

Teoria

20/05/2023

Cyber Intelligence

- prof. Maurizio Tesconi
- esame: 28 luglio, consegna progetto + discussione, gruppi di 3 persone max. Realizzazione di un semplice sistema di cyber intelligence con 3 componenti principali: raccolta dati (social, scraping web, open data), analisi (sentiment, word freq, interaction) + visualizzazione (analytics, charts, graph). Utilizzo NIFI, ElasticSearch e Kibana



PROPOSTE DI PROGETTO

- Formare i gruppi
- Pensare una proposta
- Inserirla in questo documento

<https://docs.google.com/document/d/1peh1Ezq2mFQX1xFu-r6fOQW2D6bkgOUH1kyYR4g9qfA/edit>

Progetti Master Cyber Security

Corso Cyber Intelligence 2022-2023

NOTA: **1 progetto per ogni pagina**, in nero i progetti DA CONSEGNARE, in blu i progetti CONSEGNATI

Copiare il template seguente e riempirlo con i dati del progetto

Titolo progetto	
Team	
Abstract	

2

Intelligence

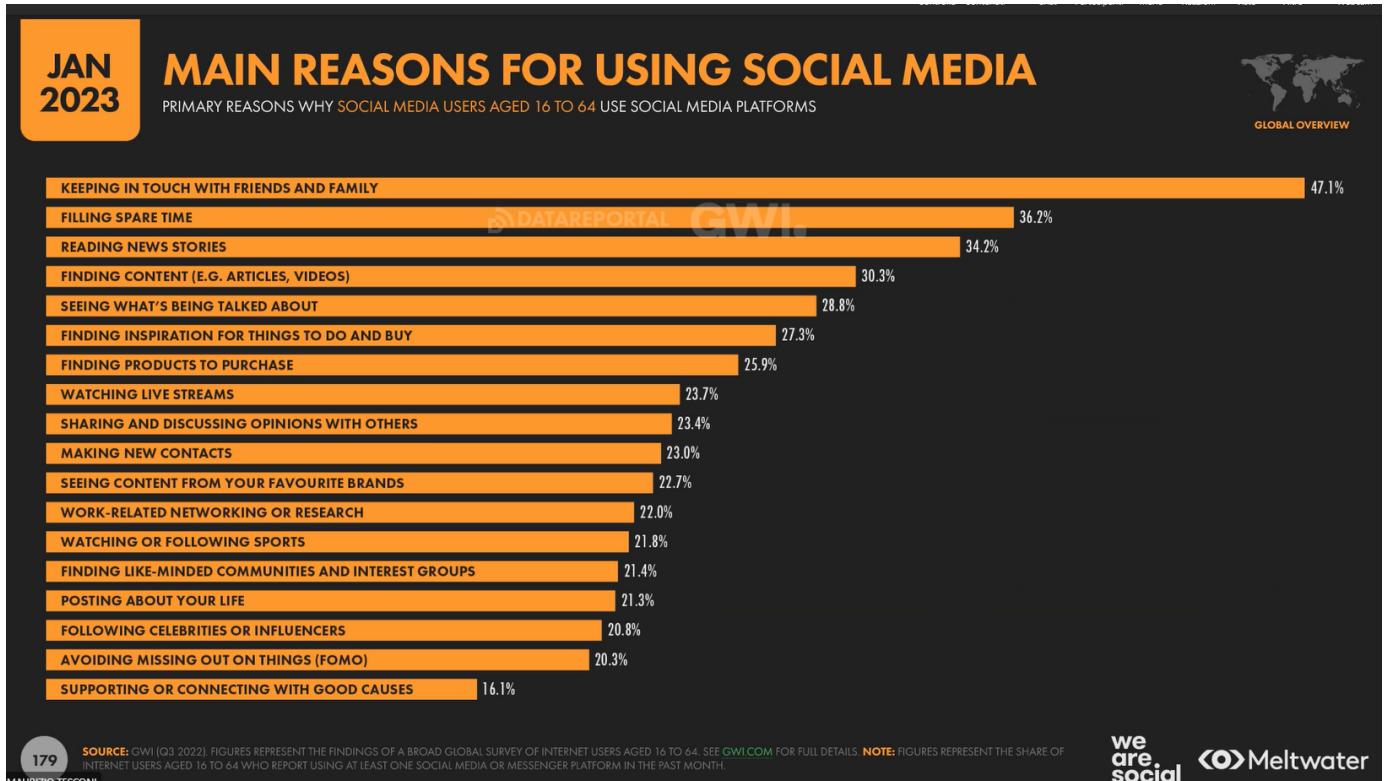
- Servizio (per lo più segreto o riservato) di raccolta di informazioni su persone o enti. (Treccani)
 - Servizio di spionaggio e di controspionaggio. (Corriere)
 - Insieme delle funzioni, delle attività e degli organismi coinvolti nel processo di pianificazione, ricerca, elaborazione e disseminazione di informazioni di interesse per la sicurezza nazionale. (Glossario di Intelligence)
-
- Raccolta e la successiva analisi di notizie e dati dalla cui elaborazione ricavare [informazioni](#) utili al processo decisionale [militare](#), nonché a quello relativo alla [sicurezza nazionale](#) ed alla prevenzione di attività destabilizzanti di qualsiasi natura. (Wikipedia)
 - Il prodotto dell'elaborazione di una o più notizie di interesse per la sicurezza nazionale. (Glossario di Intelligence)

Cyber intelligence

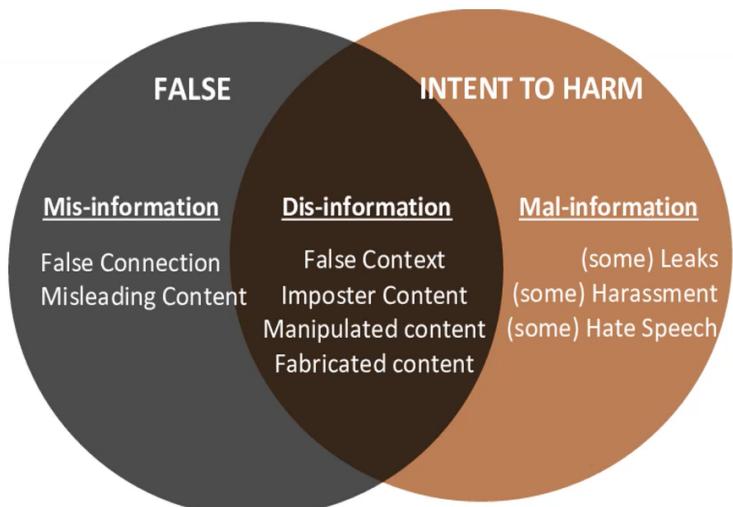
- OSINT: Open Source Intelligence – Ricerca ed elaborazione di notizie da *fonti aperte*.
- VHUMINT: Virtual HUMINT – Raccolta di informazioni di intelligence attraverso operatori virtuali, avatar e identità online
- SOCMINT: strumenti di raccolta e soluzioni che permettono di monitorare canali social e conversazioni, rispondere a segnali sociali e sintetizzare social data point in trend e analisi basate sui bisogni degli utenti.
Termine inventato da David Omand, esperto di intelligence, nel 2012

Il VHUMINT si tratta di un inserimento all'interno di altre reti sociali usando account fake impersonificando quella identità digitale per potersi inserire in gruppi per fare ascolto, fare controffensiva, molto similmente a quello che in modo fisico viene fatto con l'agente sotto copertura.

? Le principali ragioni per cui gli utenti usano i social network?

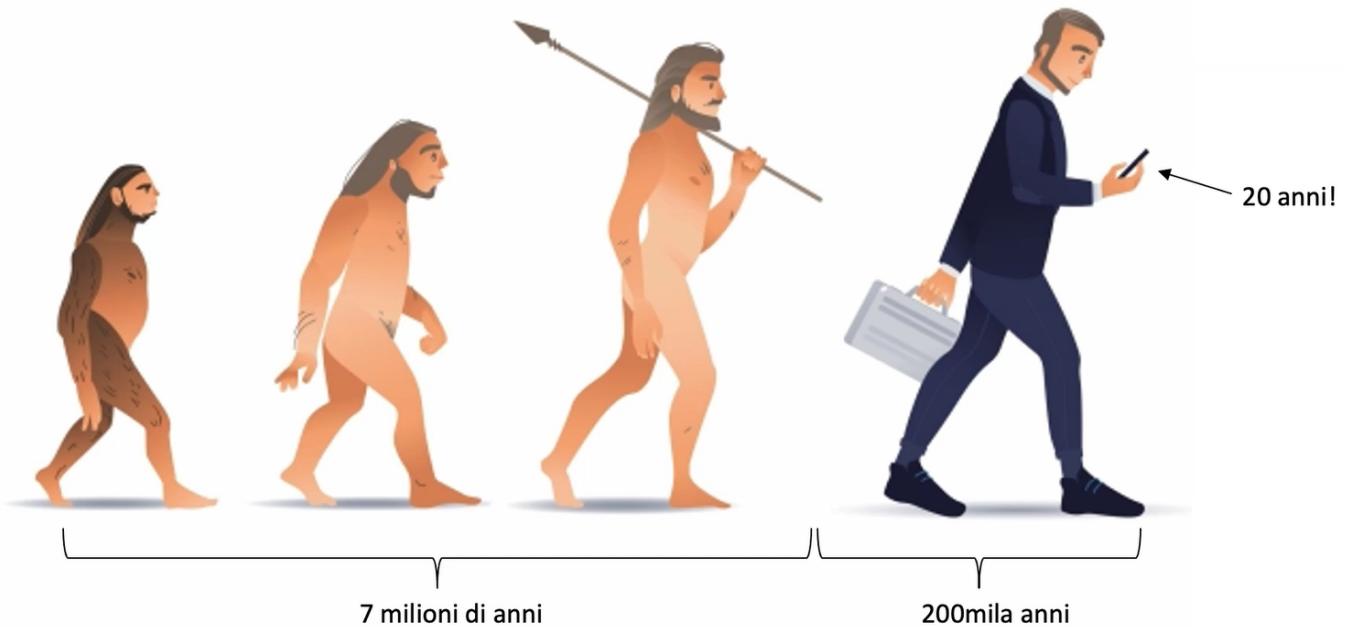


Information Disorder



I bias cognitivi del cervello umano unito a come i social network sono stati progettati creano un terreno molto fertile per quello che prende il nome con Information Disorder.

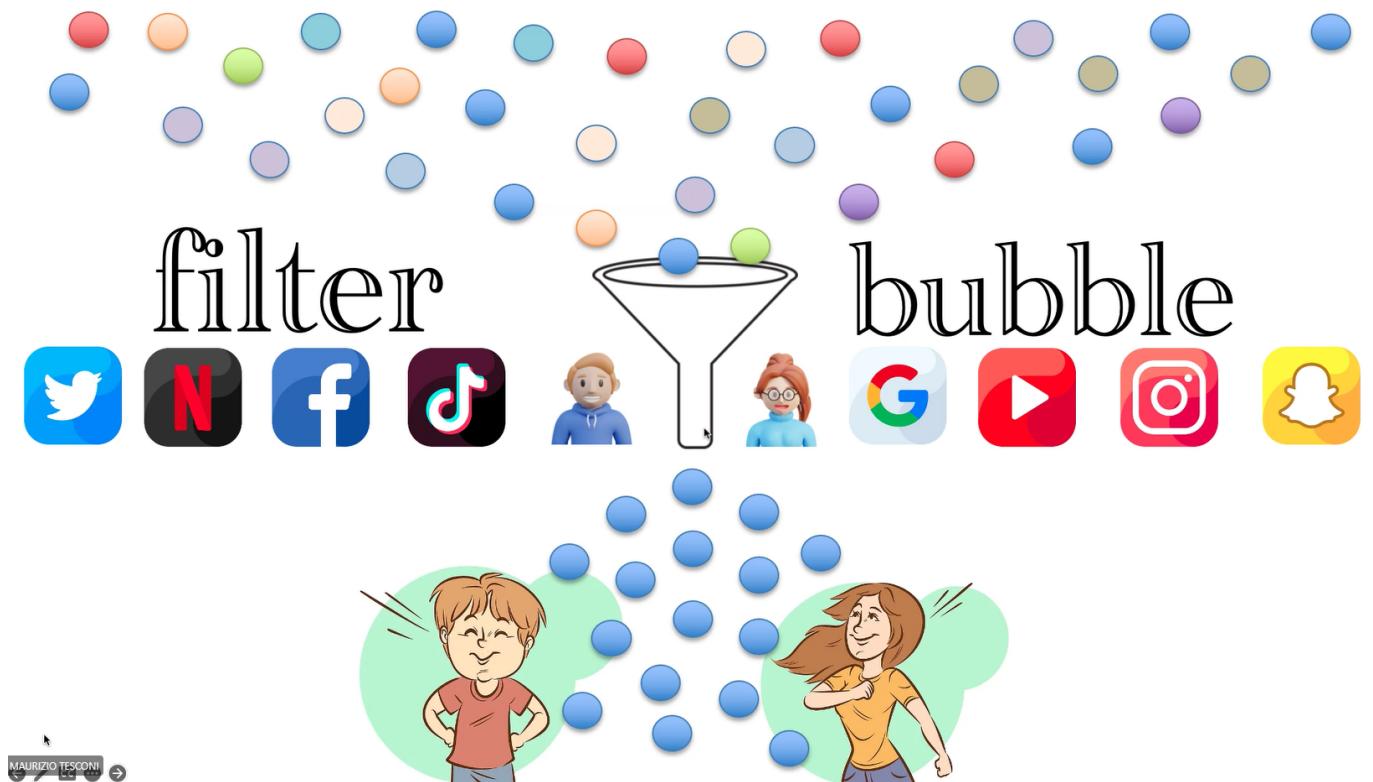
Il mondo delle informazioni false e l'intento di danneggiare si uniscono in un diagramma di Venn come in figura. La disinformazione invece prevede di creare delle informazioni false con l'intento di danneggiare.



Esistono diversi bias cognitivi:

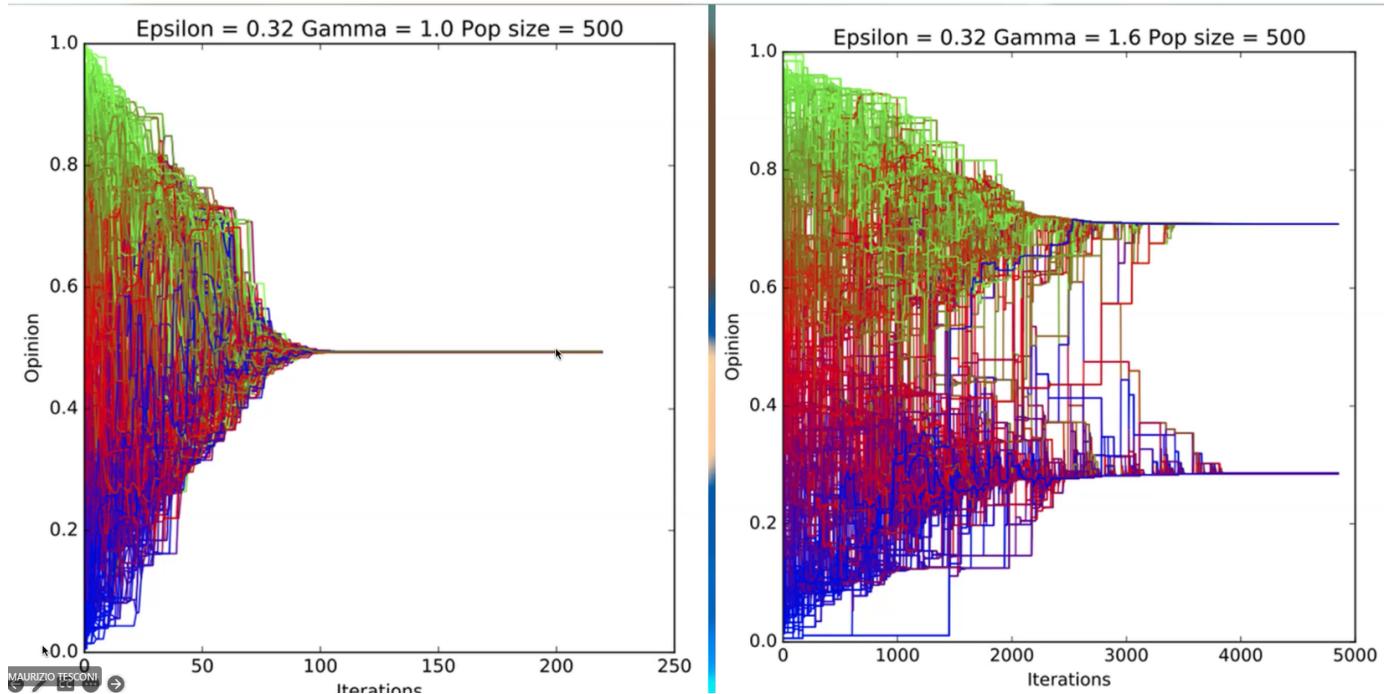
Cognitive biases

Actor-observer bias	Bandwagon effect	Confirmation bias	The Dunning-Kruger effect
False consensus effect	Hindsight bias	Misinformation effect	Sunk cost effect



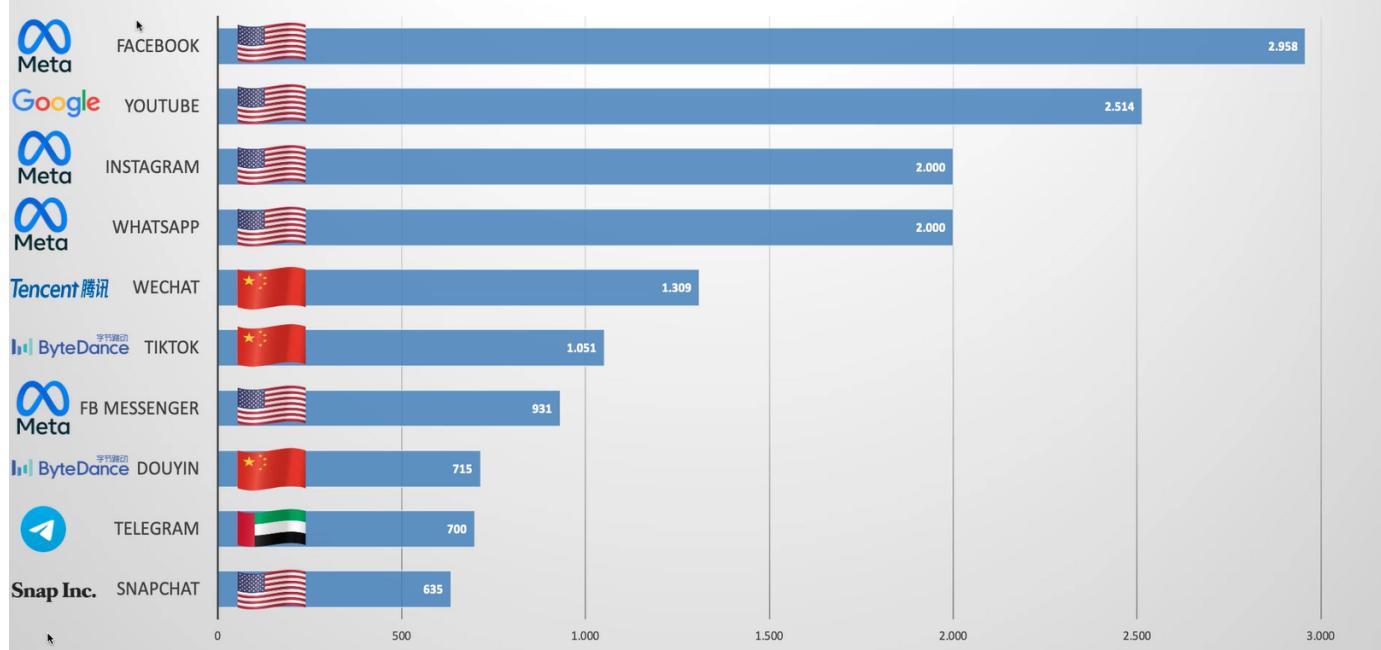
MAURIZIO TESCONI

Simulazione di Filter Bubble



Questa è la classifica dei social media più utilizzati al mondo a gennaio 2023 (espresso in milioni di utenti attivi mensilmente):

Social Media più utilizzati al mondo (gennaio 2023)



Il mondo social oramai ha 2 padroni: America e Cina, semplicemente perchè in Cina la popolazione non può usare le piattaforme occidentali. Alcune piattaforme cinesi sono però prenicate nel mondo occidentale, come ad esempio Douyin che è diventato il TikTok in occidente. Inoltre la maggior parte dei social americani sono gestiti da un'unica azienda Meta.

facebook

Accetta le nostre Condizioni aggiornate per continuare a usare Facebook

Abbiamo aggiornato le nostre Condizioni per spiegare meglio il nostro servizio e cosa chiediamo a ogni persona che usa Facebook.

Ora puoi controllare in modo più semplice le tue impostazioni di protezione, sulla privacy e sui dati in un unico posto e in qualsiasi momento nelle impostazioni. Abbiamo anche aggiornato la nostra Normativa sui dati e Normativa sui cookie per riflettere le nuove funzioni che abbiamo sviluppato e per spiegare meglio in che modo creiamo un'esperienza personalizzata per te. Gli aggiornamenti saranno i seguenti:

- Nuove funzioni come Marketplace, gli effetti della fotocamera e gli strumenti per l'accessibilità
- Maggiori dettagli sull'elaborazione da parte dei nostri sistemi dei contenuti che condividi, come testo, foto e video
- Un resoconto su come condividiamo informazioni, sistemi e tecnologie tra i prodotti delle aziende di Facebook, inclusi WhatsApp, Instagram e Oculus

Cliccando su Accetto, accetti le Condizioni aggiornate. Se non vuoi accettare le Condizioni, Scopri le opzioni a tua disposizione.

Accetto

Facebook © 2018

Normativa sui dati

Ricerca e innovazione per il bene della società

Usiamo le informazioni in nostro possesso (inclusi le informazioni dei partner di ricerca con cui collaboriamo) per effettuare e supportare la ricerca e l'innovazione su argomenti relativi al benessere sociale generale, ai progressi tecnologici, all'interesse pubblico, alla salute e al benessere. Ad esempio, analizziamo le informazioni di cui disponiamo relative ai percorsi di migrazione durante le emergenze per aiutare i soccorsi. Scopri di più sui nostri programmi di ricerca.

Questo è un esempio basato sulle condizioni di Facebook in cui dichiarano di usare i dati delle persone per condurre degli esperimenti:

Esperimenti sulle persone

Esperimento fatto da **ricercatori di Facebook** su **61 milioni di persone** che ha dimostrato che grazie alla modifica di un post FB sono andate a votare circa **340.000 persone in più** per le elezioni del Congresso degli Stati Uniti del 2010

- 98% hanno ricevuto questo messaggio nella parte superiore dei loro feed di notizie
- 1% nessun messaggio
- 1% senza queste foto

Today is Election Day

Find your polling place on the U.S. Politics Page and click the "I Voted" button to tell your friends you voted.

I Voted

0 1 1 5 5 3 7 6 People on Facebook Voted

Jaime Settle, Jason Jones, and 18 other friends have voted.

MAURIZIO TESCONI



Cambridge Analytica ha raccolto i dati personali di 87 milioni di account Facebook senza il loro consenso e li ha usati per scopi di propaganda politica.



Christopher Wylie



Aleksandr Kogan

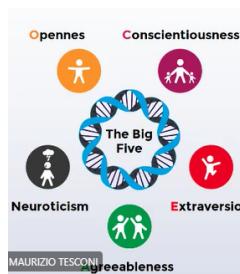


Alexander Nix



Microtargeting politico

Tecnologie usate da CA per influenzare il popolo americano targetizzato



MAURIZIO TESCONI



Alcune definizioni...



Bot

Software costantemente in esecuzione che formulano decisioni, agiscono secondo le precedenti senza l'intervento umano (Tsvetkova et al. 2017).

Social bot



Agenti progettati per adempiere ad uno scopo specifico mediante una comunicazione mono o multi direzionale online (Grimme et al. 2017).

Troll

persona reale che usa profili fake per provocare, insultare, creare confusione



Social Botnet

Insieme di social bot che operano in modo coordinato per raggiungere un obiettivo



¶

Esperimento condotto dal prof.Tesconi per capire come avviene il riconoscimento di account Twitter reali o meno:

Quiz su larga scala



Domanda: dato un account Twitter, sei in grado di riconoscere se è **reale** o **spambot**?

