Al supermercato con Azure Data Explorer

Sponsors

With the support of:

About me

Nicola Paro

Cloud Solutions Engineer

Linkedin: linkedin.com/in/nicolaparo

Github: github.com/nicolaparo

Twitter: @nicola_paro

DATA SATURDAYS

inn0va community

DATA SATURDAYS

A Real Life Story

A Real Life Story

Creerò degli scraper per leggere i prezzi dei prodotti giorno per giorno.

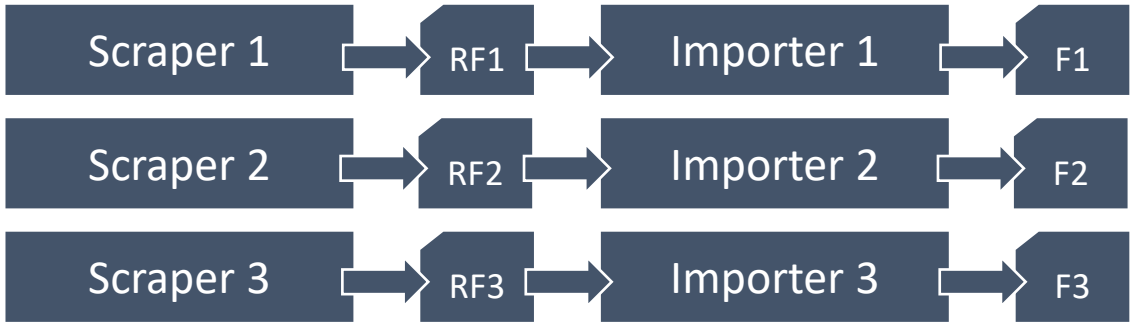
- Scraper 1
- Scraper 2
- Scraper 3

A Real Life Story

Salvo i dati estratti in alcuni RawFiles in formato TXT

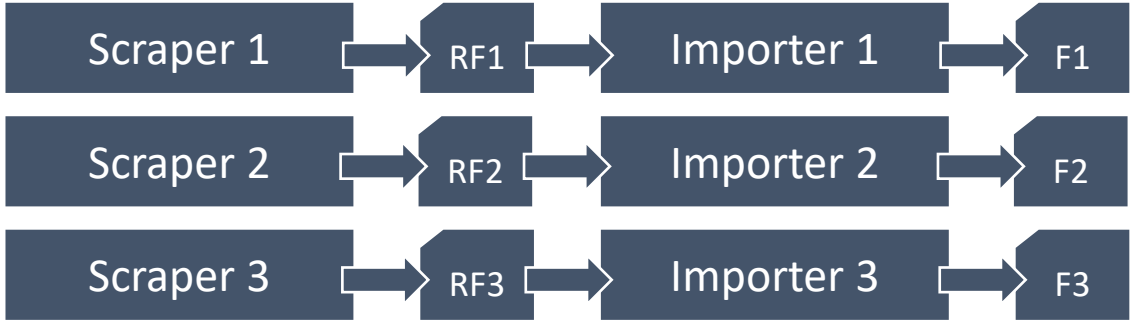
A Real Life Story

I miei importer estrarranno i dati rilevanti e li persisteranno in un formato condiviso

