RdS
RICERCA DI SISTEMA

ENEA
AGENZIA NAZIONALE PER LE NUOVE TECNOLOGIE, L'ENERGIA E LO SVILUPPO ECONOMICO SOSTENIBILE

MINISTERO DELL'AMBIENTE E DELLA SICUREZZA ENERGETICA

Ricerca di Sistema elettrico

# Studio e definizione di componenti per l'analisi dei flussi di informazioni relativi alla cybersecurity e predisposizione di sistemi di continuous intelligence/stream analytics (LA3.3)

Angelo Mariano

RdS_PTR 22-24_PR 2.1_LA3.3_069

# STUDIO E DEFINIZIONE DI COMPONENTI PER L'ANALISI DEI FLUSSI DI INFORMAZIONI RELATIVI ALLA CYBERSECURITY E PREDISPOSIZIONE DI SISTEMI DI CONTINUOUS INTELLIGENCE/STREAM ANALYTICS (LA3.3)

Angelo Mariano (ENEA)

Giugno 2023

Report Ricerca di Sistema Elettrico

Accordo di Programma Ministero dell'Ambiente e della Sicurezza Energetica - ENEA
Piano Triennale di Realizzazione 2022-2024

Obiettivo: *Decarbonizzazione/Digitalizzazione ed evoluzione delle reti*
Progetto: *Tema di ricerca 2.1, Progetto integrato cyber security dei sistemi energetici*
Linea di attività: *LA 3.3*
Responsabile del Progetto: Maria Valenti, ENEA
Responsabile Linea di Attività: Angelo Mariano, ENEA
Mese inizio previsto: M1
Mese inizio effettivo: M1
Mese fine previsto: M18
Mese fine effettivo: M18

# Indice

# 1 Risultati attesi

Lista dei risultati attesi come da capitolato vigente

Si riporta di seguito la lista dei risultati attesi come da capitolato vigente:

- Dettagli relativi alla fase di definizione della piattaforma di stream analytics.
- Progetto della piattaforma di continuous intelligence/stream analytics.

# 2 Risultati ottenuti

Lista dei risultati ottenuti *(Evidenziare in che misura il risultato è stato ottenuto ed il beneficio per il sistema elettrico nazionale e i suoi utenti. Aggiungere eventuali risultati ottenuti non previsti nel capitolato)*
Di seguito è riportato l'elenco dei risultati ottenuti:

- **Dettagli relativi alla fase di definizione della piattaforma di stream analytics**
  Definire una piattaforma di stream analytics per la cybersecurity richiede una pianificazione attenta, che tenga in dovuto conto i seguenti elementi:
  - Analisi dei requisiti per definire chiari obiettivi e requisisti di sicurezza
  - Fonti di dati: identificazione e valutazioni delle sorgenti di dati, come log dei firewall o traffico generato da sensori
  - Architettura: scelta di infrastruttura dedicata con attenzione alla scalabilità
  - Tecnologie e strumenti: supporto agli strumenti per l'analisi in tempo reale e per l'uso di tecniche di machine learning
  - Integrazione: facile integrazione con i sistemi esistenti e flessibilità della soluzione individuata
  - Privacy: attenzione alla conformità alle normative sulla privacy
  - Aggiornamenti continui: la piattaforma deve essere pensata per agire in un'ottica MLOps con un aggiornamento continuo dei modelli di machine learning

  La fase di definizione della piattaforma di stream analytics per la cybersecurity è fondamentale per garantire un sistema robusto e reattivo contro le minacce informatiche in evoluzione, che impattano sul sistema elettrico ed i suoi componenti, integrati nella rete dati.

- **Progetto della piattaforma di continuous intelligence/stream analytics**
  Nell'era digitale, la nostra piattaforma di continuous intelligence/stream analytics mira a fornire un'analisi in tempo reale dei dati, rivelando rischi e anomalie. Scalabile e sicura, utilizza tecnologie avanzate per migliorare la risposta alle minacce e fornire insight cruciali per decisioni ottimali. Il progetto mira a disegnare una risorsa indispensabile per organizzazioni che cercano innovazione nell'ambito della gestione delle informazioni in tempo reale, legate alla cybersecurity e soprattutto in connessione al sistema elettrico che trasmette dati nella rete. Il progetto è meglio dettagliato nei paragrafi successivi.