RISULTATI

UMANI

TWITTER

spambot tradizionali	91%	60% 
spambot sociali	24%	4% 
account reali	92%	



3 spambot sociali su 4 **non vengono riconosciuti!**

S Cresci, R Di Pietro, M Petrocchi, A Spognardi, M Tesconi - The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race
http://www.csie.ntu.edu.tw/~cis99999/paper/26th/wwwcom2017.pdf of the 26th international conference on world wide web companion, 2017

Botometer fa un analisi comportamentale e del linguaggio, per individuare i bot dei social:

Tecniche di bot detection

Diversi tipi di approcci:

- analisi comportamentale;
- analisi dei contenuti e del linguaggio;
- crowdsourcing;
- honeypots.

Botometer

Il sistema analizza alcuni gruppi di feature:

- network feature;
- user feature;
- friends feature;
- temporal feature;
- content feature;



MAURIZIO TESCONI

The Rise of Social Bots - E Ferrara, O Varol, C Davis, F Menczer, A Flammini , Communications of the ACM

Botnets detection: Digital DNA

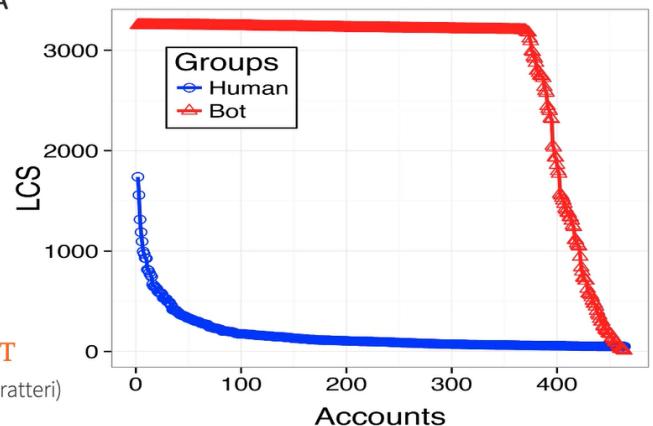
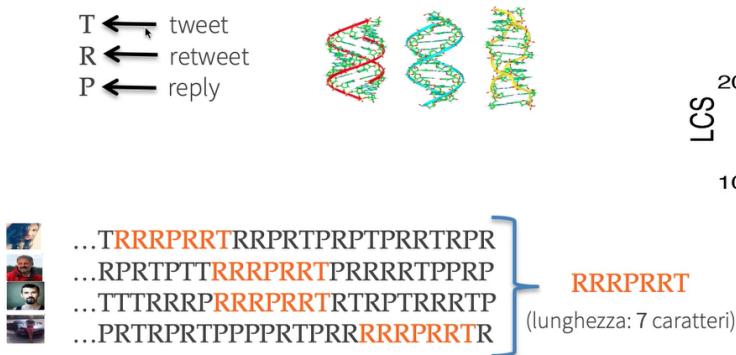
Intuizione

Accounts automatizzati (spambots) hanno sequenze di DNA simili



LCS (longest common substring)

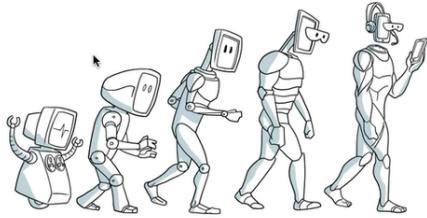
Più lunga sottosequenza comune tra N sequenze di DNA



MAURIZIO TESCONI
Spatiotemporal online behavioral modeling and its application to spambot detection - S Cresci, R Di Pietro, M Petrocchi, A Spognardi, M Tesconi, IEEE Intelligent Systems

Chi fa la parte di social bot ovviamente si tiene aggiornato sulle nuove ricerche in questo ambito al fine di aggiornarsi. Inoltre ci sono altri fenomi come i cyborgs che sono gestiti a metà tra persone e bot automatici oppure i troll che gestiscono manualmente tanti account falsi:

Sfide aperte nella bot detection



Evoluzione dei Social Bot



Cyborgs



Mancanza di dataset annotati



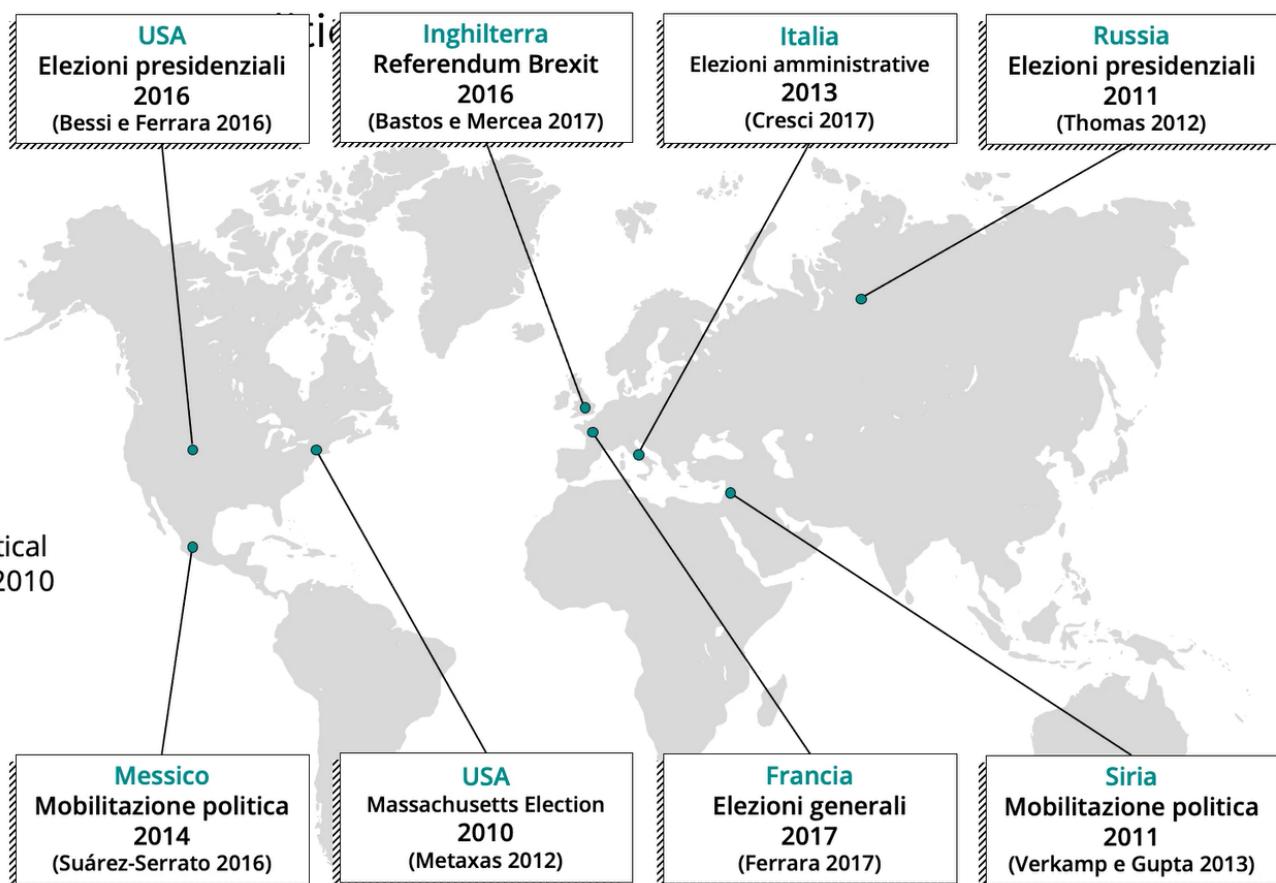
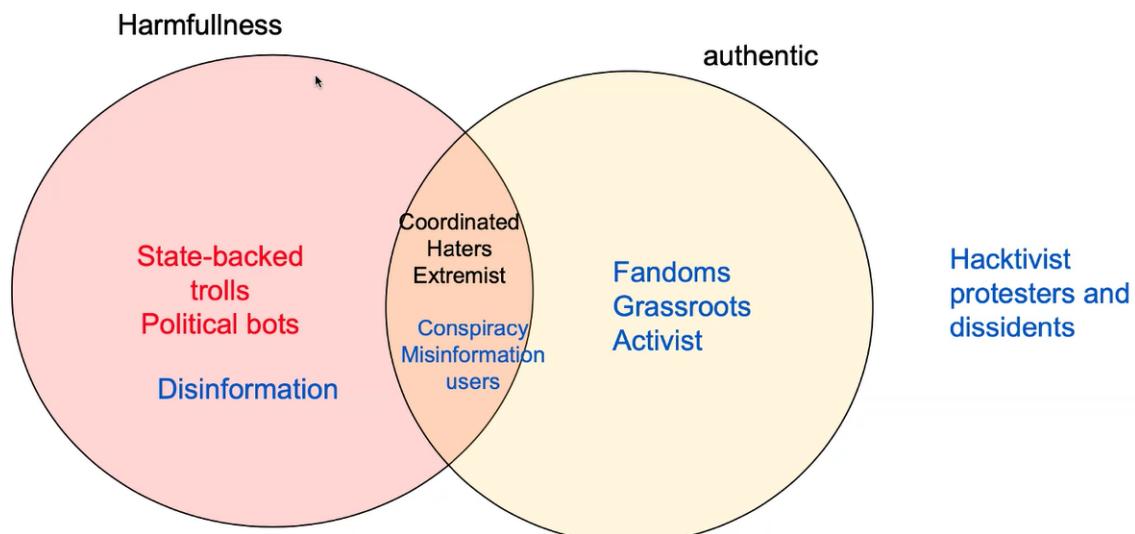
Trolls

La ricerca sta spostando su: **coordinated inauthentic behavior**

Il trend di questi ultimi anni è studiare il comportamento coordinato autentico:

Coordinated Behavior

"groups of users performing synergic actions
in pursuit of a common intent"



Ambiente di sviluppo Dockerizzato

<https://github.com/tizfa/cyber-intelligence-2022>

1. Scaricate e installate [Docker Desktop](#)
2. All'interno di una cartella del vostro computer lanciate:

```
git clone https://github.com/tizfa/cyber-intelligence-2022.git
```
3. Eseguite il comando: docker-compose up

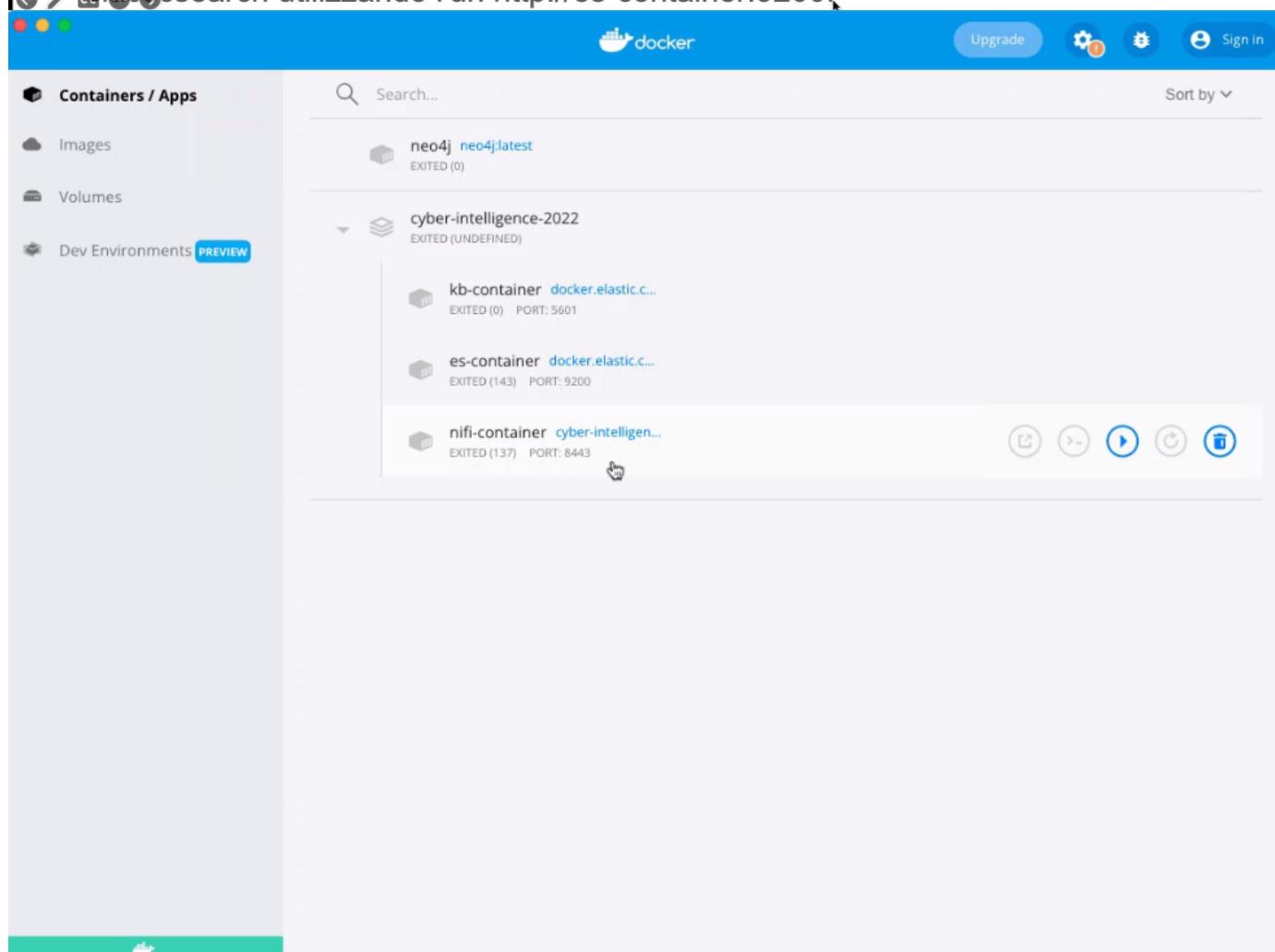
Verranno scaricate e opportunamente create le varie immagini Docker (una per Elastic Search, una per Kibana e una per Nifi)

Risultato:

- o nifi-container -> Nifi: <https://localhost:8443/nifi/> (user:cyberintelligence)
- o es-container -> Elastic Search: <http://localhost:9200/>
- o kb-container -> Kibana: <http://localhost:5601/app/home/>

Da ciascuna di queste macchine si può comunicare con le altre riferendosi con il nome del container riportato sopra, ad esempio nel processore PutElasticsearchHttp di Nifi posso connettermi al web service esposto da

elasticsearch utilizzando l'url <http://es-container:9200>,

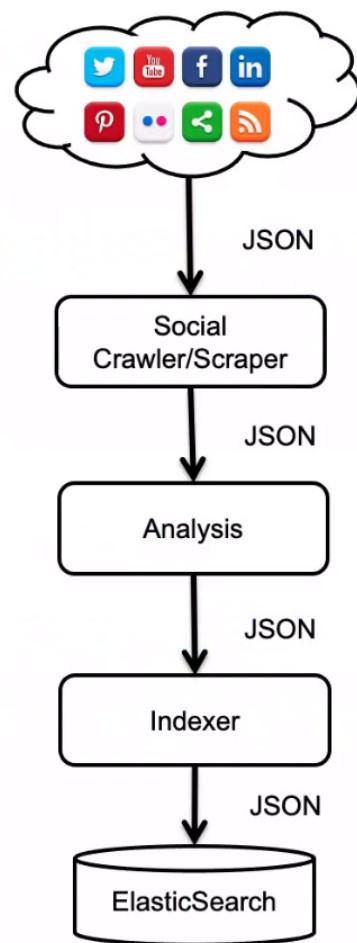


Useremo questa tecnologia (NIFI) perchè permette di lavorare con i big data in modo semplice ed esplorabile:



Social Media Intelligence

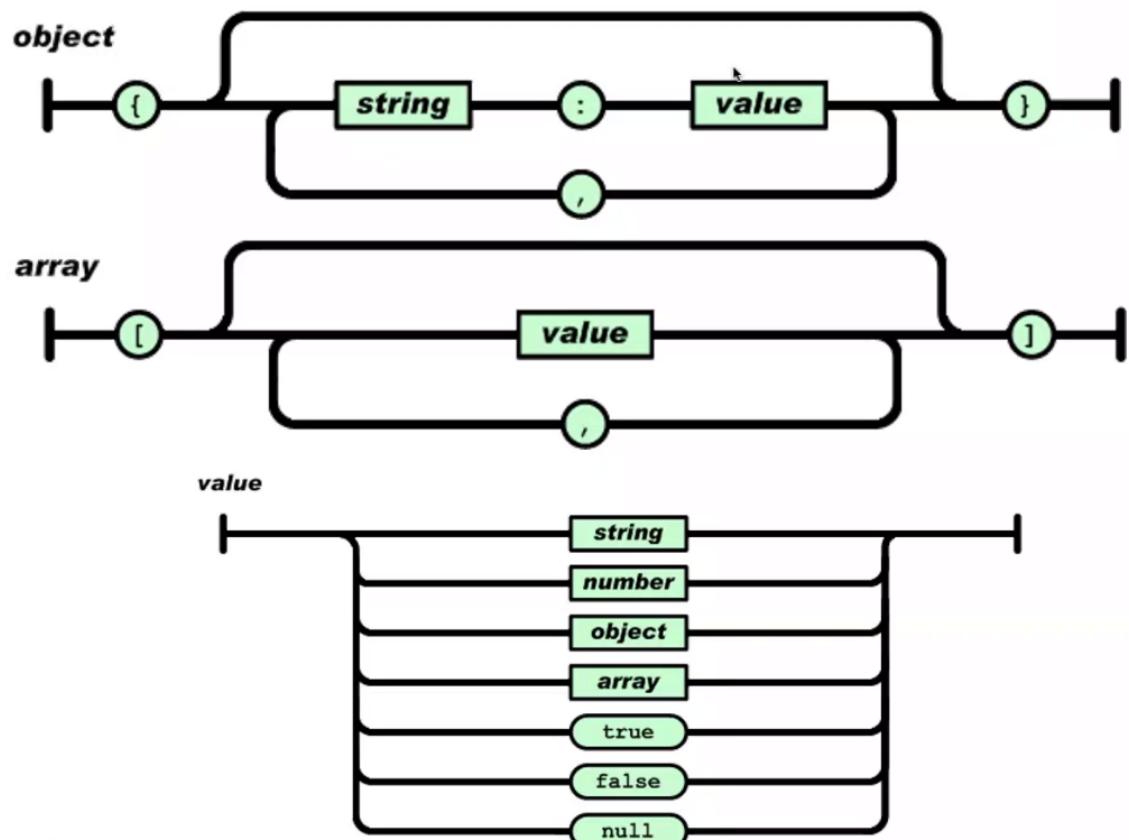
- Esempio di pipeline di analisi Social ->
- Lo standard “de facto” per invocare servizi/componenti si basa su chiamate REST
 - <http://www.acme.com/phonebook/UserDetails/12345>
 - <http://www.acme.com/phonebook/UserDetails?firstName=John&lastName=Doe>
- Il formato standard di interscambio dati tra tutti i componenti/servizi è il **JSON**



3

NIFI consentirà di gestire tutto il flusso di analisi agganciando insieme i diversi moduli che si occupano di effettuare le analisi. Successivamente questi dati vengono indicizzati tramite data storage come ad esempio ElastiSearch per poi rappresentarli con Kibana ad esempio.

Sintassi JSON



← ↎ CC ... →

5

NiFi

Che cosa è Apache NiFi?

Problema: Internet ha portato ad una **crescita esponenziale** dei dati potenzialmente utili per analisi ma questi dati sono **sparsi, isolati** gli uni dagli altri e memorizzati in **formati diversi** => **Integrazione è molto difficile!**

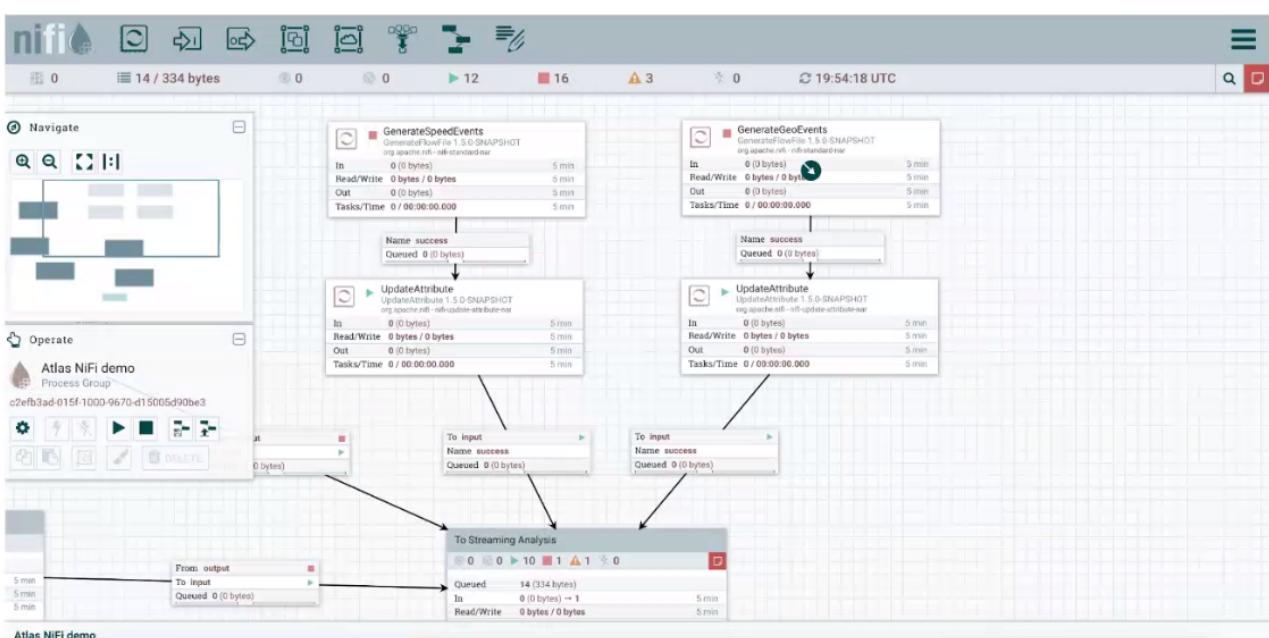
Definizione

“Apache NiFi was built to automate and manage the flow of data between systems and address the global enterprise dataflow issues. It provides an end-to-end platform that can collect, curate, analyze and act on data in real-time, on-premises, or in the cloud with a **drag-and-drop visual interface.**”

Disponibile su <https://nifi.apache.org/>

6

NiFi: gestione dataflow attraverso una UI drag-and-drop



Dataflow = processo di analisi che definisce come le informazioni fluiscono e vengono processate all'interno del task di elaborazione dei dati.

