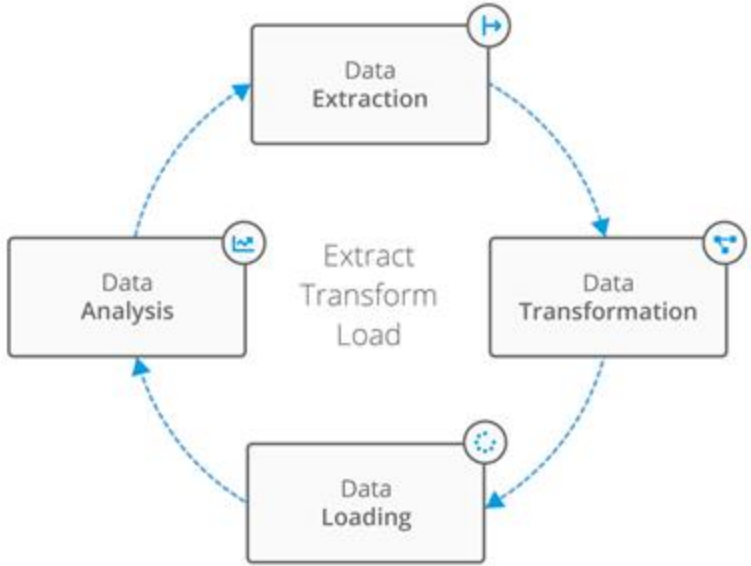
- Raccogliere dati da diverse fonti, interne ed esterne → **Data Integration**
- Organizzare i dati in modo strutturato e accessibile → **Data Governance**
- Assicurare la qualità dei dati, eliminando errori e incongruenze → **Data quality**
- Proteggere i dati da accessi non autorizzati e perdite → **Data Security**
- Utilizzare i dati per analisi, reportistica e decisioni strategiche → **Data Visualization**

2.1 I processi ETL

Cosa sono i processi ETL?

I processi ETL sono dei processi di trasformazione del dato che si possono suddividere in questi step :

- **Extract**: Raccolta dati grezzi da diverse fonti (database, file, ecc.)
- **Transform**: "Pulizia" e conversione dei dati in un formato utilizzabile (correzione errori, uniformazione formati, aggregazione, ecc.)
- **Load**: Inserimento dei dati trasformati in un sistema di destinazione (data warehouse, data lake, ecc.)



2.1.1 ETL vs ELT

ETL

I dati vengono trasformati prima di essere caricati nel data warehouse.

Utile quando:

- Le risorse del data warehouse sono limitate.
- È necessario un elevato controllo sulla qualità dei dati.
- I dati devono essere conformi a schemi rigidi.

