SMART SHAPED

SOFTWARE

www.smartshaped.com

SMART SHAPED
SOFTWARE

SMART SHAPED SOFTWARE

Smart Shaped s.r.l. è un'azienda nata nel **2015** specializzata nella **progettazione, sviluppo e manutenzione di soluzioni software altamente tecnologiche** e nella **digitalizzazione dei processi.** Grazie alla sua forte esperienza offre supporto in outsourcing ai clienti garantendo un apporto di alta qualità ai progetti a prezzi competitivi. Il nostro team di professionisti IT qualificati affiancano costantemente i clienti per tutto il life-cycle del progetto.

La ricerca e lo sviluppo applicati alle nuove nuove tecnologie sono il cuore pulsante della nostra azienda: Smart Shaped ha partnership aperte con l'Università de L'Aquila, l'Università di Chieti-Pescara, il Politecnico di Milano, **Università del Sannio (AI e Forecasting per il Cambiamento Climatico)**, l'Università di Eindhoven e l'Università di Tilburg.

SmartShaped ha ricevuto le prestigiose certificazioni **HappyIndex®AtWork2024** e **WeImpactIndex®2024** di ChooseMyCompany posizionandosi al **secondo posto delle migliori aziende dove lavorare in Italia** (segmento 25-99 collaboratori).

FRAMEWORK

CLIMATE CHANGE AI: SFIDE PRINCIPALI

ETEROGENEITÀ E VOLUME DEI DATI: Dati climatici provenienti da fonti molteplici (sensori IoT, satelliti, archivi storici), con formati e granularità differenti, rendono complesso il processo di integrazione.

QUALITÀ E AFFIDABILITÀ: Raccogliere e validare dati accurati, completi e aggiornati è un compito continuo che richiede controlli rigorosi e metodologie standardizzate.

TRASPARENZA E TRACCIABILITÀ: Comprendere la provenienza e le trasformazioni dei dati è fondamentale per creare fiducia nelle previsioni e supportare decisioni informate.

BARRIERE TECNOLOGICHE ED ECONOMICHE: La necessità di infrastrutture scalabili, costi di avvio o utilizzo contenuti, e semplicità d'uso limita l'adozione su larga scala, soprattutto in contesti con limitate risorse tecniche o economiche.

ADATTABILITÀ ALLE MUTAZIONI DEL CLIMA: Il framework deve poter evolvere costantemente per integrare nuove fonti, modelli e strategie, rispondendo a fenomeni climatici in continua trasformazione.

FRAMEWORK: PRINCIPI DI DESIGN

TRASPARENZA E TRACCIABILITÀ: Ogni passaggio di trasformazione dei dati è documentato, così che gli utenti possano comprendere l'origine e l'evoluzione delle informazioni utilizzate.

QUALITÀ E AFFIDABILITÀ DEI DATI: Il framework integra controlli sistematici per garantire che i dati siano coerenti, completi e accurati, fornendo una base solida per le analisi predittive.

ACCESSIBILITÀ NO-CODE: Un design che riduce o elimina la necessità di programmazione, permettendo a utenti non tecnici di configurare pipeline e modelli AI con (relativa) semplicità.

MODULARITÀ E SCALABILITÀ: Un'architettura flessibile che può crescere con le esigenze dell'utenza, integrando nuove fonti di dati, modelli e componenti senza stravolgere il sistema.

INTEROPERABILITÀ E COLLABORAZIONE: Supporto a formati e API standardizzati per favorire l'integrazione con strumenti esterni, promuovendo uno scambio di conoscenze e metodologie tra diversi attori.

ARCHITETTURA LAMBDA E ADATTAMENTI PER CLIMATE CHANGE AI

LAMBDA ARCHITECTURE CLASSICA:

- **Batch Layer:** Elabora dati storici su larga scala, generando viste consolidate e stabili nel tempo.
- **Speed Layer:** Gestisce i flussi di dati in tempo reale, fornendo insight immediati ma potenzialmente non consolidati.
- **Serving Layer:** Unisce i risultati di batch e speed, offrendo all'utente finale una visione integrata e aggiornata.