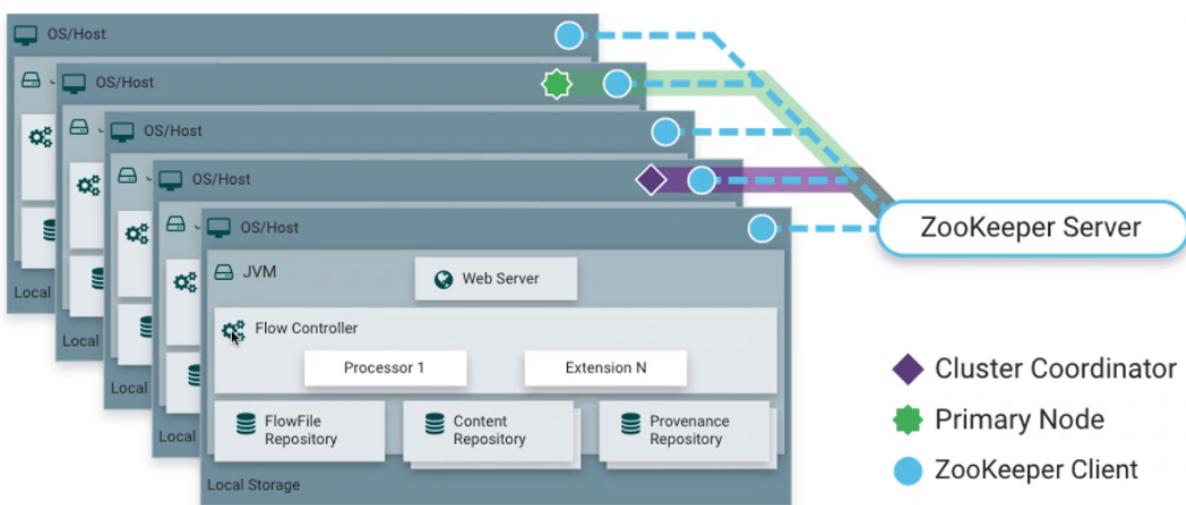
7

NiFi: principali features

- Interfaccia Web amichevole per la costruzione e l'esecuzione di dataflow anche complessi.
- Più di 300 processori disponibili e già pronti per l'uso per filtrare/arricchire/trasformare i dati.
- Possibilità di implementare e utilizzare processori custom.
- Scalabile e tollerante ai guasti.
- Utilizzo di code asincrone con possibilità di gestire automaticamente la *back-pressure* sui dati.
- Possibilità di gestire la priorità dei dati sulle code.
- Possibilità di configurare il dataflow per dare maggiore priorità al *throughput* rispetto alla *latenza* di processamento o viceversa.
- *Data provenance* su tutti i dati transiti dal dataflow.
- Possibilità di creare dataflow templates per riutilizzare facilmente task di processamento su differenti istanze di NiFi.

8

NiFi: architettura logica



- **Web Server:** serve a gestire il dataflow di interesse.
- **Flow controller:** gestisce tutta la computazione definita nel dataflow. La computazione è eseguita attraverso i processori nativi e le estensioni custom.
- Il sistema tiene traccia di tutti i FlowFile attivi attraverso il **FlowFile Repository** e il **Content Repository**.
- La provenienza dei dati è invece tracciata tramite il **Provenance Repository**.
- NiFi può essere eseguito su una singola macchina o un cluster di computer.

9

Principali concetti su NiFi: FlowFile

FlowFile = un messaggio di informazione che si muove all'interno del sistema. Il FlowFile può essere in generale qualsiasi cosa processabile, ad esempio un file su disco, un tweet, una riga di testo, la risposta ad una chiamata di web service, ecc.

Il FlowFile è caratterizzato da:

- *Payload*: è il puntatore al dato originale, ad esempio il contenuto di un file.
- Un dizionario di **attributi** chiave/valore in cui è possibile aggiungere meta-informationi collegate al dato originale.
- Riferimenti per la data provenance dell'informazione associata.

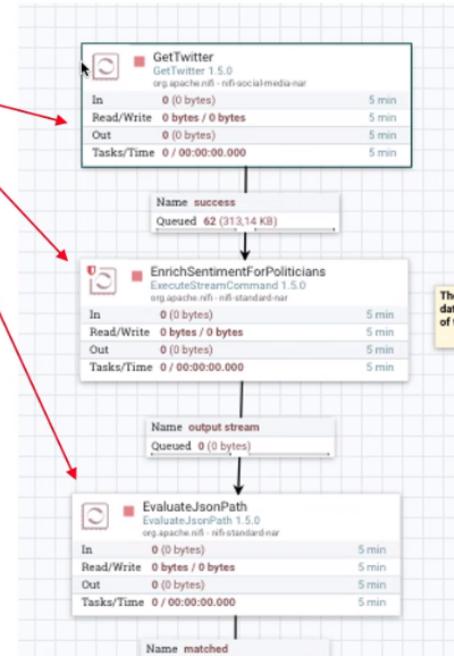
NiFi processor

Principali concetti su NiFi: processor

Processor = E' un componente che processa uno o più FlowFile adottando una opportuna logica.

Attraverso l'uso di funzioni di filtro, arricchimento e trasformazione sui dati di input è in grado di produrre uno o più FlowFile in output.

- Può funzionare da input data source (ad esempio il processore GetFTP o GetTwitter) oppure lavorare in streaming su un insieme di dati di input in arrivo in altro modo dal sistema.
- Può arricchire un FlowFile di metainformazioni attraverso l'uso di attributi specifici (ad esempio per un file potrebbe essere il creation_date, il filename, ecc.)
- Può definire uno o più connessioni di uscita sulle quali distribuire i contenuti processati.



11

NiFi: aggiungere un processore

Dalla toolbar
premere il
pulsante



e trascinare
dove di vuole
posizionare il
processore

Add Processor

Source Displaying 309 of 309

Type ▲ Version Tags

Source	Type	Version	Tags
all groups	AttributeRollingWindow	1.16.0	rolling, data science, Attribute ...
amazon attributes	AttributesToCSV	1.16.0	flowfile, csv, attributes
avro aws azure	AttributesToJson	1.16.0	flowfile, json, attributes
cloud consume	Base64EncodeContent	1.16.0	encode, base64
database delete	CalculateRecordStats	1.16.0	stats, record, metrics
fetch get ingest	CaptureChangeMySQL	1.16.0	cdc, jdbc, mysql, sql
json listen logs	CompareFuzzyHash	1.16.0	fuzzy-hashing, hashing, cyber...
message	CompressContent	1.16.0	lzma, snappy-hadoop, deflate, ...
microsoft pubsub	ConnectWebSocket	1.16.0	subscribe, consume, listen, We...
put record	ConsumeAMQP	1.16.0	receive, amqp, rabbit, get, cons...
restricted source	ConsumeAzureEventHub	1.16.0	cloud, streaming, streams, eve...
storage text	ConsumeEWS	1.16.0	FWS, Exchange, Email, Consu...
update			

AttributeRollingWindow 1.16.0 org.apache.nifi - nifi-stateful-analysis-nar

Track a Rolling Window based on evaluating an Expression Language expression on each FlowFile and add that value to the processor's state. Each FlowFile will be emitted with the count of FlowFiles and total aggregate value of values processed in the current time window.

CANCEL ADD 15

NiFi: configurare un processore (2)

Configure Processor | GetFile 1.16.0

Stopped

SETTINGS SCHEDULING PROPERTIES RELATIONSHIPS COMMENTS

Scheduling Strategy ?
Timer driven

Strategia scheduling

Concurrent Tasks ?
1

Execution ?
All nodes

Numero di task di questo tipo attivi simultaneamente

Run Schedule ?
0 sec

Ogni quanto tempo il processore deve essere schedulato (Timer driven)

Indica quanto tempo il processore può occupare la CPU. Un processore quando rilascia la CPU deve poi aggiornare il repository dei FlowFile processati (processo costoso da spalmare su 1 o n pacchetti!). Per minimizzare la latenza dei pacchetti impostare il selettore tutto a sinistra, per massimizzare il throughput invece spostarlo tutto a destra.

CANCEL APPLY

17

NiFi: configurare un processore (3)

Configure Processor | GetFile 1.16.0

Stopped

SETTINGS SCHEDULING PROPERTIES RELATIONSHIPS COMMENTS

Required field

Property	Value
Input Directory	? /home/cint/tweets
File Filter	? [^\\].*
Path Filter	? No value set
Batch Size	? 10
Keep Source File	? false
Recurse Subdirectories	? true
Polling Interval	? 0 sec
Ignore Hidden Files	? true
Minimum File Age	? 0 sec
Maximum File Age	? No value set
Minimum File Size	? 0 B
Maximum File Size	? No value set

Opzioni specifiche per ogni processore

CANCEL APPLY

18

NiFi: configurare un processore (4)

Configure Processor | EvaluateJsonPath 1.16.0

⚠ Invalid

SETTINGS **SCHEDULING** **PROPERTIES** **RELATIONSHIPS** **COMMENTS**

Automatically Terminate / Retry Relationships ?

failure

terminate retry

FlowFiles are routed to this relationship when the JsonPath cannot be evaluated against the content of the FlowFile; for instance, if the FlowFile is not valid JSON

matched

terminate retry

FlowFiles are routed to this relationship when the JsonPath is successfully evaluated and the FlowFile is modified as a result

unmatched

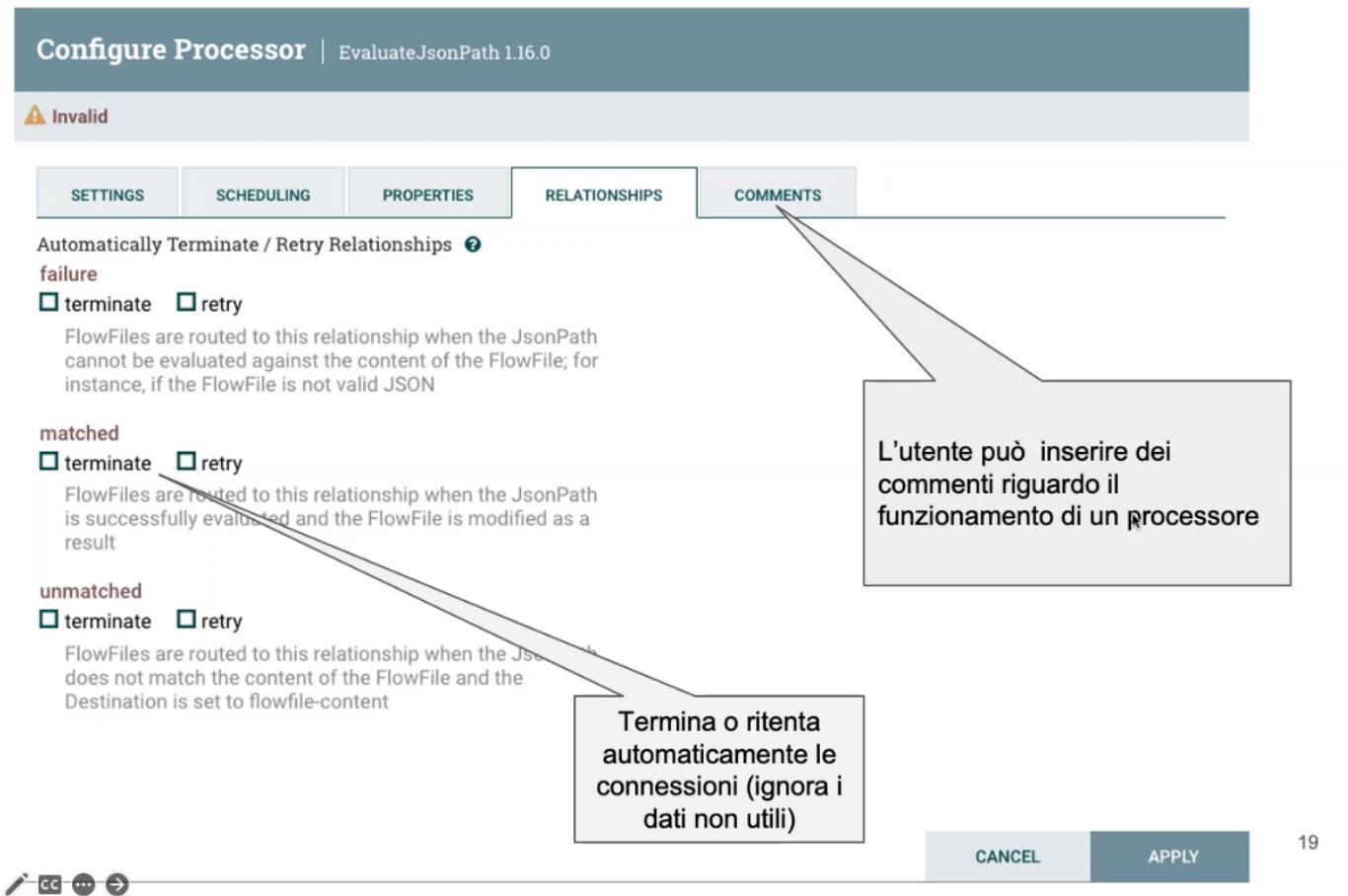
terminate retry

FlowFiles are routed to this relationship when the JsonPath does not match the content of the FlowFile and the Destination is set to flowfile-content

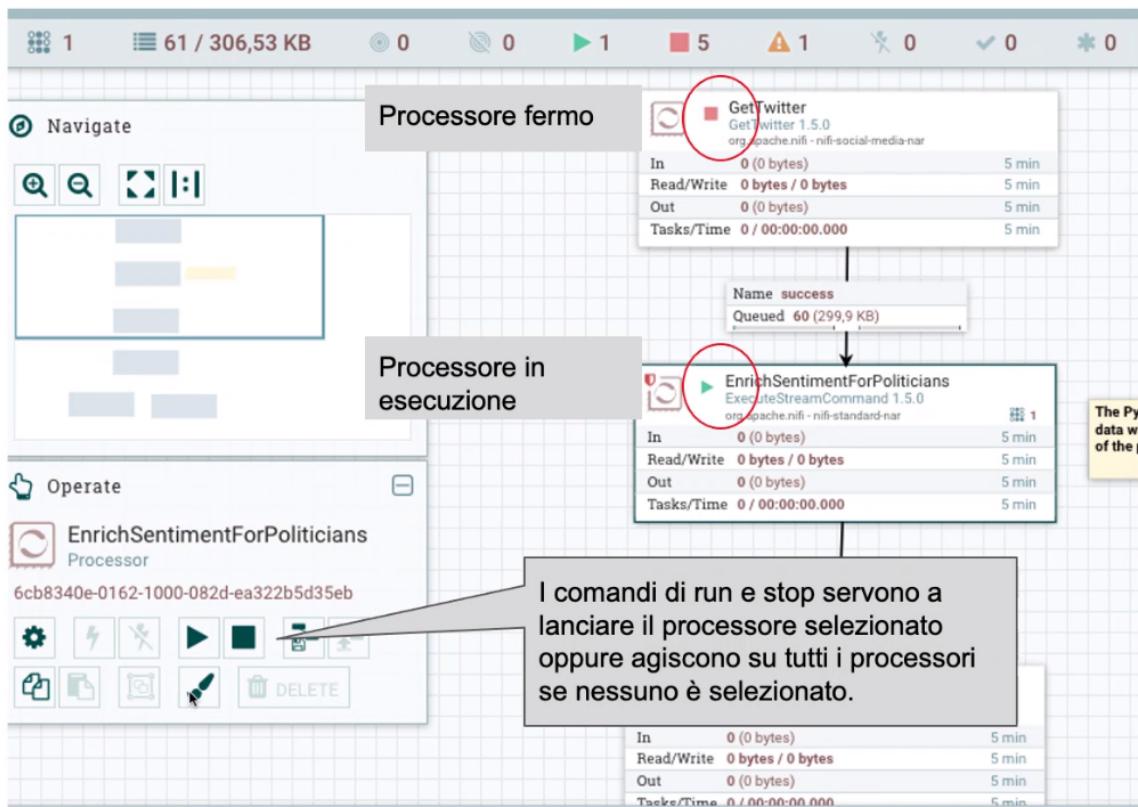
L'utente può inserire dei commenti riguardo il funzionamento di un processore

Termina o ritenta automaticamente le connessioni (ignora i dati non utili)

CANCEL **APPLY**



NiFi: eseguire e fermare un processore



20

Per gestire al meglio la back-pressure:

NiFi: configurare una connessione tra processori

Cliccare tasto destro sulla connessione, quindi sulla voce "Configure".

Configure Connection

DETAILS SETTINGS

Name:

Available Prioritizers: **NewestFlowFileFirstPrioritizer**, **OldestFlowFileFirstPrioritizer**, **PriorityAttributePrioritizer**

Selected Prioritizers: **FirstInFirstOutPrioritizer** X

Id: e3886840-0180-1000-b01a-477d4bad2965

FlowFile Expiration: 0 sec

Back Pressure Object Threshold: 10000

Size Threshold: 1 GB

Load Balance Strategy: Do not load balance

Si può scegliere come dare eventuale priorità ai FlowFile processati.

Tempo di scadenza dei FlowFile da processare. Se uno di questi non viene processato in tempo viene automaticamente eliminato.

CANCEL APPLY

21



NiFi: configurare la back-pressure su una connessione

Configure Connection

DETAILS SETTINGS

Name:

Available Prioritizers: **PriorityAttributePrioritizer**

Selected Prioritizers: **FirstInFirstOutPrioritizer**

Id: e3886840-0180-1000-b01a-477d4bad2965

FlowFile Expiration: 0 sec

Back Pressure Object Threshold: 10000

Size Threshold: 1 GB

Load Balance Strategy: Do not load balance

Imposta il limite sul numero di FlowFile in coda da processare prima che il processore sorgente sia reso non schedulabile.

Imposta il limite sulla dimensione max data dai FlowFile in coda da processare prima che il processore sorgente sia reso non schedulabile.

Quando la back-pressure è attivata, si può vedere al volo sul dataflow qual'è lo stato di riempimento della coda rispetto ai due parametri di back-pressure

Name: success
Queued: 44 (206 bytes)

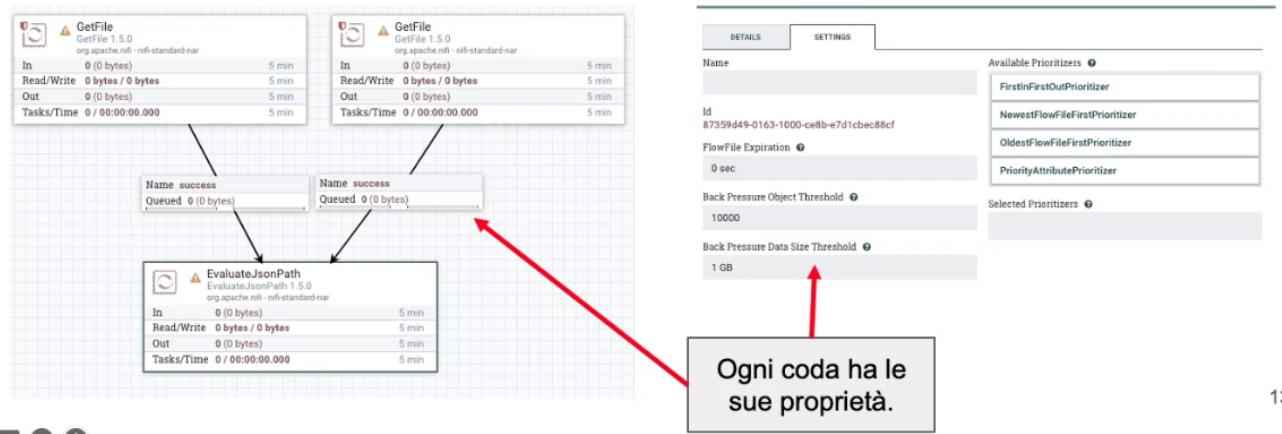
Queue is 82% full based on Back Pressure Data Size Threshold

22



Principali concetti su NiFi: connection (2)

- Le code in uscita sono associate al singolo processore e il loro numero e la semantica dipende dal tipo di processore considerato.
- Un processore può processare in generale simultaneamente n code di input.
- La comunicazione tra due processori è possibile associando una coda di uscita di un processore con l'ingresso di un'altro processore.
- Le proprietà di funzionamento di un coda sono specifiche per coda.

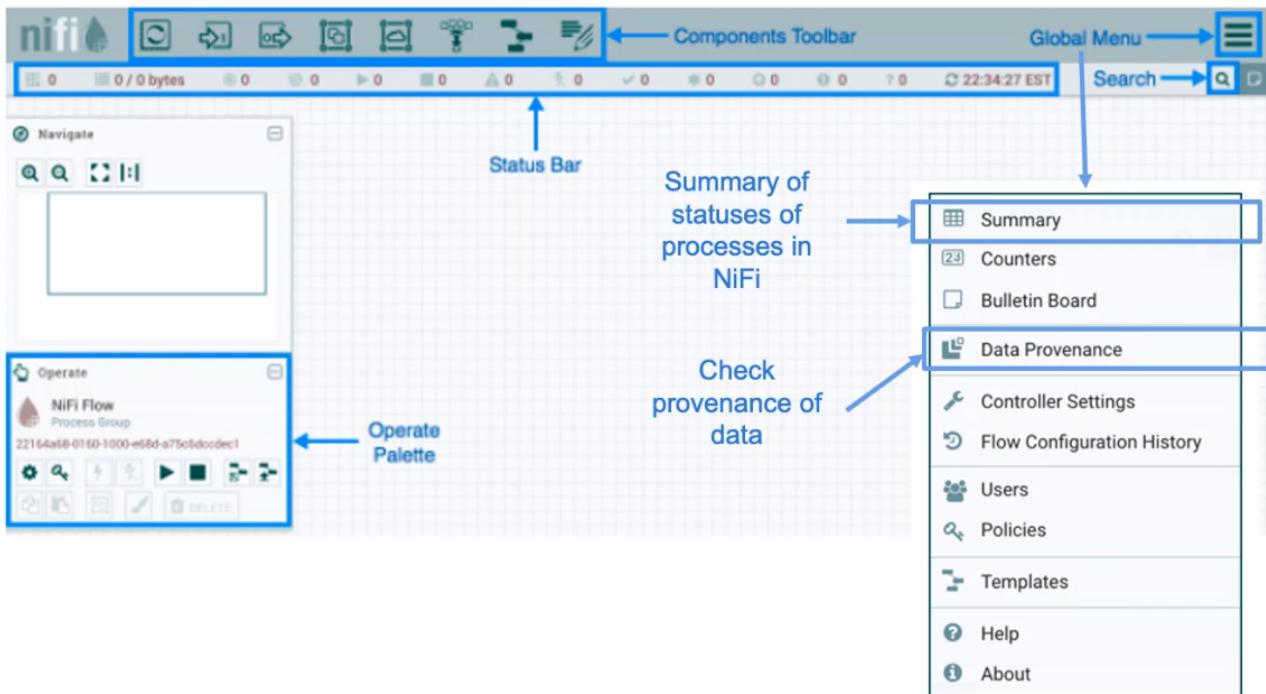


13



NiFi web interface

NiFi: interfaccia Web



14

NiFi expression language

NiFi: expression language

Per interagire con gli **attributi** di un FlowFile si può usare la sintassi speciale \${...} per ottenere il dato originale o trasformato da una sequenza di operazioni.

Alcuni esempi:

- Controllare se un filename contiene una sottostringa:
 `${filename:contains('NiFi')}`
- Aggiungere una sottodirectory ad un path esistente:
 `${path:append('/new_directory')}`

- Fare una operazione matematica:
 `${amount_owed:minus(5)}`
- Confrontare più variabili:
 `${variable_one:gt(${variable_two})}`

L'expression language si usa direttamente come valore della proprietà

Configure Processor

Sul tooltip di aiuto viene indicato se su una proprietà si può usare expression language

Specifies how to determine which relationships to use when evaluating the Expression Language

Required field

Property

Default value: Route to Property name
Supports expression language: false

Routing Strategy

Value

LowFollowedUsers

Route to Property name

`${numFollowers:gt(250)}`

`${numFollowers:le(250)}`

23



XPath e JSONPath

NiFi: uso di XPath e JSONPath per accesso a doc XML e JSON

Su NiFi XML e JSON sono cittadini di prima classe!
Attraverso i processori EvaluateXPath e EvaluateJsonPath si possono processare direttamente i dati in tali formati.

