**Pipeline Dettagliata per il Progetto di Rilevamento delle Fake News**

**1. Raccolta dei Dati**

**Google Fact Check Tools**:

* **Scopo**: Ottenere dati di fact-checking da una fonte autorevole.
* **Motivazione**: Google Fact Check Tools fornisce risultati di fact-checking aggiornati e affidabili, essenziali per verificare l'autenticità delle notizie.

**Flask App**:

* **Scopo**: Creare un'interfaccia web per raccogliere i titoli delle notizie dagli utenti e inviare i dati a Kafka.
* **Motivazione**: Flask è un micro framework per Python che permette di creare rapidamente applicazioni web. È leggero e facile da configurare, ideale per prototipi e applicazioni di piccole dimensioni.
* **Funzionamento**:
  + L'utente inserisce un titolo di notizia tramite un modulo web.
  + L'app Flask invia il titolo a Kafka per l'elaborazione.
  + L'app Flask interroga anche Elasticsearch per verificare se il titolo è già stato verificato in precedenza.

**2. Ingestione dei Dati**

**Kafka**:

* **Scopo**: Trasferire i dati in tempo reale tra diversi componenti della pipeline.
* **Motivazione**: Kafka è una piattaforma di streaming distribuita che garantisce alta disponibilità e scalabilità. È ideale per gestire flussi di dati in tempo reale.
* **Funzionamento**:
  + Kafka riceve i titoli delle notizie dall'app Flask e li mette in un topic chiamato fact-check-request.
  + I consumatori di Spark leggono i dati da questo topic per l'elaborazione.

**3. Elaborazione dei Dati**

**Spark App per Topic Modeling**:

* **Scopo**: Leggere i dati dal topic Kafka, eseguire il topic modeling utilizzando Spark NLP e LDA, e scrivere i risultati su Elasticsearch.
* **Motivazione**: Spark è un motore di elaborazione dati distribuito che permette di eseguire analisi complesse su grandi volumi di dati in modo efficiente. L'uso di Spark NLP e LDA consente di identificare i topic nei contenuti delle notizie, fornendo una comprensione più approfondita dei dati.
* **Funzionamento**:
  + Il consumatore Spark legge i titoli delle notizie dal topic Kafka.
  + Utilizza Spark NLP per tokenizzare, normalizzare e lemmatizzare i testi.
  + Applica LDA (Latent Dirichlet Allocation) per identificare i topic nei contenuti delle notizie.
  + Scrive i risultati del topic modeling su Elasticsearch.

**Spark App per Fake News Detection**:

* **Scopo**: Leggere i dati dal topic Kafka, interagire con l'API di Google Fact Check Tools per verificare se un titolo è una fake news, e scrivere i risultati su Elasticsearch.
* **Motivazione**: L'uso di Spark per la rilevazione delle fake news permette di elaborare i dati in tempo reale e su larga scala. L'integrazione con l'API di Google Fact Check Tools fornisce una verifica autorevole e aggiornata delle notizie.
* **Funzionamento**:
  + Il consumatore Spark legge i titoli delle notizie dal topic Kafka.
  + Interroga l'API di Google Fact Check Tools per verificare se il titolo è una fake news.
  + Scrive i risultati della verifica su Elasticsearch.

**4. Memorizzazione dei Dati**

**Elasticsearch**:

* **Scopo**: Memorizzare e cercare i dati elaborati.
* **Motivazione**: Elasticsearch è un motore di ricerca e analisi che permette di eseguire ricerche veloci e analisi avanzate sui dati. È altamente scalabile e ottimizzato per gestire grandi volumi di dati.
* **Funzionamento**:
  + I risultati del topic modeling e della rilevazione delle fake news vengono memorizzati in Elasticsearch.
  + Elasticsearch permette di eseguire ricerche veloci e analisi avanzate sui dati memorizzati.

**5. Visualizzazione dei Dati**

**Kibana**:

* **Scopo**: Creare dashboard e visualizzare i dati memorizzati in Elasticsearch.
* **Motivazione**: Kibana è uno strumento di visualizzazione che si integra perfettamente con Elasticsearch. Permette di creare visualizzazioni interattive e dashboard per monitorare e analizzare i dati.
* **Funzionamento**:
  + Kibana legge i dati da Elasticsearch e permette di creare visualizzazioni interattive e dashboard.
  + Gli utenti possono utilizzare Kibana per monitorare e analizzare i dati delle fake news.

**6. Interazione con l'Utente**

**Flask App**:

* **Scopo**: Mostrare una pagina web per l'inserimento dei titoli delle notizie, recuperare i risultati da Elasticsearch e Google Fact Check Tools, e visualizzarli all'utente.
* **Motivazione**: Fornire un'interfaccia utente semplice e intuitiva per interagire con il sistema di rilevamento delle fake news.
* **Funzionamento**:
  + L'utente inserisce un titolo di notizia tramite un modulo web.
  + L'app Flask interroga Elasticsearch per verificare se il titolo è già stato verificato.
  + Se il titolo non è presente in Elasticsearch, l'app Flask interroga l'API di Google Fact Check Tools.