# 3 Prodotti attesi

Non ci sono prodotti hardware/software attesi per la LA.

# 4 Prodotti sviluppati

Lista dei prodotti hardware/software eventualmente sviluppati nella LA, illustrando, per il software, le modalità di accesso per gli utenti *(Aggiungere eventuali prodotti sviluppati non previsti nel capitolato)*

Non ci sono prodotti hardware/software sviluppati nella LA.

# 5 Analisi degli scostamenti su attività e risultati

(8000 caratteri max)
Descrivere le motivazioni di eventuali scostamenti tecnici/economici rispetto al preventivo e criticità riscontrate (*Evidenziare il contenuto in riferimento al piano di rischi presentato*)

Le attività non hanno registrato scostamenti o criticità tecnici. In termini di rendiconto economico, si registra una minima differenza di costo dovuta a:

- differenza nel costo di personale tra preventivo e rendiconto per effetto del passaggio di livello contrattuale che ha comportato, per i dipendenti ENEA interessati, il passaggio alla fascia di costo superiore a decorrere dal 1° gennaio 2023.
- necessità di coinvolgimento di personale tecnico e di un minore numero di ore di personale di profilo MEDIO e ALTO. Lo sviluppo della piattaforma di continuous intelligence e stream analytics ha richiesto verifiche supplementari, rispetto a quanto pianificato, sulla piena compatibilità dei software di programmazione e di comunicazione, che verranno sviluppati nel progetto, con gli strati software già presenti nell'infrastruttura ENEA con cui vengono gestiti i server di calcolo, i firewall e il database centrale. Pertanto, per assicurare la massima efficienza e affidabilità della piattaforma, è stato necessario il coinvolgimento di personale tecnico per eseguire le suddette verifiche.

# 6 Sintesi delle attività svolte

(1000 caratteri max)

Le attività hanno riguardato una piattaforma di stream analytics e continuous intelligence. L'obiettivo principale è quello di elaborare in tempo reale gli eventi che compromettono la sicurezza delle reti elettriche e adottare le opportune misure di reazione utilizzando algoritmi di intelligenza artificiale. La piattaforma è stata appositamente progettata per gestire grandi quantità di dati provenienti dalle reti, consentendo un monitoraggio costante e una risposta tempestiva alle eventuali anomalie o minacce alla sicurezza. La soluzione è stata disegnata in maniera distribuita, garantendo una scalabilità ottimale del sistema, per far fronte anche a situazioni di elevato carico di lavoro. In conclusione, la piattaforma disegnata rappresenta un importante tassello per la gestione della cybersecurity, grazie all'utilizzo di algoritmi di intelligenza artificiale e alla capacità di elaborare i dati in tempo reale, identificando prontamente minacce e avviando azioni immediate per mitigare i rischi soprattutto con un focus sulle reti elettriche.

# 7 Dettaglio delle attività svolte

(15000 caratteri max)
Descrivere in dettaglio le attività svolte nella LA (*Evidenziare come si sono ottenuti i risultati. Descrivere brevemente anche le attività, per le quali si sono spese delle risorse, che tuttavia non hanno portato all'ottenimento dei risultati previsti al fine di permettere la corretta valutazione di congruità e pertinenza dei costi* rendicontati.)

L'era digitale ha portato numerosi vantaggi e opportunità, ma ha anche dato luogo a nuove minacce e vulnerabilità informatiche. La sicurezza informatica, o cybersecurity, è diventata una priorità ineludibile per individui, aziende e istituzioni, in particolare per le reti energetiche. La crescente complessità delle minacce richiede approcci sempre più sofisticati per la protezione delle informazioni e dei sistemi. In questo contesto, l'analisi dei flussi di informazioni gioca un ruolo cruciale nell'identificazione tempestiva di potenziali minacce e vulnerabilità. Obiettivo di questo rapporto tecnico è illustrare i risultati attesi e le attività collegate alla definizione dei componenti per l'analisi dei flussi di informazioni relativi alla cybersecurity ed il progetto di implementazione di sistemi di continuous intelligence e stream analytics per affrontare queste sfide in modo efficace.