Esempio di JSON

```
{
  "store": {
    "book": [
      {
        "category": "reference",
        "author": "Nigel Rees",
        "title": "Sayings of the Century",
        "price": 8.95
      },
      {
        "category": "fiction",
        "author": "Evelyn Waugh",
        "title": "Sword of Honour",
        "price": 12.99
      },
      {
        "category": "fiction",
        "author": "Herman Melville",
        "title": "Moby Dick",
        "isbn": "0-553-21311-3",
        "price": 8.99
      },
      {
        "category": "fiction",
        "author": "J. R. R. Tolkien",
        "title": "The Lord of the Rings",
        "isbn": "0-395-19395-8",
        "price": 22.99
      }
    ],
    "bicycle": {
      "color": "red",
      "price": 19.95
    }
  }
}
```

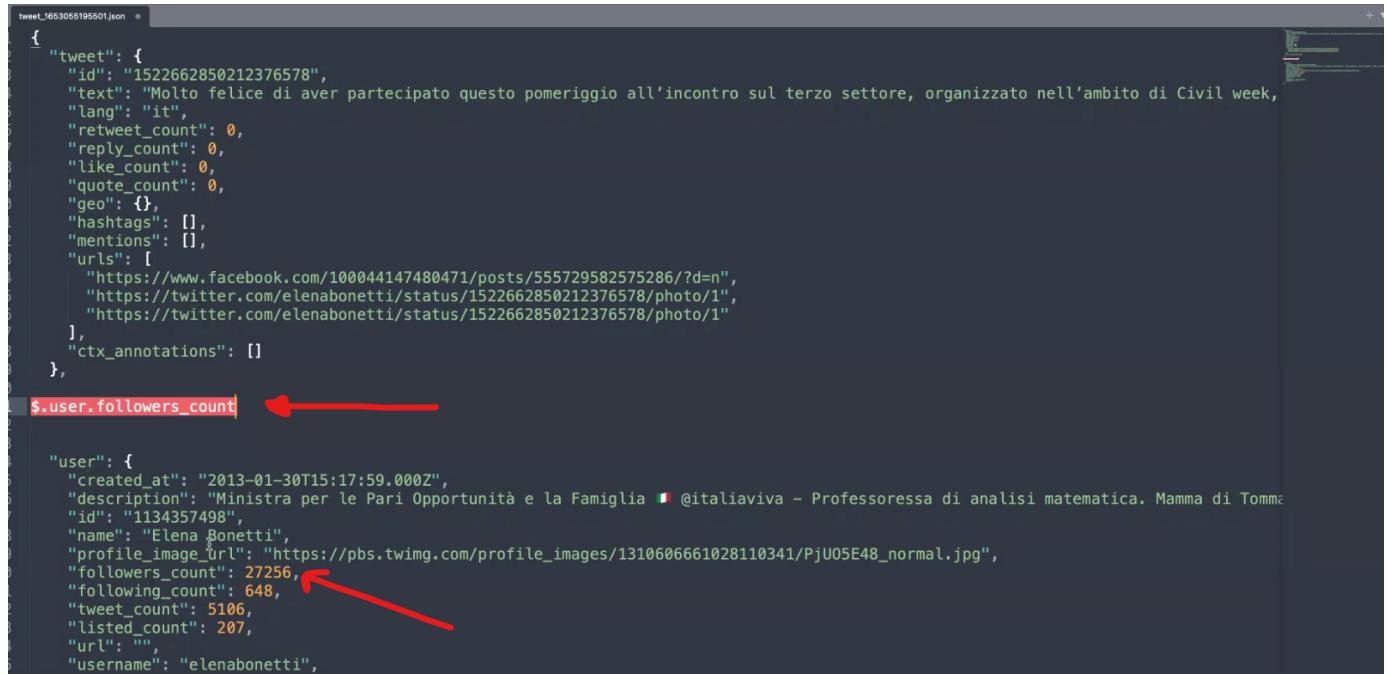
24

NiFi: uso di XPath e JSONPath per accesso a doc XML e JSON (2)

XPath	JSONPath	Result
/store/book/author	\$.store.book[*].author	the authors of all books in the store
//author	\$..author	all authors
/store/*	\$.store.*	all things in store, which are some books and a red bicycle.
/store//price	\$.store..price	the price of everything in the store.
//book[3]	\$..book[2]	the third book
//book[last()]	\$..book[(@.length-1)] \$..book[-1:]	the last book in order.
//book[position() <3]	\$..book[0,1] \$..book[:2]	the first two books
//book[isbn]	\$..book[?(@.isbn)]	filter all books with isbn number
//book[price<10]	\$..book[?(@.price<10)]	filter all books cheaper than 10
//*	\$...*	all Elements in XML document. All members of JSON structure.

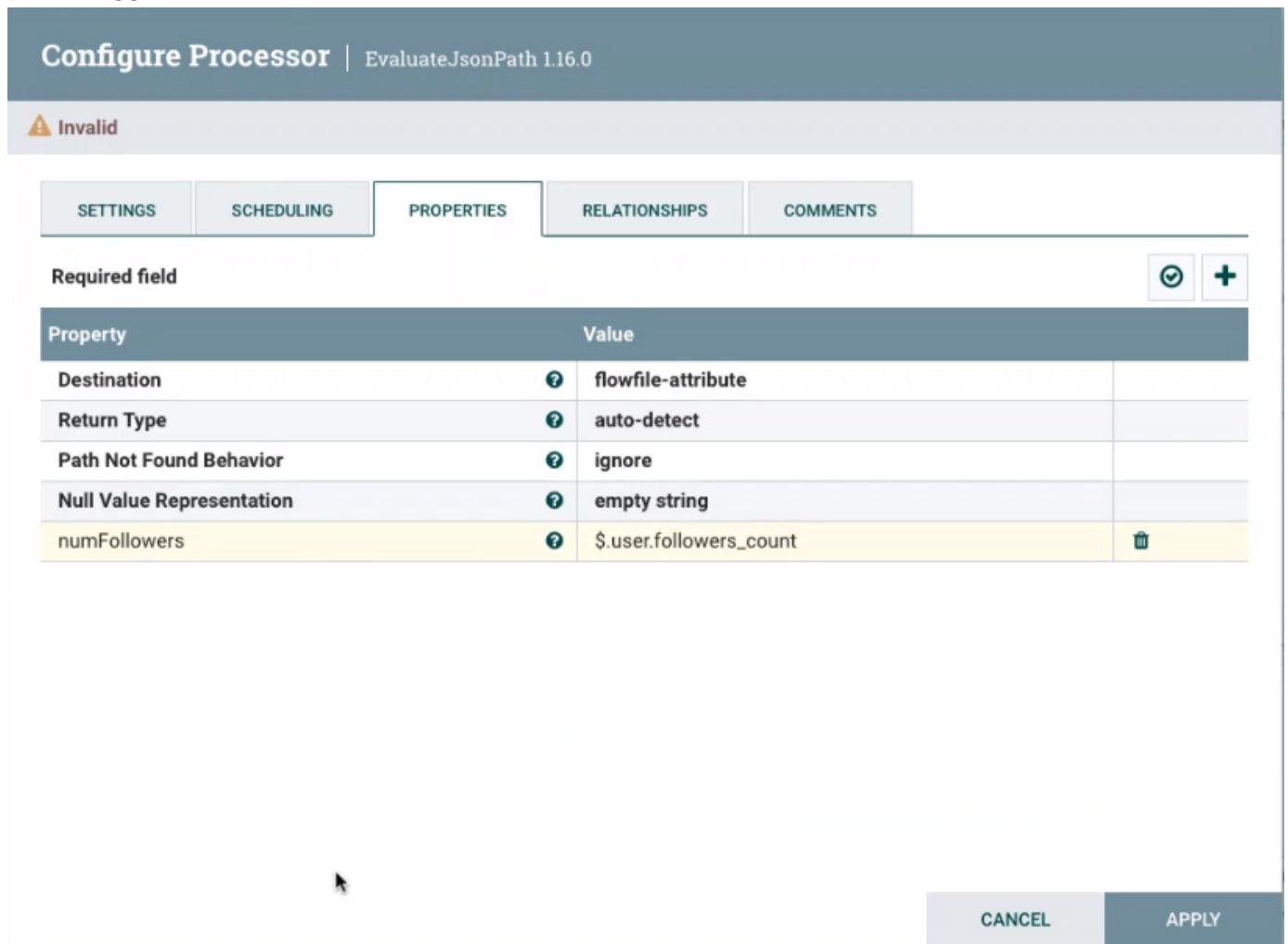
25

Con questa sintassi:



```
tweet_1653055195501.json
{
  "tweet": {
    "id": "1522662850212376578",
    "text": "Molto felice di aver partecipato questo pomeriggio all'incontro sul terzo settore, organizzato nell'ambito di Civil week,",
    "lang": "it",
    "retweet_count": 0,
    "reply_count": 0,
    "like_count": 0,
    "quote_count": 0,
    "geo": {},
    "hashtags": [],
    "mentions": [],
    "urls": [
      "https://www.facebook.com/100044147480471/posts/555729582575286/?d=n",
      "https://twitter.com/elenabonetti/status/1522662850212376578/photo/1",
      "https://twitter.com/elenabonetti/status/1522662850212376578/photo/1"
    ],
    "ctx_annotations": []
  },
  "$.user.followers_count" →
  "user": {
    "created_at": "2013-01-30T15:17:59.000Z",
    "description": "Ministra per le Pari Opportunità e la Famiglia 🇮🇹 @italiaviva – Professoressa di analisi matematica. Mamma di Tommaso e Giacomo",
    "id": "1134357498",
    "name": "Elena Bonetti",
    "profile_image_url": "https://pbs.twimg.com/profile_images/1310606661028110341/PjU05E48_normal.jpg",
    "followers_count": 27256, →
    "following_count": 648,
    "tweet_count": 5106,
    "listed_count": 207,
    "url": "",
    "username": "elenabonetti",
  }
}
```

posso leggermi l'attributo all'interno del JSON indicato con la freccia.



Configure Processor | EvaluateJsonPath 1.16.0

⚠ Invalid

Property	Value
Destination	flowfile-attribute
Return Type	auto-detect
Path Not Found Behavior	ignore
Null Value Representation	empty string
numFollowers	\$.user.followers_count

Required field ✖ ✚

CANCEL APPLY

NiFi: data provenance

è possibile ispezionare per un dato le operazioni effettuate:



NiFi: data provenance

Dal menu in alto a destra, selezionare la voce “Data provenance”:

NiFi Data Provenance

Displaying 95 of 95 04/26/2018 16:28:24 CEST

by component name

Type	FlowFile Uuid	Size	Component Name	Component Type
16:29:02.162 CE.. DROP	de2cf9a7-931a-423a-a763-8c...	5,04 KB	HighFollowedDir	PutFile
04/26/2018 16:29:02.161 CE.. SEND	de2cf9a7-931a-423a-a763-8c...	5,04 KB	HighFollowedDir	PutFile
04/26/2018 16:29:02.148 CE.. ROUTE	de2cf9a7-931a-423a-a763-8c...	5,04 KB	RouteOnAttribute	RouteOnAttribute
04/26/2018 16:29:02.140 CE.. ATTRIBUTES_MODIFIED	de2cf9a7-931a-423a-a763-8c...	5,04 KB	EvaluateJsonPath	EvaluateJsonPath
04/26/2018 16:29:02.133 CE.. DROP	ce4caf8b-9cf0-43e6-a35e-a0...	4,77 KB	EnrichSentimentForPoliticians	ExecuteStreamCommand
04/26/2018 16:28:57.310 CE.. DROP	4d17c461-71ae-409c-9492-7...	ce4caf8b-9cf0-43e6-a35e-a0...	HighFollowedDir	PutFile
04/26/2018 16:28:57.309 CE.. SEND	4d17c461-71ae-409c-9492-7...	3,53 KB	HighFollowedDir	PutFile
04/26/2018 16:28:57.298 CE.. ROUTE	4d17c461-71ae-409c-9492-7...	3,53 KB	RouteOnAttribute	RouteOnAttribute
04/26/2018 16:28:57.288 CE.. ATTRIBUTES_MODIFIED	4d17c461-71ae-409c-9492-7...	3,53 KB	EvaluateJsonPath	EvaluateJsonPath
04/26/2018 16:28:57.283 CE.. FORK	ce4caf8b-9cf0-43e6-a35e-a0...	4,77 KB	EnrichSentimentForPoliticians	ExecuteStreamCommand
04/26/2018 16:28:57.272 CE.. DROP	c96673ca-1065-4fs2-b8b6-f4...	3,28 KB	EnrichSentimentForPoliticians	ExecuteStreamCommand
04/26/2018 16:28:55.019 CE.. RECEIVE	af94ddff-2e67-459a-87b4-4e...	5,17 KB		
04/26/2018 16:28:55.016 CE.. RECEIVE	ccf132ca-f685-459e-916b-c6...	2,19 KB		
04/26/2018 16:28:54.013 CE.. RECEIVE	ce4caf8b-9cf0-43e6-a35e-a0...	4,77 KB		
04/26/2018 16:28:52.442 CE.. DROP	35a29bea-1c34-4a18-ab2-2...	6,57 KB		
04/26/2018 16:28:52.442 CE.. SEND	35a29bea-1c34-4a18-ab2-2...	6,57 KB	HighFollowedDir	PutFile
04/26/2018 16:28:52.430 CE.. ROUTE	35a29bea-1c34-4a18-ab2-2...	6,57 KB	RouteOnAttribute	RouteOnAttribute
04/26/2018 16:28:52.420 CE.. ATTRIBUTES_MODIFIED	35a29bea-1c34-4a18-ab2-2...	6,57 KB	EvaluateJsonPath	EvaluateJsonPath
04/26/2018 16:28:52.417 CE.. FORK	c96673ca-1065-4fs2-b8b6-f4...	3,28 KB	EnrichSentimentForPoliticians	ExecuteStreamCommand

Showing the most recent events.

Dettaglio evento

Tipo di evento

Nome componente su cui l'evento è stato generato

Lineage grafico del dato

Vai direttamente al componente che ha generato l'evento

Last updated: 16:29:07 CEST

26

Icons: refresh, edit, CC, comments, arrow

NiFi: dettaglio evento durante la data provenance

Provenance Event

Dettagli principali sul FlowFile

Attributi definiti nel FlowFile

Contenuto del FlowFile. Possibilità di ispezionare sia l'input che l'output.

Time: 04/26/2018 16:29:25.492 CEST
Event Duration: < 1ms
Lineage Duration: 00:00:11.457
Type: SEND
FlowFileUuid: d97ecff1-47cb-4324-8b49-5479ce5ea99d
File Size: 2,53 KB
ComponentId: 01621022-55a9-130d-724b-bcf86c898469
ComponentName: LowFollowedDir
ComponentType:
filename: 1615642784741098.json
mime type: application/json
numFollowers: 89
path: /

Provenance Event

Attribute Values:

- execution.command: /Users/tiziano/git/cyberintelligence/venv/bin/python
- execution.command.args: /Users/tiziano/git/cyberintelligence/CheckPoliticians.py
- execution.error: Empty string set
- execution.status: 0

Provenance Event

Input Claim:

- Container: default
- Section: 2
- Identifier: 1524752929000-2
- Offset: 58711
- Size: 2,53 KB

Output Claim:

- Container: default
- Section: 2
- Identifier: 1524752929000-2
- Offset: 58711
- Size: 2,53 KB

Replay
Connection Id: 01621023-55a9-130d-7787-fc863b6cf6b2f

27



NiFi: navigazione tra i livelli (1)

Processore primitivo

Process Group usato come black box. L'input è ottenuto dal componente GetFile mentre l'output generato viene processato dal componente PutFile

Barra di navigazione, indica il livello attuale

Cyberintelligence » ZipBigFileExample

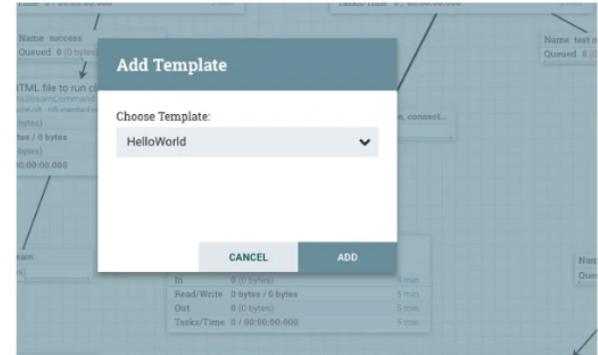
The screenshot shows a NiFi canvas with a process group named "ZipBigFileExample". Inside the group, there is a "GetFile" primitive processor connected to a "To InputFiles" relationship, which then connects to a "ZipBigFile" process group. This group contains a "PutFile" primitive processor connected to a "From OutputFiles" relationship, which then connects back to another "PutFile" primitive processor. The "PutFile" processor is also connected to a "To InputFiles" relationship. The "PutFile" processor has a "Name success" attribute. The "PutFile" processor is highlighted with a red border, indicating it is the active component. A navigation bar at the bottom left shows the current level as "ZipBigFileExample".

29

NiFi: dataflow template

NiFi: i dataflow template

- I DFM (DataFlow managers) possono costruire dataflow arbitrariamente complessi usando opportunamente i componenti di base (Processor, Funnel, Input/Output Port, Process Group, and Remote Process Group).
- Un meccanismo per riutilizzare più volte la logica di un dataflow esistente si rende necessario, anche su istanze di NiFi diverse!

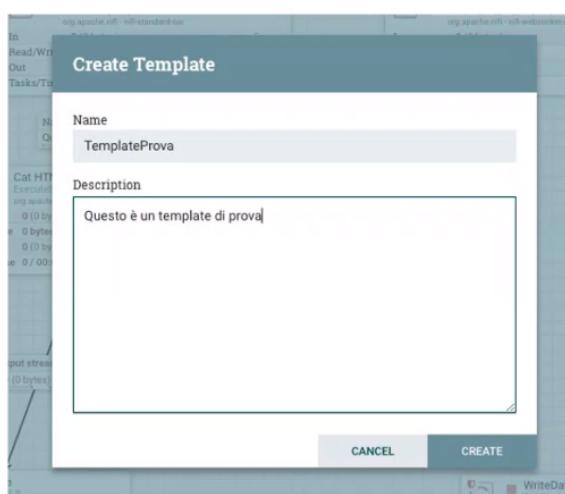


I **template** permettono di riutlizzare dataflow o parti di esso sulla stessa istanza o altre istanze di NiFi.

31

NiFi: creare un nuovo template

- Seleziona tutti i componenti che vuoi includere nel template. Se non selezioni nulla, tutti i componenti nel Process Group attuale saranno inclusi nel template.
- Premere il pulsante “Salva template” per aggiungere un nuovo template
- Riempite la form con i dettagli del template.



32

I diversi tipi di processore

GetFile

NiFi: il processore GetFile



Scopo: data una directory di input, il componente crea uno stream di FlowFile corrispondenti ai file individuati nella cartella di input.

Configure Processor | GetFile 1.16.0

Stopped

SETTINGS SCHEDULING PROPERTIES RELATIONSHIPS COMMENTS

Required field

Property	Value
Input Directory	/home/cint/tweets
File Filter	[^\\.]*
Path Filter	No value set
Batch Size	10
Keep Source File	false
Recurse Subdirectories	true
Polling Interval	0 sec
Ignore Hidden Files	true
Minimum File Age	0 sec
Maximum File Age	No value set
Minimum File Size	0 B
Maximum File Size	No value set

CANCEL APPLY

PutFile

NiFi: il processore PutFile



Scopo: scrive il contenuto di un FlowFile in una cartella su filesystem.

Configure Processor | PutFile 1.16.0

⚠ Invalid

SETTINGS SCHEDULING PROPERTIES RELATIONSHIPS COMMENTS

Required field

Property	Value
Directory	/home/cint/output
Conflict Resolution Strategy	fail
Create Missing Directories	true
Maximum File Count	No value set
Last Modified Time	No value set
Permissions	No value set
Owner	No value set
Group	No value set

CANCEL APPLY 38



vedremo che la cartella di output potrebbe essere dinamica in base all'attributo della lingua del tweet ad esempio

GetTwitter

NiFi: il processore GetTwitter

Scopo: ottiene aggiornamenti di stato da Twitter sfruttando la Streaming API.

Configure Processor

SETTINGS	SCHEDULING	PROPERTIES	COMMENTS
Required field			
Property	Value		
Twitter Endpoint	Filter Endpoint		
Consumer Key	Ukob5T8hRBBGWDn4w6FYPmMxc		
Consumer Secret	Sensitive value set		
Access Token	189161498-oA5FwfwqzjxtNpf1ur9ZXfZ5ZWlut6G5qpmFWneO		
Access Token Secret	Sensitive value set		
Languages	it		
Terms to Filter On	salvini,renzi,grillo,berlusconi		
IDs to Follow	No value set		
Locations to Filter On	No value set		

GetHTTP

NiFi: il processore GetHTTP

Scopo: Permette di interrogare un server HTTP, utile per l'accesso a Web service di tipo REST.

Configure Processor

SETTINGS	SCHEDULING	PROPERTIES	COMMENTS																												
Required field																															
<table border="1"> <thead> <tr> <th>Property</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>URL</td> <td>https://min-api.cryptocompare.com/data/pricemultifull?fsyms=BTC</td> </tr> <tr> <td>Filename</td> <td>btceth_\${now():toNumber()}</td> </tr> <tr> <td>SSL Context Service</td> <td>StandardSSLContextService</td> </tr> <tr> <td>Username</td> <td>No value set</td> </tr> <tr> <td>Password</td> <td>No value set</td> </tr> <tr> <td>Connection Timeout</td> <td>30 sec</td> </tr> <tr> <td>Data Timeout</td> <td>30 sec</td> </tr> <tr> <td>User Agent</td> <td>No value set</td> </tr> <tr> <td>Accept Content-Type</td> <td>No value set</td> </tr> <tr> <td>Follow Redirects</td> <td>false</td> </tr> <tr> <td>Redirect Cookie Policy</td> <td>default</td> </tr> <tr> <td>Proxy Host</td> <td>No value set</td> </tr> <tr> <td>Proxy Port</td> <td>No value set</td> </tr> </tbody> </table>				Property	Value	URL	https://min-api.cryptocompare.com/data/pricemultifull?fsyms=BTC	Filename	btceth_\${now():toNumber()}	SSL Context Service	StandardSSLContextService	Username	No value set	Password	No value set	Connection Timeout	30 sec	Data Timeout	30 sec	User Agent	No value set	Accept Content-Type	No value set	Follow Redirects	false	Redirect Cookie Policy	default	Proxy Host	No value set	Proxy Port	No value set
Property	Value																														
URL	https://min-api.cryptocompare.com/data/pricemultifull?fsyms=BTC																														
Filename	btceth_\${now():toNumber()}																														
SSL Context Service	StandardSSLContextService																														
Username	No value set																														
Password	No value set																														
Connection Timeout	30 sec																														
Data Timeout	30 sec																														
User Agent	No value set																														
Accept Content-Type	No value set																														
Follow Redirects	false																														
Redirect Cookie Policy	default																														
Proxy Host	No value set																														
Proxy Port	No value set																														
<div style="border: 1px solid black; padding: 5px; margin-top: 10px;"> Importante configurare il contesto SSL per accesso a servizi in https </div>																															

40

anche su configurazioni SSL:

NiFi: il processore GetHTTP, configurazione SSL

1. Creare un controller di tipo StandardSSLContextService.
2. Configurarla correttamente.
3. Abilitarla.

Screenshot of the NiFi Controller Services interface showing the configuration of a StandardSSLContextService.

The interface has two tabs: GENERAL and CONTROLLER SERVICES. The CONTROLLER SERVICES tab is selected, showing a list of controllers:

Name	Type	Bundle	State	Scope
JettyWebSocketServer	JettyWebSocketServer 1.5.0	org.apache.nifi - nifi-websocket-s...	Disabled	Cyberintelligence
JettyWebSocketServer	JettyWebSocketServer 1.5.0	org.apache.nifi - nifi-websocket-s...	Disabled	Cyberintelligence
StandardHttpContextMap	StandardHttpContextMap 1.5.0	org.apache.nifi - nifi-http-context...	Enabled	Cyberintelligence
StandardHttpContextMap	StandardHttpContextMap 1.5.0	org.apache.nifi - nifi-http-context...	Disabled	Cyberintelligence
StandardRestrictedSSLContextS...	StandardRestrictedSSLContextS...	org.apache.nifi - nifi-ssl-context-s...	Enabled	Cyberintelligence
StandardSSLContextService	StandardSSLContextService 1.5.0	org.apache.nifi - nifi-ssl-context-s...	Enabled	BitcoinPrice

A callout box points to the 'Enabled' button for the StandardSSLContextService, with the text: "Abilita o disabilita il controller".

Below the table, a large red rectangle highlights the row for 'StandardSSLContextService'. A cursor arrow is pointing at the 'Enabled' status indicator in that row.

At the bottom, a 'Controller Service Details' panel is shown with three tabs: SETTINGS, PROPERTIES, and COMMENTS. The SETTINGS tab is selected, showing a table of required fields:

Property	Value
Keystore Filename	No value set
Keystore Password	No value set
Key Password	No value set
Keystore Type	No value set
Truststore Filename	/Library/Java/JavaVirtualMachines/dk1.8.0_121.jdk/Co...
Truststore Password	Sensitive value set
Truststore Type	JKS
TLS Protocol	TLS

41

ReplaceText

NiFi: il processore ReplaceText

Scopo: sostituisce il contenuto di un FlowFile o parte di esso mediante un valore definito dall'utente. Il processore è molto configurabile per operare in modo molto diverso a seconda delle esigenze.

