UNIVERSITY OF PADUA
UNIVERSITA' DEGLI STUDI DI PADOVA

Creazione di una Data Pipeline per il trattamento dei dati con Apache Kafka e Apache Druid

Dipartimento di Matematica "Tullio Levi Civita"

Corso di Laurea in Informatica

Esame di Laurea - 22 Settembre 2023

Laureando: Marco Brugin - Matricola n. 2010012

Relatrice: Prof.ssa Ombretta Gaggi

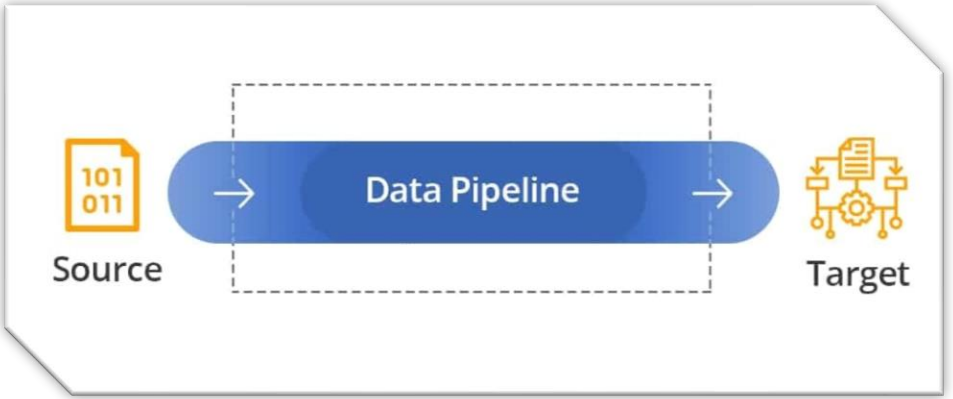
DIPARTIMENTO MATEMATICA

L’azienda

- Sviluppo e consulenza IT
- Certificazioni **ISO LL-C**
- Business Innovation

- Studio e sperimentazione sulle tecnologie utilizzate in una Data Pipeline

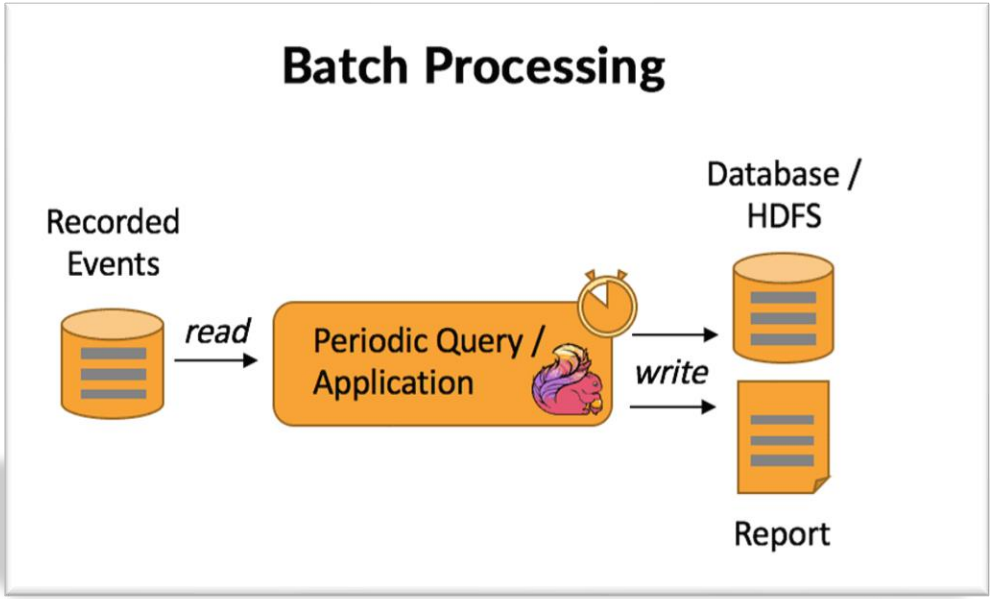
- Creazione di un prototipo in grado dimostrarne le potenzialità e le prestazioni