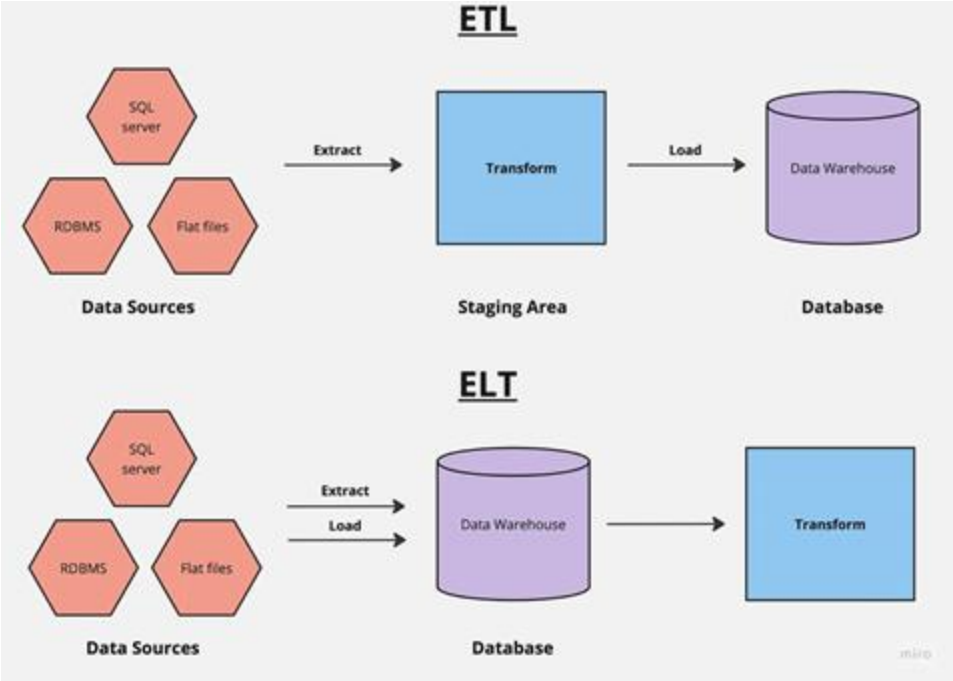
ELT

I dati vengono caricati nel data warehouse prima di essere trasformati.

Utile quando:

- Si dispone di un data warehouse scalabile (cloud).
- Si gestiscono grandi volumi di dati non strutturati.
- Si necessita di flessibilità nell'analisi dei dati.

2.1.1 ETL vs ELT



https://medium.com/@c.chai/main-architecture-types-in-data-engineering-5e2242cace15

2.1.2 Gestione a Batch e Gestione streaming

L'elaborazione batch e l'elaborazione in streaming sono due approcci fondamentali per l'elaborazione dei dati, ognuno con i suoi pro e contro.

La scelta migliore dipende dalle esigenze specifiche dell'applicazione.

2.1.2.1 Elaborazione Batch

Pro:

- **Efficiente per grandi volumi di dati**: L'elaborazione batch è ideale per elaborare grandi quantità di dati storici, in quanto può essere eseguita in modo efficiente su larga scala.
- **Costo-efficacia:** L'elaborazione batch può essere più conveniente rispetto all'elaborazione in streaming, in quanto può utilizzare risorse di elaborazione meno costose.
- **Semplicità**: L'elaborazione batch è generalmente più semplice da implementare e gestire rispetto all'elaborazione in streaming.

Contro:

- **Latenza elevata**: L'elaborazione batch introduce un ritardo significativo tra l'acquisizione dei dati e la disponibilità dei risultati, il che la rende inadatta per applicazioni in tempo reale.
- **Difficoltà nell'aggiornamento dei dati**: L'aggiornamento dei dati elaborati in batch può essere complesso e richiedere molto tempo.

2.1.2.2 Elaborazione Streaming

Pro:

- **Latenza bassa**: L'elaborazione in streaming consente di elaborare i dati in tempo reale, il che la rende ideale per applicazioni che richiedono risposte immediate, come il monitoraggio delle frodi etc etc.
- **Aggiornamenti in tempo reale**: L'elaborazione in streaming consente di aggiornare continuamente i dati e i risultati, fornendo una visione aggiornata dell'andamento.

Contro:

- **Complessità**: L'elaborazione in streaming è generalmente più complessa da implementare e gestire rispetto all'elaborazione batch.
- **Costo**: L'elaborazione in streaming può essere più costosa rispetto all'elaborazione batch, in quanto richiede risorse di elaborazione dedicate e una maggiore capacità di gestione dei dati.
- **Sfide nella gestione degli errori**: La gestione degli errori nell'elaborazione in streaming può essere più complessa, in quanto i dati vengono elaborati in modo continuo.

3.1 Databricks

Databricks è una piattaforma cloud che unifica **Data Analysis**, **Data Engineering** e **Data Science**. Offre un ambiente collaborativo in cui diversi profili professionali possono lavorare insieme su progetti di dati, sfruttando strumenti e funzionalità per:

- **Data Analysis**: esplorare, analizzare e visualizzare i dati per estrarre informazioni utili al business.
- **Data Engineering**: costruire e gestire pipeline di dati affidabili e scalabili per l'elaborazione di grandi volumi di informazioni.
- **Data Science**: sviluppare, addestrare e distribuire modelli di machine learning per ottenere previsioni e automatizzare processi.