L'analisi dei flussi di informazioni è una componente essenziale della strategia di cybersecurity. Essa consiste nell'osservare, raccogliere e analizzare dati provenienti da varie fonti, come log di sistemi, traffico di rete, attività degli utenti e dati dei dispositivi. L'obiettivo è individuare schemi anomali o comportamenti sospetti che potrebbero indicare un attacco informatico in corso o una violazione della sicurezza. Questo approccio proattivo consente di prendere misure preventive prima che il danno si verifichi. I componenti chiave che fanno parte dell'analisi dei flussi di informazioni sono i seguenti:

1. raccolta dei dati: è essenziale raccogliere dati da fonti variegate, come server, dispositivi endpoint e sensori di rete. Questi dati possono includere registri di accesso, metriche di sistema, informazioni sul traffico di rete e altro ancora;
2. normalizzazione dei dati: i dati raccolti spesso provengono da fonti eterogenee. La normalizzazione dei dati è cruciale per renderli comparabili e correlabili. La correlazione consente di individuare relazioni tra eventi apparentemente separati, rivelando possibili attacchi complessi;
3. analisi dei comportamenti anomali: utilizzando algoritmi avanzati di machine learning e intelligenza artificiale, è possibile identificare pattern di comportamento anomali che potrebbero suggerire attività malevole. Questa analisi richiede un modello di riferimento dei comportamenti normali;
4. risposta automatica e/o manuale: in base ai risultati dell'analisi, possono essere attivate risposte automatiche o richieste interventi manuali. Le risposte possono includere il blocco di indirizzi IP sospetti, l'isolamento di dispositivi compromessi o l'avvio di procedure di risposta agli incidenti.

La continuous intelligence è un approccio che integra l'analisi in tempo reale dei dati con le decisioni automatizzate. Questo consente di rilevare le minacce in tempo reale e prendere decisioni rapide per contrastarle. Lo stream analytics, d'altra parte, si concentra sull'analisi dei flussi di dati in movimento, consentendo di trarre informazioni significative da grandi volumi di informazioni in tempo reale. Fra i vantaggi della continuous intelligence e delle piattaforme di stream analytics possiamo elencare:

1. rilevazione tempestiva delle minacce: la velocità è un fattore chiave nella cybersecurity. I sistemi di continuous intelligence e stream analytics possono individuare minacce e violazioni in pochi secondi o addirittura millisecondi, riducendo il tempo di reazione;
2. adattamento dinamico: questi sistemi di intelligenza artificiale possono apprendere dai dati in tempo reale e adattare i modelli di rilevamento alle nuove minacce e alle mutevoli tattiche degli attacchi;
3. riduzione dei falsi positivi: l'analisi in tempo reale consente di affinare i modelli di rilevamento, riducendo il numero di falsi allarmi e focalizzandosi su minacce effettive, magari puntando su approcci di recente sviluppo come il RLHF (reinforcement learning from human feedback)

L'architettura progettata ad alto livello si basa prevalentemente su Apache Kafka, il quale costituisce la spina dorsale dell'intero sistema, grazie ai suoi diversi componenti. L'obiettivo è quello di acquisire e processare in tempo reale uno stream di dati prodotti da sensori e/o log di sistemi informatici, con l'intento di riconoscere automaticamente tentativi di intrusione malevola nei sistemi monitorati.

Per tale obiettivo è necessario definire un'architettura di tipo stream processing, ovvero in grado di elaborare un flusso continuo di dati, man mano che viene prodotto. Il flusso di dati è rappresentato da metriche e/o misurazioni acquisite da sensoristica IoT di varia natura, oppure dai log prodotti da sistemi informatici monitorati. Possiamo dunque indicare sostanzialmente due data type:

1. Time series per i dati IoT
2. Key-value per i log

Al fine di raccogliere e trasmettere in tempo reale in modo efficiente questi dati, è necessario implementare un pattern di tipo pub/sub che consente di gestire, per ogni produttore di dati in modo isolato, alcuni aspetti critici per l'affidabilità complessiva del sistema, quali ad esempio la consistenza, l'asincronicità di un flusso rispetto all'altro, l'unidirezionalità del flusso di informazioni. Grazie all'utilizzo di tali protocolli, per ogni sensore e log si possono definire *topic* differenti ai quali i *producer* di dati invieranno il dato pacchettizzato inserendolo in coda, e differenti *consumer* potranno, sottoscrivendo il topic di interesse, prelevare il dato e passarlo alla fase successiva. Per questo tipo di architetture e pattern di comunicazione, la soluzione open source più flessibile, potente e anche diffusa è sicuramente rappresentata da Apache Kafka. Grazie ai diversi componenti software che può vantare ed alla estrema possibilità di configurazione, offre la soluzione ideale per modellare l'architettura che meglio risponde alle esigenze di progetto.