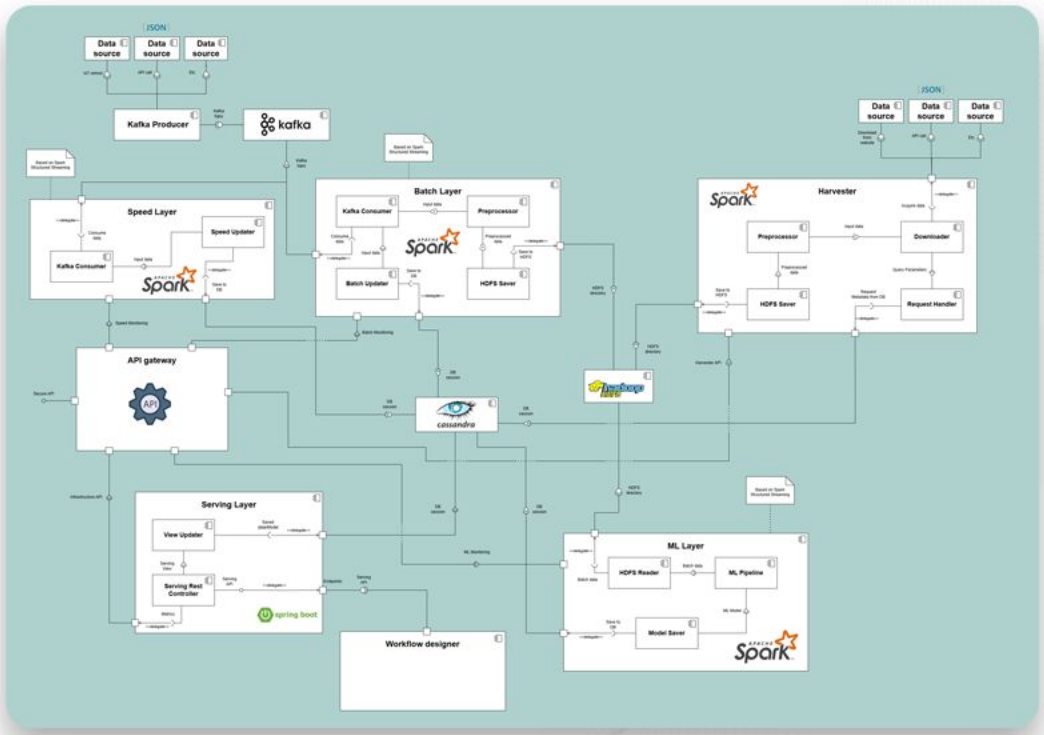
LIMITAZIONI ORIGINALI:

- Complessità nella gestione di formati eterogenei (es. dati geospaziali, immagini, sensori IoT).
- Difficoltà nell'assicurare tracciabilità e qualità dei dati end-to-end.
- Barriera tecnica per utenti non esperti, costretti a intervenire sul codice.

PRINCIPALI ADATTAMENTI PER IL FRAMEWORK:

- Progettazione e integrazione dell'**Harvester** per acquisire, normalizzare e arricchire dati da fonti eterogenee.
- Superamento dell'accoppiamento **Batch + Speed** come unico modo di gestire la data ingestion.
- Introduzione di un **Workflow Designer** No-Code per abbassare le barriere tecniche e permettere a utenti non specializzati di definire pipeline.
- Adozione di tecnologie geospaziali (es. **Spark + Sedona**) per gestire formati specifici del clima e collegare i dati a coordinate geografiche reali.
- Miglioramenti nella Data Governance (controlli di qualità, lineage) per garantire trasparenza e fiducia nei risultati.

ARCHITETTURA FRAMEWORK



BATCH LAYER: Elabora dati storici su larga scala utilizzando framework distribuiti (Spark su HDFS), generando statistiche e aggregazioni di lungo periodo.

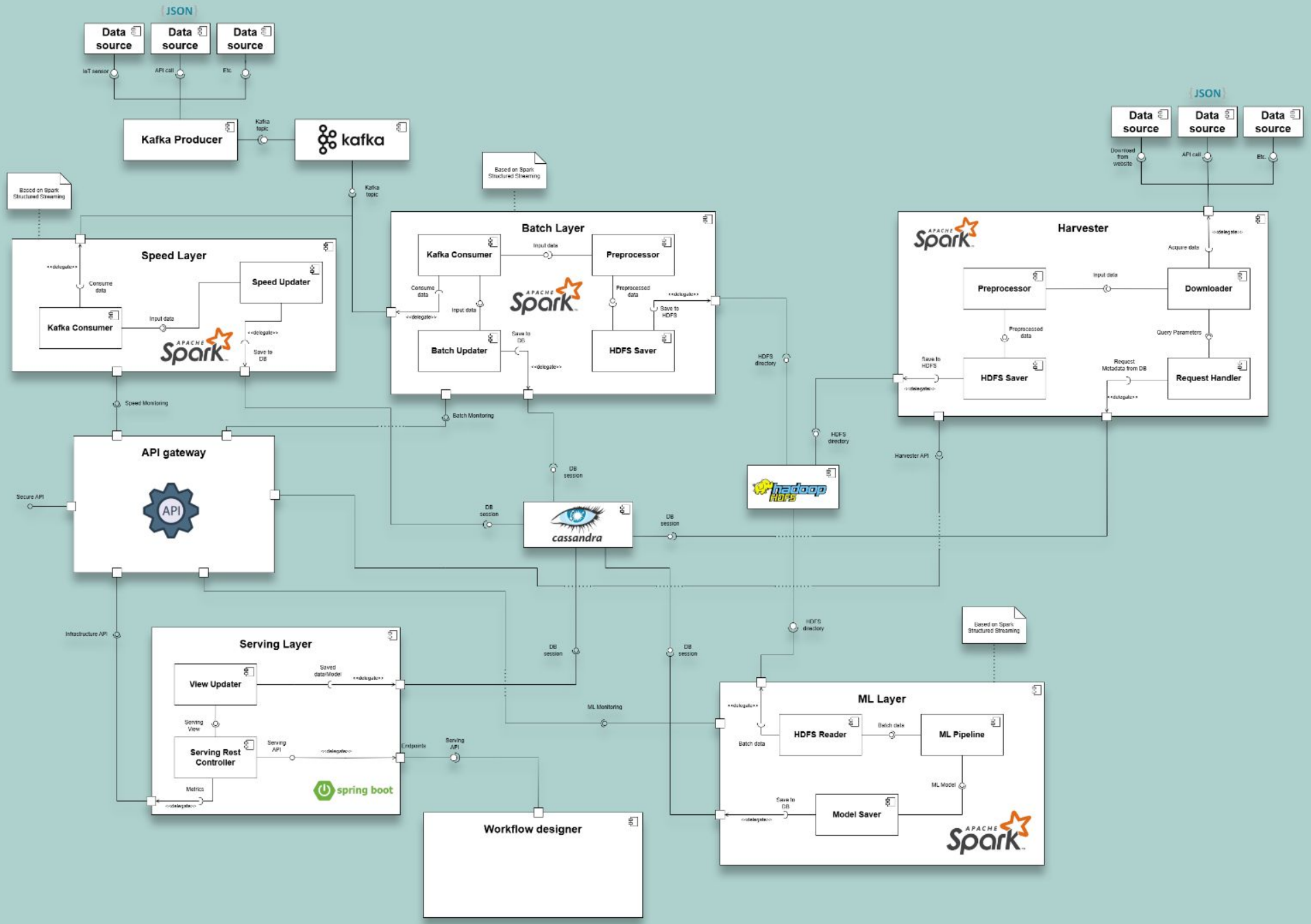
SPEED LAYER: Gestisce flussi di dati in tempo reale per fornire insight immediati e aggiornati, elemento chiave per segnalazioni rapide di eventi climatici estremi.