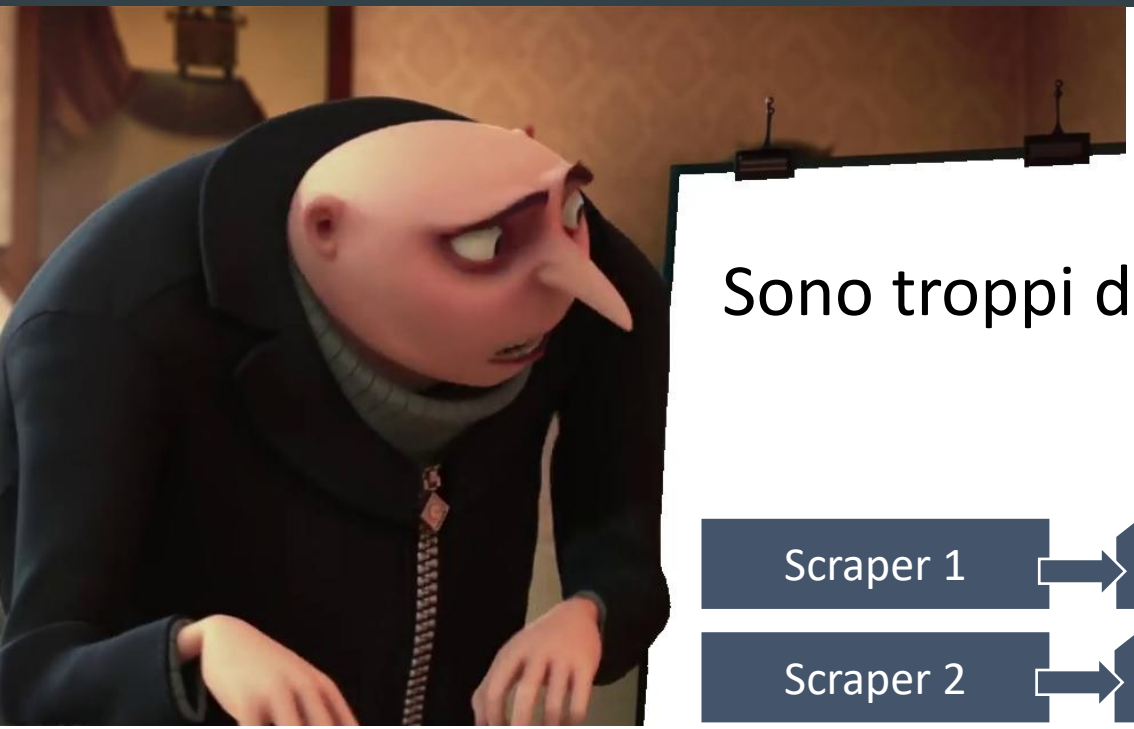


A Real Life Story

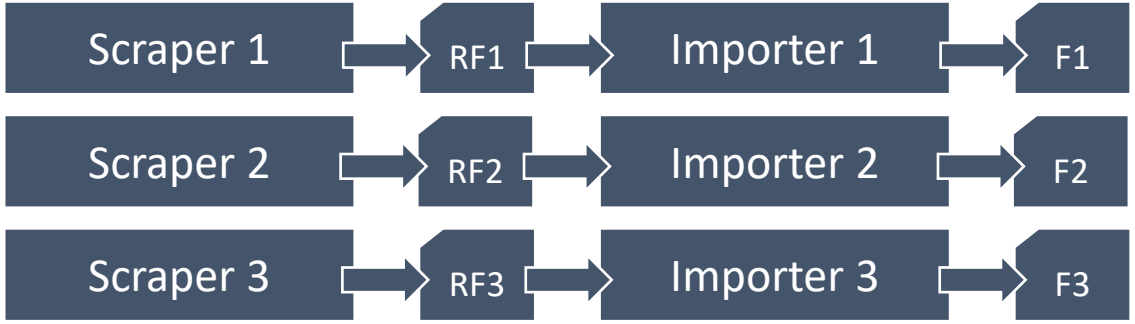
Sono troppi dati e non riesco a leggerli.



A Real Life Story

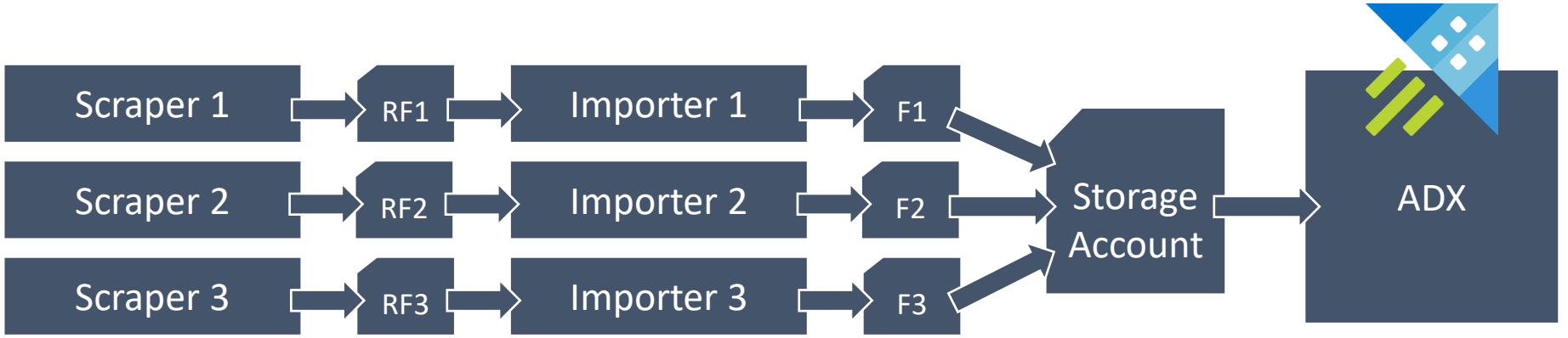


Sono troppi dati e non riesco a leggerli.



A Real Life Story

Caricherò i files su uno Storage Account e li analizzerò con Azure Data Explorer



Prima di fare Web Scraping

- **Controlla prima i termini di licenza** del sito web
- Non fare scraping di dati protetti da copyright
- Non fare scraping di contenuti protetti da login
  - Non fare scraping dopo aver fatto una login
- Analizza solamente informazioni disponibili pubblicamente
- Controlla il **/robots.txt** per essere sicuro di poter analizzare un determinato path
- Controlla la **/sitemap.xml** per un elenco agevolato delle pagine da analizzare

DATA SATURDAYS

Azure Data Explorer

Azure Data Explorer

Azure Data Explorer è un servizio di analisi dati veloce e totalmente gestito per l'analisi in tempo reale di grandi volumi di dati, provenienti da applicazioni, siti web, dispositivi IoT ecc…

Azure Data Explorer

Features Principali

- Fast Data Ingestion
- Interactive Data Exploration
- Real Time Analytics on Streaming data
- Scalability
- Integration with other Azure Services

Azure Data Explorer

Quando ha senso utilizzare ADX?

Sliding Window of data	Tante letture	Tanti Insert / Append	Poche Delete	NESSUN Update