Possiamo definire una prima bozza del flusso di analisi ed elaborazione (**Errore. L'origine riferimento non è s tata trovata.**), utilizzando alcuni componenti di Apache Kafka. Seguendo dunque il flusso dei dati da quando vengono prodotti, questi possono essere raccolti ed inviati su code/topic differenti mediante il componente Kafka Connect. Utilizzando Kafka Streams come fosse un filtro, si possono prelevare questi dati e produrne una prima elaborazione, scartandone la quota che risulta essere definibile matematicamente/statisticamente come "normale", ed inviando il rimanente alle fasi di elaborazioni successive. Tutta questa prima catena possiamo definirla la componente di Data Injection, poiché si occupa proprio di raccogliere i dati e prepararli per essere elaborati.

Lo step successivo rappresenta il cuore del sistema, in cui si andrà a definire la vera e propria logica intelligente per l'individuazione delle possibili frodi. Quello che nella **Errore. L'origine riferimento non è stata t rovata.** è definito Intelligent Awarness System Engine, costituito dall'unione di ksqlDB, che consente di operare interrogazioni con linguaggio SQL sul flusso di dati e Tensorflow/PyTorch (o altro framework di ML)

per mezzo del quale definire l'intelligenza vera e propria, con metodi, modelli e reti di apprendimento profondo che possono operare in inferenza sul dato letto in tempo reale.

Il risultato relativo agli eventi potenzialmente malevoli viene spedito ad un sistema di alert, a sistemi e strumenti di analisi utili e raccolti in un archivio che li storicizza e che può essere utilizzato anche per il fine tuning dei modelli di reti adottati.