Configure Processor

SETTINGS	SCHEDULING	PROPERTIES	COMMENTS
Required field			
Property	Value		
Search Value	?	(?s)(^.*\$)	
Replacement Value	?	\${currency.timestamp};\${eth.eur.price};\${eth.usd.price}	
Character Set	?	UTF-8	
Maximum Buffer Size	?	1 MB	
Replacement Strategy	?	Always Replace	
Evaluation Mode	?	Entire text	

Esercizio

- leggere tutti i tweet dalla directory di input
- sostituire la stringa «user» con «utente»

→

NiFi: il processore SplitJson

Scopo: Divide un file JSON in più FlowFiles separate, uno per ogni elemento di un array specificato da un'espressione JsonPath. Ogni FlowFile generato è composto da un elemento dell'array specificato e viene trasferito alla relazione 'split', mentre il file originale viene trasferito alla relazione 'original'

Configure Processor | SplitJson 1.16.0

Stopped

SETTINGS SCHEDULING PROPERTIES RELATIONSHIPS COMMENTS

Required field

Property	Value
JsonPath Expression	\$.tweet.hashtags
Null Value Representation	empty string

+5

MergeContent

NiFi: il processore MergeContent

Scopo: Unisce un gruppo di FlowFiles insieme in base a una strategia definita dall'utente e li raggruppa in un singolo FlowFile. È consigliabile configurare il Processor con una sola connessione in entrata, poiché il gruppo di FlowFiles non verrà creato dai FlowFiles in connessioni diverse.

Property	Value
Merge Strategy	Bin-Packing Algorithm
Merge Format	Binary Concatenation
Attribute Strategy	Keep Only Common Attributes
Correlation Attribute Name	No value set
Minimum Number of Entries	1
Maximum Number of Entries	1000
Minimum Group Size	0 B
Maximum Group Size	No value set
Max Bin Age	No value set
Maximum number of Bins	5
Delimiter Strategy	Text
Header	No value set

46

Configure Processor | MergeContent 1.16.0

⚠ Invalid

SETTINGS SCHEDULING PROPERTIES RELATIONSHIPS COMMENTS

Required field

Property	Value
Merge Strategy	Bin-Packing Algorithm
Merge Format	Binary Concatenation
Attribute Strategy	Keep Only Common Attributes
Correlation Attribute Name	No value set
Minimum Number of Entries	1
Maximum Number of Entries	1000
Minimum Group Size	0 B
Maximum Group Size	No value set
Max Bin Age	No value set
Maximum number of Bins	10
Delimiter Strategy	Text
Header	No value set
Footer	No value set
Demarcator	;

CANCEL APPLY

NiFi: il processore RouteOnAttribute

Scopo: Permette di creare una o più connection di routing sulla base delle condizioni verificate sugli attributi del FlowFile di input.

Configure Processor

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field +

Property	Value
Routing Strategy	Route to Property name
HighFollowedUsers	\${numFollowers:gt(250)}
LowFollowedUsers	\${numFollowers:le(250)}

49

Esercizio

- leggere tutti i tweet dalla directory di input
- estrarre il numero di followers dell'utente che ha prodotto il tweet
- salvare i tweets in 2 cartelle diverse
 - *influencer*: num_follower > 1000
 - *normal*: num_follower < 1000

NiFi: il processore ExecuteStreamCommand

Scopo: Permette di eseguire un comando esterno (ad esempio uno script o un comando di sistema) sul contenuto di un FlowFile e ottenere il risultato del comando in forma di FlowFile.

Configure Processor

SETTINGS	SCHEDULING	PROPERTIES	COMMENTS																							
Required field <table border="1"> <thead> <tr> <th>Property</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>Command Arguments</td> <td>?</td> <td>No value set</td> </tr> <tr> <td>Command Path</td> <td>?</td> <td>No value set</td> </tr> <tr> <td>Ignore STDIN</td> <td>?</td> <td>false</td> </tr> <tr> <td>Working Directory</td> <td>?</td> <td>No value set</td> </tr> <tr> <td>Argument Delimiter</td> <td>?</td> <td>;</td> </tr> <tr> <td>Output Destination Attribute</td> <td>?</td> <td>No value set</td> </tr> <tr> <td>Max Attribute Length</td> <td>?</td> <td>256</td> </tr> </tbody> </table>				Property	Value	Command Arguments	?	No value set	Command Path	?	No value set	Ignore STDIN	?	false	Working Directory	?	No value set	Argument Delimiter	?	;	Output Destination Attribute	?	No value set	Max Attribute Length	?	256
Property	Value																									
Command Arguments	?	No value set																								
Command Path	?	No value set																								
Ignore STDIN	?	false																								
Working Directory	?	No value set																								
Argument Delimiter	?	;																								
Output Destination Attribute	?	No value set																								
Max Attribute Length	?	256																								

51

Se prendiamo la data di creazione di un utente:

```
tweet_1653056195501.json
1 {
2   "tweet": {
3     "id": "1522662850212376578",
4     "text": "Molto felice di aver partecipato questo pomeriggio all'incontro sul terzo settore, organizzato nell'ambito del progetto #CittàSana. Grazie a tutti coloro che hanno partecipato e contribuito alla discussione! #CittàSana #Incontro #Pomeriggio #Organizzazione #Progetto #Felicità #Partecipazione #Discussione #Grazie #TuttiColoroCheHannoPartecipato #TerzoSettore #CittàSana #Incontro #Pomeriggio #Organizzazione #Progetto #Felicità #Partecipazione #Discussione #Grazie #TuttiColoroCheHannoPartecipato #TerzoSettore",
5     "lang": "it",
6     "retweet_count": 0,
7     "reply_count": 0,
8     "like_count": 0,
9     "quote_count": 0,
10    "geo": {},
11    "hashtags": [],
12    "mentions": [],
13    "urls": [
14      "https://www.facebook.com/100044147480471/posts/555729582575286/?d=n",
15      "https://twitter.com/elenabonetti/status/1522662850212376578/photo/1",
16      "https://twitter.com/elenabonetti/status/1522662850212376578/photo/1"
17    ],
18    "ctx_annotations": []
19  },
20  "user": {
21    "created_at": "2013-01-30T15:17:59.000Z",
22    "description": "Ministra per le Pari Opportunità e la Famiglia 🇮🇹 @italiaviva - Professoressa di analisi matematica",
23    "id": "1134357498",
24    "name": "Elena Bonetti",
25    "profile_image_url": "https://pbs.twimg.com/profile_images/1310606661028110341/PjU05E48_normal.jpg",
26    "followers_count": 27256,
27    "following_count": 648,
28    "tweet_count": 5106,
29    "listed_count": 207,
30    "url": "",
31    "username": "elenabonetti",
32    "verified": true
33  }
}
```

e il numero di tweet fatti possiamo sapere quanti tweet ha prodotto al giorno tramite uno script python:

Esercizio

- leggere tutti i tweet dalla directory di input
- calcolare per ogni tweet:

```
activity_avg = num_tweets / num_days_account_exists
```

- salvare i tweets arricchiti in una cartella enriched

52

The screenshot shows a code editor with a dark theme. At the top, there are tabs for 'activity_avg.py' and 'activity_avg.py'. Below the tabs, the Python code is displayed:

```

1 import json
2 from datetime import datetime
3
4 # Read data coming from stdin.
5 line = input()
6
7 # Parse JSON data.
8 parsedJson = json.loads(line)
9
10 created_at = parsedJson["user"]["created_at"]
11 tweet_count = parsedJson["user"]["tweet_count"]
12
13 timestamp = datetime.strptime(created_at, "%Y-%m-%dT%H:%M:%S.%fZ")
14 delta = datetime.now() - timestamp
15 activity_avg = round(tweet_count / delta.days, 2)
16
17 parsedJson["user"]["activity_avg"] = activity_avg
18 enrichedJson = json.dumps(parsedJson)
19 print(enrichedJson)
20
21

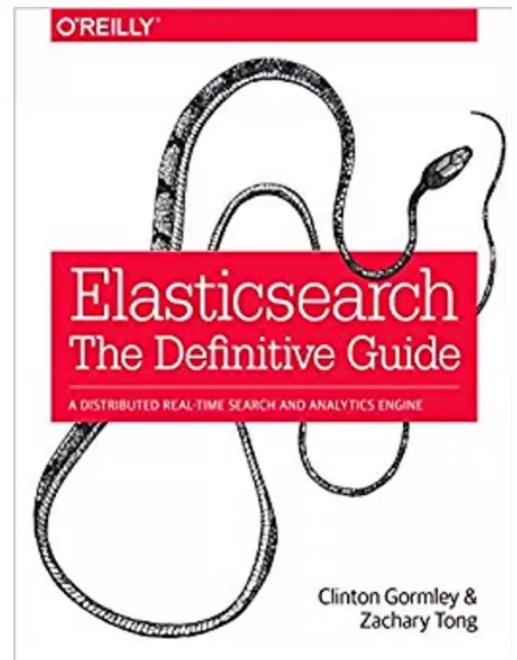
```

The code reads a single JSON line from standard input, parses it, calculates the average tweets per day since the account was created, adds this as a new field ('activity_avg') to the JSON object, converts it back to a string using json.dumps(), and then prints the enriched JSON.

ElasticSearch

ELASTICSEARCH

- “Elasticsearch is a real-time distributed search and analytics engine”
- motore di ricerca basato su Apache Lucene
- open source
- versione 7.17.3
- fa parte di Elastic Stack



Nasce come motore di ricerca real time per poter fare analytics ovvero fare dei calcoli e delle analisi sui dati. E' un progetto basato su Apache Lucene. Fa parte di Elastic Stack:

ELASTIC STACK



con Elasticsearch che è il motore di ricerca, Logstash che consentiva storicamente di caricare i GB di dati di log. E' stato poi costruito Kibana, in quanto era necessario mostrare queste informazioni con delle grafiche accattivanti e interattive.

Caratteristiche principali

CARATTERISTICHE PRINCIPALI



- basato su REST API (http & Json)
- near real-time
- full text search
- scalabile
- eventually consistent
- analisi linguaggio
- query DSL
- versioning

è near real-time ovvero pensato per gestire e interrogare una quantità di dati spaventosa. Eventually consistent nel senso che quando si parla di big data bisogna sacrificare la consistenza del dato il quale potrebbe non sempre esserci e non sempre aggiornato.

Definizioni

DEFINIZIONI

- *document*: singolo oggetto json memorizzato
- (*type*: identifica una classe di documenti dello stesso tipo o che hanno dei campi comuni – *deprecato*: adesso si usa sempre `_doc` e ogni indice ha un solo tipo)
- *index*: collezione di documenti con caratteristiche simili
- *node*: singolo server che ospita i dati e partecipa alle operazioni di indicizzazione e di ricerca del cluster
- *cluster*: collezione di nodi che nell'insieme contengono tutti i dati e forniscono le funzionalità di indicizzazione e ricerca
- *shards & replicas*: istanza Lucene che contiene una porzione di dati dell'indice.

https://www.elastic.co/guide/en/elasticsearch/reference/current/_basic_concepts.html

non viene salvato più un record ma un documento (un json). Storicamente Elasticsearch identificava la classe di documenti con una tipologia, questo è stato però deprecato. Tutti i documenti hanno un tipo standard di tipo `_doc` e ogni indice usa un solo tipo. Un indice va inteso come una sorta di database, dunque UNA COLLEZIONE DI DOCUMENTI.

Il cluster è un insieme di nodi (server), dove ci sarà un nodo primario che si interfaccia con l'utente.

ElasticSearch vs Apache Solar

ELASTICSEARCH VS SOLR



ElasticSearch e Apache Solr hanno performance molto simili, entrambe forniscono API per la maggior parte dei linguaggi di programmazione ma:

- ES ha sposato in pieno il paradigma REST
- la sintassi per le query DSL è molto flessibile e potente. Solr NON ha qualcosa di analogo
- ES riesce a fare il mapping dei dati in modo semplice e automatico
- l'installazione è semplice e non richiede l'utilizzo di complicati tools
- ES ha un sacco di funzionalità analitiche non ancora presenti in SOLR e un layer per fare visualizzazioni molto potente (Kibana) completamente integrato

ElasticSearch vs Open Search

ES VS OPEN SEARCH



Nel 2021, un gruppo di sviluppatori ha fatto un fork di Elastic Search e ha creato un nuovo progetto chiamato Open Search.

Il Fork si è originato dai seguenti cambi di licenza:

- 2014: Elastic Search è stato rilasciato sotto la licenza Apache, versione 2.0, che è una licenza open-source permissiva.
- 2018: Elastic NV, l'azienda dietro Elastic Search, ha introdotto la Elastic License, che include ulteriori restrizioni sull'uso e alla distribuzione del software.
- Nel 2021, Elastic NV ha annunciato che avrebbe cambiato la licenza di Elastic Search e di Elastic Stack con una nuova licenza chiamata Server Side Public License (SSPL)

La SSPL è simile alla licenza introdotta da MongoDB nel 2019 e include ulteriori restrizioni sull'uso del software da parte dei *cloud provider*.

Index, Node, Shard e Replica

INDEX, NODE, SHARD & REPLICA



I documenti json sono raggruppati in indici

Produce Index

```
{  
  "name": "Baby Carrots(1lb bag)",  
  "category": "Vegetables",  
  "brand": "365",  
  "price": "$0.99"  
}  
  
{  
  "name": "Clementines(3lb bag)",  
  "category": "Fruits",  
  "brand": "Cuties",  
  "price": "$4.29"  
}
```

Wine & Beer Index

```
{  
  "name": "Unanime Malbec(750ml)",  
  "brand": "Mascota Vineyards",  
  "country": "Argentina",  
  "region": "Mendoza",  
  "wine_type": "Red Wine",  
  "ABV": "14%",  
  "price": "$22.99"  
}  
  
{  
  "name": "Hazy Little Thing IPA",  
  "brand": "US",  
  "country": "California",  
  "beer_type": "Ale",  
  "beer_style": "India Pale Ale",  
  "ABV": "6.7%",  
  "price": "$14.99"  
}
```



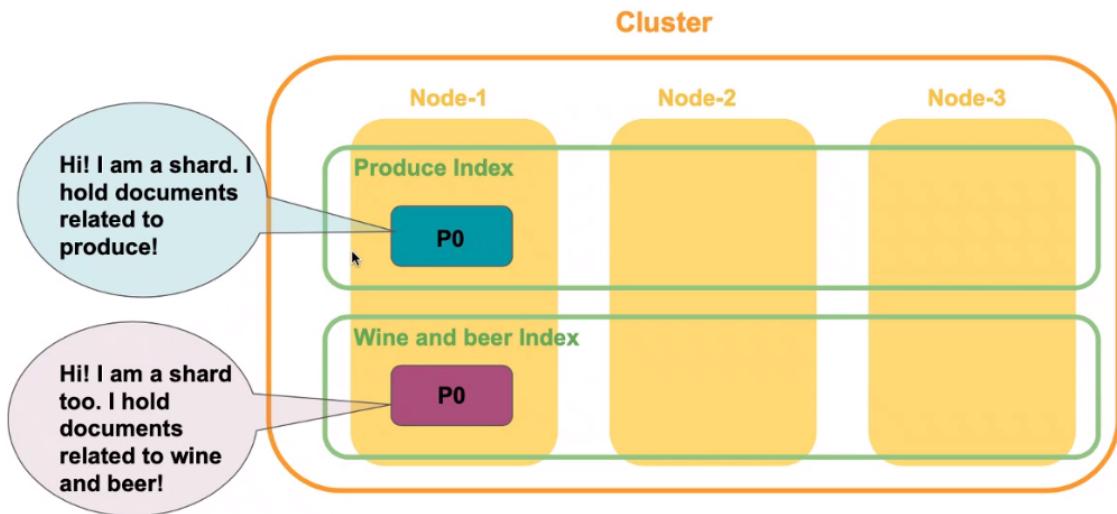
Posso creare un indice di prodotti e un indice di vino e birra ad esempio.

INDEX, NODE, SHARD & REPLICA

ES può essere installato su uno o più nodi (Server fisici o virtuali)

Un cluster è un insieme di nodi che condividono gli stessi dati

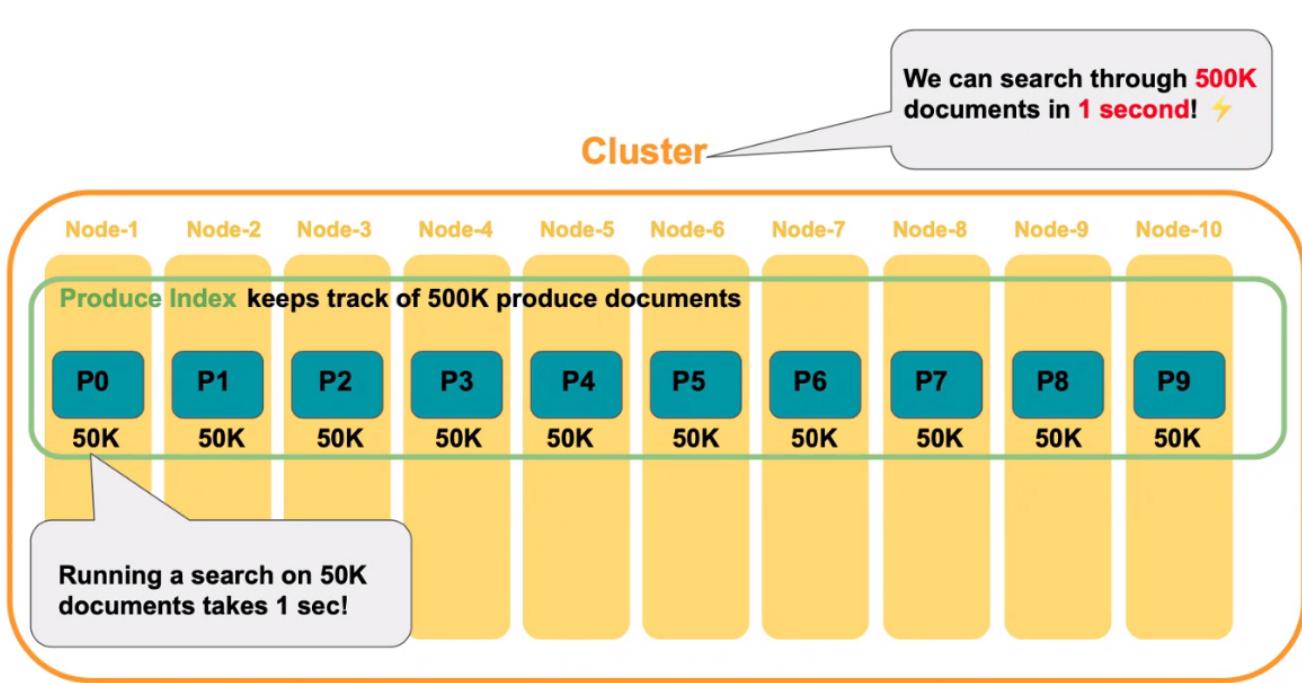
Uno shard contiene una porzione di dati dell'indice



L'indice andrà su più nodi.

INDEX, NODE, SHARD & REPLICA

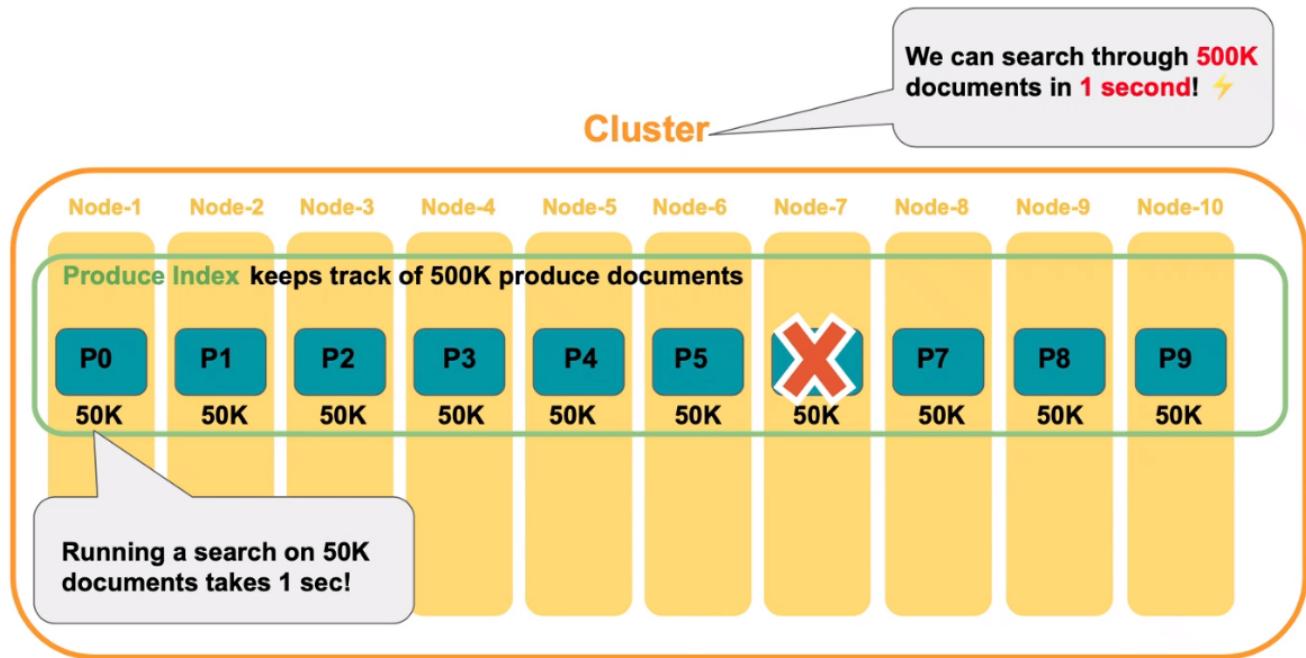
I documenti json sono raggruppati in indici



La query verrà eseguita in parallelo su tutti i nodi ed esiste un meccanismo di replica che consente di gestire questa cosa.

INDEX, NODE, SHARD & REPLICA

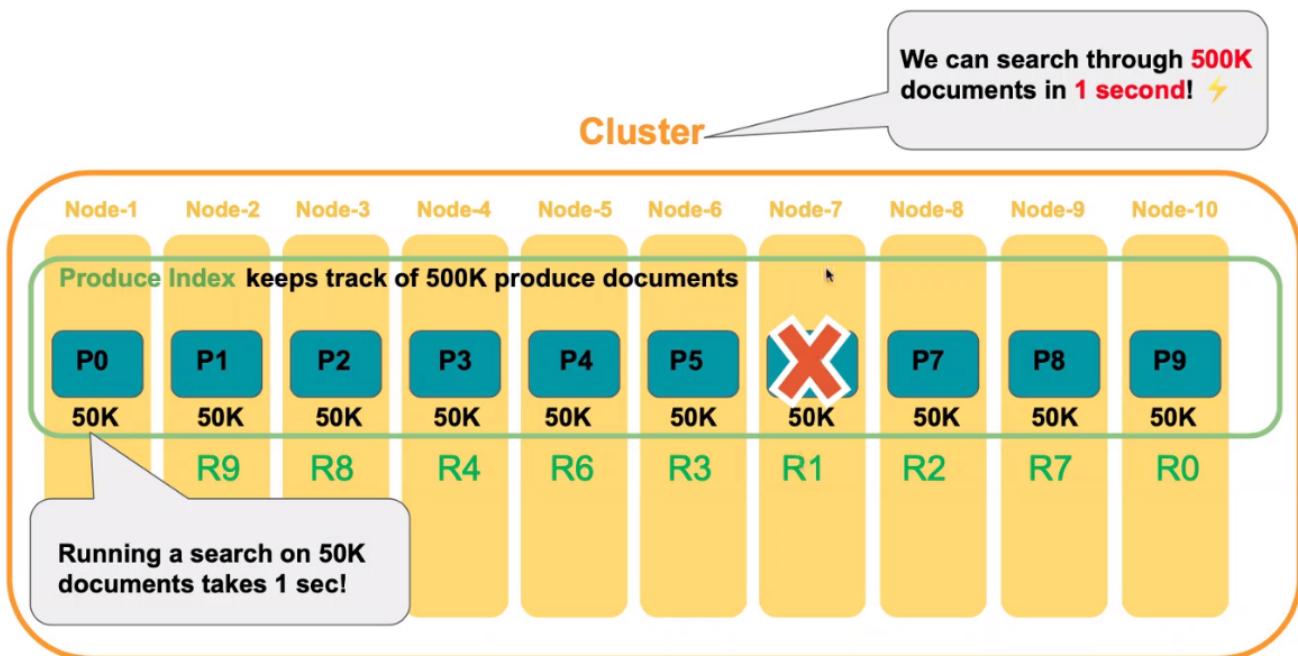
Se fallisce un nodo, è possibile rispondere lo stesso alla query utilizzando le repliche



Nel caso in cui il nodo 7 muore, su ciascun nodo ho anche le repliche di altri nodi (R9,R8,R4...):

INDEX, NODE, SHARD & REPLICA

Se fallisce un nodo, è possibile rispondere lo stesso alla query utilizzando le repliche



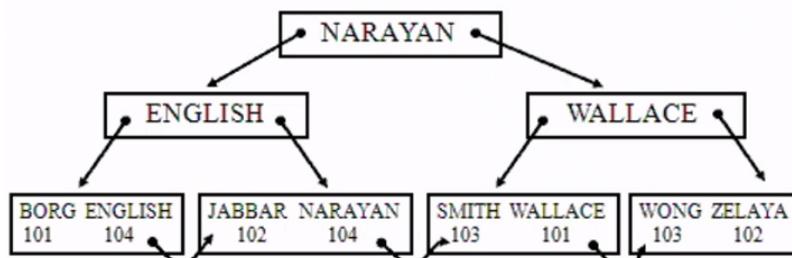
dunque il motore di elasticsearch andrà a rispondere usando la replica e non la shard che non è più accessibile.