ML LAYER: Applica modelli di Machine Learning addestrati sui dati storici, archivia i modelli su HDFS e memorizza i risultati delle inferenze in Cassandra per un accesso veloce.

HARVESTER: Raccoglie e pre-elabora dati da fonti esterne eterogenee (NASA, Copernicus), supportando formati geospaziali come GeoJSON e TIFF.

SERVING LAYER: Aggrega e rende disponibili i risultati (statistiche, previsioni) tramite API, facilitando l'integrazione con dashboard, applicazioni, o servizi esterni.

WORKFLOW DESIGNER: Uno strumento grafico che consente di creare, modificare e orchestrare le pipeline di dati e AI senza scrivere codice. Gestisce la comunicazione tra i vari componenti, fornisce metriche sulle prestazioni e consente di avviare i flussi di lavoro generati dal designer.


JSON
Data source
Data source
Data source
IoT sensor
API call
Etc.
Kafka Producer
Kafka topic
kafka
Kafka topic
Based on Spark Structured Streaming
Based on Spark Structured Streaming
Speed Layer
Consume data
Speed Updater
Kafka Consumer
Input data
Save to DB
Batch Layer
Kafka Consumer
Input data
Preprocessor
Consume data
Input data
Preprocessed data
Save to HDFS
Batch Updater
Save to DB
HDFS Saver
JSON
Data source
Data source
Data source
Download from website
API call
Etc.
Harvester
Acquire data
Preprocessor
Input data
Downloader
Preprocessed data
Query Parameters
Save to HDFS
HDFS Saver
Request Metadata from DB
Request Handler
HDFS directory
HDFS directory
Speed Monitoring
Batch Monitoring
API gateway
Secure API
API
Harvester API
DB session
DB session
DB session
cassandra
Infrastructure API
Serving Layer
View Updater
Saved data/Model
Serving View
Serving Rest Controller
Serving API
Metrics
spring boot
Endpoints
Serving API
DB session
DB session
HDFS directory
ML Monitoring
Based on Spark Structured Streaming
ML Layer
HDFS Reader
Batch data
ML Pipeline
ML Model
Batch data
Save to DB
Model Saver
Workflow designer

TECNOLOGIE PRINCIPALI

APACHE SPARK: Gestisce l'elaborazione di grandi volumi di dati in modalità batch e streaming, offrendo capacità di calcolo distribuito essenziali per analisi climatiche su larga scala.

YARN: Coordina l'allocazione delle risorse tra i diversi processi Spark, garantendo un utilizzo equilibrato di CPU e memoria nel cluster.

HDFS (HADOOP DISTRIBUTED FILE SYSTEM): Un file system distribuito affidabile e scalabile, ideale per archiviare dataset climatici di grandi dimensioni in modo da facilitarne l'accesso e l'elaborazione.