Le Data Pipeline tradizionali

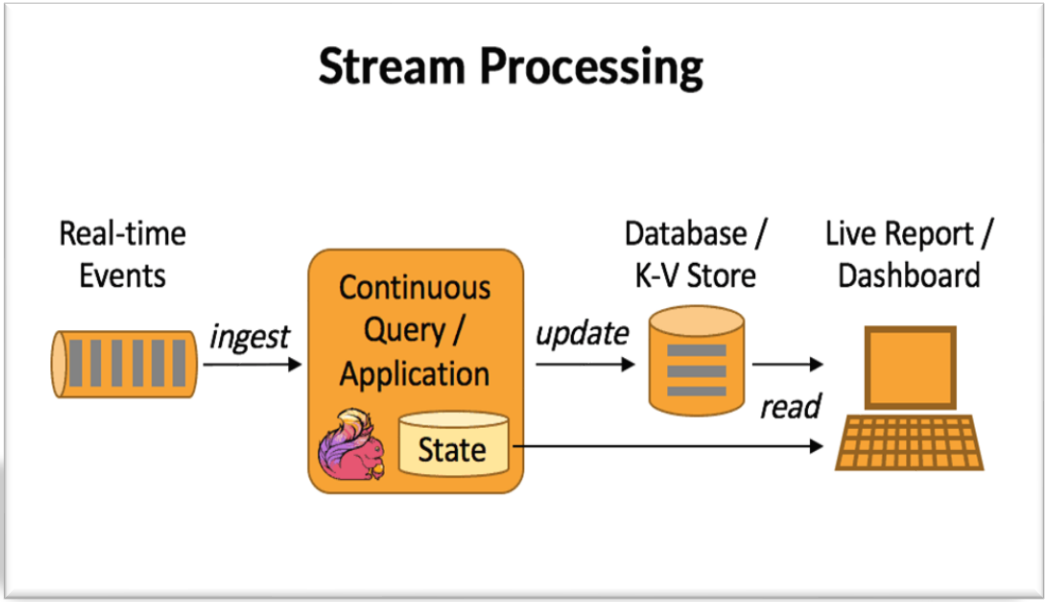
- Consumo, elaborazione, archiviazione dati in tempo reale

- SONO PROGETTATE PER ELABORARE DATI STATICI
  - Estraggono
  - Trasformano
  - Rendono disponibili i dati elaborati



Le Streaming Data Pipeline

- Grandi mole di dati
- Streaming Data Pipeline
  - **Archiviazione**
  - **Lavorazione**



Obiettivi del progetto di stage

- Realizzare un prototipo di Data Pipeline, eseguibile su più ambienti, in grado ricevere dati da un sistema di raccolta, eseguendo il Data Processing sui dati grezzi ricevuti