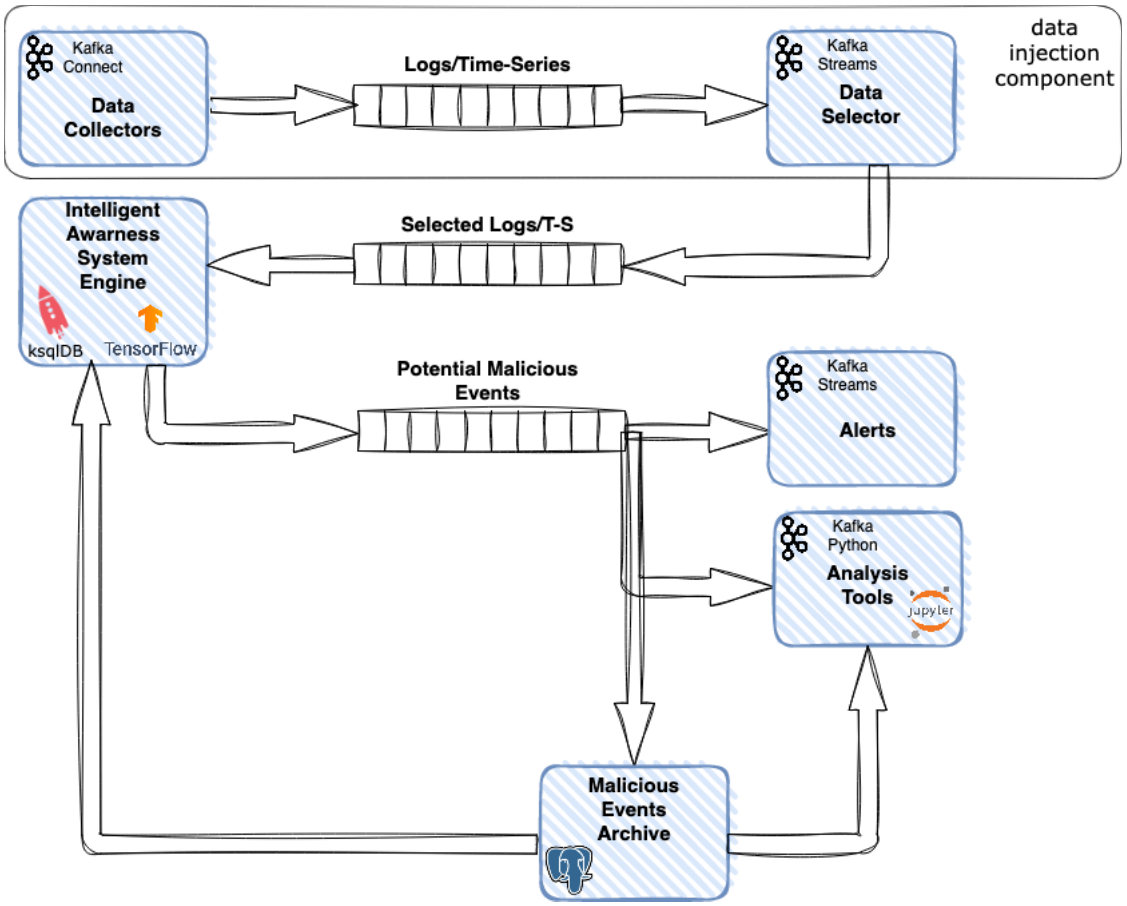


Figura 1 Flusso dei dati

L'adozione della tecnologia Kafka, oltre a consentire la gestione della complessa architettura precedentemente solo abbozzata per alcuni aspetti principali, offre anche garanzie sull'affidabilità ed accessibilità del sistema, potendo prevederne il deploy su macchine multiple per gestire la fault tolerance. Tuttavia, più che implementare una soluzione bare metal, può essere vantaggioso pensare di ospitare tutto su un cluster Kubernetes (di seguito indicato come K8s), demandando ad esso l'affidabilità e l'automazione nella gestione dei vari componenti. Anche gli altri componenti come il database per l'archiviazione dello storico degli eventi valutati dal sistema come malevoli, gli strumenti di analisi