Azure Data Explorer

Pricing

ADX Cost + **VMs Cluster Cost** + **Storage Cost**

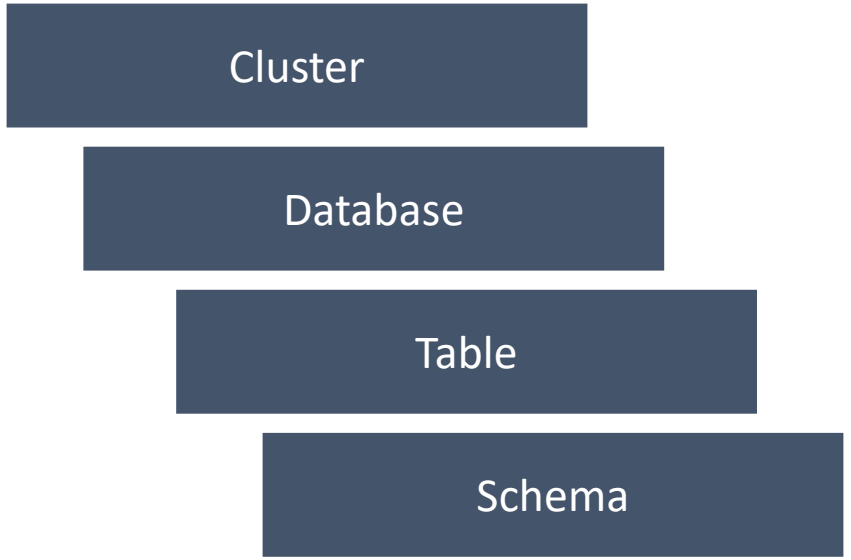
ADX Cost	VMs Cluster Cost	Storage Cost
Dev/Test **Free** (No SLA) Standard $0.11/core per hour	Varies on VM Size $0.126/hour for the smallest VM	Varies by usage

More info on https://azure.microsoft.com/en-us/pricing/details/data-explorer/

Azure Data Explorer

Organizzazione dei dati

Simile ai database relazionali (ex: SqlServer)



Organizzazione dei dati – Differenze rispetto ai RDMS

- No Primary Key
- No Unique Keys
- No Foreign Keys
- Columnstore Indexes
- Data Sharding (Extents)

Azure Data Explorer

Data Ingestion

Azure Data Explorer

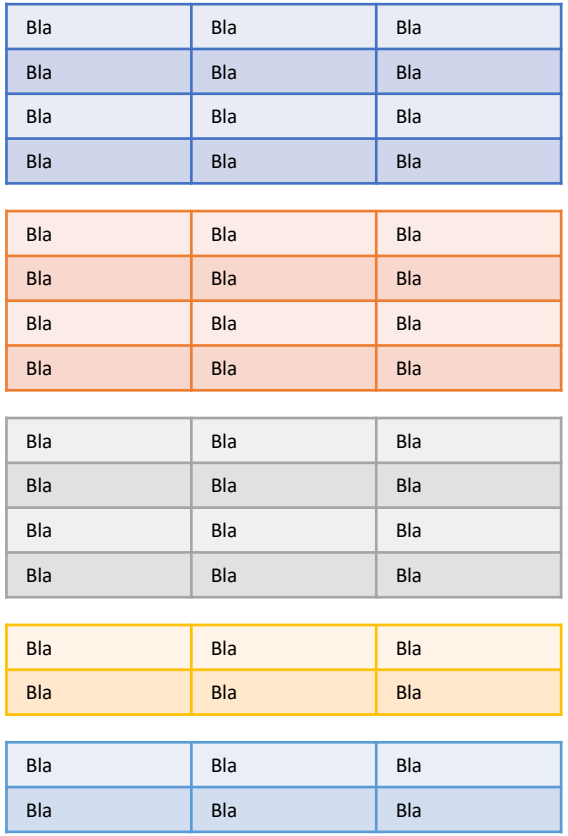
Dietro le quinte

Bla	Bla	Bla
Bla	Bla	Bla
Bla	Bla	Bla
Bla	Bla	Bla
Bla	Bla	Bla
Bla	Bla	Bla
Bla	Bla	Bla
Bla	Bla	Bla
Bla	Bla	Bla
Bla	Bla	Bla
Bla	Bla	Bla
Bla	Bla	Bla
Bla	Bla	Bla
Bla	Bla	Bla
Bla	Bla	Bla
Bla	Bla	Bla

Azure Data Explorer

Dietro le quinte

I dati delle tabelle sono divisi in extents
(aka shards, partizioni, ...)



Azure Data Explorer

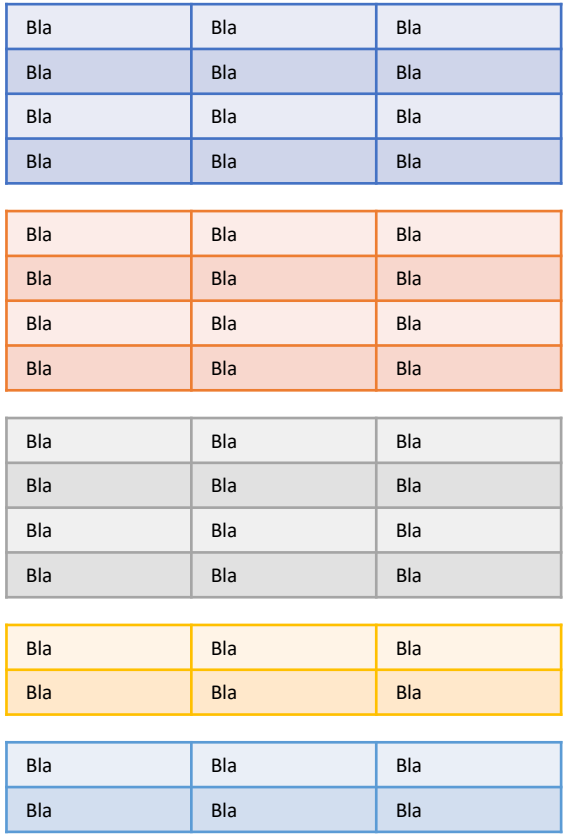
Dietro le quinte

I dati delle tabelle sono divisi in extents (aka shards, partizioni, ...)

Un extent è una mini-tabella che contiene dati e metadati.

