



BREAKFAST CHANNEL

Martedì 7 aprile 2020



Data Lake VS Data Virtualization?

Mauro Mariuzzo & Stefano Pampaloni

DATA LAKE

Molteplici fonti informative
“interne”

ERP

CRM

ECM

MES
MOM

IIoT

ECommerce

DATA LAKE

Molteplici fonti informative
“Aziendali”

Cloud

ERP

CRM

On Premise

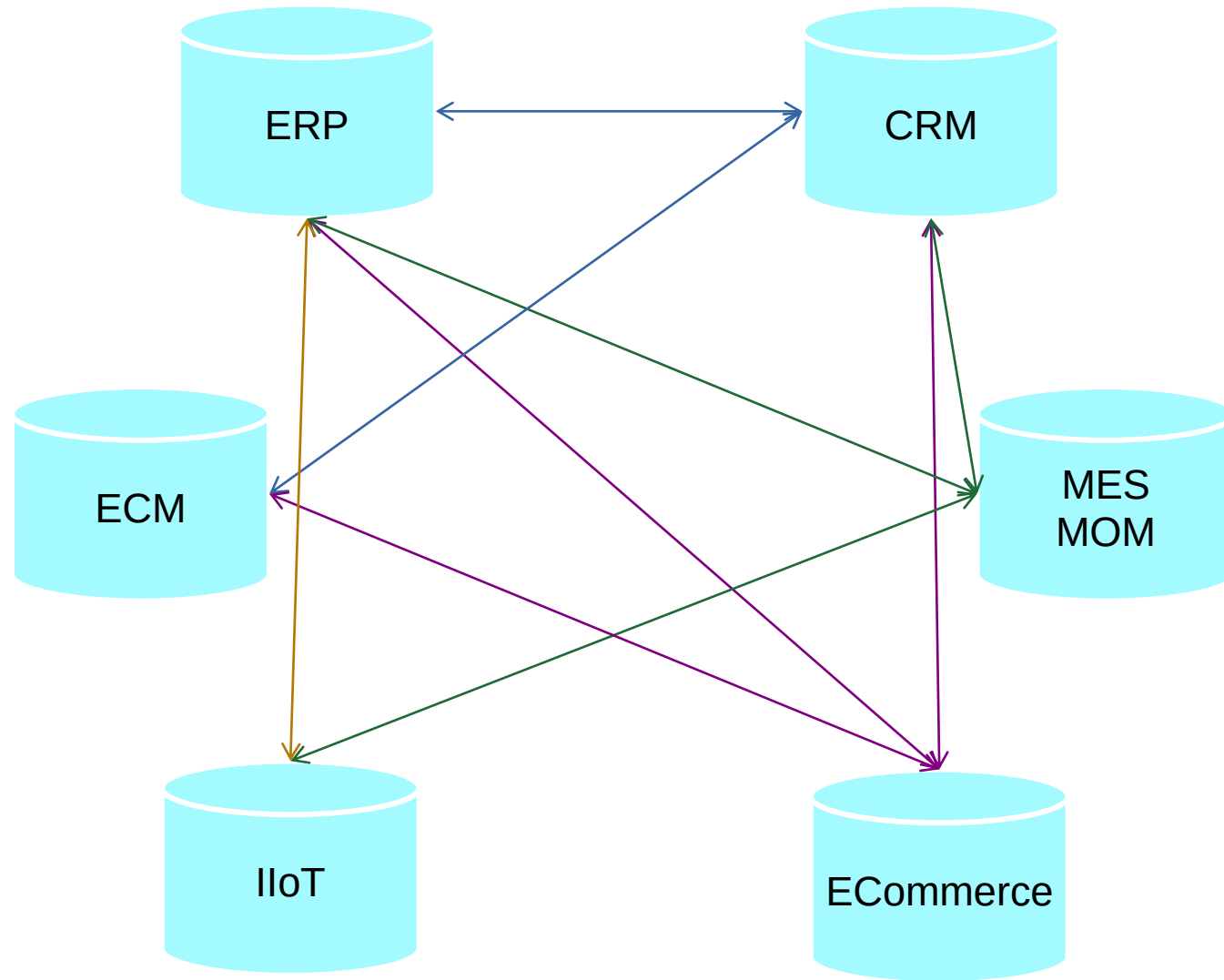
ECM

MES
MOM

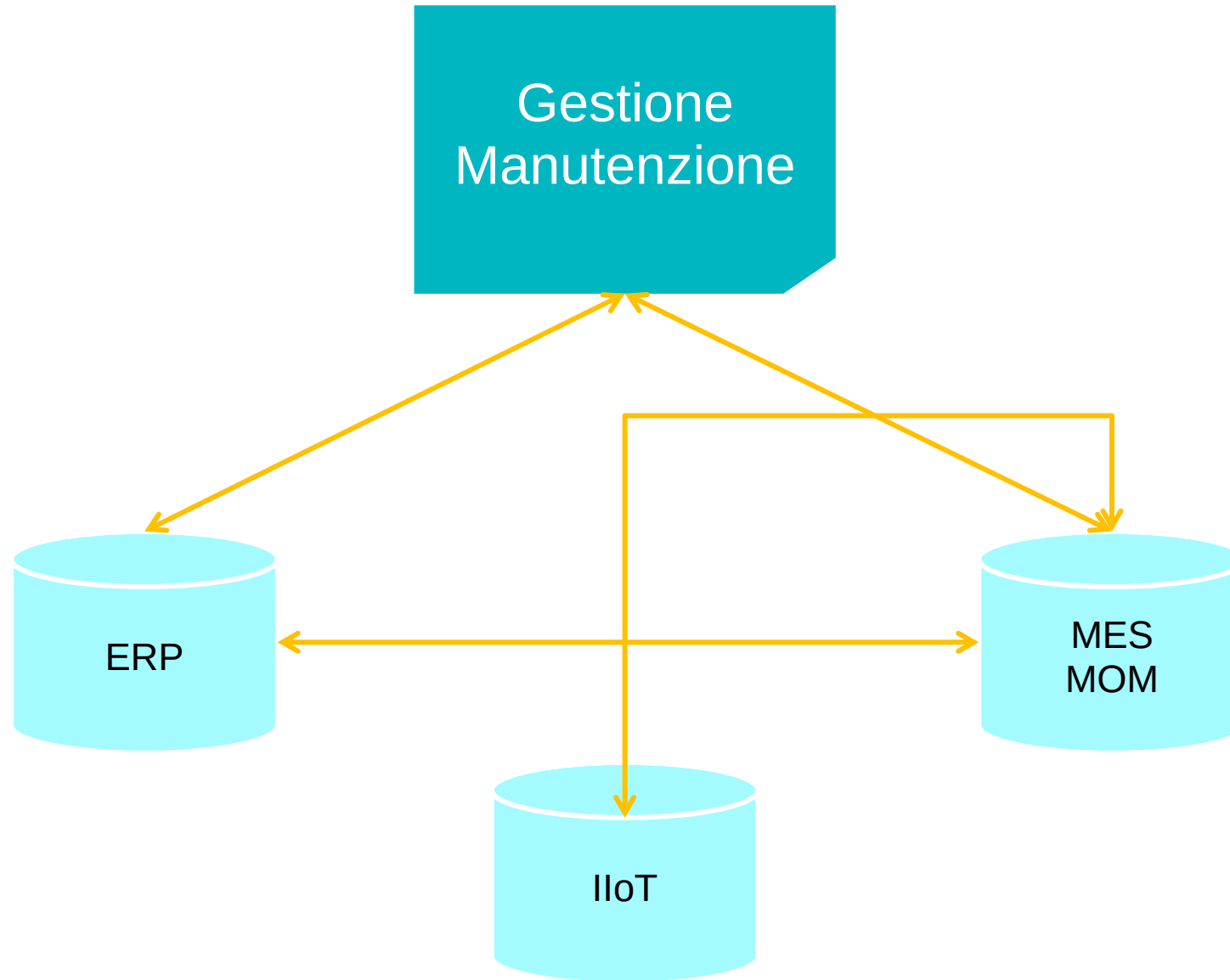
IIoT

ECommerce

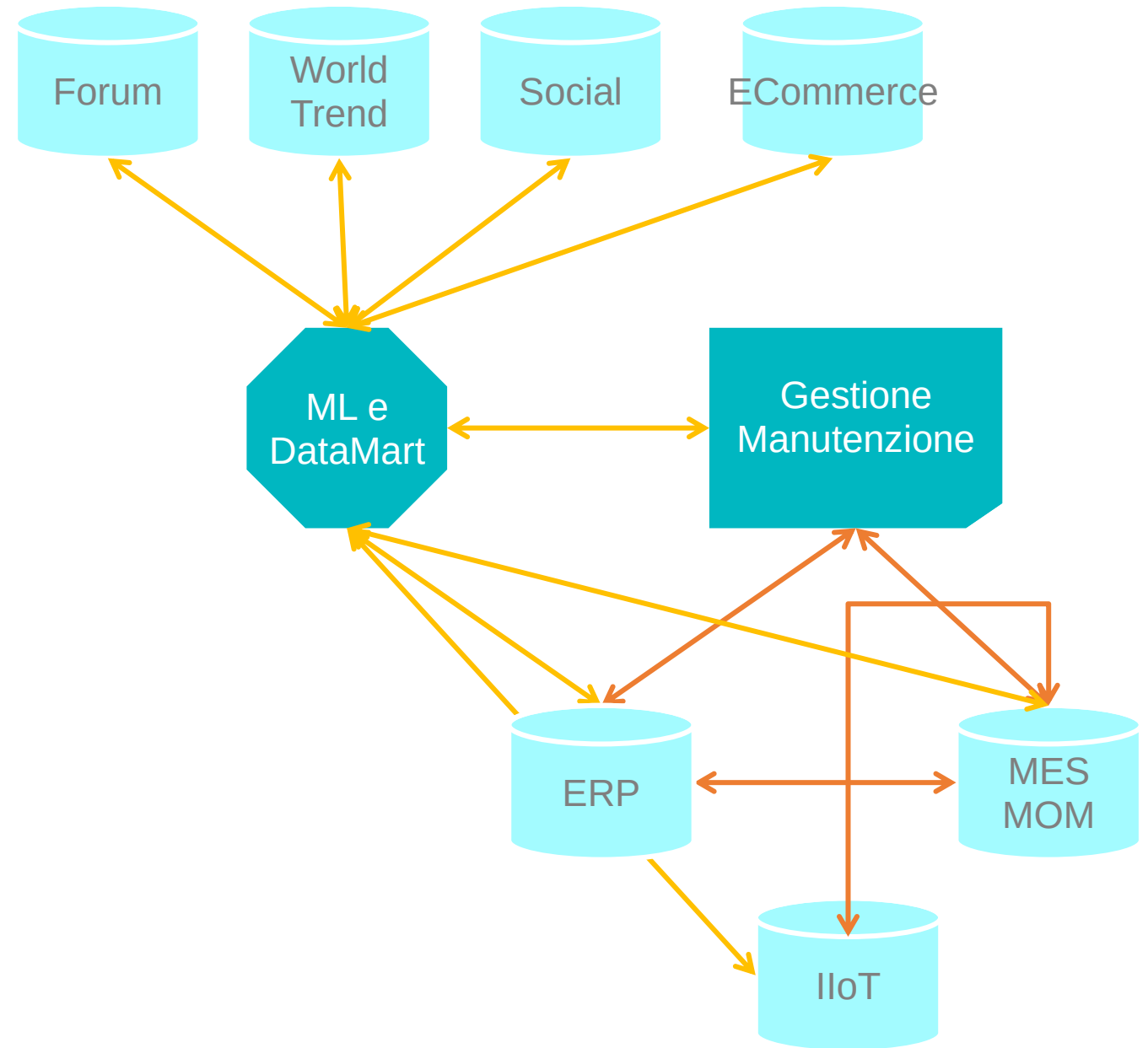
MOLTEPLICI FONTI INFORMATIVE " COLLEGATE O CORRELATE "