per data scientist (come ad esempio JupyterHub), la dashboard e/o sistemi di alert (mail server, telegram bot, …) possono essere gestiti anch'essi dalla stessa infrastruttura, come mostrato schematicamente in Figura 2.

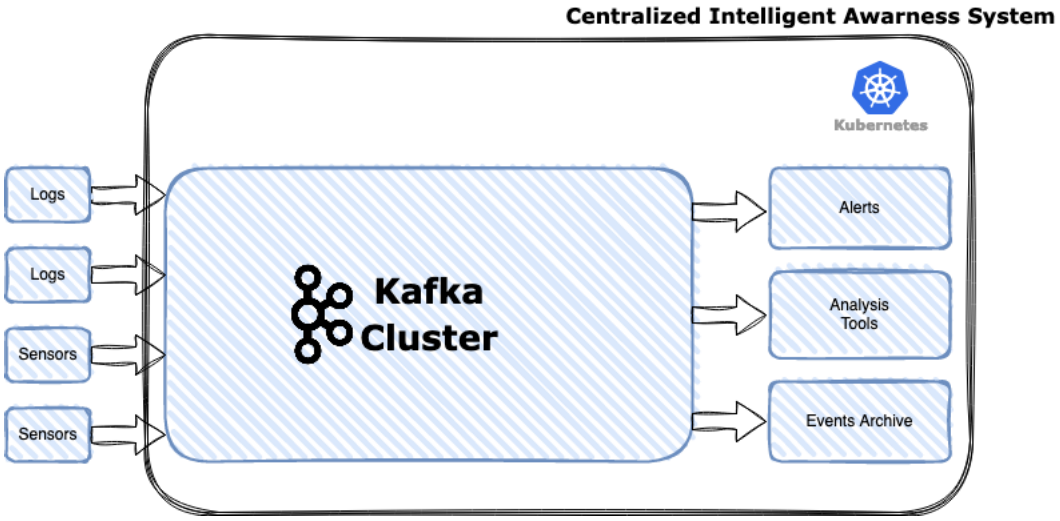


**Figura 2 - Schema ad alto livello dell'architettura dell'Intelligent Awarness System**

La finalità del progetto è quella di raccogliere le tipologie di dati individuate su più siti, fisicamente distinti e geograficamente dislocati in regioni italiane diverse. Volendo conservare un approccio centralizzato in fase di analisi, dobbiamo necessariamente pensare ad una soluzione distribuita invece per l'acquisizione dei dati, disaccoppiando i componenti di Injection dal resto dell'architettura descritta.