Un extent **non può mai essere modificato**, ma può essere cancellato

I dati sono organizzati in colonne



Dietro le quinte

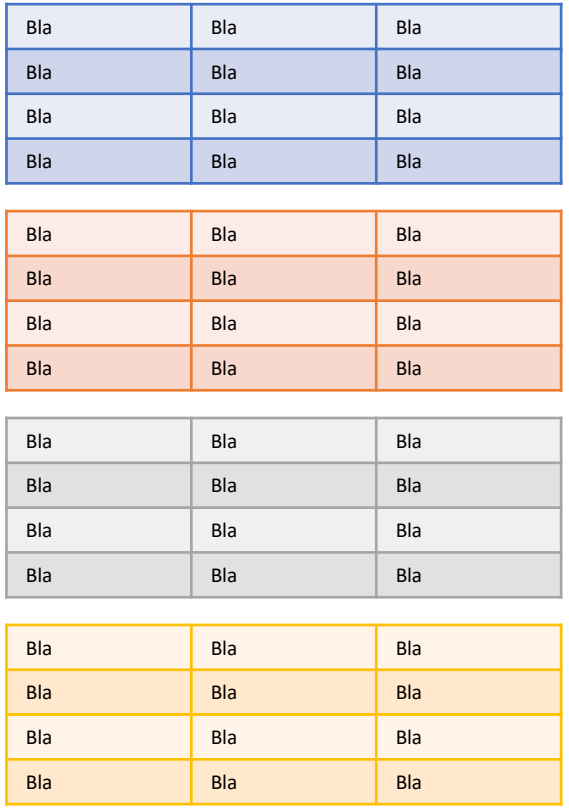
I dati delle tabelle sono divisi in extents (aka shards, partizioni, ...)

Un extent è una mini-tabella che contiene dati e metadati.