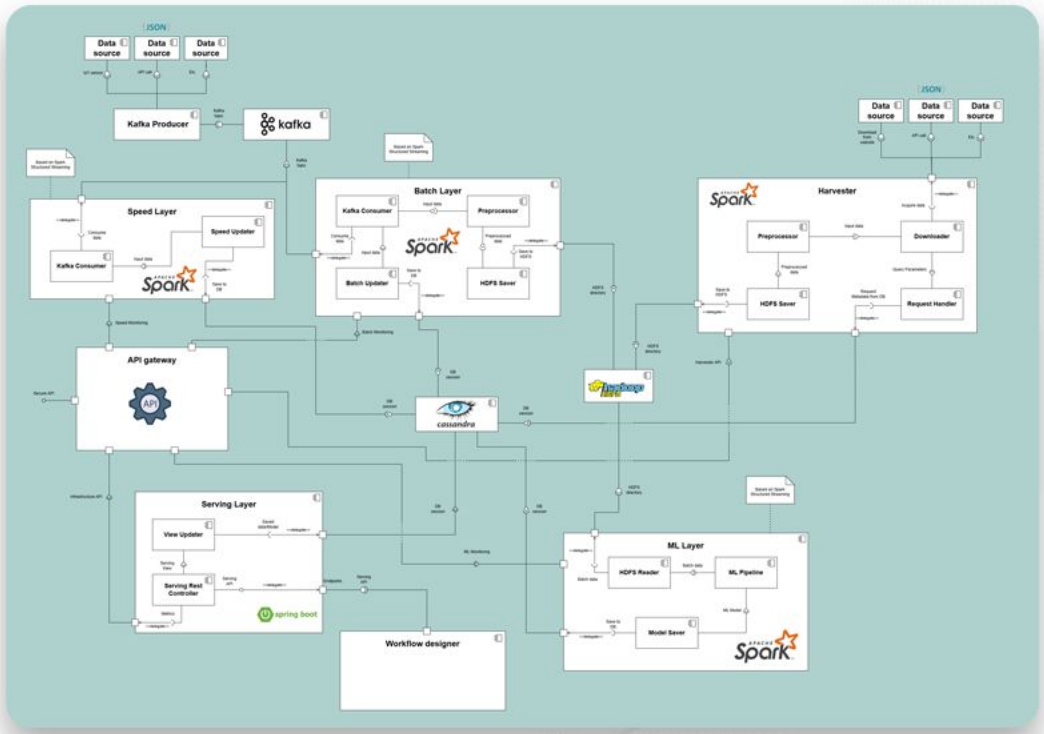
CASSANDRA: Un database NoSQL performante, in grado di gestire rapidamente i risultati delle inferenze e garantire risposte veloci alle query, anche sotto carico elevato.

SEDONA (ESTENSIONE PER SPARK): Abilita l'elaborazione geospaziale avanzata direttamente in Spark, facilitando l'analisi di dati cartografici, come immagini satellitari e coordinate geografiche.

KAFKA (OPZIONALE): Un sistema di messaging ad alta velocità per l'acquisizione, il buffering e la distribuzione di dati in tempo reale, fondamentale quando si lavora con flussi continui (es. sensori IoT).

DOCKER: Favorisce la containerizzazione dei componenti, semplificando il deployment, la portabilità e la manutenzione dell'intero ecosistema.

SVILUPPI FUTURI



SVILUPPO E INTEGRAZIONE LAYER: Sviluppare e integrare i layer mancanti, cioè Serving Layer, API Gateway e Workflow Designer.

INTEGRAZIONE DI MLOPS: Automatizzare gli aggiornamenti, i test e il deployment dei modelli, garantendo un ciclo di vita del Machine Learning più fluido e sostenibile.

SUPPORTO AUTOML: Semplificare ulteriormente la scelta, l'ottimizzazione e la messa a punto degli algoritmi di previsione, riducendo l'intervento manuale nella selezione dei modelli.