ES <-> MODELLO RELAZIONALE

- index = database
- document = tupla (record nel DB)

Block	FNAME	LNAME	EMPID	DNO	SALARY
101	JAMES	BORG	888665555	1	55000
	JENNIFER	WALLACE	987654321	4	43000
102	AHMAD	JABBAR	987987987	4	25000
	ALICIA	ZELAYA	999887777	4	25000
103	JOHN	SMITH	123456789	5	30000
	FRANKLIN	WONG	333445555	5	40000
104	JOYCE	ENGLISH	453453453	5	25000
	RAMESH	NARAYAN	666884444	5	38000

B+ Tree Index



OTTIMIZZAZIONI

- Si raccomanda di lasciare almeno il 50% della memoria per lo heap di ElasticSearch e il 50% libera
- Meglio evitare di superare i 32GB di heap
- **E' importante disabilitare lo swap completamente**
 - > sudo swapoff -a
- per lanciare ES come un servizio
 - > ./bin/elasticsearch -d -p pid
- per configurazioni particolarmente grandi e complesse ci sono 2 tools:

[Puppet](#)

[Chef](#)

GESTIONE INDICI



Fare una lista degli indici:

```
> GET /_cat/indices?v
```

Creare un indice di nome "customer"

```
> PUT /customer?pretty
```

cancellare un indice

```
> DELETE /customer?pretty
```

Per effettuare queste operazioni si può usare la shell:

```
lilpil@win10-VM:~$ curl -XPUT http://localhost:9200/prova
{"acknowledged":true,"shards_acknowledged":true,"index":"prova"}
```

health	status	index	uuid	pri	rep	docs.count	docs.deleted	store.size	pri.store.size
green	open	.geoip_databases	-HkCuzzuRK-gQPibahWlNA	1	0	42	0	39.9mb	39.9mb
green	open	.kibana_task_manager_7.17.3_001	4jWMQQKtzulgu76o3NuQ	1	0	17	11	18.2kb	18.2kb
green	open	.apm-custom-link	gISiqP95Sn-wnaiLg_kONA	1	0	0	0	226b	226b
yellow	open	prova	uOglAttPS_qOyc1togMtiQ	1	1	0	0	226b	226b
green	open	.apm-agent-configuration	5jNy-jNMRie0Wa4uMsFRvg	1	0	0	0	226b	226b
green	open	.kibana_7.17.3_001	jkjUyR64Q9aBn-mo0Mm08g	1	0	10	0	2.3mb	2.3mb

oppure usando kibana:

The screenshot shows the Elasticsearch Dev Tools interface with the 'Console' tab selected. The search bar at the top contains the query: `GET _search`. Below the search bar, there is a code editor area with the following code:

```
1 GET _search
2 {
3   "query": {
4     | "match_all": {}
5   }
6 }
7
8
9 PUT nuovo_indice
10
11 GET _cat/indices?v
```

On the right side of the interface, the results of the search are displayed in a table format. The table has columns for index name, type, status, and various metrics. The results are as follows:

Index	Type	Status	Score	Size (mb)	Size (kb)
yellow open tweets_geomapping		zvhXGHqQrGG9D3VW5uLnA	1 1 4144	0 6.1mb	6.1mb
yellow open prova		x1ZFV5uTsyonAcl12XNs	1 1 0	0 226b	226b
green open .apm-agent-configuration		815JEHsMSyU0FHITGZUUmA	1 0 0	0 226b	226b
yellow open esercitazione		jOPAG-hzSRGxskFHL78UYA	1 1 9097	0 12.5mb	12.5mb
yellow open tweets_geografici		xhSWdh-oQo04mSLG8euUxVA	1 1 497	0 914.5kb	914.5kb
yellow open pippo		UafINGFZRlKaZzT-ZgrUg	1 1 2	0 7.4kb	7.4kb
yellow open twitter_prova		Ook_2q7RQd6Id8PNIaqBQ	1 1 1641	0 2.7mb	2.7mb
green open .tasks		AeQYkEtVQY6a1HEWNQz4oQ	1 0 22	0 45kb	45kb
yellow open twitter_musk		NF3bC-yWSM2PqeBBQdmwTw	1 1 18976	0 25.7mb	25.7mb
green open .geoip_databases		R265yhYFQXmooiY_Y197_A	1 0 42	42 40.2mb	40.2mb
yellow open nuovo_indice		vzGjyAcqQhaozT_z4WRs-g	1 1 0	0 226b	226b
yellow open politici_italiani		cqq08E43RxakK-1MgtMsQ0	1 1 276	0 620.5kb	620.5kb
green open .kibana_task_manager_7.17.3_001		s19e3YQjTE2V2h0hxtCkBA	1 0 17	55698 6.8mb	6.8mb
yellow open twitter		pxhXn0595_af4F79uyrzHg	1 1 10198	0 14.1mb	14.1mb
green open .reporting-2022-05-29		PWYwhsf55ZemBpooQyR3hyQ	1 0 2	1 2mb	2mb
green open .apm-custom-link		Thm4IB3hrX6Y2wqlhc2s3A	1 0 0	0 226b	226b
green open .async-search		Zb-r6BDrqZejViio1MBQlg	1 0 0	0 3.7kb	3.7kb
yellow open geo_tweets		usULfitqSiWqJ3RzwPYdhw	1 1 66029	0 87mb	87mb
yellow open paperino		X5Zqd17VShOUFA3yOjRdhg	1 1 6	0 35.9kb	35.9kb
green open .kibana_7.17.3_001		q4KNdbr1SWej9xGX3khocA	1 0 161	4 2.6mb	2.6mb
yellow open customer		H9Z-PwZtR0enW9JFRVsUpw	1 1 4	0 5.7kb	5.7kb

REST API

La regola generale (type è sempre = «_doc»)

> <REST Verb> /<Index>/<Type>/<ID>

Inserire un documento

```
> PUT /customer/_doc/1?pretty
{
    "name": "John Doe"
}
```

se l'indice non esiste viene creato automaticamente

recuperare un documento da un certo indice

> GET /customer/_doc/1?pretty

The screenshot shows the Elasticsearch Dev Tools interface with the 'Console' tab selected. The console displays a series of Elasticsearch API requests and their responses. The requests include:

- GET _search
- PUT nuovo_indice
- GET _cat/indices?v
- DELETE nuovo_indice?pretty
- PUT prova2/_doc/1
- POST /customer/_doc/1?pretty (This is the highlighted request in the screenshot)

The response table shows the following data for the customer index:

uuid	pri	rep	docs.count	docs.deleted	store.size	pri.store.size
xT4pAnUlRwmvnGkCVWI3zQ	1	1	1	0	4.3kb	4.3kb
815jEHsMSYu0FHITGZUJnA	1	0	0	0	226b	226b
zbvhXGHq0rG9D3W5uLnA	1	1	4144	0	6.1mb	6.1mb
x1ZPV5uTSyonAcL12XNsg	1	1	0	0	226b	226b
jOPAG-hzSRGxsxkFHL78UYA	1	1	9097	0	12.5mb	12.5mb
xhSWdh-q0o04m5LGselJxVA	1	1	497	0	914.5kb	914.5kb
UofINGFZRLKAzT-ZgrIUg	1	1	2	0	7.4kb	7.4kb
Ook_zg7RQd6Id8PN1Ag0BQ	1	1	1641	0	2.7mb	2.7mb
AeQkEtVQY6a1HENWNo24oQ	1	0	22	0	45kb	45kb
NF3bC-yWSM2PaeBBQdmwTw	1	1	18976	0	25.7mb	25.7mb
R265yhYFQXmooiY197_A	1	0	42	42	40.2mb	40.2mb
cqq08E43RxaKx-1WgtMs0Q	1	1	276	0	620.5kb	620.5kb
s19e3yQjTE2V2h0hxtCkB8A	1	0	17	55908	6.8mb	6.8mb
pxhKn0S95_qf4F79uryrzHg	1	1	10198	0	14.1mb	14.1mb
THm4IB3hRX6Y2wqlhCzs3A	1	0	0	0	226b	226b
PWYwhsFS5ZembooQyR3hyQ	1	0	2	1	2mb	2mb
2b-r6BDtQZejViiolMBQLg	1	0	0	0	3.7kb	3.7kb
X5Zad17VSh0UFA3y0jRdng	1	1	6	0	35.9kb	35.9kb
usULf1tgSiWqJ3RzwPydhw	1	1	66029	0	87mb	87mb
q4KWdbRISWej9xGX3khocA	1	0	163	17	2.7mb	2.7mb
H9Z-Pw2iROenW9jFRV5uPw	1	1	4	0	5.7kb	5.7kb

Se uso la POST senza passare l'ID, ES crea un ID randomico per il documento aggiunto:

```

7 "max_score" : 1.0,
8 "hits" : [
9   {
10     "_index" : "prova2",
11     "_type" : "_doc",
12     "_id" : "1",
13     "_score" : 1.0,
14     "_source" : {
15       "name" : "maurizio",
16       "cognome" : "tesconi"
17     }
18   },
19   {
20     "_index" : "prova2",
21     "_type" : "_doc",
22     "_id" : "w4_GoigBLNmFiddheU", ←
23     "_score" : 1.0,
24     "_source" : {
25       "name" : "topolino",
26       "eta" : 33
27     }
28   }
29 ]
30 ]
31 ]
32 ]
33 ]
34 ]
35 ]
36 ]
37 ]
38 ]
39 ]
40 ]

```

200 - OK 33 ms

passare esplicitamente un ID è un qualcosa che rallenta di molto l'operazione, in quanto ogni volta che si passa un ID lui deve controllare che quell'ID non esista già altrimenti dovrebbe fare il versioning anzichè la semplice aggiunta.

Update e Delete

UPDATE & DELETE



Update di un documento

```

> POST /customer/_doc/1/_update?pretty
{
  "doc": { "name": "Jane Doe", "age": 20 }
}

```

Cancellare un documento

```

> DELETE /customer/_doc/2?pretty

```

Bulk API

BULK API

la BULK API serve per gestire più documenti con 1 sola chiamata

```
> POST /customer/_bulk?pretty
{"index": {"_id": "1"} } ← comando
{"name": "John Doe" } ← documento
{"index": {"_id": "2"} } ← comando
{"name": "Jane Doe" } ← documento
```

si possono mescolare operazioni diverse

```
> POST /customer/_bulk?pretty
{"update": {"_id": "1"} }
{"doc": { "name": "John Doe becomes Jane Doe" } }
{"delete": {"_id": "2"} }
```

Mapping

MAPPING



è il processo per definire gli schemi dei documenti indicizzati dentro ES, tramite il mapping è possibile definire:

- quali stringhe devono essere trattate come indici full text
- quali campi contengono numeri, date, punti geolocalizzati o altri tipi di dato
- quali campi devono essere inseriti nella ricerca nel campo speciale _all
- il formato dei campi
- regole personalizzate per gestire l'aggiunta dinamica di campi

In questo esempio, ci stiamo riferendo all'indice twitter_index dove con questo comando riesco a far gestire le date presenti all'interno dei tweet e fargliele indicizzare nel modo giusto:

MAPPING - ESEMPIO



```
PUT twitter_index
{
  "mappings": {
    "properties": {
      "created_at": {
        "type": "date",
        "format": "strict_date_time"
      }
    }
  }
}
```

...

Tipi di dato

TIPI DI DATO



sono gestiti campi di tipo semplice come:

- text
- keyword
- date
- long
- double
- boolean
- ip

campi di natura gerarchica come:

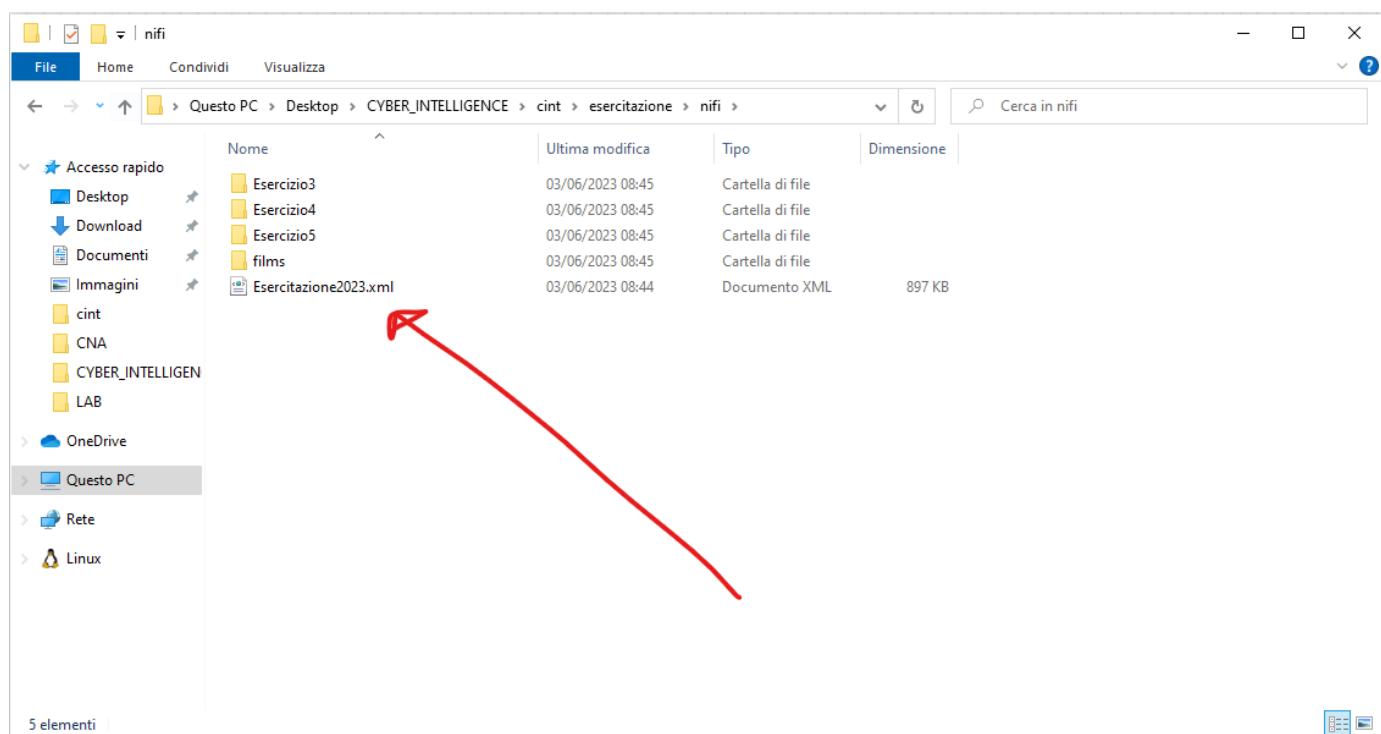
- object
- json

campi più specializzati come:

- geo_point
- geo_shape
- completion

Per inviare i dati ad elasticsearch da NiFi tramite il processor PutElasticsearchHttp 1.16.0:

16/06/2023



Questo file contiene tutti gli esercizi in formato template.

Una volta importato il template avremo i diversi DataFlow group contenuti al suo interno divisi per esercizio.

Esercizio 1

Esercizio 1

In questo esercizio impareremo a usare e configurare i seguenti processori:

1. GetHTTP per accedere senza autenticazione ai dati forniti da un web service.
2. EvaluateJSONPath per catturare alcune informazioni dal JSON di input.
3. ReplaceText per generare un FlowFile con contenuto customizzato.

Obiettivo: Periodicamente ogni 10 secondi, ottenere e salvare su due cartelle distinte il prezzo attuale in dollari e euro delle crittovalute Bitcoin e Ethereum.

Per ottenere il prezzo corrente, si utilizzerà il Web service di CryptoCompare:

<https://www.cryptocompare.com/api/>

8

Esempio di richiesta/risposta del Web service

Esempio di richiesta per risolvere l'esercizio:

<https://min-api.cryptocompare.com/data/pricemultifull?fsyms=BTC,ETH&tsyms=USD,EUR>

Estratto JSON di risposta:

```
"RAW": {  
    "BTC": {  
        "USD": {  
            "TYPE": "5",  
            "MARKET": "CCCAGG",  
            "FROMSYMBOL": "BTC",  
            "TOSYMBOL": "USD",  
            "FLAGS": "4",  
            "PRICE": 6758.39,  
            "LASTVOLUME": 1.82501163,  
            "..."  
        },  
        "EUR": {  
            "TYPE": "5",  
            "MARKET": "CCCAGG",  
            "FROMSYMBOL": "BTC",  
            "TOSYMBOL": "EUR",  
            "FLAGS": "4",  
            "PRICE": 5758.39,  
            "LASTVOLUME": 1.82501163,  
            "..."  
        }  
    }  
}
```

9

NiFi: il processore GetHTTP

Scopo: Permette di interrogare un server HTTP, utile per l'accesso a Web service di tipo REST.

Configure Processor

SETTINGS	SCHEDULING	PROPERTIES	COMMENTS
Required field			
Property	Value		
URL	https://min-api.cryptocompare.com/data/pricemultifull?fsyms=BTC&tsyms=EUR		
Filename	btceth_\${now():toNumber()}		
SSL Context Service	StandardSSLContextService		
Username	No value set		
Password	No value set		
Connection Timeout	30 sec		
Data Timeout	30 sec		
User Agent	No value set		
Accept Content-Type	No value set		
Follow Redirects	false		
Redirect Cookie Policy	default		
Proxy Host	No value set		
Proxy Port	No value set		

Importante configurare
il contesto SSL per
accesso a servizi in
https

Per l'HTTPS bisogna settare:

SETTINGS PROPERTIES COMMENTS

Required field

Property	Value
Keystore Filename	No value set
Keystore Password	No value set
Key Password	No value set
Keystore Type	No value set
Truststore Filename	/usr/lib/jvm/default-java/lib/security/cacerts
Truststore Password	Sensitive value: /usr/lib/jvm/default-java/lib/security/cacerts
Truststore Type	JKS
TLS Protocol	TLS

CANCEL APPLY



Esercizio 1: traccia su come realizzare il dataflow

1. Configurate il processore GetHttp per accedere al WebService tramite l'URL indicato.
2. Impostate il processore GetHttp affinchè sia schedulato ogni 10 secondi.
3. Processate i JSON provenienti dal Web service tramite il processore EvaluateJSONPath ed inserire dei nuovi attributi con il prezzo di BTC e ETH.
4. Splittate il flusso in due sottoflussi, ognuno corrispondente a ciascuna crittovaluta.
 - a. Usate il processore ReplaceText per riscrivere i dati nel formato voluto, ad esempio CSV.
 - b. Scrivete i dati su una cartella di output tramite il processore PutFile.

Esercizio 2

Esercizio 2

Problemi con soluzione Esercizio 1:

- Ogni flusso ha il suo processore PutFile
- Ogni file su disco contiene una sola misurazione.

Obiettivo: modificare il flusso di Esercizio 1 affinchè da ciascun sottoflusso escano FlowFile contenenti 5 misurazioni ciascuno e ognuno di questo sia salvato su una opportuna cartella sulla base del valore dell'attributo "CryptoCurrency".

Nuovi processori:

- UpdateAttribute per aggiungere un nuovo attributo.
- MergeContent per unire i contenuti

14

Esercizio 2: traccia su come realizzare il dataflow

1. Partire da una copia dell'Esercizio 1.
2. In ciascun sottoflusso, dopo avere generato il FlowFile con formato custom, andare a inserire mediante il processore UpdateAttribute un nuovo attributo "CryptoCurrency" contenente il valore "BTC" o "ETH".
3. Nello stesso sottoflusso, utilizzare il processore MergeContent per unire 5 FlowFile differenti in un nuovo FlowFile. Il processore MergeContent deve essere impostato a:
 - "Merge Strategy" = "Bin-Packing algorithm"
 - "Merge Format" = "Binary concatenation"
 - "Delimiter strategy" = "Text"
 - "Demarcator" = newline (premere Shift + Invio)
4. Utilizzare una unica istanza del processore PutFile per ricevere i FlowFile aggregati dai 2 sottoflussi e sfruttare l'attributo "CryptoCurrency" per scrivere i dati in cartelle diverse.

[Qui](#) trovate maggiori dettagli sul processore MergeContent

15

DA COMPLETARE! VEDERE REGISTRAZIONE!**

17/06/2023

Configure Processor | RouteOnAttribute 1.16.0

■ Stopped

SETTINGS SCHEDULING PROPERTIES RELATIONSHIPS COMMENTS

Required field

Property	Value
Routing Strategy	Route to Property name
BigFile	$\$\{fileSize:gt(1000000)\}$
SmallFile	$\$\{fileSize:le(1000000)\}$

CANCEL APPLY

devono essere definite delle proprietà per poi instradare il flow file in base a queste proprietà. Non è detto che debbano essere mutuamente esclusive, in generale lo decide chi crea il flusso.
Sulla BigFile voglio mandarci tutti i flow file con dimensione maggiore di un milione di byte.
la `filesize` è una proprietà che esiste sempre sui flow file.

Reddit: how to register a new app

You need valid credentials to use Reddit API!

- After logged in, go to
<https://www.reddit.com/prefs/apps>
- Click are you a developer? create an app...

Note: [API's Terms of Usage are changing](#)

Esercizio 4

Esercizio 4

Nell'esercizio sfrutteremo i processori:

- RedditListenerProcessor per ottenere un flusso di commenti da tutti i subreddit tramite l'API di Reddit.
- ExecuteStreamCommand per eseguire uno script Python custom.

Obiettivo: Mettersi in ascolto sullo stream di Reddit che da accesso ai nuovi commenti postati sul social e processarli tramite lo script EnrichSentiment.py (disponibile nella cartella Esercizio4). Lo script arricchisce i FlowFile con la polarità del sentiment di ciascun commento. Sulla base del valore di questo sentiment, si scrivono i FlowFile finali in tre cartelle distinte (sentiment negativo, sentiment neutro e sentiment positivo).

Esercizio 6

Esercizio 6

Obiettivo

Riadattare i dataflow degli Esercizi 4 e 5 in modo tale da incapsularli in 2 componenti riutilizzabili. Sfruttando questi 2 nuovi componenti, scrivere un dataflow facente le seguenti cose:

1. Si mette in ascolto sulle nuove submission di Reddit provenienti da /r/all .
2. Le submission recuperate vengono processate e arricchite dal componente “Esercizio 4”.
3. Le submission arricchite dal componente “Esercizio 4” (solo quelli con polarità positiva o negativa) sono quindi processati dal componente “Esercizio 5”.
4. Le submission provenienti da componente “Esercizio 5” (sia con contenuto Web sia senza) sono memorizzate su Elasticsearch nell’indice “esercizio6”.
5. Le submission aventi polarità neutra oppure senza link a contenuto Web sono raccolte e memorizzate in una cartella di output.