Un extent **non può mai essere modificato**, ma può essere cancellato

I dati sono organizzati in colonne

Extent più piccoli possono essere uniti in extent più grandi



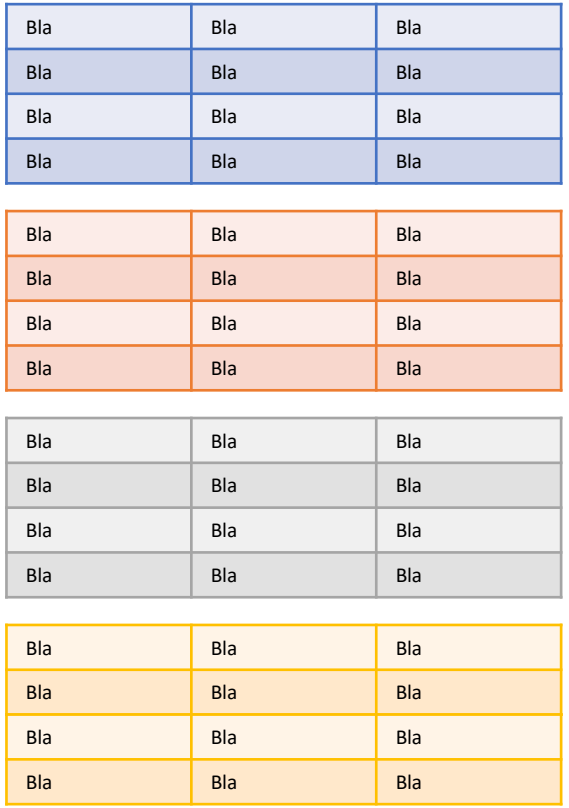
Dietro le quinte

Gli extent sono creati durante le operazioni di inserimento

Un extent è unito ad altri

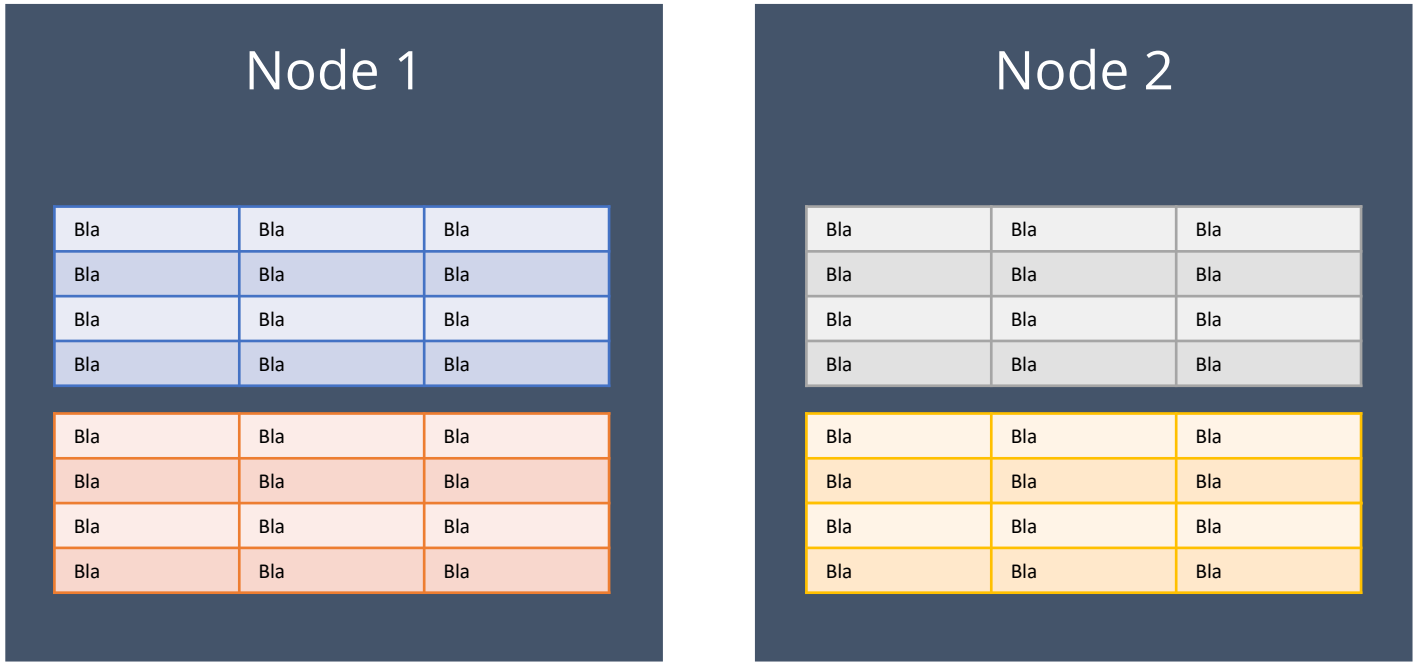
- Shard rebuild
- Shard merge

Un extent può essere cancellato con una retention-policy



Dietro le quinte

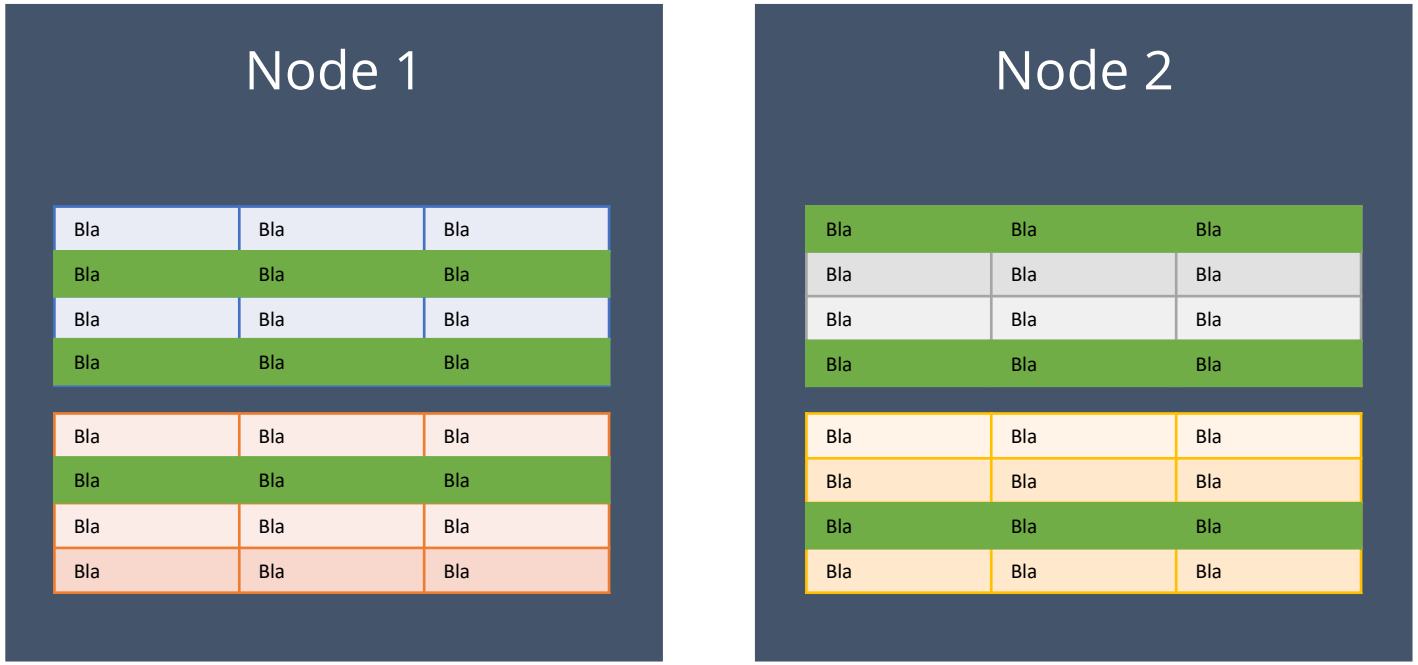
Gli shards sono distribuiti tra i nodi del cluster



Una “semplice query” nel cluster

Logs

| where Timestamp > ago(1h)