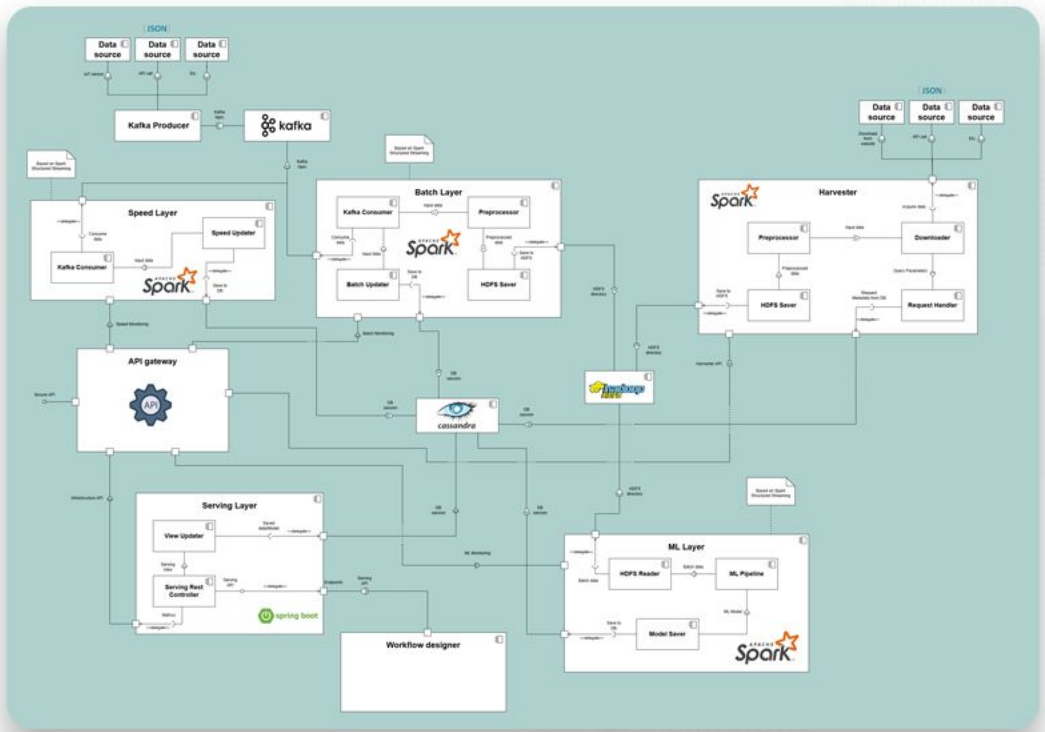
ESPANSIONE DEI FORMATI DATI & API: Integrare nuove sorgenti e tipologie di dati (es. sensori IoT più eterogenei, dati da missioni satellitari specifiche) per coprire scenari climatici ancor più vari e complessi.

SCALABILITÀ OPERATIVA: Validare la piattaforma in contesti di produzione su larga scala, gestendo milioni di dati in tempo reale, testando resilienza, prestazioni e costi operativi.

CASI D'USO SETTORIALI: Applicazione della piattaforma a casi concreti (ad es. catalogazione del rischio da flash floods) per dimostrarne l'effettivo valore sul campo.

DEMO: BATCH LAYER

BATCH LAYER DEMO

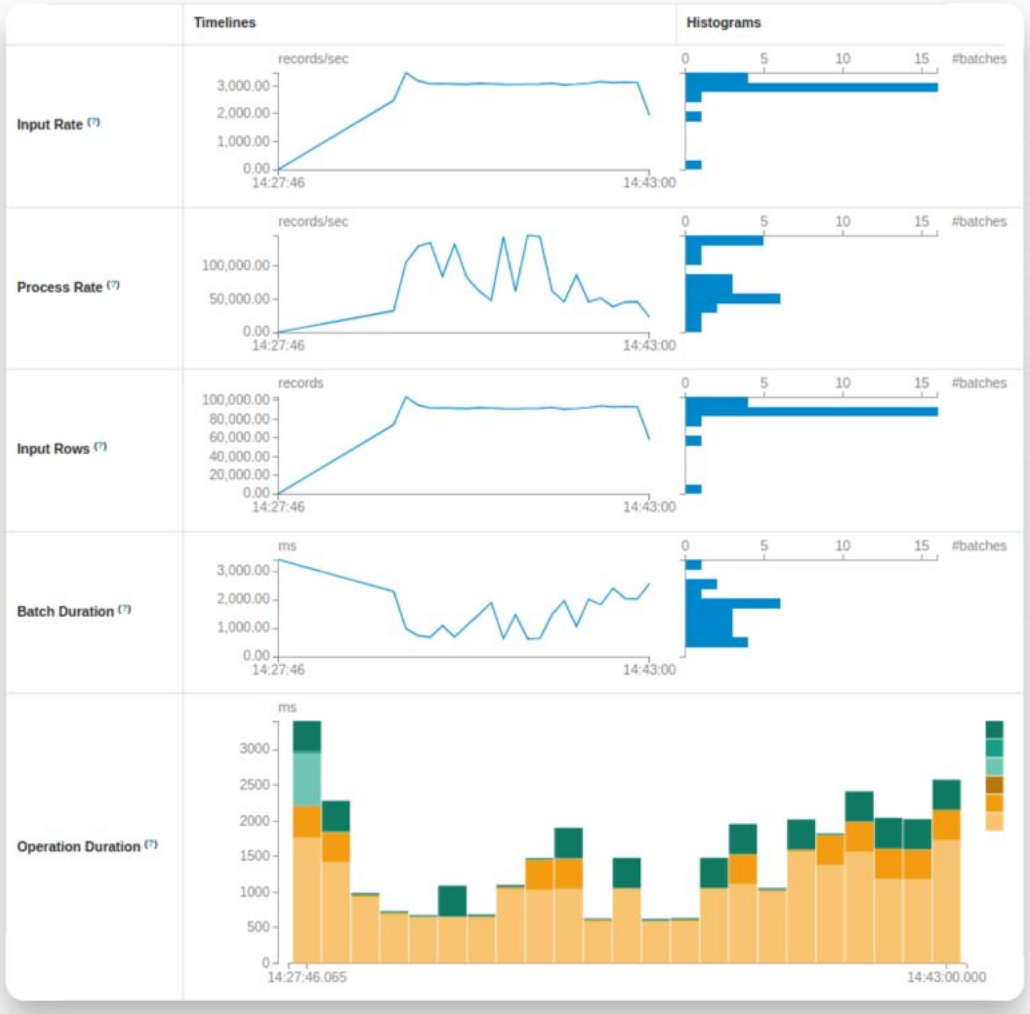


Il dataset di partenza, scaricabile dal codice QR, contiene le recensioni dei prodotti venduti su Amazon, in formato json.