Su Kibana bisogna creare un index pattern di cui vogliamo visualizzare l’indice:

The screenshot shows the Elasticsearch Kibana Management interface at the URL `localhost:5601/app/management/kibana/indexPatterns`. The left sidebar contains navigation links for Management, Ingest Pipelines, Data (Index Management, Lifecycle Policies, Snapshot and Restore, Rollup Jobs, Transforms, Remote Clusters), Alerts and Insights (Rules and Connectors, Reporting, Machine Learning Jobs), Kibana (Index Patterns, Saved Objects, Tags, Search Sessions, Spaces, Advanced Settings), and Stack (License Management, Upgrade Assistant). The main area is titled "Index pattern" and "Create index pattern". It includes fields for "Name" (set to "esercizio6"), "Timestamp field" (set to "none"), and a "Source" section listing "esercizio6" and "prova" with "Index" selected. A "Create index pattern" button is at the bottom right. The top bar shows the Elasticsearch logo and the URL.

Dopo aver creato l’indice recarsi su Analytics -> Discover:

localhost:5601/app/discover/?_g=(filters:[],refreshInterval:(pause:0,value:0),time:(from:now-15m,to:now))&_a=(columns:[],filters:[],interval:auto,query:(language:kuery,query:""),sort:[],time:(from:now-15m,to:now))

Importa preferiti GitHub - tiziano/cyber... NiFi Flow Elasticsearch Kibana preferenze (reddit...)

elastic Search Elastic Options New Open Share Inspect Save Refresh

Analytics

- Overview
- Discover** (highlighted with a red arrow)
- Dashboard
- Canvas
- Maps
- Machine Learning
- Visualize Library

Enterprise Search

- Overview
- App Search
- Workplace Search

Observability

- Overview
- Alerts
- Cases
- Logs
- Metrics
- APM
- Uptime

+ Add integrations

23 hits

Document

> sentiment_polarity: 1 submission.author.comment_karma: 2,044 submission.author.id: sg8qjh0s submission.author.is_employee: false submission.author.is_mod: false submission.author.name: SixEightAKS submission.created_utc: 1,687,589,500 submission.id: 14hlug submission.subreddit_id: t5_zubbb submission.subreddit_name: r/PlayStationPlus submission.title: Hey guys, can u recommend me must play ps5 games in ps plus extra? submission.url: https://www.reddit.com/r/PlayStationPlus/comments/14hlug/hey_guys_can_u_recommend_me_must_play_ps5_games/_id: e4UwTgB4SSG_VpL2y0-_index: esercizio6 _score: 1 _type: doc

> sentiment_polarity: 1 submission.author.comment_karma: 18,007 submission.author.id: 7xdw9 submission.author.is_employee: false submission.author.is_mod: true submission.author.name: ortofon88 submission.created_utc: 1,687,589,500 submission.id: 14hlul1 submission.subreddit_id: t5_zsdef submission.subreddit_name: r/manprovement submission.title: I've been listening to affirmations on YouTube every morning and I think it's a good habit to try out submission.url: https://www.reddit.com/r/manprovement/comments/14hlul1/ive_been_listening_to_affirmations_on_youtube/_id: fIuW7IgB4SSG_VpL2y2z _index: esercizio6 _score: 1 _type: doc

> sentiment_polarity: 1 submission.author.comment_karma: 657 submission.author.id: vn0c7sq5 submission.author.is_employee: false submission.author.is_mod: false submission.author.name: Otherwise_Hippo6885 submission.created_utc: 1,687,589,500 submission.id: 14hluk submission.subreddit_id: t5_3345f submission.subreddit_name: r/h3h3productions submission.title: While the crew was eating a dan sandwich, Richard Wolff was eating his words on Russian supremacy submission.url: https://abcnews.go.com/International/wagner-mercenary-chief-calls-army-rebellion-russian-military/story?id=100335756 webpage.content: The head of Russia's Wagner mercenary group appears to be threatening an armed rebellion against Russia's military leadership, after accusing it of deliberately shelling his forces on Friday.Wagner's founder Yevgeny Prigozhin in an audio message on Friday claimed his forces would now punish Russia's defense minister and chief of

> sentiment_polarity: -1 submission.author.comment_karma: 70 submission.author.id: uxpefig1 submission.author.is_employee: false submission.author.is_mod: true submission.author.name: reservedoperator292 submission.created_utc: 1,687,589,500 submission.id: 14hlv7 submission.subreddit_id: t5_6nm64v submission.subreddit_name: r/EuropeanForum submission.title: Putin in crisis: Wagner chief Prigozhin declares war on Russian military leadership submission.url: https://www.politico.eu/article/putin-in-crisis-as-wagner-chief-prigozhin-declares-war-on-russian-military-leadership/ webpage.content: Vladimir Putin is facing a major military crisis after Russian mercenary leader Yevgeny Prigozhin declared war on Moscow's own defense ministry, claiming Kremlin officials had killed thousands of his soldiers. In a statement issued Friday night, the FSB security agency said it had "legally and reasonably begun criminal proceedings" against the

> sentiment_polarity: 1 submission.author.comment_karma: 39,960 submission.author.id: 9rjgte4t submission.author.is_employee: false submission.author.is_mod: false submission.author.name: ExileForever submission.created_utc: 1,687,589,500 submission.id: 14hltym submission.subreddit_id: t5_ztezb submission.subreddit_name: r/FinalFantasyVII submission.title: Who would you say is your favorite female character: Tifa or Aerith and why? submission.url: https://www.reddit.com/r/FinalFantasyVII/comments/14hltym/who_would_you_say_is_your_favorite_female/_id: ciUwTgB4SSG_VpL1zt _index: esercizio6 _score: 1 _type: doc

> sentiment_polarity: -1 submission.author.comment_karma: 10,060 submission.author.id: d4nubh submission.author.is_employee: false submission.author.is_mod: true submission.author.name: Jazafalara

Selezionando l'indice di interesse sarà possibile visualizzare i dati salvati:

localhost:5601/app/discover/?_a=(columns:[],index:9a28fce0-125d-11ee-9b16-035beb1fa5b6,interval:auto,query:(language:kuery,query:""),sort:[],time:(from:now-15m,to:now))

Importa preferiti GitHub - tiziano/cyber... NiFi Flow Elasticsearch Kibana preferenze (reddit...)

elastic Search Elastic Options New Open Share Inspect Save Refresh

Search

+ Add filter (highlighted with a red arrow)

esercizio6*

23 hits

Document

> sentiment_polarity: 1 submission.author.comment_karma: 2,044 submission.author.id: sg8qjh0s submission.author.is_employee: false submission.author.is_mod: false submission.author.name: SixEightAKS submission.created_utc: 1,687,589,500 submission.id: 14hlug submission.subreddit_id: t5_zubbb submission.subreddit_name: r/PlayStationPlus submission.title: Hey guys, can u recommend me must play ps5 games in ps plus extra? submission.url: https://www.reddit.com/r/PlayStationPlus/comments/14hlug/hey_guys_can_u_recommend_me_must_play_ps5_games/_id: e4UwTgB4SSG_VpL2y0-_index: esercizio6 _score: 1 _type: doc

> sentiment_polarity: 1 submission.author.comment_karma: 18,007 submission.author.id: 7xdw9 submission.author.is_employee: false submission.author.is_mod: true submission.author.name: ortofon88 submission.created_utc: 1,687,589,500 submission.id: 14hlul1 submission.subreddit_id: t5_zsdef submission.subreddit_name: r/manprovement submission.title: I've been listening to affirmations on YouTube every morning and I think it's a good habit to try out submission.url: https://www.reddit.com/r/manprovement/comments/14hlul1/ive_been_listening_to_affirmations_on_youtube/_id: fIuW7IgB4SSG_VpL2y2z _index: esercizio6 _score: 1 _type: doc

> sentiment_polarity: 1 submission.author.comment_karma: 657 submission.author.id: vn0c7sq5 submission.author.is_employee: false submission.author.is_mod: false submission.author.name: Otherwise_Hippo6885 submission.created_utc: 1,687,589,500 submission.id: 14hluk submission.subreddit_id: t5_3345f submission.subreddit_name: r/h3h3productions submission.title: While the crew was eating a dan sandwich, Richard Wolff was eating his words on Russian supremacy submission.url: https://abcnews.go.com/International/wagner-mercenary-chief-calls-army-rebellion-russian-military/story?id=100335756 webpage.content: The head of Russia's Wagner mercenary group appears to be threatening an armed rebellion against Russia's military leadership, after accusing it of deliberately shelling his forces on Friday.Wagner's founder Yevgeny Prigozhin in an audio message on Friday claimed his forces would now punish Russia's defense minister and chief of

> sentiment_polarity: -1 submission.author.comment_karma: 70 submission.author.id: uxpefig1 submission.author.is_employee: false submission.author.is_mod: true submission.author.name: reservedoperator292 submission.created_utc: 1,687,589,500 submission.id: 14hlv7 submission.subreddit_id: t5_6nm64v submission.subreddit_name: r/EuropeanForum submission.title: Putin in crisis: Wagner chief Prigozhin declares war on Russian military leadership submission.url: https://www.politico.eu/article/putin-in-crisis-as-wagner-chief-prigozhin-declares-war-on-russian-military-leadership/ webpage.content: Vladimir Putin is facing a major military crisis after Russian mercenary leader Yevgeny Prigozhin declared war on Moscow's own defense ministry, claiming Kremlin officials had killed thousands of his soldiers. In a statement issued Friday night, the FSB security agency said it had "legally and reasonably begun criminal proceedings" against the

> sentiment_polarity: 1 submission.author.comment_karma: 39,960 submission.author.id: 9rjgte4t submission.author.is_employee: false submission.author.is_mod: false submission.author.name: ExileForever submission.created_utc: 1,687,589,500 submission.id: 14hltym submission.subreddit_id: t5_ztezb submission.subreddit_name: r/FinalFantasyVII submission.title: Who would you say is your favorite female character: Tifa or Aerith and why? submission.url: https://www.reddit.com/r/FinalFantasyVII/comments/14hltym/who_would_you_say_is_your_favorite_female/_id: ciUwTgB4SSG_VpL1zt _index: esercizio6 _score: 1 _type: doc

> sentiment_polarity: -1 submission.author.comment_karma: 10,060 submission.author.id: d4nubh submission.author.is_employee: false submission.author.is_mod: true submission.author.name: Jazafalara

Esercizio 1

Obiettivo

Sfruttando il template presente sul file “FilmsMapping.json” in “cint/esercitazione/elastic/mapping_films”, creare un nuovo mapping per ES che consideri anche i campi di arricchimento definiti negli Esercizi 4 e 5 su NiFi. I campi nuovi fanno riferimento a questi dati:

```
"sentiment_polarity": "-1",
"webpage": {
    "url": "https://t.co/6wixF7WHyq",
    "title": "PesNew Era su Twitter...",
    "content": "Ora sei..."
}
```

Sfruttare <https://www.elastic.co/guide/en/elasticsearch/reference/current/mapping-types.html>

2

```
1 PUT films
2 {
3     "settings": {
4         "analysis": {
5             "filter": {
6                 "italian_elision": {
7                     "type": "elision",
8                     "articles": [
9                         "c", "l", "all", "dall", "dell",
10                        "nell", "sull", "coll", "pell",
11                        "gl", "agl", "dagl", "degl", "negl",
12                        "sugl", "un", "m", "t", "s", "v", "d"
13                    ],
14                    "articles_case": true
15                },
16                "italian_stop": {
17                    "type": "stop",
18                    "stopwords": "_italian_"
19                }
20            },
21            "analyzer": {
22                "italian_custom_analyzer": {
23                    "tokenizer": "standard",
24                    "filter": [
25                        "italian_elision",
26                        "lowercase",
27                        "italian_stop"
28                    ]
29                }
30            }
31        },
32        "mappings": {
33            "properties": {
34                "anno": {
35                    "type": "date"
36                },
37                "attori": {
38                    "type": "text",
39                    "fieldData": true,
40                }
41            }
42        }
43    }
44 }
```

There's an update available: Visual Studio Code 1.60.0

Install

l'analizzatore per la lingua italiana

Esercizio

Sfruttando la “match” query, cercare tutti i film che verificano le query:

1. frase “rapina banca” con parole in “and” sul campo contenente la descrizione del film.
2. frase “terroismo banca” con parole in “or” sul campo contenente la descrizione del film.
3. frase “bova morante” con parole in “or” sul campo che fa riferimento agli attori di un film.

Full text queries: “match phrase” query

GET /_search

```
{
  "query": {
    "match_phrase": {
      "message": {
        "query": "this is a test",
        "slop": 0
      }
    }
  }
}
```

Cerca il matching per la frase specificata nel campo indicato. Con il parametro “slop” si può indicare quanto preciso può essere fatto il matching considerando l’ordine delle parole.

Il parametro “slop” cerca di rispondere a questa domanda

“By how far apart we mean how many times do you need to move a term in order to make the query and document match?”

<https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-match-query-phrase.html>

<https://www.elastic.co/guide/en/elasticsearch/guide/current/slop.html>

6



Full text queries: Multi Match query

GET /_search

```
{
  "query": {
    "multi_match": {
      "query": "brown fox",
      "type": "best_fields",
      "fields": [ "subject^3", "message" ],
    }
  }
}
```

Stessa semantica della “match” query ma con la possibilità di utilizzare più campi di ricerca e combinare gli score.

<https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-multi-match-query.html>

10

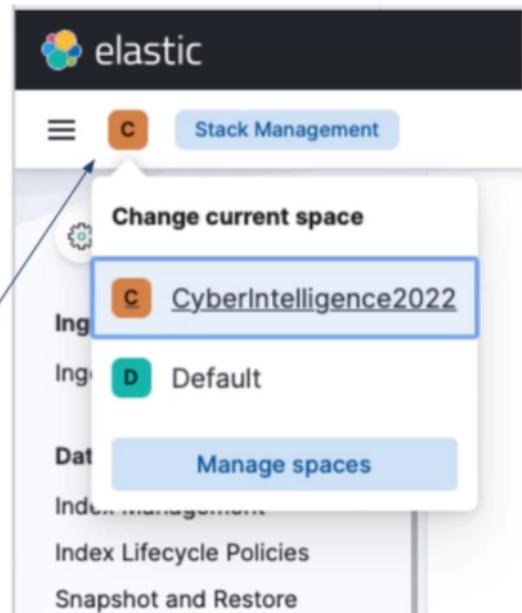
Gli space non condividono gli index pattern ma solo gli indici:



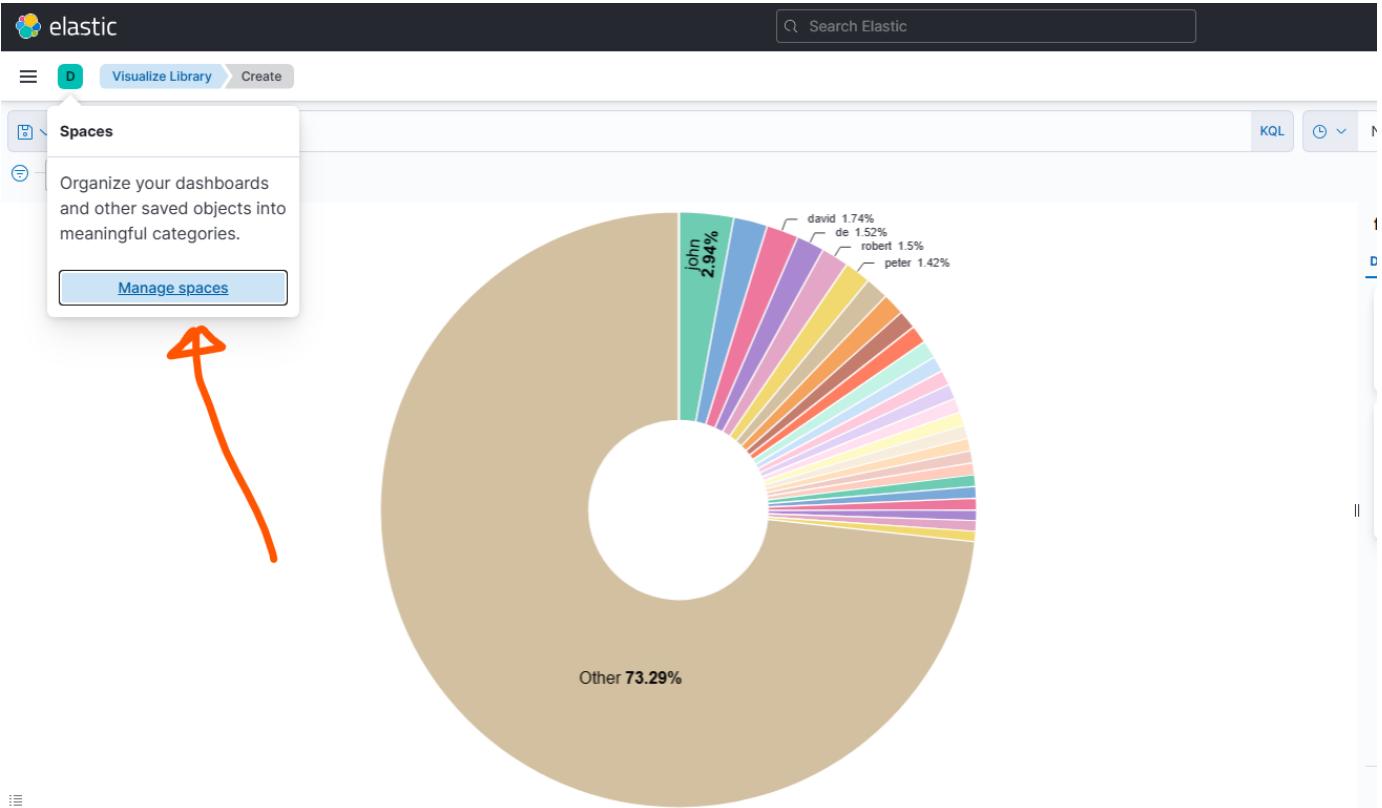
Creazione e gestione Space su Kibana

Gli Spaces permettono di organizzare i contenuti su Kibana creando dei contenitori logici sotto cui raggruppare “Index Patterns”, viste tabellari, visualizzazioni e dashboard.

Per creare e gestire uno Space andare qui



Gli indici sono sempre condivisi tra gli Space ma gli Index Patterns no!



Create space

Organize your dashboards and other saved objects into meaningful categories.

General

Describe this space

Give your space a name that's memorable.

Name

Description

Optional

The description appears on the space selection screen.

URL Identifier

You can't change the URL identifier once created.

Create an avatar

Choose how your space avatar appears across Kibana.



Avatar type

Initials Image

Initials

Enter up to two characters.

Background color

#DA8B45

Features

Per importare la visualizzazione di Kibana:

The screenshot shows the Kibana 'Saved Objects' interface. At the top right, there are buttons for Refresh, Import (highlighted with a red arrow), Export 2 objects, and other actions. Below the header is a search bar and a table with columns for Type, Title, Tags, and Actions. A single row is visible in the table, representing an object named 'film*'. At the bottom left, there's a 'Rows per page' dropdown set to 50, and at the bottom right, there are navigation arrows.

Import saved objects

X

Select a file to import



[kibana-esercitazione.ndjson](#)

[Remove](#)

Import options



Check for existing objects

ⓘ

Automatically overwrite conflicts

Request action on conflict



Create new objects with random IDs

ⓘ

[Cancel](#)

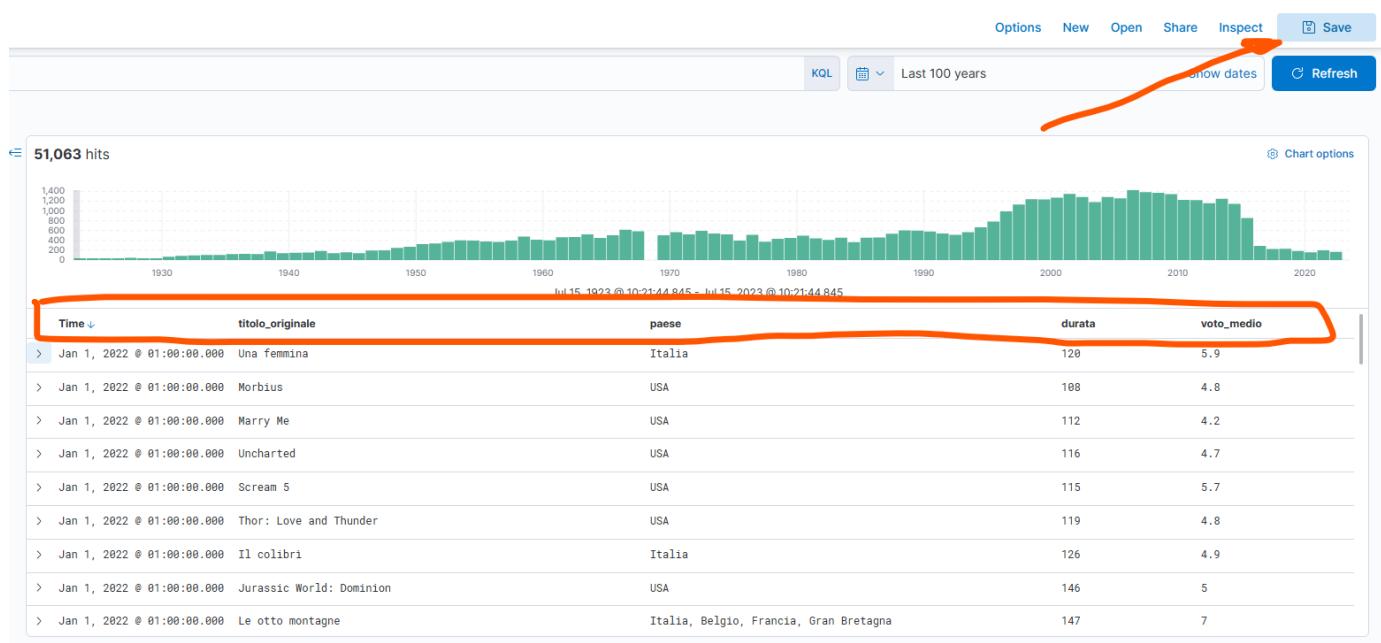
[Import](#)

Query salvate

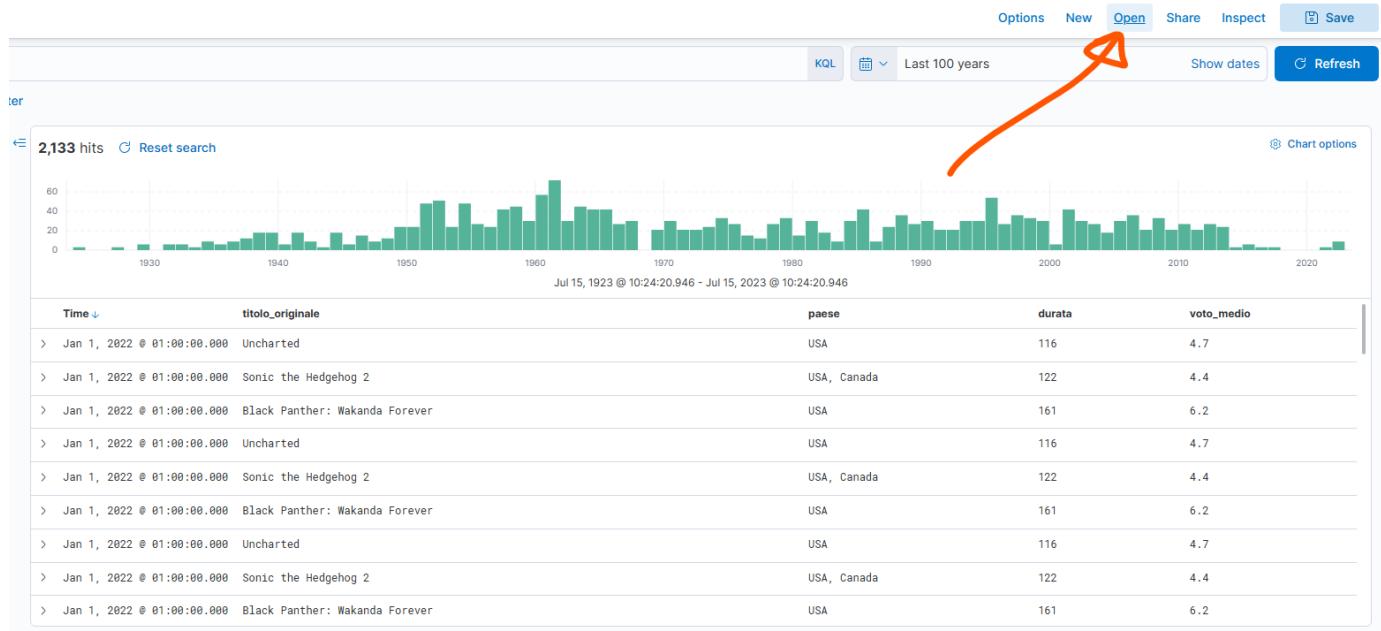
Dal tab “Discover” è possibile creare viste tabellari specifiche che visualizzano i dati con eventuali filtri applicati.