LE ESIGENZE INFORMATIVE

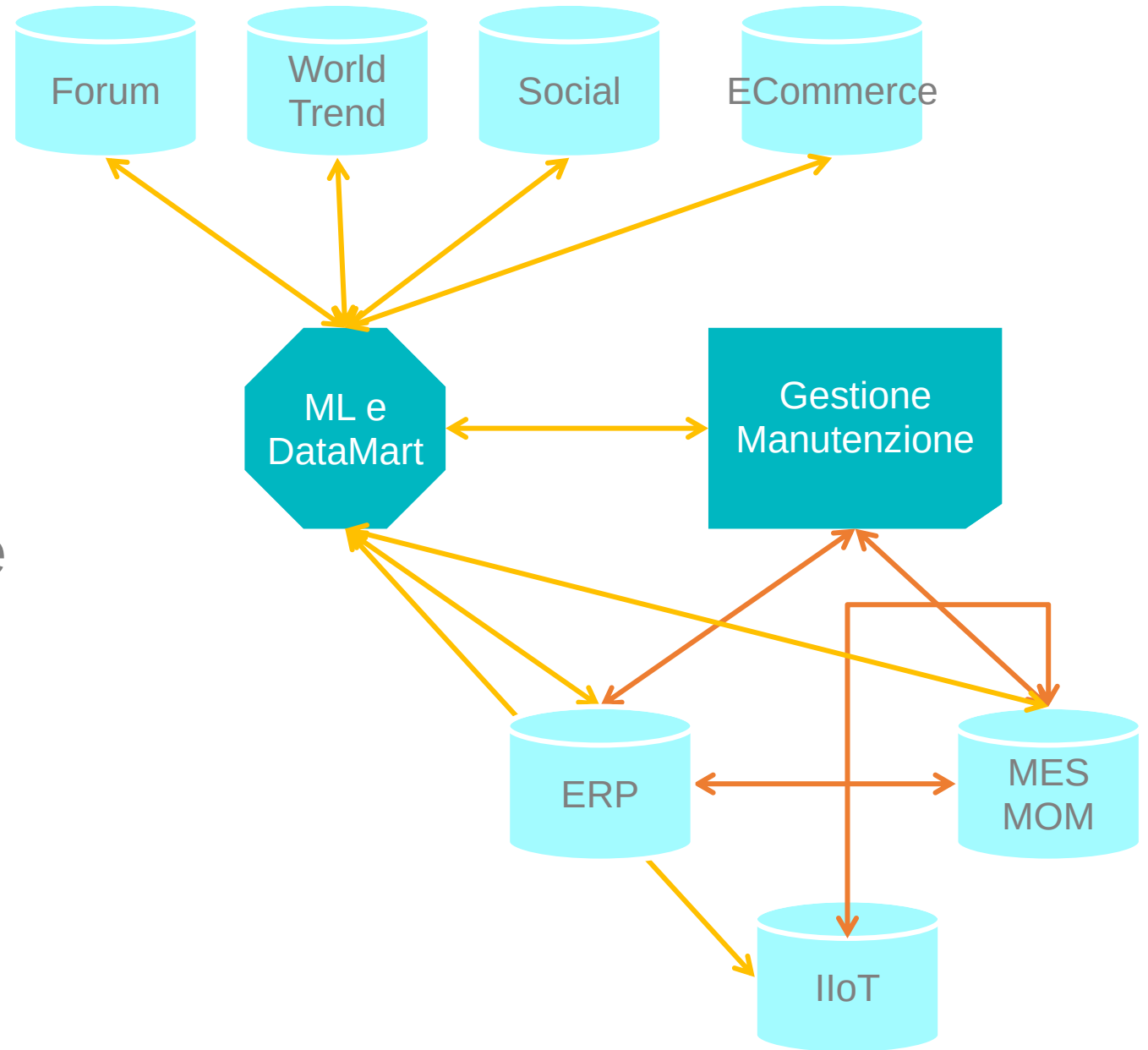


LE ESIGENZE INFORMATIVE