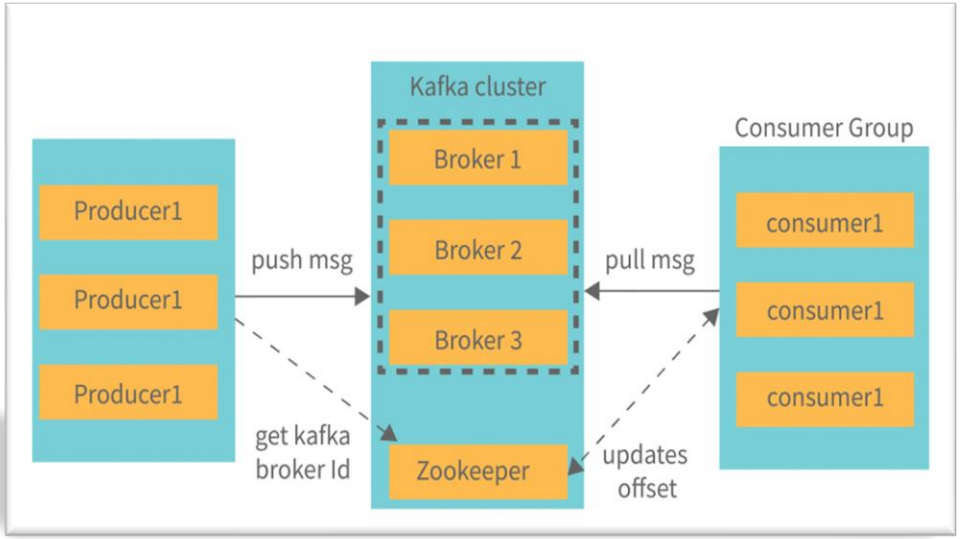
- Studio delle funzionalità offerte e delle differenze con le tecnologie tradizionali, in particolare con i **RDBMS**

Requisiti

- Realizzazione di una Streaming Data Pipeline
- Utilizzo di Apache Kafka come strumento di raccolta dati
- Utilizzo di Apache Druid come strumento **OLAP**
- Configurazione di tutto l'ambiente con **Docker Compose**
- Verifica delle funzionalità e prestazioni offerte da tali strumenti rispetto alle tecnologie tradizionali

Apache Kafka

- Message broker in grado di
  - Pubblicare e sottoscrivere flussi di messaggi attraverso **topic**
  - Archiviarli in modo duraturo
  - Elaborare tali flussi in modo retrospettivo

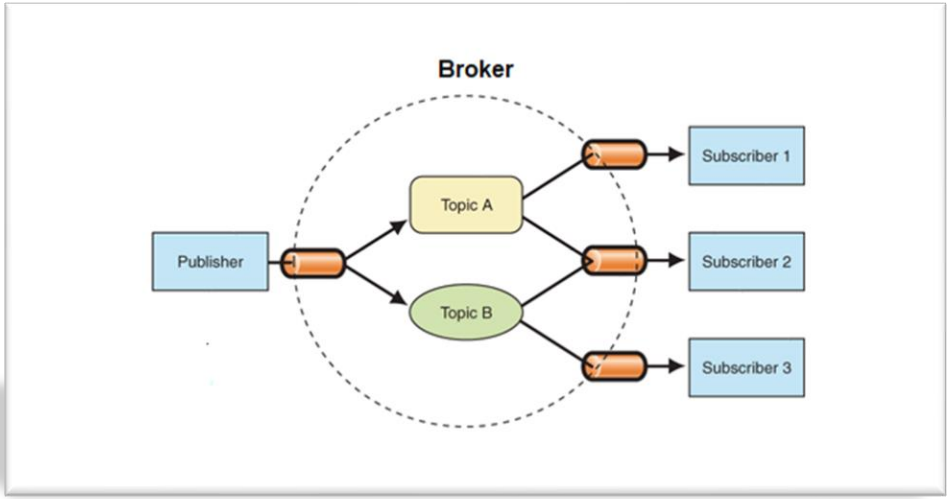


- Architettura distribuita che opera su nodi
- Server Kafka
  - Kafka Broker
  - Kafka Connect
- Client Kafka
  - **Produttori**
  - **Consumatori**