- **View1:** Creare una vista che riporta “titolo_originale”, “paese”, “durata” e “voto_medio”
- **View2:** Creare una vista che riporta “titolo_italiano”, “voti_totali”, “registi” e “attori” sui dati prefiltrati che contengono solo film con votazione media > 7.

23



Per richiamare una vista salvata bisogna cliccare su "Open":



Visualizzazione Tag Cloud

Visualizzazione 1a e 1b

Componente “Tag Cloud”

Visualizzare i 25 registi più attivi e le 25 parole più usate nei titoli italiani dei film di genere “Thriller”

a)



b)



24

- ☰ Te Discover test_save
- Home
- Recently viewed
 - test_save
- Analytics
 - Overview
 - Discover
 - Dashboard
 - Canvas
 - Maps
 - Machine Learning
 - Visualize Library
- Enterprise Search

Overview

App Search

Workplace Search

 Add integrations

New visualization

drop editor. Switch between visualization types at any time. *Recommended for most users.*

TSVB
Perform advanced analysis of your time series data.

Custom visualization
Use Vega to create new types of visualizations. *Requires knowledge of Vega syntax.*

Aggregation based
Use our classic visualize library to create charts based on aggregations.
[Explore options →](#)

Tools

- Text**: Add text and images to your dashboard.
- Controls**: Add dropdown menus and range sliders to your dashboard.

Want to learn more? [Read documentation ↗](#)

Visualizzare1

Visualizzazione1a

Visualizzazione1b

Visualizzazione1c

Visualizzazione2

Visualizzazione3

Visualizzazione4

Visualizzazione5

Visualizzazione6

Visualizzazione7

Visualizzazione7b

Create visualization

Title

Tags

Actions

New visualization

Emphasize the data between an axis and a line.

Display data in rows and columns.

Show the status of a metric.

Goal

Track how a metric progresses to a goal.

Heat map

Shade data in cells in a matrix.

Horizontal bar

Present data in horizontal bars on an axis.

Line

Display data as a series of points.

Metric

Show a calculation as a single number.

Pie

Compare data in proportion to a whole.

Tag cloud

Display word frequency with font size.

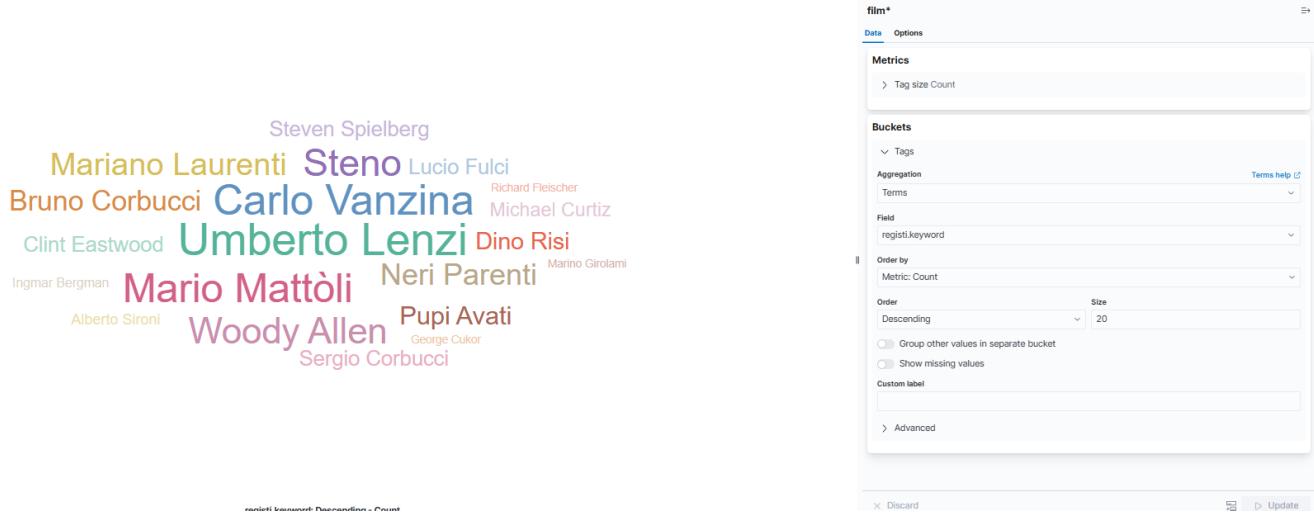
Timelion

Show time series data on a graph.

Vertical bar

Present data in vertical bars on an axis.

<campo>.keyword prende il testo completo senza essere analizzato
mentre <campo> sono le singole parole del testo



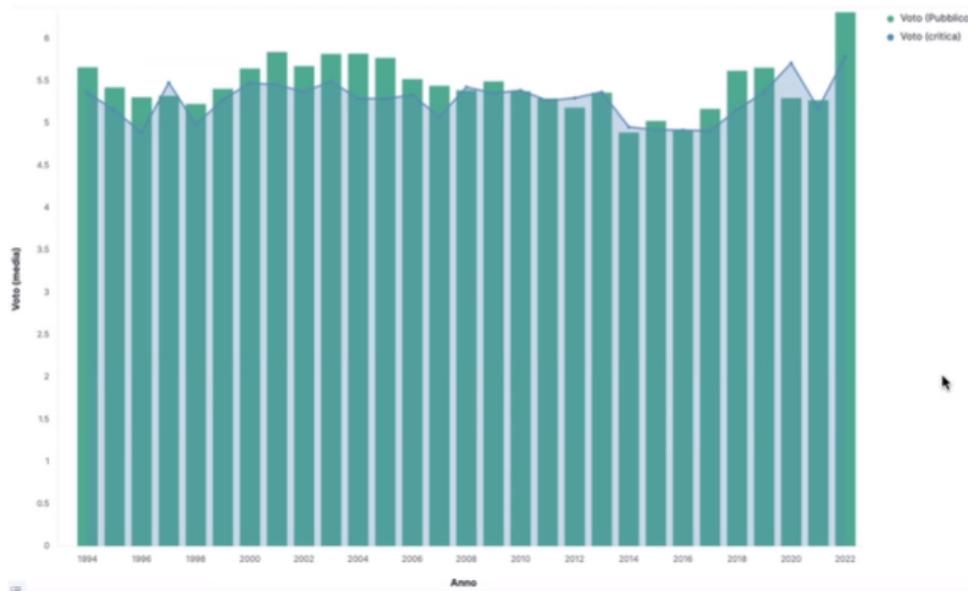
Visualizzazione Vertical Bar



Visualizzazione 2

Componente “Vertical Bar”

Visualizzare la media del voto del pubblico e la media del voto della critica di tutti i film italiani considerando gli ultimi 30 anni



26

L'idea di questa visualizzazione è di capire come si pone il voto della critica rispetto al voto del pubblico.

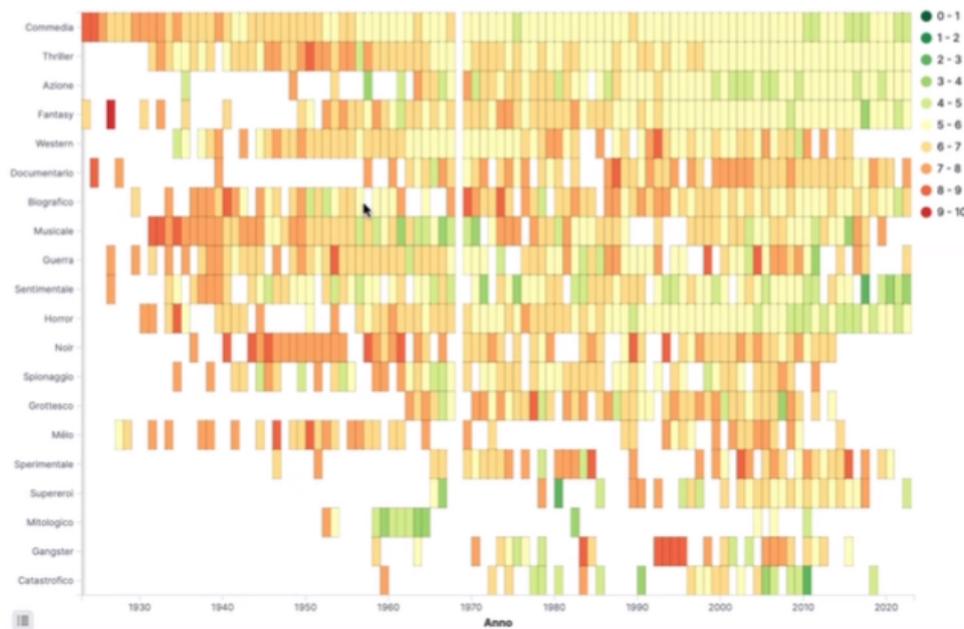
Create new visualization -> Aggregation Based -> Vertical Bar

Visualizzazione Heat Map

Visualizzazione 3

Componente “Heat Map”

Realizzare una heatmap che visualizzi nel corso del tempo il voto medio rispetto al genere di film.



27

Visualizzazione Map

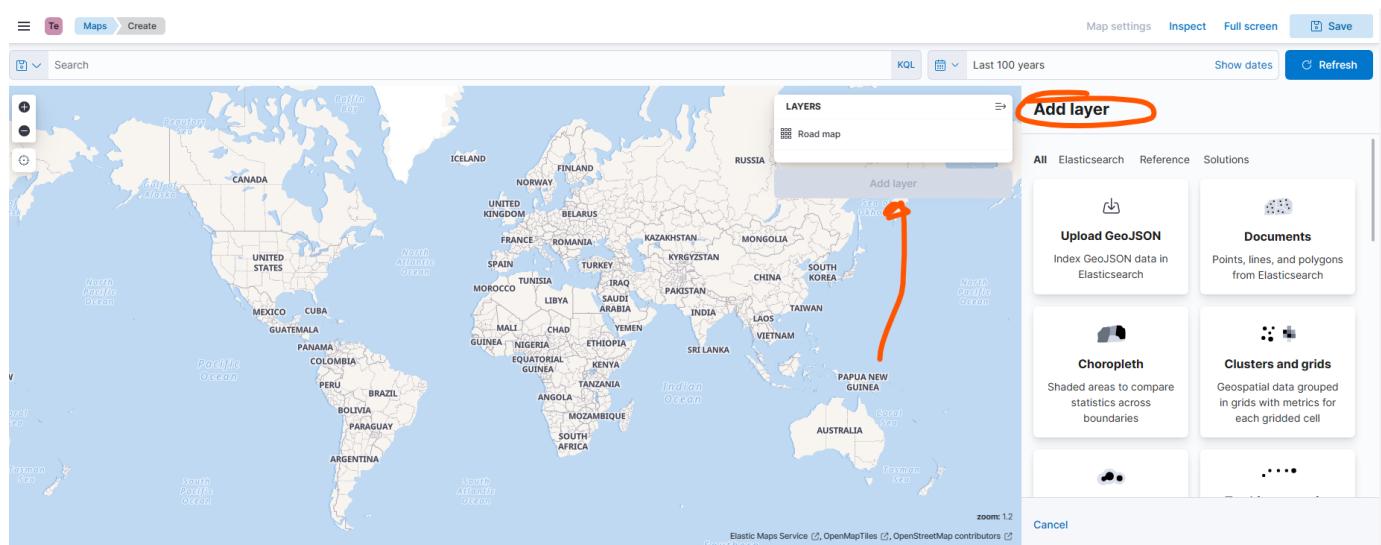
Visualizzazione 4

Componente “Map”

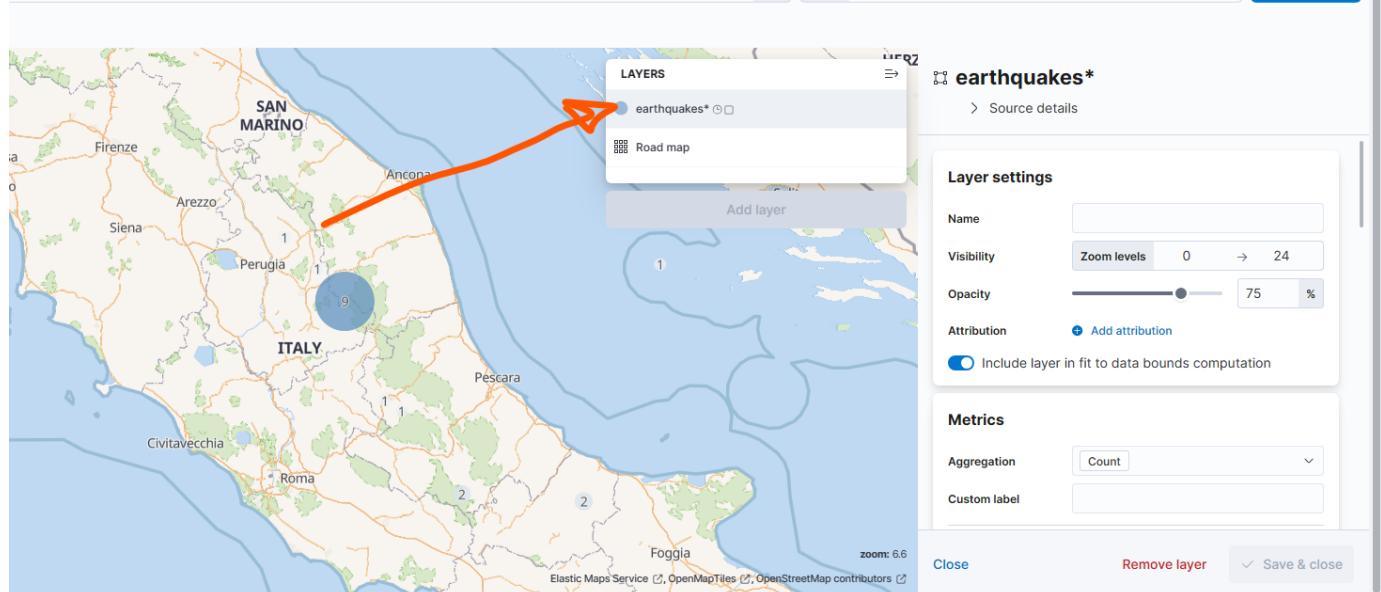
Utilizzando l'indice “earthquakes” creato con il flusso “EarthquakesESIngestor”, visualizzare i terremoti occorsi nel corso degli anni sfruttando i layer per “Clusters and grid”, “Heatmaps” e “Documents”



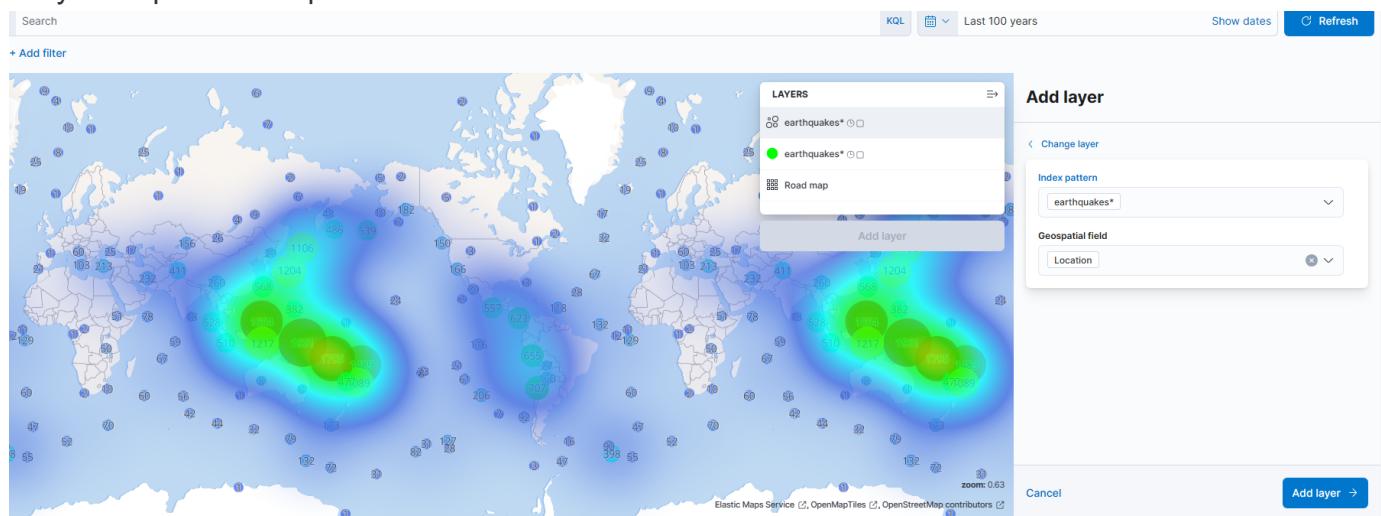
28



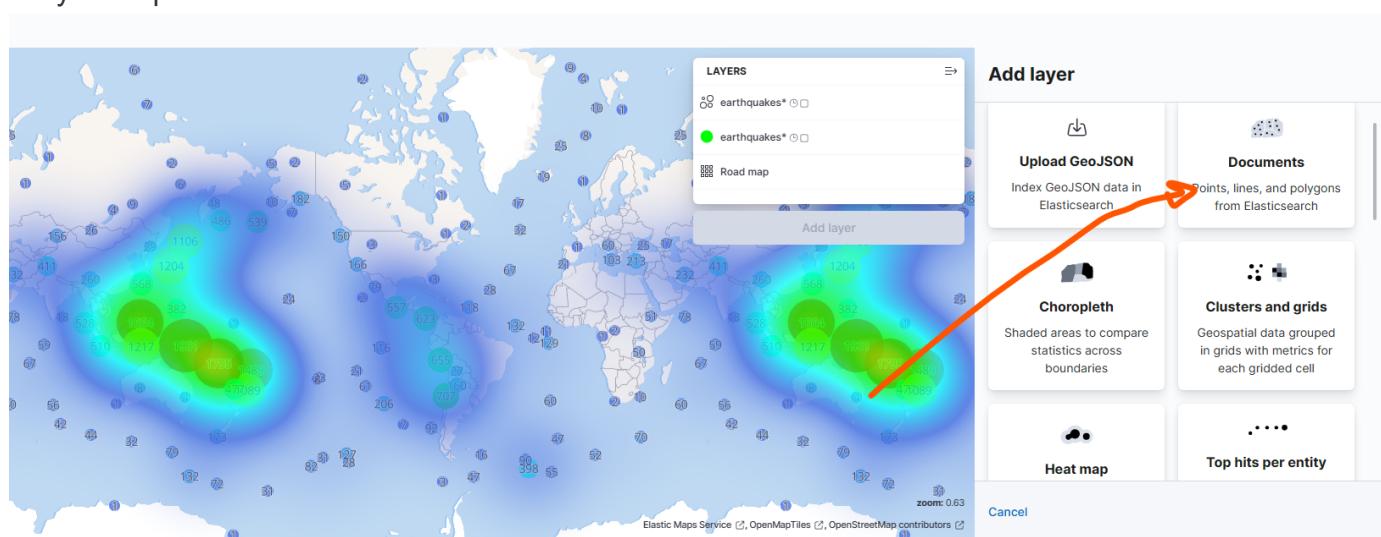
Il layer viene aggiunto in modo definitivo alla mappa:

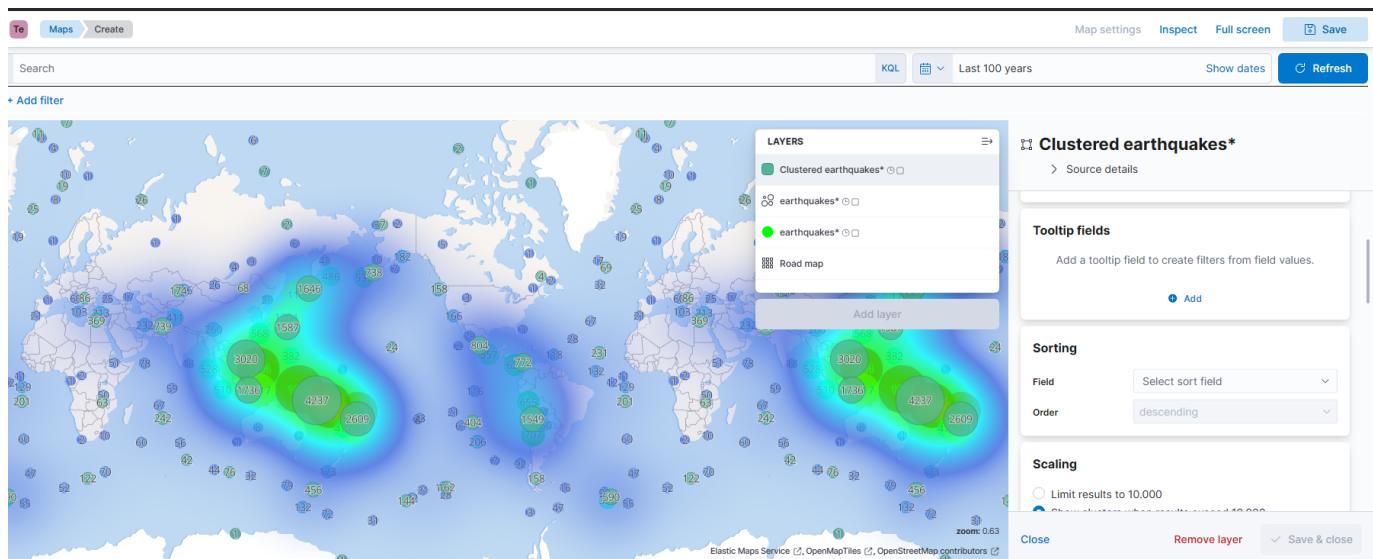


Il layer di tipo Heat Map:



Il layer di tipo Documents:





Visualizzazione Goal

Visualizzazione 5



Componente “Goal”

Film di tipo “drammatico” suddivisi per voto medio



Film drammatici suddivisi per voto medio

29

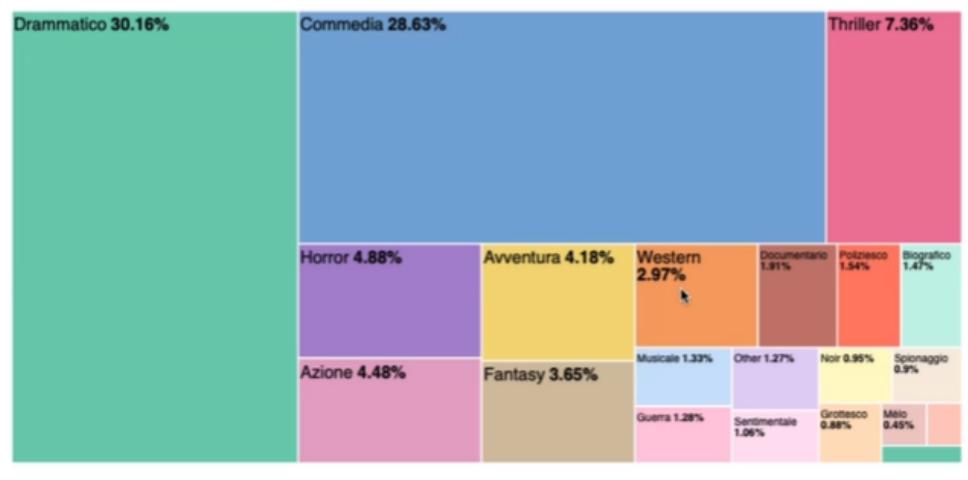
Aggregation -> Goal

Visualizzazione Pie

Visualizzazione 7b

Componente “Pie”

Usare la modalità “Lens” per visualizzare la distribuzione dei film all’interno dei vari generi di pellicole. Usare la visualizzazione Treemap a tale scopo.



32

Create visualization -> Lens