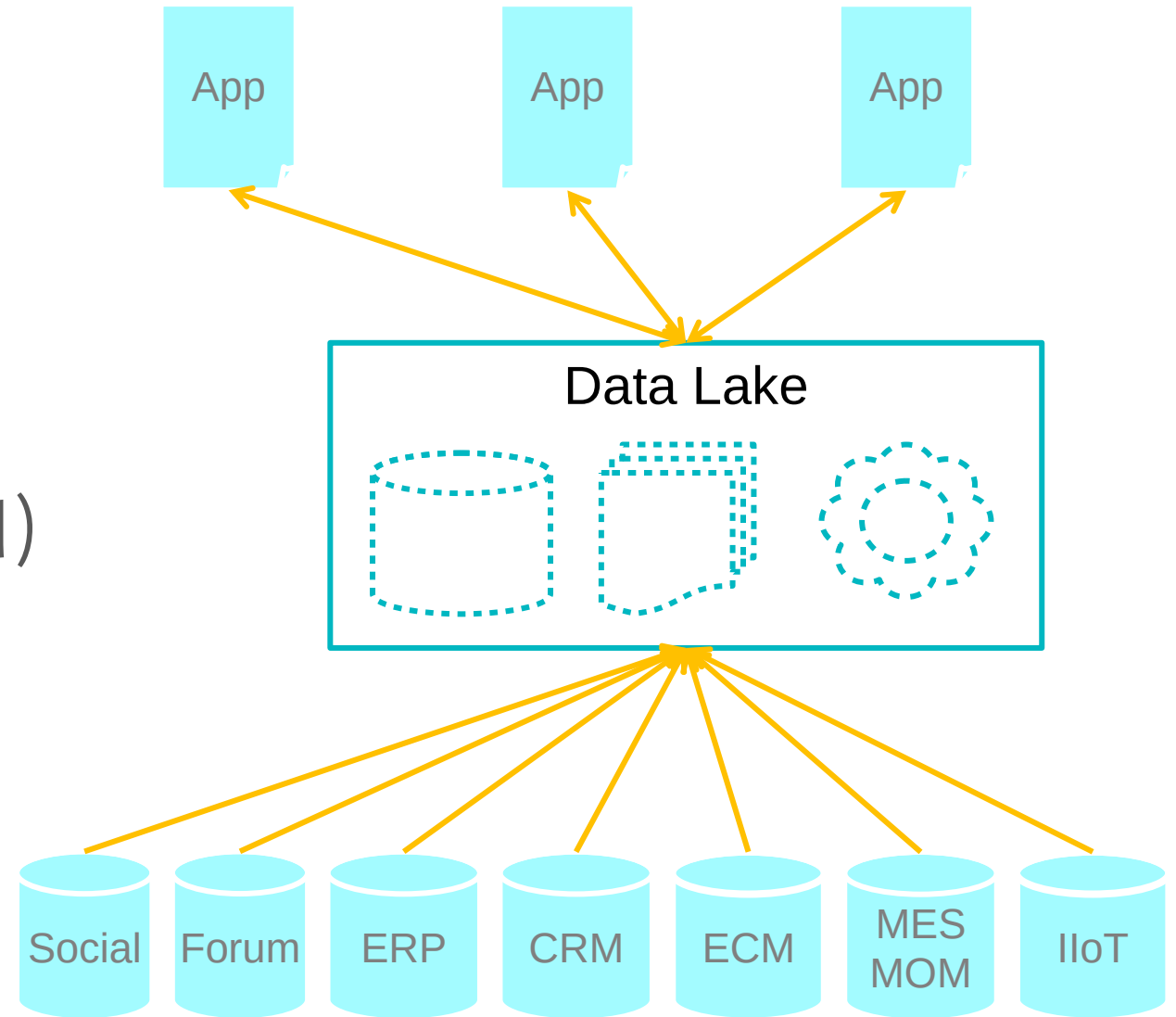


LE ESIGENZE INFORMATIVE

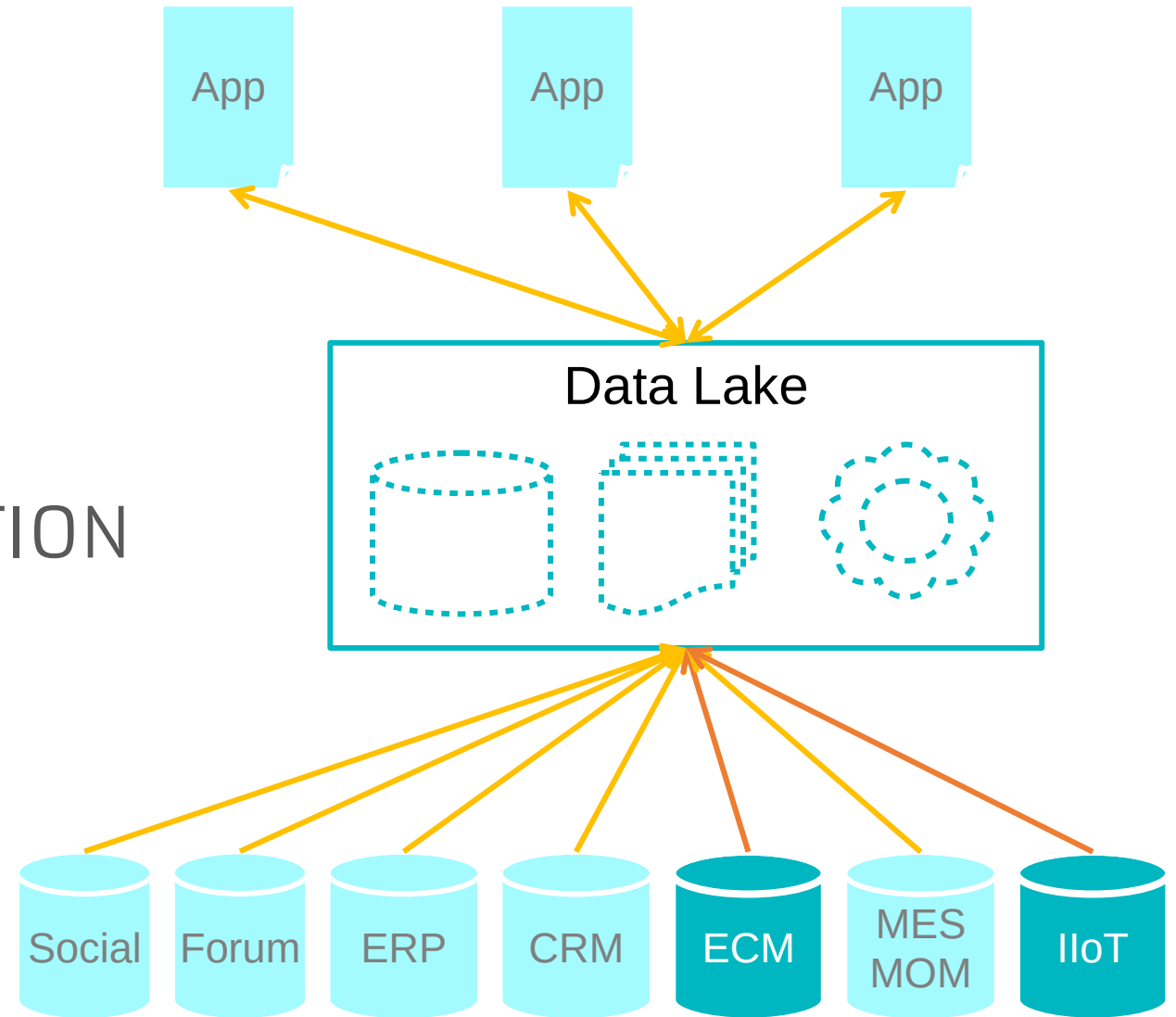