Pattern Publisher-Subscriber

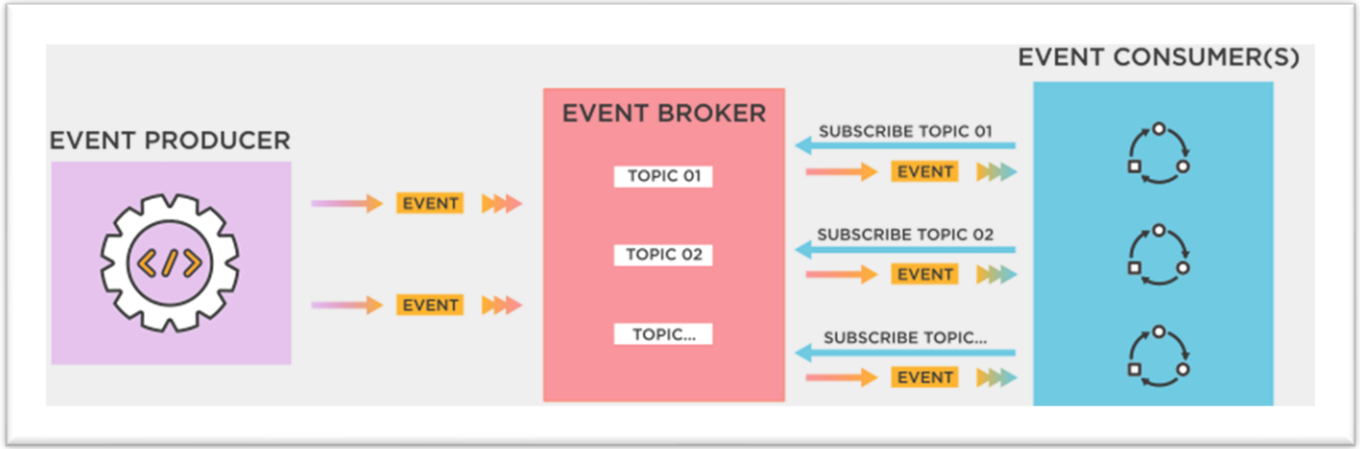
- Sistemi distribuiti
- Comunicazione asincrona
- Ridimensionabile **dinamica**

- Vantaggi
  - Debole **accoppiamento** tra le componenti
  - elevata **scalabilità**
  - utilizzo della comunicazione **asincrona** ad eventi
  - indipendenza dal protocollo di comunicazione



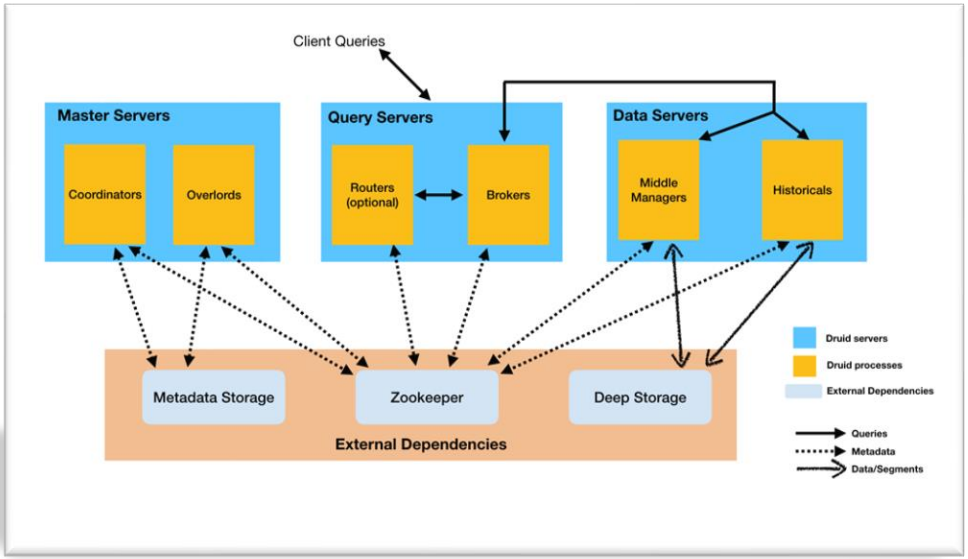
Apache Kafka e l'architettura EDA

- L'**E**ven **D**riven **A**rchitecture è un pattern in cui gli agenti coinvolti sono in grado e di ricevere tali eventi
- Per utilizzare Apache Kafka in una architettura **EDA** la chiave è andare a sfruttare il disaccoppiamento
  - Evitare il **polling**
  - Attendere il verificarsi di un evento