Il primo passo nella costruzione della demo consiste nell'inviare al sistema una porzione di questi dati, corrispondente alle recensioni di un intervallo temporale pari a 5 anni;

Nella pratica è stato definito uno script python che, tramite la libreria "kafka-python-ng", invia le righe del file su un topic kafka a cui la batch-app e la speed-app hanno poi accesso.

BATCH LAYER DEMO



L'applicazione spark del batch layer utilizza le funzionalità di spark structured streaming per leggere i messaggi dal topic kafka e trasformarli in righe di un dataset.

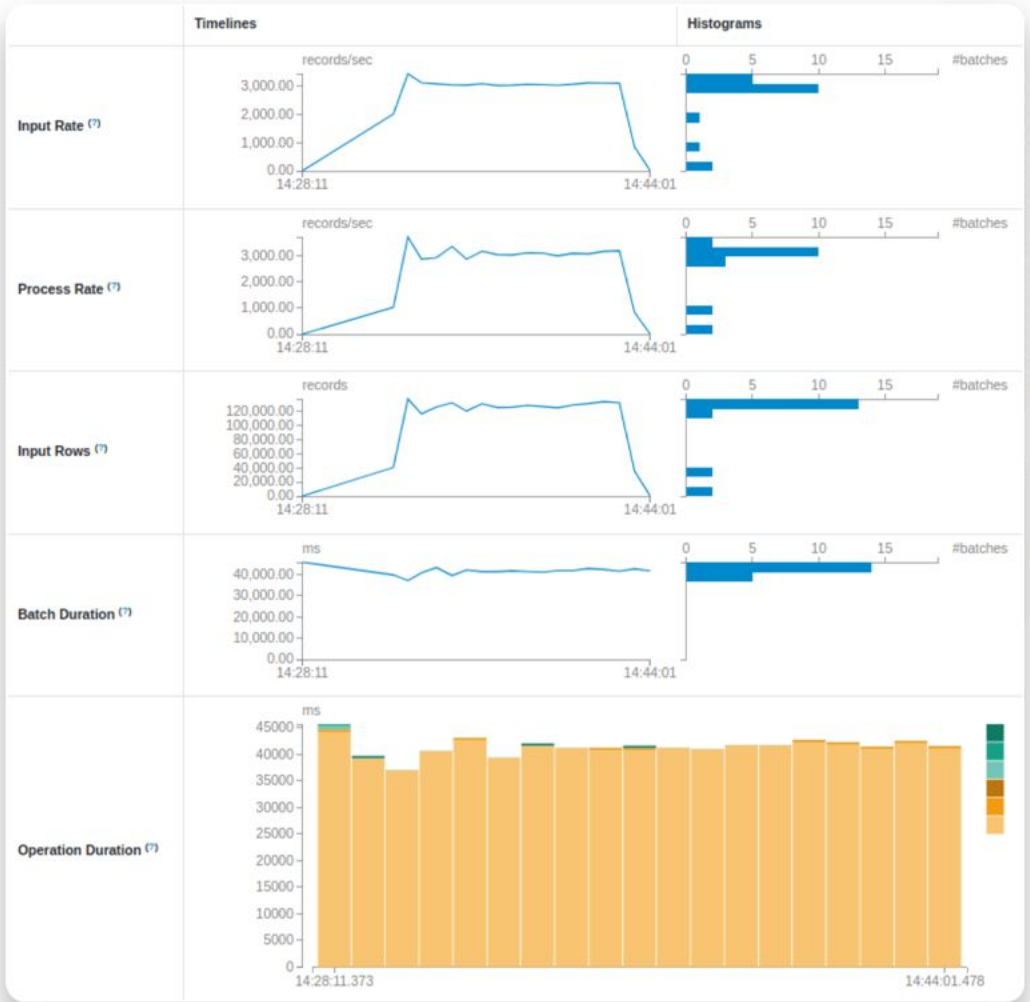
Il dataset ottenuto da questa lettura è caratterizzato da un campo "value" di tipo stringa dal quale vengono estratte le informazione del json grazie ad un apposito Preprocessor.

Questa app ha un duplice obiettivo: da un lato processare i dati ricevuti per scriverli su HDFS, dall'altro usare il BatchUpdater per estrarre, ad ogni fase di batch, i 10 prodotti con il più alto valore medio di recensione e scriverli sul DB Cassandra.

Per eseguire quest'ultimo passaggio vengono considerati ad ogni fase di batch tutti i dati già letti.

DEMO: SPEED LAYER

SPEED LAYER DEMO



Analogamente alla batch-app, anche la speed-app si mette in ascolto sul topic kafka su cui vengono scritti i dati delle recensioni.