Schema On-write



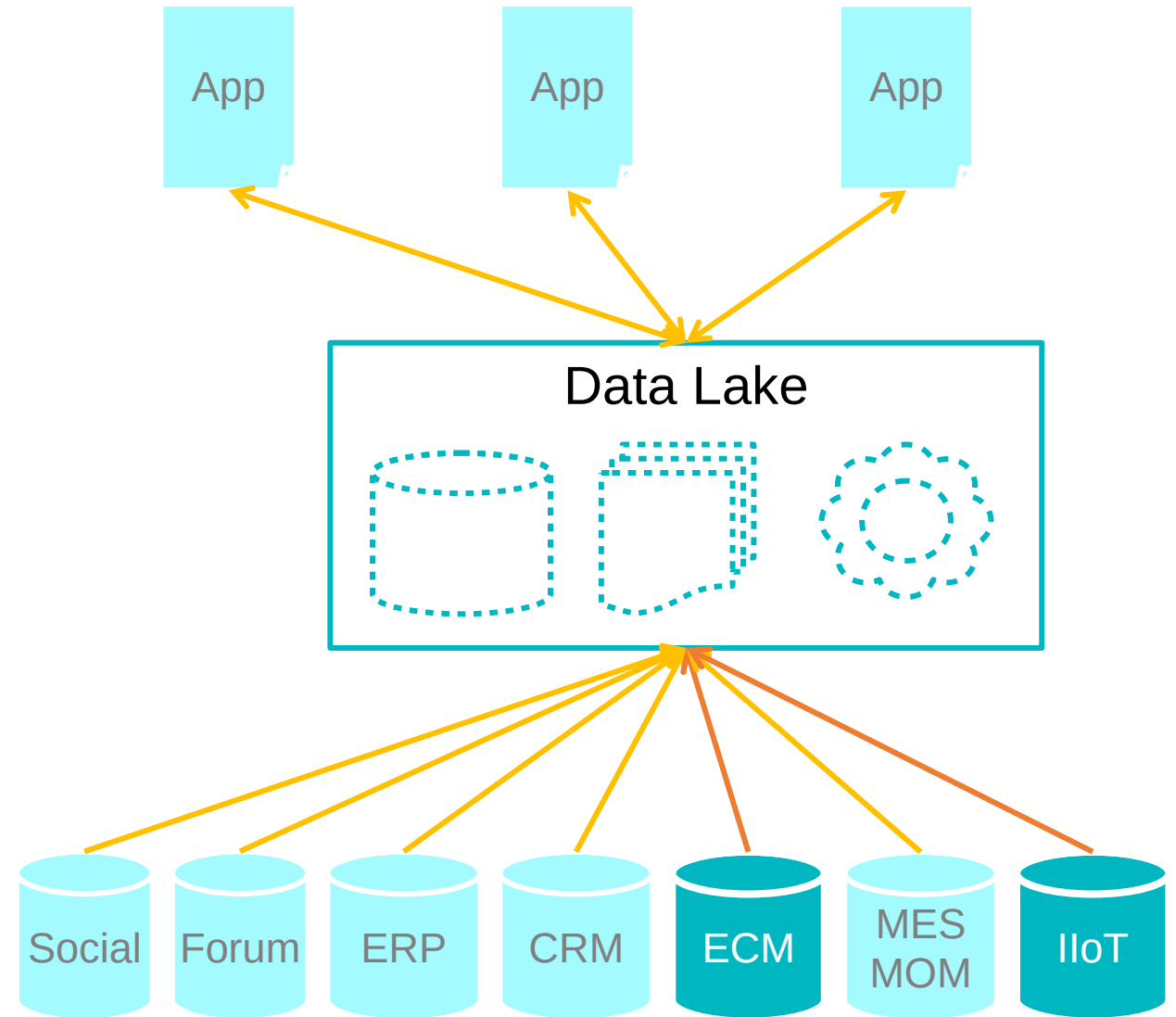
Data Lake (Schema on Read)