Illustriamo dunque brevemente come rimodulare il sistema intelligente centrale ed i vari sistemi di acquisizione locali, alla luce di questa osservazione.

Si tratta sostanzialmente di sostituire il data injection component con un *data injection layer*, il quale raccoglie come consumer i dati provenienti dai diversi nodi locali, ognuno dei quali con il proprio pool di topic, dove vengono pubblicati log e time series già preselezionati dai rispettivi nodi di acquisizione locali.

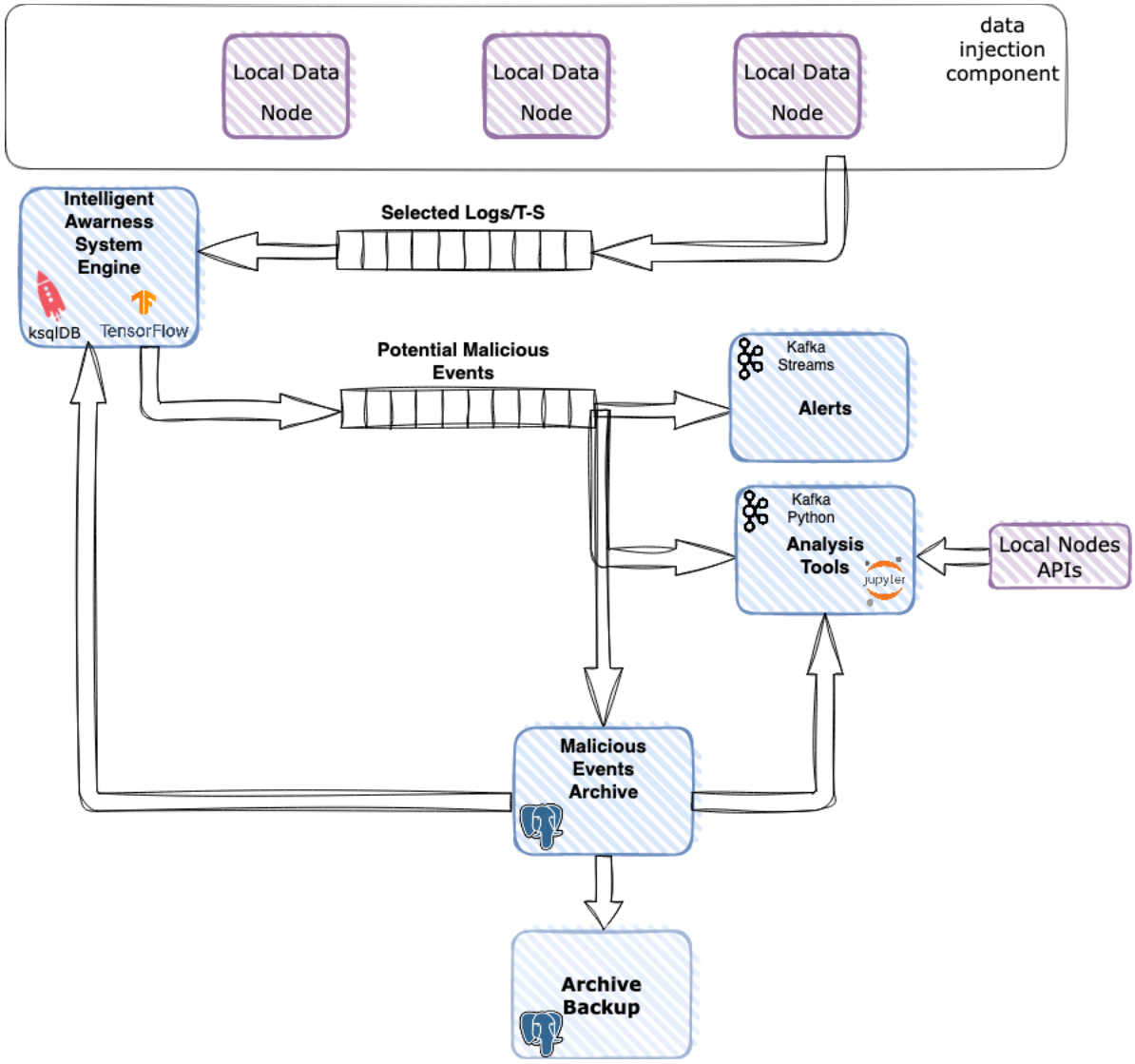


**Figura 3 - Rimodulazione del flusso di dati nel componente centralizzato dell'infrastruttura distribuita**

A questo va aggiunto un sistema di backup per l'archivio degli eventi, esterno al cluster K8s per maggior sicurezza, e la possibilità di contattare delle API ospitate sui nodi locali, con le quali reperire il dato completo prodotto dal sistema locale, per eventuali analisi approfondite. Un ulteriore dettaglio su queste API sarà illustrato nella sezione successiva.

Ogni singolo nodo (Figura 4), istallato localmente per acquisire tutti i dati del relativo sito, dovrà gestire la logica del data injection component descritto in precedenza. Ci sono però alcune modifiche che si possono pensare nell'architettura e nelle tecnologie coinvolte. Dal punto di vista architetturale, può aver senso disaccoppiare il message broker che raccoglie i dati dal sistema che li pre-processa. Questo perché, nell'ottica di far gestire anche il nodo locale da K8s, magari come single node, possiamo far scalare i componenti separatamente, considerando che prima e dopo il pre-processing, il volume di dati subisce un considerevole ridimensionamento. Possiamo inoltre definire un servizio di archiviazione di tutti i dati, con policies di data retention opportunamente dimensionate, ed un servizio che esponga delle API per interrogare tale archivio. Questo consentirebbe, in caso di evento malevolo di particolare interesse, di risalire all'intera serie temporale che lo ha preceduto e seguito, per eventuali analisi direttamente dalla piattaforma che ospita gli strumenti di data analysis (JupyterHub).

Per quanto riguarda le tecnologie coinvolte, mentre per i componenti di pre-processing e pubblicazione sui topic che poi verranno sottoscritti dal sistema centrale, si potrebbe pensare di usare sempre Kafka Streams. Per il message broker ed il db con data retention, si potrebbe usare una soluzione di più facile configurazione (rispetto a Kafka), rappresentata da Redis ad esempio. Si potrebbero usare anche altre soluzioni come, ad esempio, RabbitMQ per il message broker. Redis, tuttavia può essere usato come message broker e come un db memcache, oltre a consentire di salvare anche dati su disco. Dunque, potremmo gestire, con un'unica tecnologia, entrambe i componenti.