La principale funzione di questo layer, che non legge nè scrive su HDFS, risiede nell'uso dello SpeedUpdater per il calcolo di statistiche riassuntive dei dati forniti, da salvare sul DB Cassandra.

Nel nostro caso specifico, la speed-app fa uso di concetti di spark quali il watermark e le window trattare dati in streaming e ricavare analitiche da essi senza disporre del dataset completo, come invece accade nella batch-app.

L'obiettivo finale consiste nel salvare sul DB Cassandra I prodotti con punteggio medio più alto in ciascuna finestra temporale definita.

DEMO: ML LAYER

ML LAYER DEMO

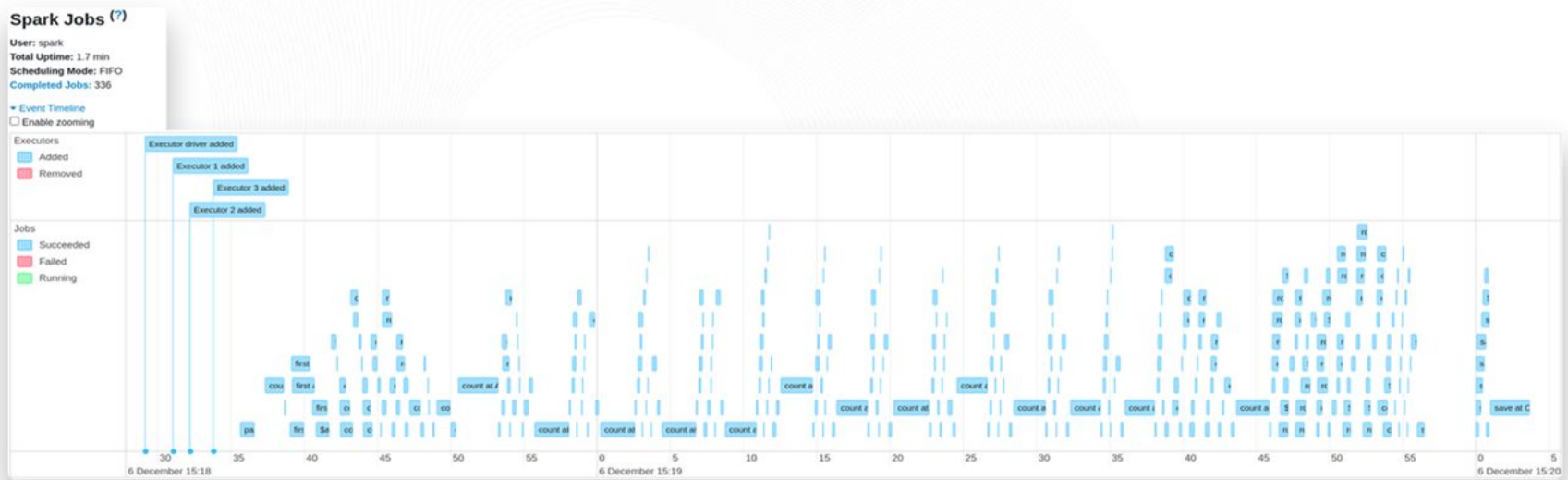
Il **layer di ML** ha come scopo principale l'impiego dei dati finora raccolti tramite il batch layer per l'addestramento di un modello di machine learning ed il suo utilizzo per effettuare predizioni.

I componenti principali sono tre:

- **HdfsReader**, deputato a leggere ed eventualmente elaborare i dati da un singolo path del file system di Hadoop; nel caso della nostra specifica ml-app, trattandosi di un singolo path da cui prelevare i dati, anche l'HdfsReader è uno solo, senza particolari necessità di processing.
- **Pipeline**, si occupa di utilizzare per il machine learning i dataset provenienti dai vari HdfsReader.
- **ModelSaver**, responsabile del salvataggio del risultato finale della predizione sul DB Cassandra e del file contenente il modello su HDFS.

La pipeline di ML è la parte più consistente della demo ed è costituita da una fase di processing del dataset, in cui quest'ultimo viene ripulito dai dati "rumorosi", ovvero le recensioni di utenti occasionali e viene effettuata la divisione in train e test set.
Per quanto riguarda invece le trasformazioni applicate ai singoli campi, le più significative sono state la scalatura dei punteggi nel campo "rating" e la conversione degli id di utenti e prodotti da campi testuali a numerici.

Il modello usato si basa sull'algoritmo ALS del package recommendation di spark ML, a cui si applica una semplice grid-search per il fine-tuning.