Data Lake DATA VIRTUALIZATION



Data Lake DATA ENGINE



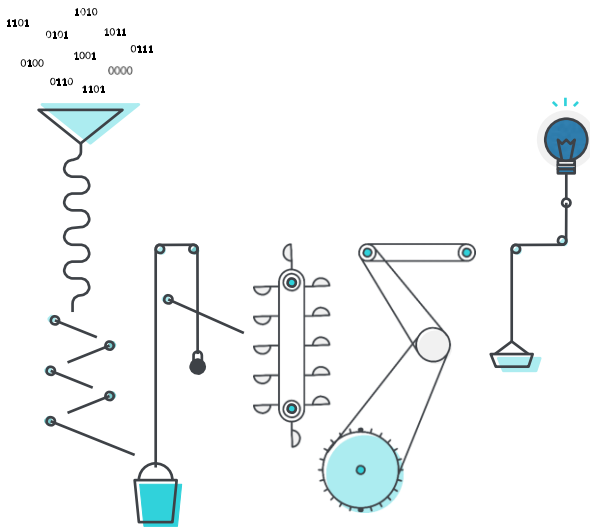
Data Lake vs Data Virtualization



Data Lake vs Data Virtualization

Le sfide dalla Data Analytics

Complessità senza precedenti



Aumento della domanda

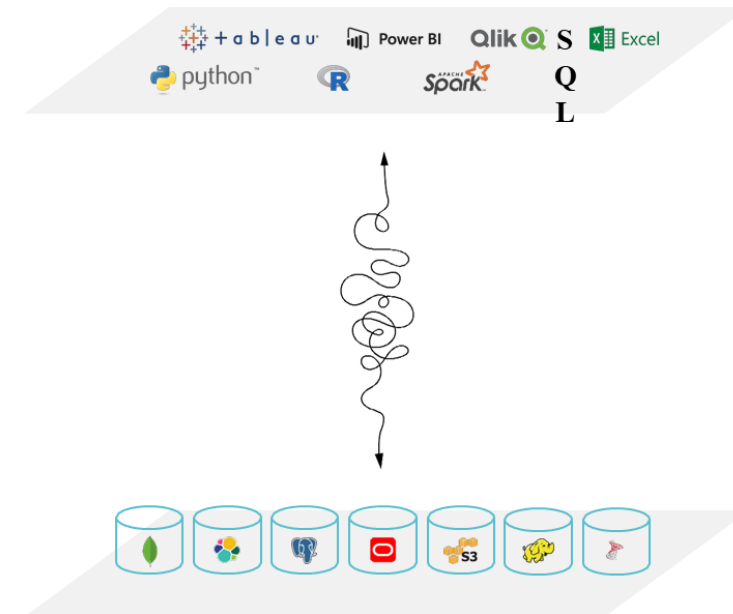
La necessità di disporre dei dati è cresciuta esponenzialmente



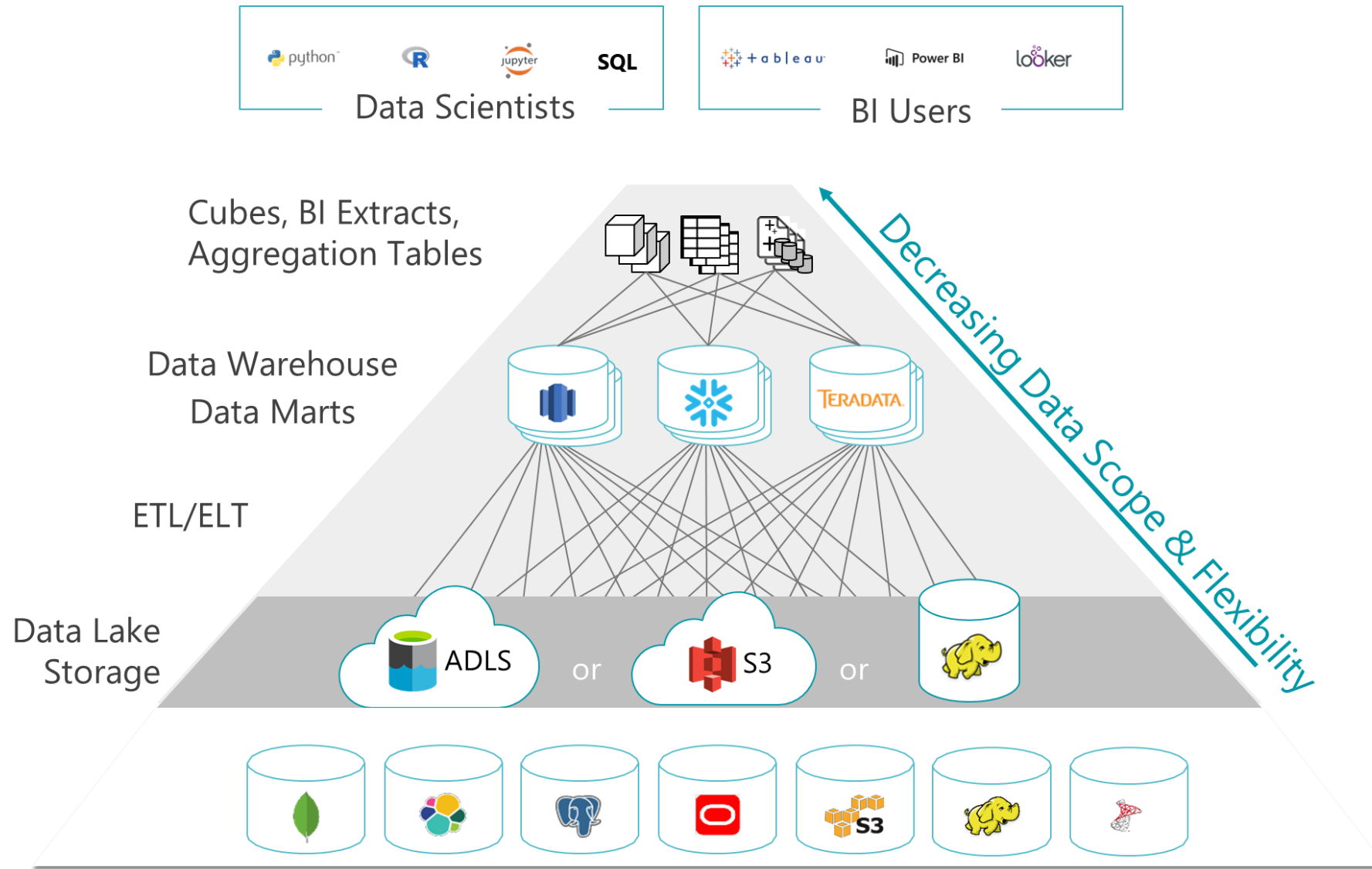
- Analisi predittive
- Rilevamento minacce
- Business Intelligence
- Nuovi prodotti
- Machine Learning
- Fraud Detection
- Prevenzione
- Reportistica