Apache Druid

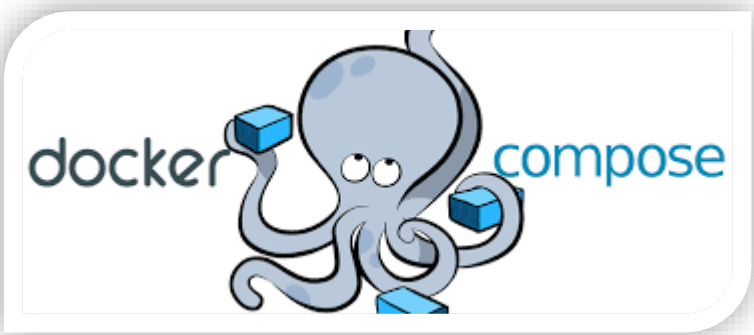
- Strumento di analisi **OLAP**
- Distribuito
- Scalabile
- Tollerante ai guasti
- Compatibile con il cloud



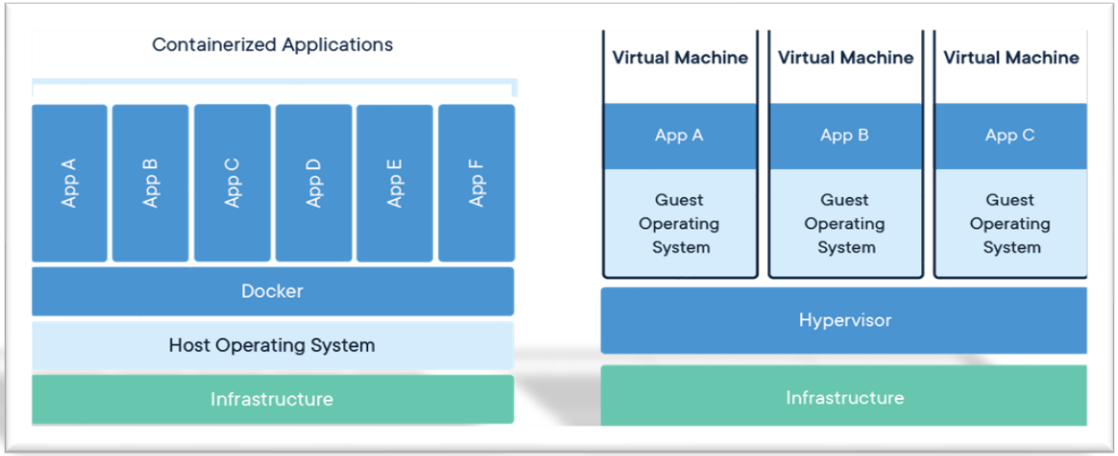
- Server principali
- Server dati
- Server d'interrogazione