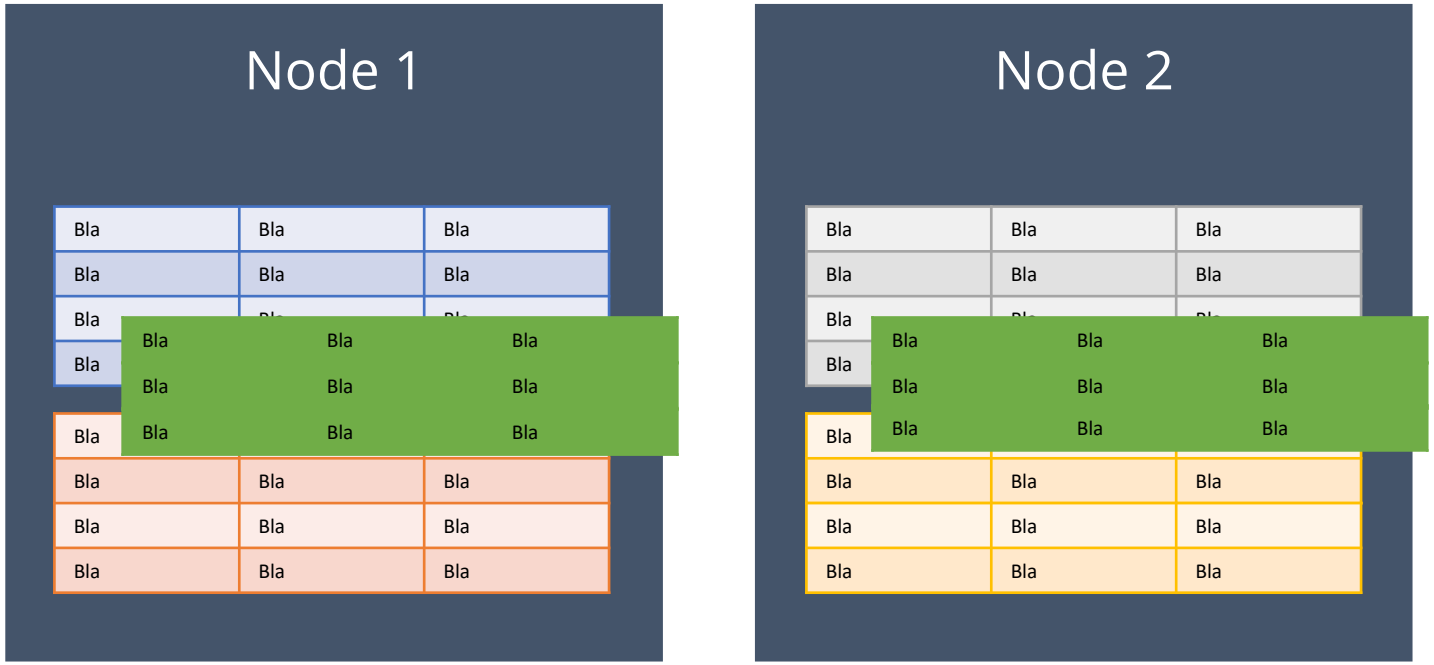


Azure Data Explorer

Una “semplice query” nel cluster

Logs

| where Timestamp > ago(1h)

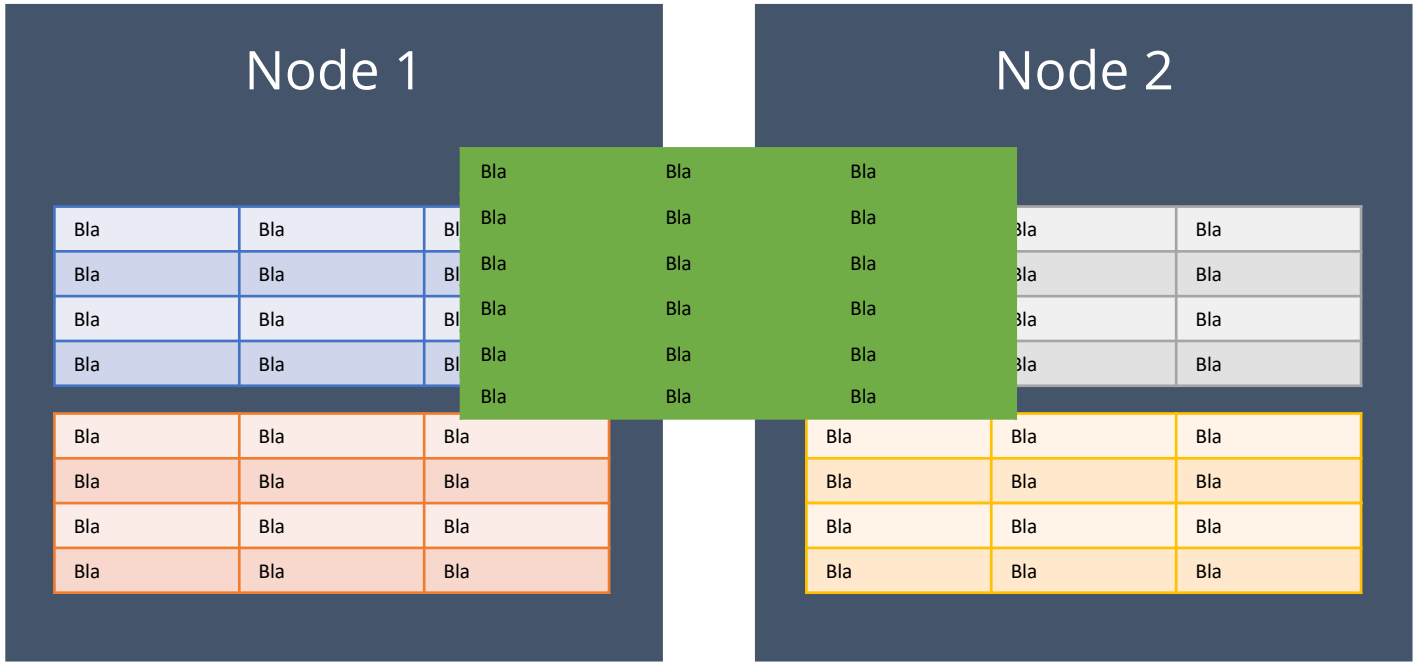


Azure Data Explorer

Una “semplice query” nel cluster

Logs

| where Timestamp > ago(1h)



Kusto Query Language

Una query Kusto è una richiesta in sola lettura per il processamento dei dati e la produzione di risulati.

La richiesta è effettuata tramite testo, utilizzando un modello di data-flow che è semplice da leggere, scrivere ed automatizzabile.

Le query Kusto sono composte di una o più istruzioni.

Kusto Query Language

Una query Kusto è una richiesta in sola lettura per il processamento dei dati e la produzione di risulati.

La richiesta è effettuata tramite testo, utilizzando un modello di data-flow che è semplice da leggere, scrivere ed automatizzabile.

Le query Kusto sono composte di una o più istruzioni.