Fame di dati

Analisti e Data Scientist sono "affamati" di dati



Data Lake vs Data Virtualization



Data Lake vs Data Virtualization

The Data Lake is Exploding, Yet The Data Remains Inaccessible

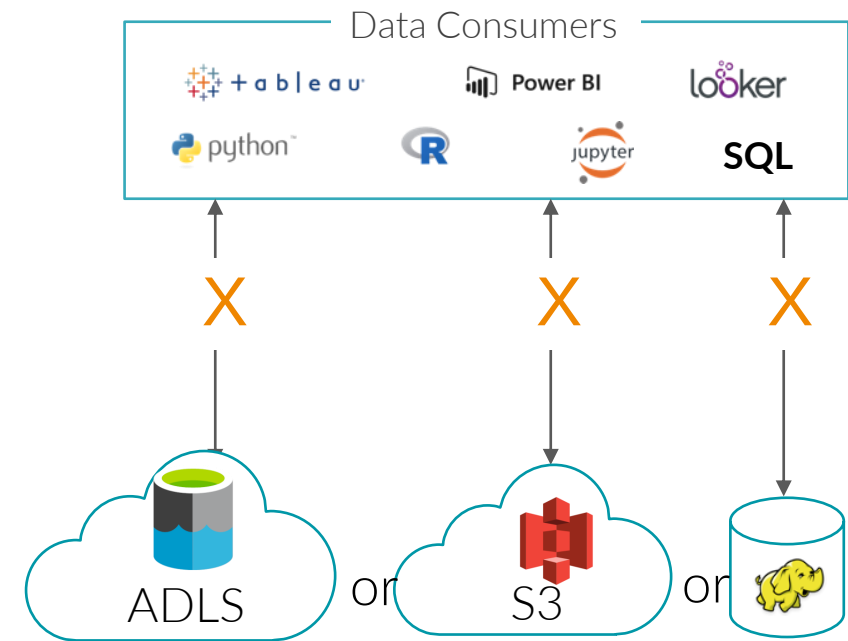
I Data Lakes stanno diventando il repository principali per i dati



But

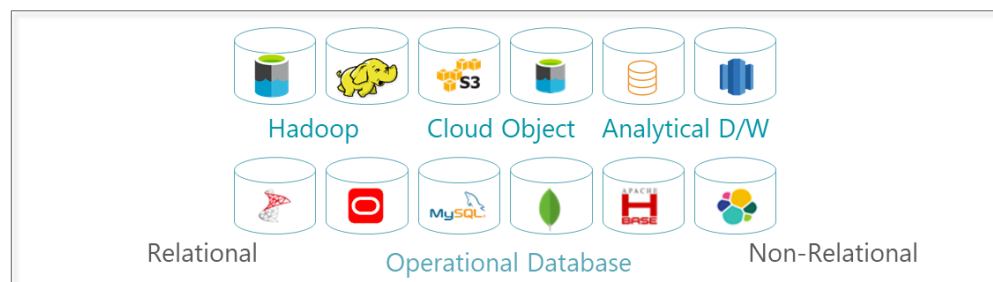
.....