1. **Google Fact Check Tools** → **Flask** → **Kafka**
2. **Kafka** → **Spark (Topic Modeling)** → **Elasticsearch**
3. **Kafka** → **Spark (Fake News Detection)** → **Elasticsearch**
4. **Elasticsearch** → **Kibana**
5. **Flask** ↔ **Elasticsearch** e **Google Fact Check Tools**

**Perché Logstash Legge i Dati da Kafka per poi Inoltrarli a Elasticsearch in Questo Progetto?**

1. **Trasformazione dei Dati**: Rimuove campi non necessari dai dati.
2. **Integrazione con Altre Fonti**: Può unificare dati da altre fonti future.
3. **Gestione degli Errori**: Gestisce errori e retry per garantire che i dati non vengano persi.
4. **Monitoraggio e Debugging**: Facilita il monitoraggio inviando i dati alla console.

**Perché Flask Prende i Dati sia dall'API che da Elasticsearch in Questo Progetto?**

1. **Completezza dei Dati**: Garantisce accesso ai dati più completi e aggiornati.
2. **Performance**: Riduce la latenza e il carico sulle API esterne.
3. **Ridondanza e Affidabilità**: Se un servizio non è disponibile, l'altro può essere utilizzato.
4. **Storico dei Dati**: Elasticsearch memorizza i risultati delle verifiche precedenti per analisi storiche.

Queste scelte garantiscono una pipeline robusta, scalabile e in grado di fornire risultati accurati e tempestivi agli utenti.

Il **topic modeling** è una tecnica di elaborazione del linguaggio naturale (NLP) utilizzata per scoprire i temi o "topic" nascosti all'interno di un insieme di documenti. È particolarmente utile per analizzare grandi raccolte di testi non strutturati e identificare i principali argomenti trattati.

Un **bigramma** è una sequenza di due parole consecutive in un testo. È un concetto utilizzato nell'elaborazione del linguaggio naturale (NLP) per analizzare e modellare le relazioni tra parole in un corpus di testo.

**News Topics Index (news-topics-index)**: In questo caso, l'indice contiene informazioni sul titolo, sul contenuto, sui bigrammi generati e sulla distribuzione dei topic tramite LDA. La creazione dei bigrammi e l'applicazione di LDA possono produrre un numero maggiore di documenti, specialmente se ogni elaborazione produce una distribuzione di topic diversa

**Ogni articolo potrebbe produrre più rappresentazioni**: LDA genera una distribuzione dei topic per ogni articolo. Ogni volta che un articolo viene processato, potrebbe essere associato a una combinazione diversa di topic. Questo potrebbe portare a una scrittura più complessa (più dati per articolo). A ogni esecuzione di LDA, l'output (la distribuzione dei topic) potrebbe essere leggermente diverso, a causa della natura statistica del modello

1. **Cloud topic distribution**:
   * **Come è fatto**: È una nuvola di parole dove ogni numero rappresenta un topic e la dimensione del numero indica la frequenza relativa del topic.
   * **A cosa serve**: Serve a visualizzare rapidamente quali topic sono più frequenti nel dataset.
   * **Cosa si evince**: Si evince che il topic con valore 0.016 è il più frequente, seguito da altri valori come 0.103 e 0.1.
2. **Distribuzione dei titoli sui record**:
   * **Come è fatto**: È un istogramma che mostra il conteggio dei record per le prime 5 parole chiave nei titoli.
   * **A cosa serve**: Serve a mostrare la distribuzione dei record per le principali parole chiave nei titoli, aiutando a capire quali argomenti sono più comuni.
   * **Cosa si evince**: Si nota che la parola chiave "Vaccines contain microchips to track people" ha il maggior numero di record (5), mentre altre parole chiave come "Climate change" hanno 2 record ciascuna.
3. **Distribuzione dei topic sui record**:
   * **Come è fatto**: È un istogramma che mostra il conteggio dei record per diversi valori di topicDistribution.
   * **A cosa serve**: Serve a visualizzare la distribuzione dei topic sui record, aiutando a identificare quali topic sono più frequenti.
   * **Cosa si evince**: Si osserva che il topic con valore 0.02 è il più frequente, con 6 record, mentre altri topic come 0.1 hanno 2 record ciascuno.
4. **Distribuzione notizie per etichette**:
   * **Come è fatto**: È un istogramma che mostra il conteggio dei record per le principali etichette di valutazione delle fake news.
   * **A cosa serve**: Serve a mostrare la distribuzione delle notizie per etichette di valutazione delle fake news, aiutando a capire quali tipi di valutazioni sono più comuni.
   * **Cosa si evince**: Si evince che l'etichetta "False" è la più frequente, con 4 record, seguita da altre etichette come "Flawed Paper" e "Misleading" con 2 record ciascuna.
5. **Distribuzione risultati analisi sui record**:
   * **Come è fatto**: È un grafico a torta che mostra la percentuale di ciascuna etichetta di valutazione nel dataset.
   * **A cosa serve**: Serve a visualizzare la distribuzione dei risultati dell'analisi sui record, aiutando a identificare la proporzione di ciascuna etichetta di valutazione.
   * **Cosa si evince**: Si nota che la maggior parte dei record è etichettata come "False" (57.14%), mentre le altre etichette come "Flawed Paper", "Misleading" e "Missing Context" hanno ciascuna il 14.29%.
6. **Heat map topic distribution vs bigrammi**:
   * **Come è fatto**: È una heat map che mostra la distribuzione dei topic rispetto ai bigrammi, con diverse tonalità di colore che rappresentano la frequenza.
   * **A cosa serve**: Serve a mostrare la distribuzione dei topic rispetto ai bigrammi, aiutando a identificare le associazioni tra i topic e i bigrammi.
   * **Cosa si evince**: Si osserva che i valori di topicDistribution 0.103 e 0.1 sono associati a bigrammi con valori elevati (≥ 2.6).