Azure Data Explorer

Kusto Query Language

SQL	KQL
SELECT	**project**, extend, project-away, project-keep …
WHERE	**where**, search, …
JOIN	**join kind=inner**
UNION	**union**
GROUP BY	**summarize**
ORDER BY	**sort by**, order by, top by
TOP, LIMIT	**take**

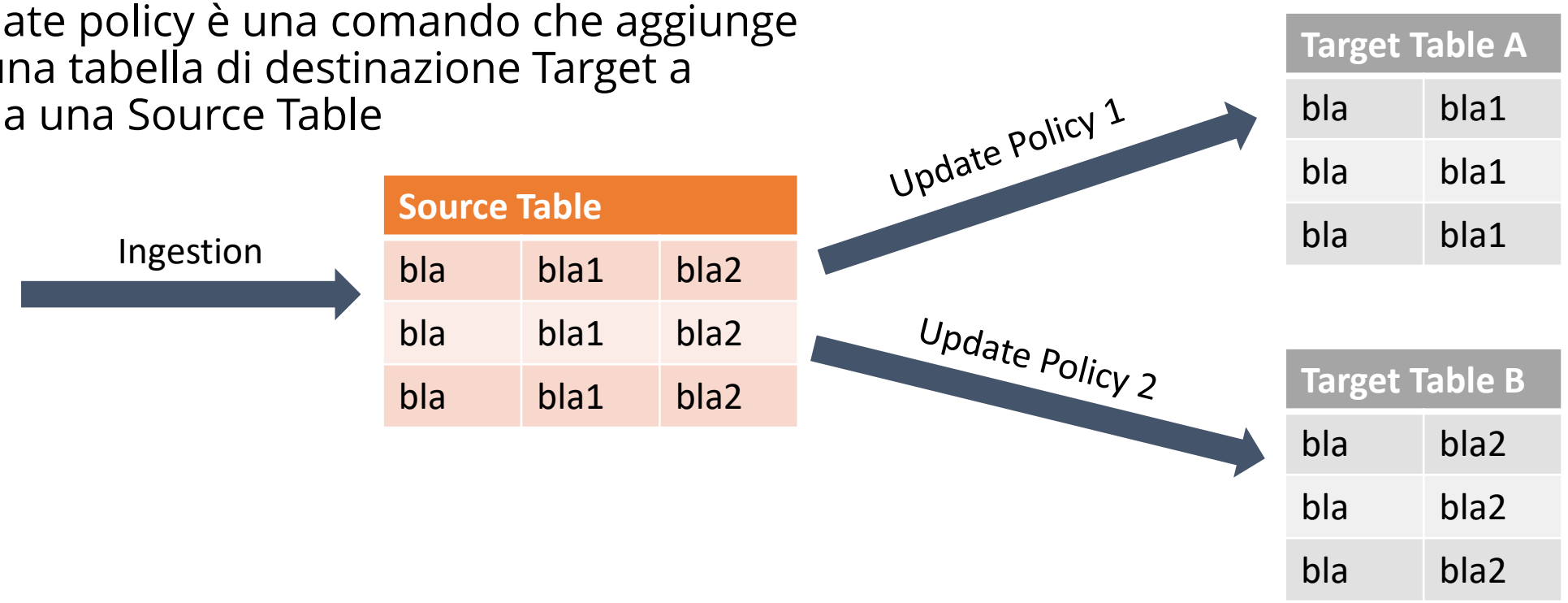
More on https://learn.microsoft.com/en-us/azure/data-explorer/kusto/query/

DATA SATURDAYS

Demo

Update Policies

Una update policy è una comando che aggiunge dati ad una tabella di destinazione Target a partire da una Source Table



Materialized View

Una Materialized View è una vista aggregate sui dati di una tabella ADX. I dati sono materializzati anche su disco.

Vantaggi nell'adozione delle Materialized View

- Performance Improvement
- Data Freshness
- Cost Reduction

Azure Data Explorer

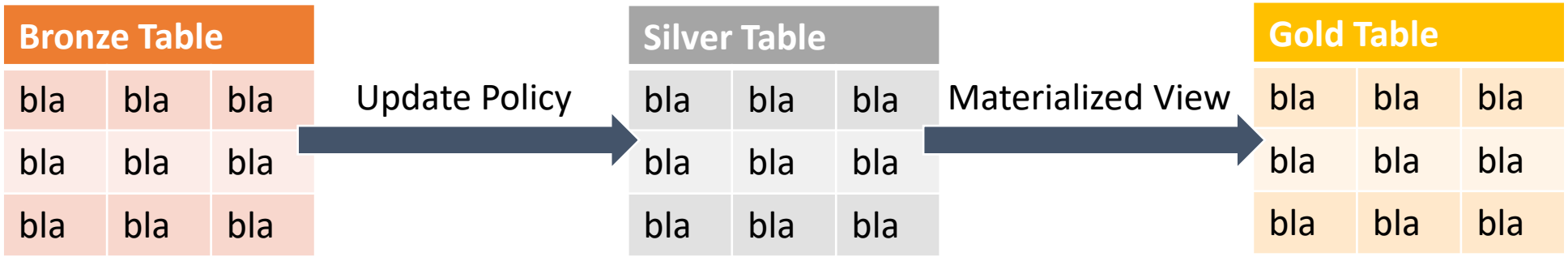
Update Policy o Materialized View?