L'accesso ai dati rimane lento e difficoltoso



Data Lake vs Data Virtualization

DREMIO: l'approccio Self-Service Data



Data Lake vs Data Virtualization

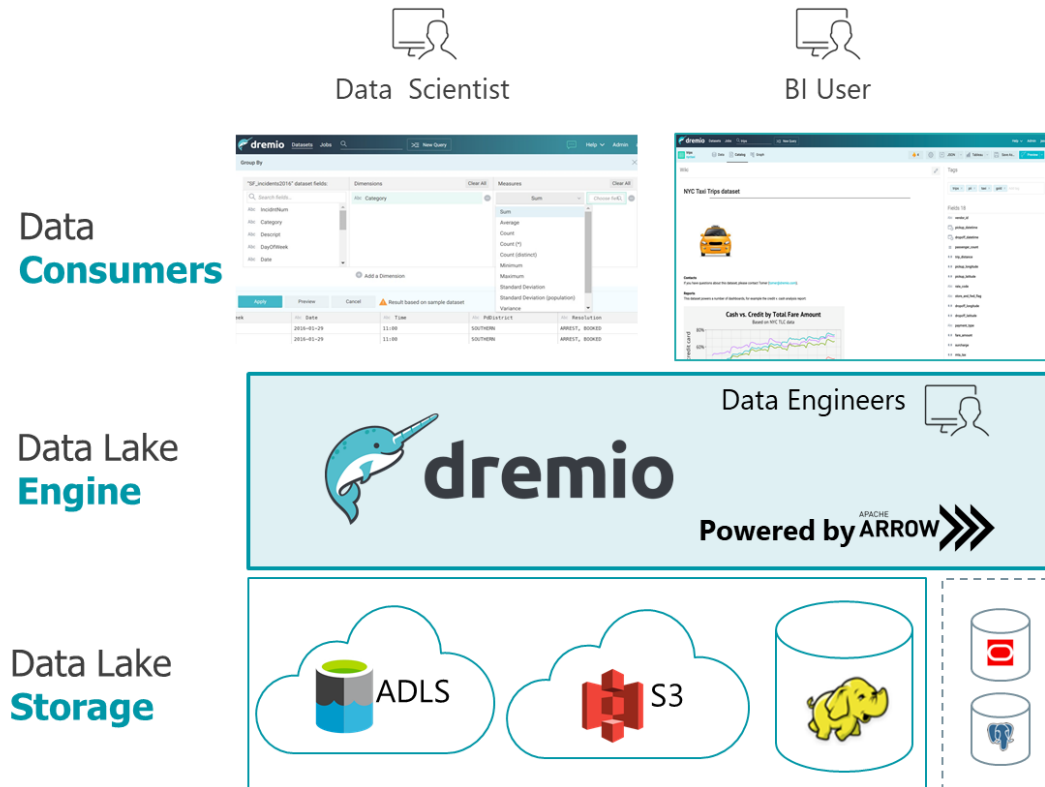
La complessità multi layer del dato



Data Lake vs Data Virtualization

Self Service Layer Enabling Virtual Data Sets, with an Easy-to-Use UI

Simplified Data Access



Data Consumers

- ✓ Consistent business logic & KPIs
- ✓ No more waiting for IT
- ✓ Easy data discovery & sharing

Data Engineers & Architects

- ✓ Centralize data security & governance
- ✓ No more reactive, tedious work
- ✓ Easy collaboration with data consumers

Data Lake vs Data Virtualization

La complessità multi layer del dato

Cubes, BI
Extracts &
Aggregation
Tables



Data
Warehouse



Data
Staging



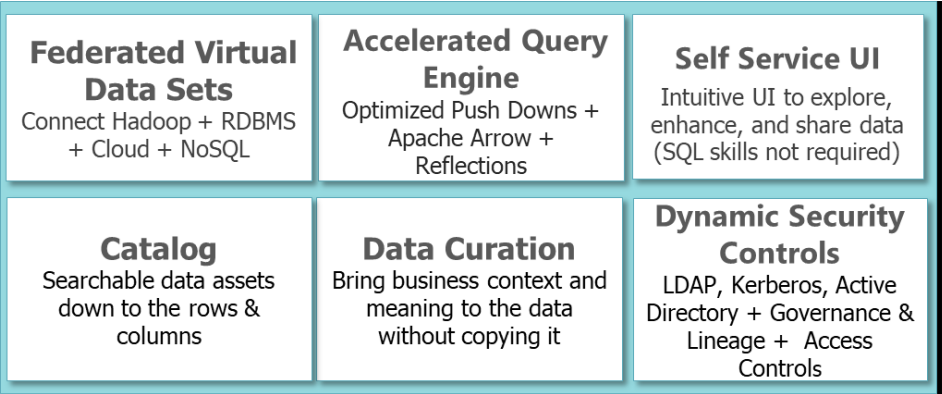
Data Lake vs Data Virtualization

Dremio's Data-as-a-Service Platform

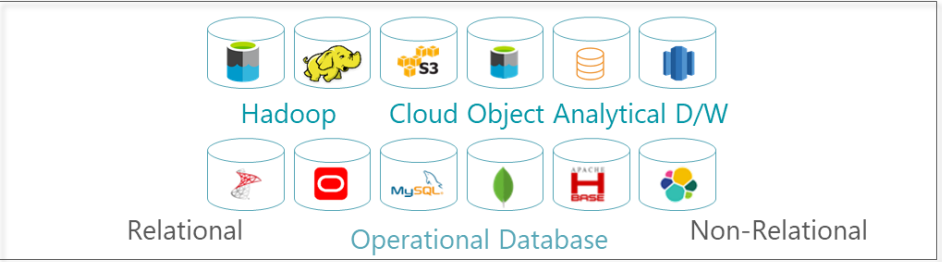
Data Consumers



Data Engineering



Data Sources



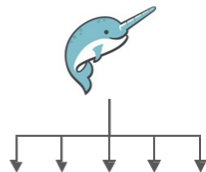
Data Lake vs Data Virtualization

**Dremio ottimizza i dati e velocizza le Query
in modo automatico, con incrementi di
prestazioni fino a 1000x**



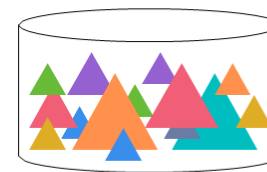
Apache Arrow Execution

Da 1 a 1000 nodi, è attivo su infrastruttura dedicata attraverso YARN



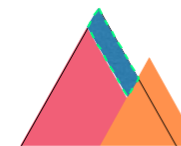
Native Push-Downs

Query semantiche ottimizzate per ogni sorgente dati: strutture relazionali, NoSQL, HDFS, Amazon S3 e molto altro



Data Reflections™

Strutture dati fisiche ottimizzate per operazioni di aggregazione



Universal Relational Algebra

Il pianificatore di Query esegue automaticamente aggiustamenti per migliorare l'uso della cache relazionale



BREAKFAST CHANNEL

Grazie



Segui @RIOS_opensource



<http://www.reteitalianaopensource.net>