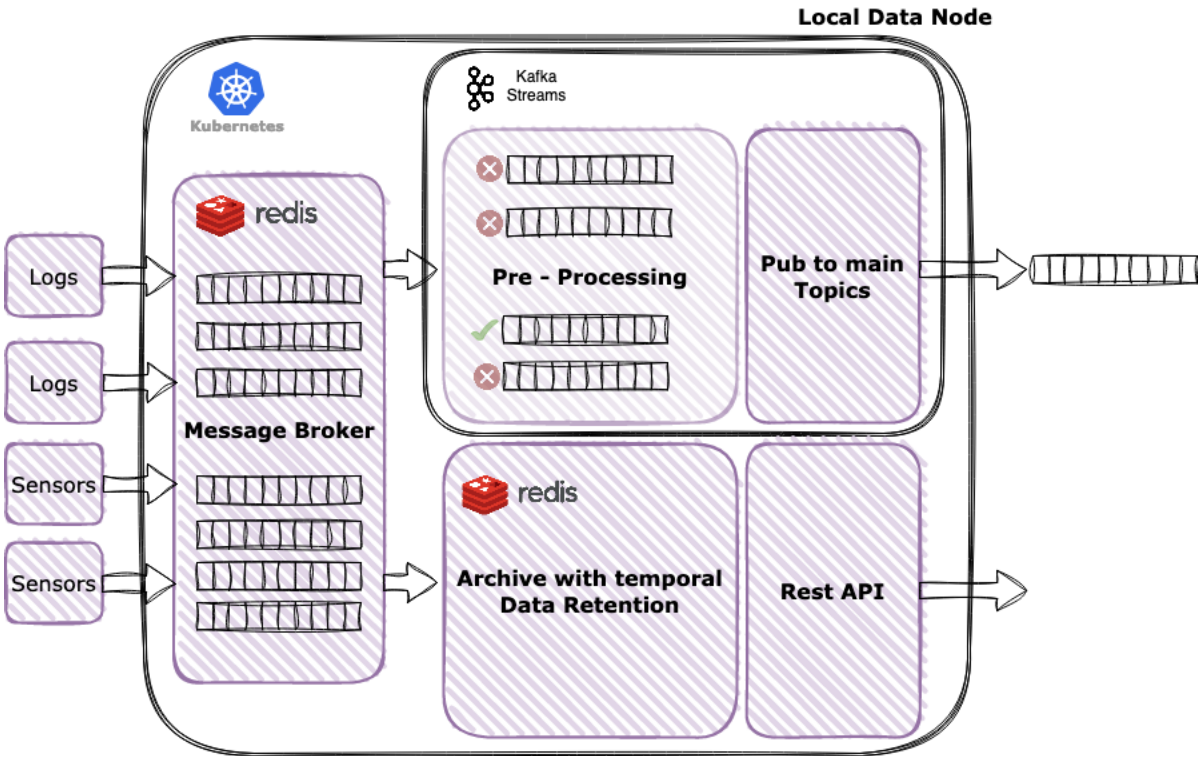


Figura 4 - Schema dell'architettura del singolo nodo locale

A quanto descritto, dovremmo aggiungere un discorso separato per ciascun tipo di produttore di dato. Al momento non vi sono dettagli utili alla definizione di questi componenti. Supponendo ragionevolmente di rimanere nello spettro di soluzioni standard sia per l'IoT che per i logger, possiamo pensare a sistemi che integrino Redis nella loro definizione dei protocolli di comunicazione, per poter pubblicare i dati da essi prodotti, sui topic del nodo locale.

Quanto esposto è stato pensato per poter essere pronto a scalare orizzontalmente (tra le altre cose, senza interruzioni di servizio). Kubernetes si occupa infatti, oltre all'orchestrazione dei componenti software (definiti pods), anche a scalare le prestazioni a seconda della richiesta. Oltre a questo, K8s stesso è una tecnologia distribuita pensata per scalare. Qualora fosse necessario, il cluster può essere espanso aggiungendovi altri nodi (sempre in numero dispari).

Come soluzione di database è stato scelto PostgreSQL, ma si può pensare ad altre soluzioni, anche non necessariamente SQL, tipo mongoDB. Questa soluzione avrebbe il vantaggio di poter essere configurata in modalità Shard all'interno di K8s, offrendo ulteriori vantaggi in termini di scalabilità e ridondanza del dato. In realtà però il dato gestito dall'archivio non ha particolari requisiti nell'accesso o nella scrittura, dato che l'obiettivo dell'archivio è prevalentemente mantenere lo storico dei dati. Soprattutto, per la persistenza del

dato, il componente di riferimento è il backup, esterno al cluster K8s. In quest'ottica, dunque, si può pensare comunque all'utilizzo di mongoDB nella sua versione Community, lasciando che sia K8s a gestirne la ridondanza e gestendo in modo più agevole i backup.

Da notare che l'archivio degli eventi malevoli ha potenzialmente un valore commerciale non indifferente.

In un mondo sempre più interconnesso e digitalizzato, la sicurezza informatica è un imperativo. L'analisi dei flussi di informazioni attraverso sistemi di continuous intelligence e stream analytics offre un'opportunità di rilevare e affrontare le minacce informatiche in modo proattivo. L'evoluzione continua delle minacce richiede una costante innovazione e adattamento dei metodi di analisi e difesa. Attraverso l'integrazione di tecnologie avanzate e strategie di analisi sofisticate, come quelle analizzate e proposte nei precedenti paragrafi, le organizzazioni possono affrontare le sfide della cybersecurity nell'ambito dell'interazione con sistemi elettrici con maggiore efficacia, proteggendo i propri dati e sistemi critici.

# 8 Contributo delle eventuali consulenze alle attività sopra descritte

Non ci sono contributi di eventuali consulenze.

# 9 Pubblicazioni scientifiche

Non ci sono state pubblicazioni scientifiche.

# 10 Eventi di disseminazione

Non ci sono stati eventi di disseminazione.