DEMO: HARVESTER (PREVIEW)

HARVESTER DEMO (PREVIEW)

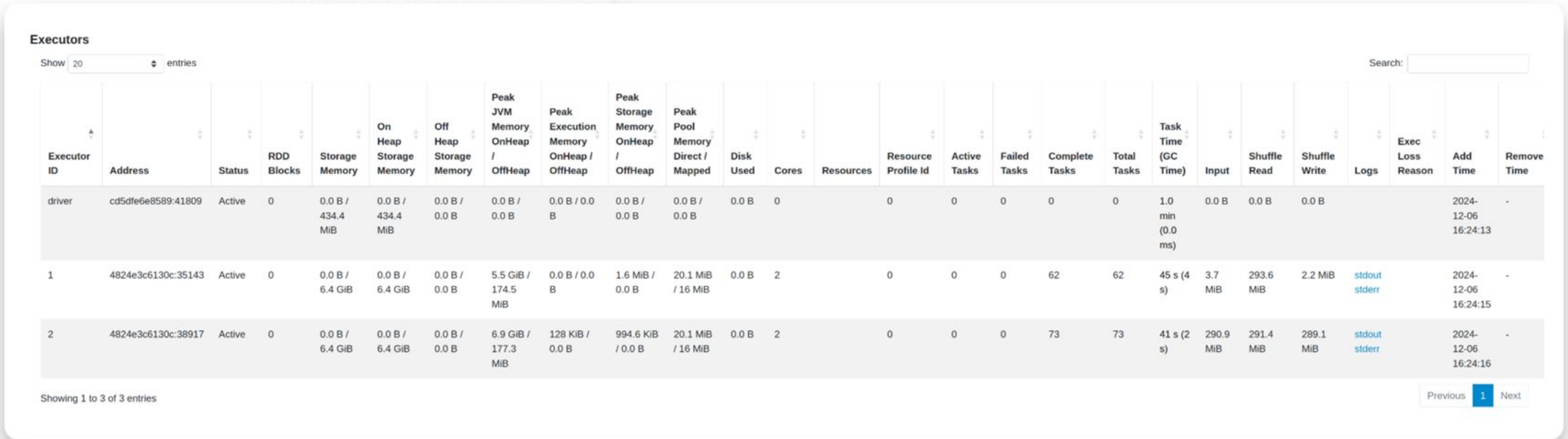
ACQUISIZIONE MULTI-SORGENTE: Raccoglie dati da API esterne (es. json NASA, TIF/TIFF Copernicus), archivi storici, e altre fonti eterogenee.

NORMALIZZAZIONE E PULIZIA: Converte formati diversi in strutture dati omogenee, verifica la qualità e rimuove duplicati o valori mancanti.

ARRICCHIMENTO E RICONCILIAZIONE DEI DATASET: Integra dati geospaziali (es. coordinate, immagini satellitari) e metadati utili per l'analisi climatica.

PROVENIENZA TRACCIATA: Registra l'origine e le trasformazioni effettuate, garantendo trasparenza e facilità di audit.

BASE SOLIDA PER LE FASI SUCCESSIVE: Fornisce dati "puliti" e coerenti al sistema, in particolare all'ML Layer, semplificando e migliorando l'intero flusso di lavoro.



Executors

Show 20 entries

Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	On Heap Storage Memory	Off Heap Storage Memory	Peak JVM Memory OnHeap / OffHeap	Peak Execution Memory OnHeap / OffHeap	Peak Storage Memory OnHeap / OffHeap	Peak Pool Memory Direct / Mapped	Disk Used	Cores	Resources	Resource Profile Id	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs	Exec Loss Reason	Add Time	Remove Time
driver	cd5dfe6e8589:41809	Active	0	0.0 B / 434.4 MiB	0.0 B / 434.4 MiB	0.0 B / 0.0 B	0.0 B / 0.0 B	0.0 B / 0.0 B	0.0 B / 0.0 B	0.0 B / 0.0 B	0.0 B	0		0	0	0	0	0	1.0 min (0.0 ms)	0.0 B	0.0 B	0.0 B			2024-12-06 16:24:13	-
1	4824e3c6130c:35143	Active	0	0.0 B / 6.4 GiB	0.0 B / 6.4 GiB	0.0 B / 0.0 B	5.5 GiB / 174.5 MiB	0.0 B / 0.0 B	1.6 MiB / 0.0 B	20.1 MiB / 16 MiB	0.0 B	2		0	0	0	62	62	45 s (4 s)	3.7 MiB	293.6 MiB	2.2 MiB	stdout stderr		2024-12-06 16:24:15	-
2	4824e3c6130c:38917	Active	0	0.0 B / 6.4 GiB	0.0 B / 6.4 GiB	0.0 B / 0.0 B	6.9 GiB / 177.3 MiB	128 KiB / 0.0 B	994.6 KiB / 0.0 B	20.1 MiB / 16 MiB	0.0 B	2		0	0	0	73	73	41 s (2 s)	290.9 MiB	291.4 MiB	289.1 MiB	stdout stderr		2024-12-06 16:24:16	-

Showing 1 to 3 of 3 entries

Previous 1 Next

Corso Gabriele Manthoné, 69 65127 Pescara (PE)

info@smartshaped.com

+39 375 516 5842

www.smartshaped.com

facebook.com/smartshaped/

linkedin.com/company/smart-shaped-software

twitter.com/SmartShaped

medium.com/@smart_shaped

instagram.com/smartshaped/

threads.net/@smartshaped