Docker Compose

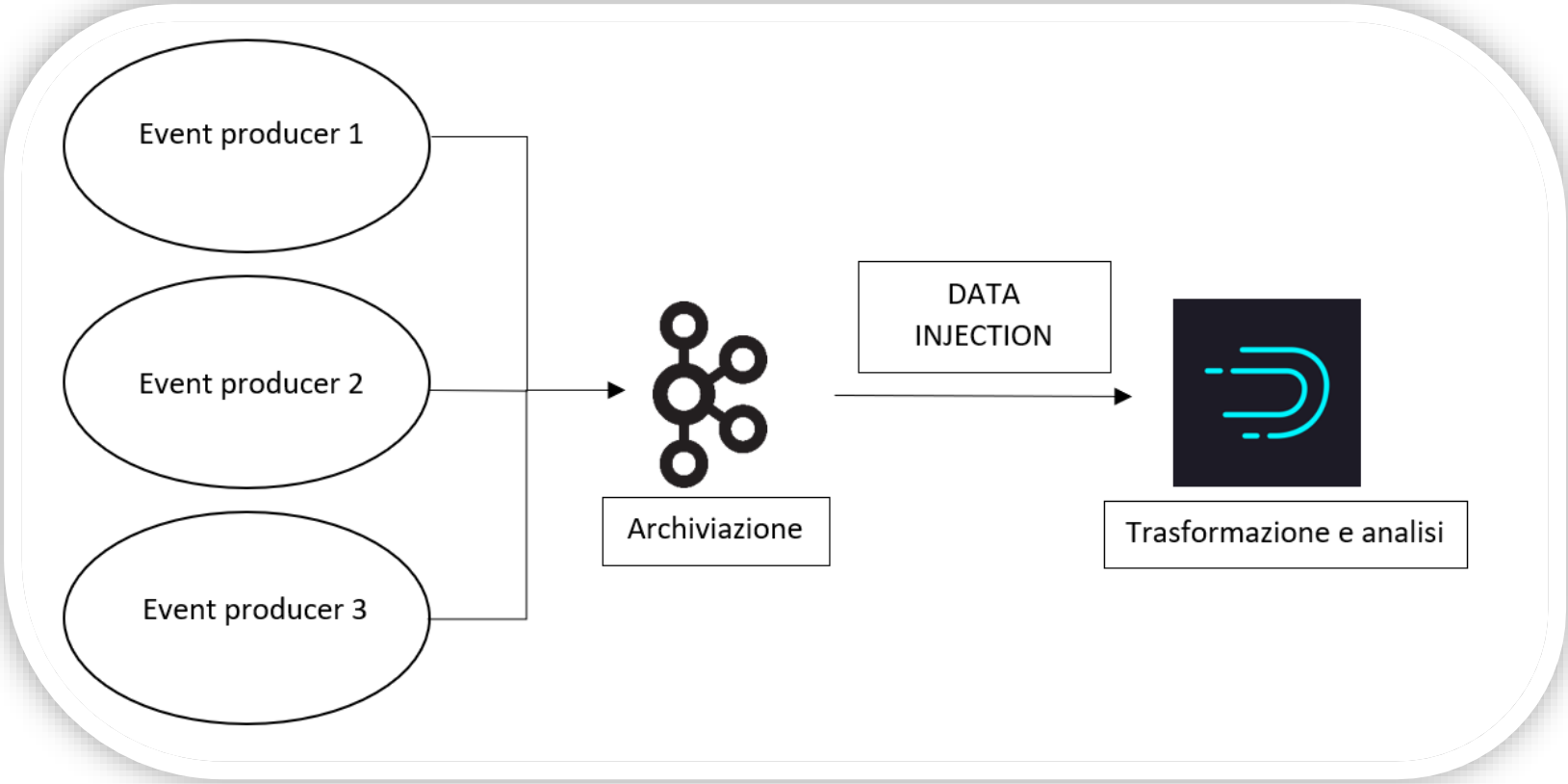
- Gestione di **multi-container**



- Permette di definire
  - Relazioni tra i container
  - Configurazioni di rete
  - Volumi utilizzati dai container
  - Variabili d'ambiente



Il prototipo



Testing

- Performance
  - Verifica delle prestazioni di esecuzione di Apache Druid all'interno di un cluster Docker rispetto ad un **RDBMS**
  - Verifica dei miglioramenti dati dalla funzionalità di **rollup**
- Funzionalità
  - Tabelle di **lookup**

Conclusioni

- Obiettivi raggiunti
  - Realizzazione di una Streaming Data Pipeline che utilizzi Apache Kafka e Apache Druid, eseguibile con Docker Compose
  - Studio approfondito delle funzionalità e prestazioni offerte da tali tecnologie
- Competenze apprese
  - Configurazione di un' architettura distribuita sviluppata all'interno di container
  - Sviluppo di un prototipo all'interno di un team
  - Esperienza in ambito Data Analyst e **OLAP**