Update Policy

- Data Transformation
- Data Enrichment

Materialized View

- Data Aggregation

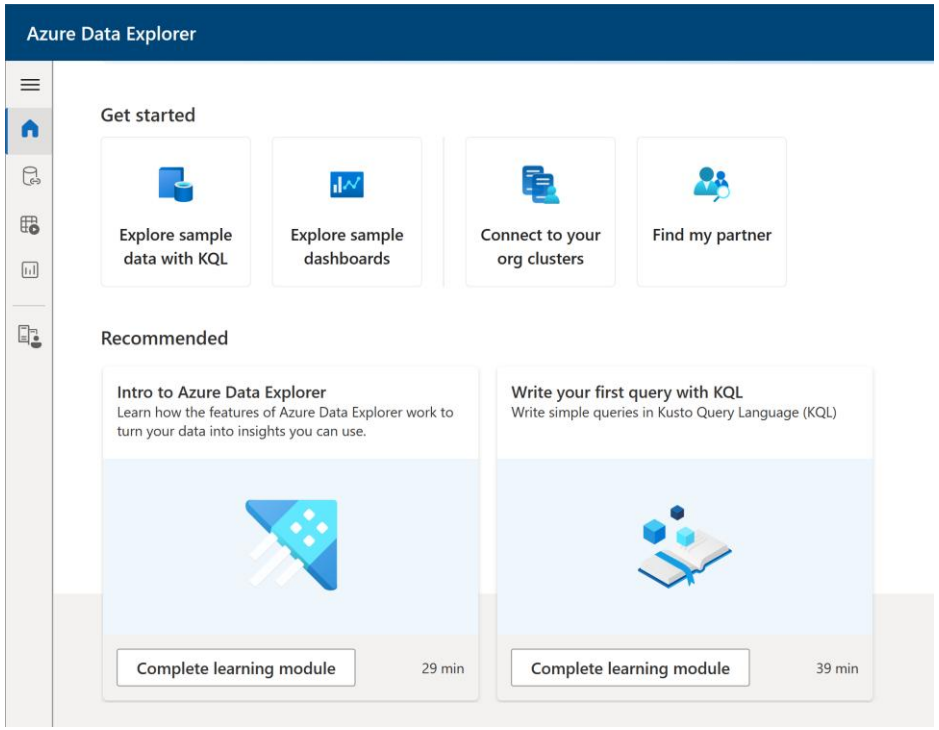


Azure Data Explorer

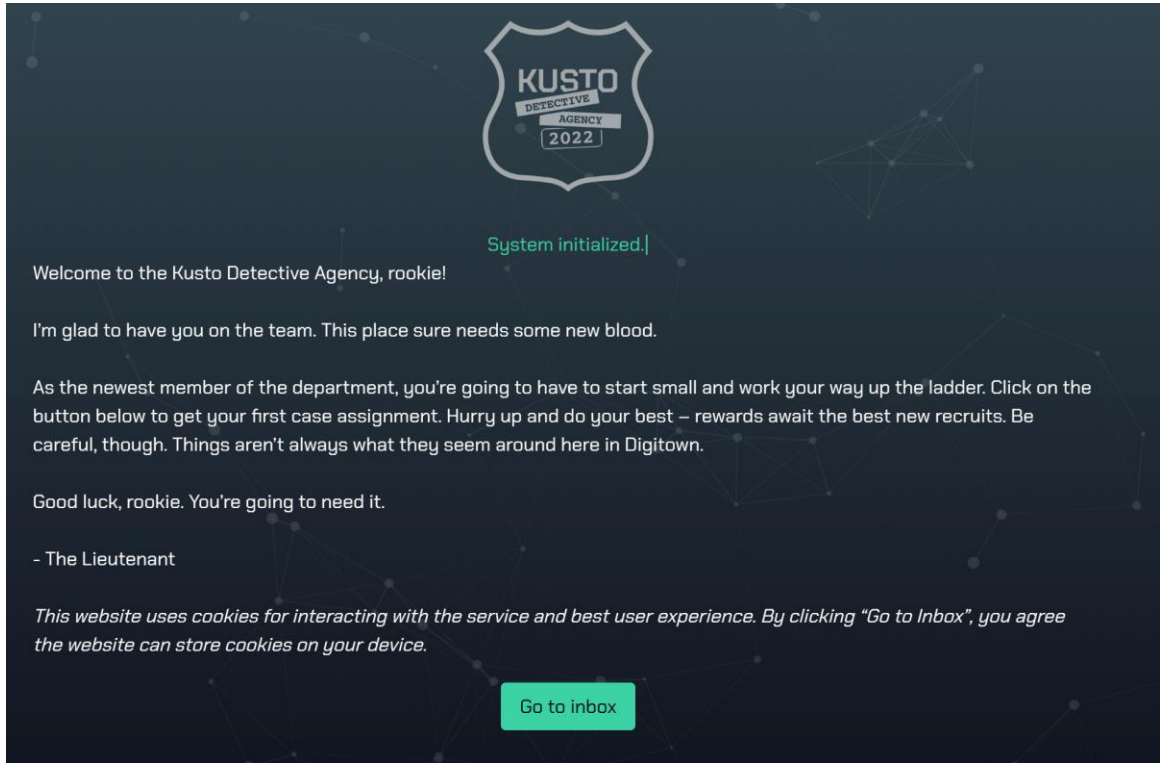
Devo pagare un cluster per fare pratica con Kusto?

Azure Data Explorer supporta dei database “sample” gratuiti su cui è possibile effettuare delle interrogazioni per provare

https://dataexplorer.azure.com/home



Devo pagare un cluster per fare pratica con Kusto?



Kusto Detective Agency: una gamification per imparare ad usare kusto.

https://detective.kusto.io/

DATA SATURDAYS

Q & A

DATA SATURDAYS

Thank You!

DATA SATURDAYS

Link e Contatti

Nicola Paro

Cloud Solutions Engineer

With the support of:

inn0va community · altitudo · beanTech IT moves your business · fluentis the ERP ready to go live! · QUANTUMDATIS · [stesi] Powered by Innovation · CREDEM · CREDEMTEL · packt