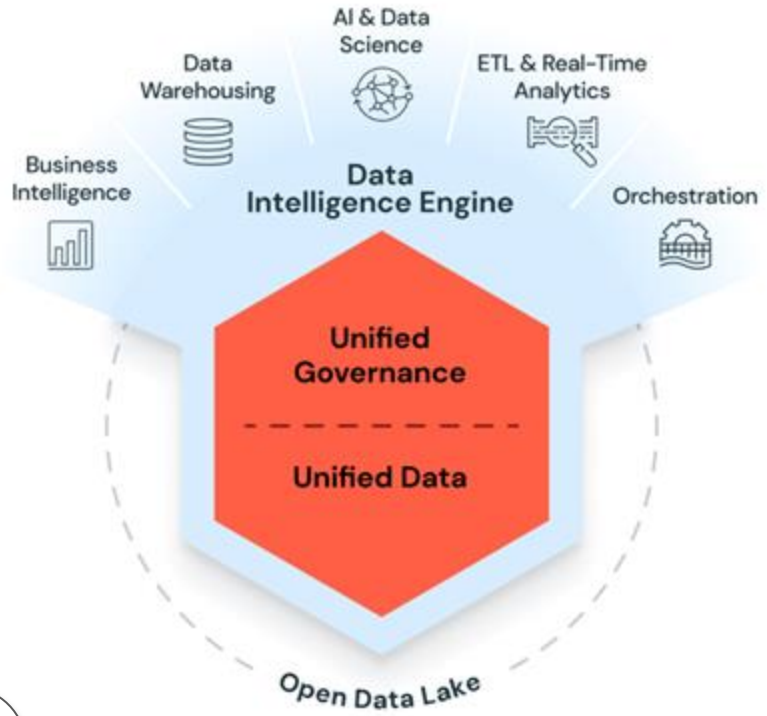
https://www.databricks.com/

3.1 Databricks

Databricks si basa su **Apache Spark**, offre un ambiente collaborativo e scalabile per:

- Analizzare grandi quantità di dati con velocità e flessibilità.
- Costruire pipeline di dati affidabili e automatizzate.
- Sviluppare e distribuire modelli di machine learning.

3.1 Databricks



3.2 Databricks : Apache Spark

What is Apache Spark™?

Apache Spark™ is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters.

https://spark.apache.org/

3.2 Databricks : Apache Spark

Cos’è Apache Spark?

Un framework open-source per l'elaborazione distribuita di grandi quantità di dati.

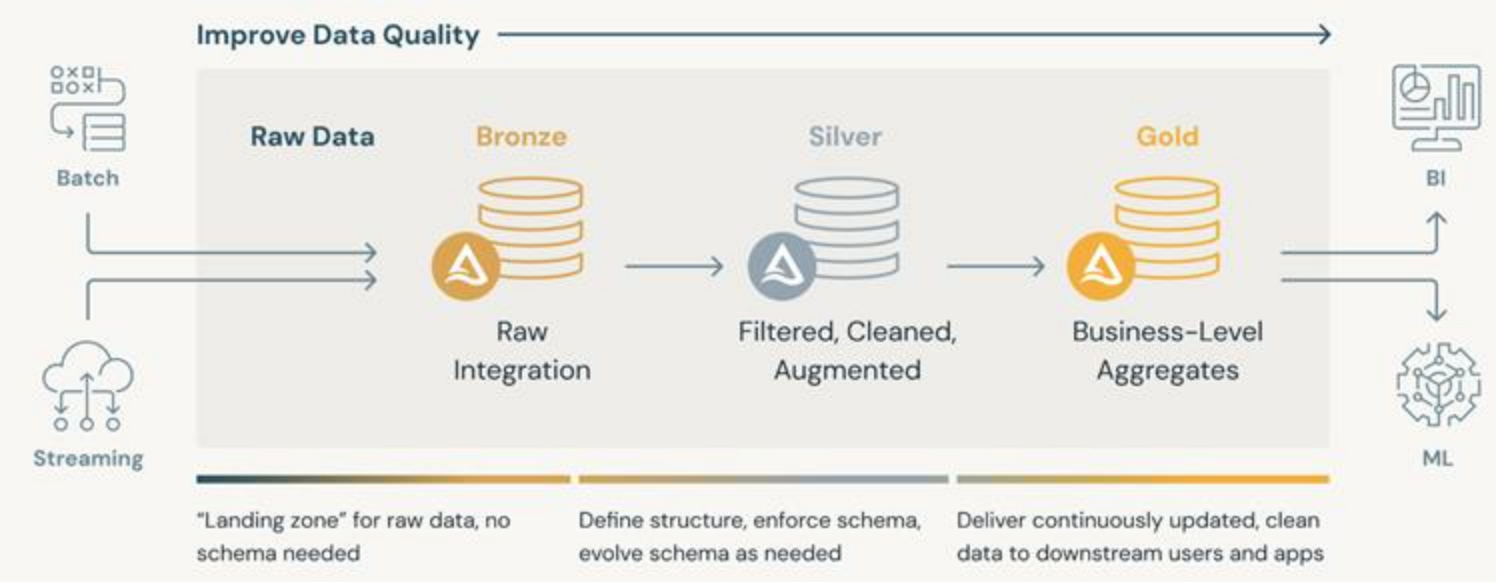
Caratteristiche principali:

- **Velocità**: Esegue calcoli in memoria per prestazioni elevate.
- **Scalabilità**: Può essere eseguito su cluster di macchine, da piccole a grandi.
- **Facilità d'uso**: Supporta API in **Java**, **Scala**, **Python** e **R**.
- **Versatilità**: Può essere utilizzato per analisi batch, elaborazione di flussi, machine learning e SQL.

Esempi di architettura per la trasformazione del dato

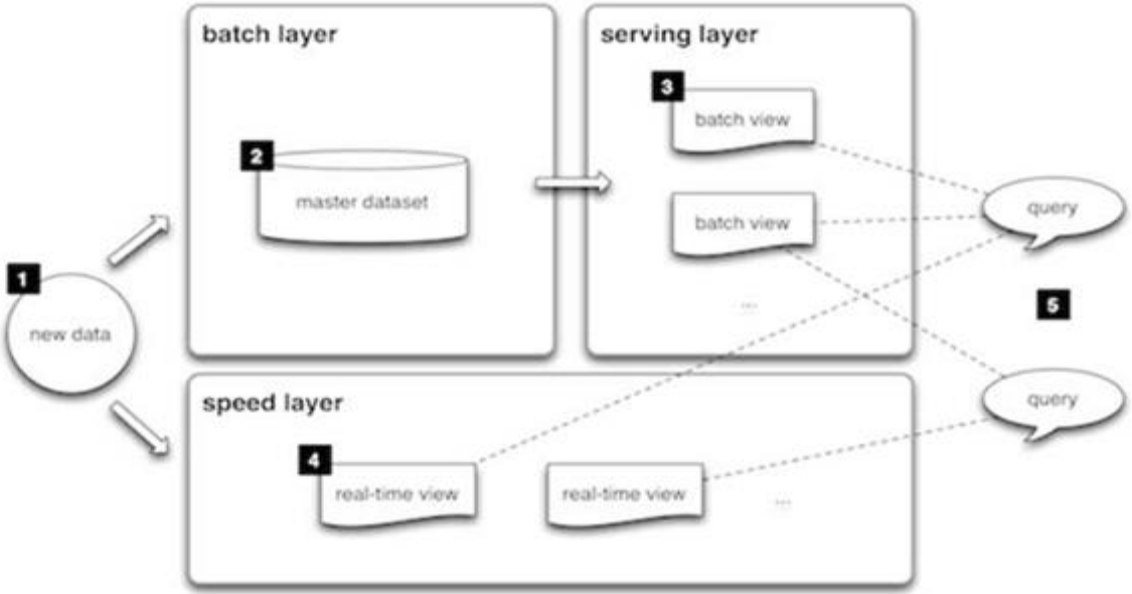
- L'architettura **Medallion** è un approccio tradizionale che categorizza i dati in tre livelli: bronzo, argento e oro, in base alla qualità e all'affidabilità dei dati.
- Architetture basate su **streaming**: Elaborano i dati in tempo reale, il che è cruciale per applicazioni come il rilevamento delle frodi e l'analisi del sentiment dei clienti.
- Architetture **lambda**: Combinano l'elaborazione batch e in streaming per gestire sia i dati storici che quelli in tempo reale.
- Kappa, Data Vault Architecture & more more others

Medallion Architecture



https://www.databricks.com/glossary/medallion-architecture

Lambda Architecture



https://www.databricks.com/it/glossary/lambda-architecture

4. Use Case

Realizziamo un **processo** per la gestione di transazioni per avere dei KPI di performance che possano permettere di prendere decisioni strategiche importanti per il business.

Grazie

Next event

17/04/2025