DOI: 10.26314/GARR-Conf24-proceedings-07

La Piattaforma di big data in Istat nei processi statistici e di sperimentazione

Francesco Altarocca, Marco Polizzi

Istat

Abstract. Negli ultimi anni, in conformità con i Piani Triennali per l'informatica interni, l'Istat ha applicato una politica per garantire l'evoluzione continua, la razionalizzazione e il consolidamento delle piattaforme IT, nell'ottica di ottimizzare il supporto tecnologico alla ricerca e ai processi di produzione statistica. In particolare, l'Istat ha adottato soluzioni IT on premise, per la cui gestione ed evoluzione è stato necessario coinvolgere figure specializzate (sistemisti, architetti, analisti), in grado di interfacciarsi con esperti di dominio e sviluppatori in soluzioni big data. Il presente contributo descrive l'esperienza Istat nell'acquisizione di una piattaforma big data, sulla quale sono stati implementati importanti processi di produzione statistica. La piattaforma ha consentito di svolgere alcune sperimentazioni su algoritmi computazionali particolarmente complessi, che hanno permesso ai ricercatori di sviluppare metodologie statistiche prima difficilmente implementabili

Keywords. Big data, tecnologie abilitanti, data lake, basi di dati a grafo

Introduzione

Nell'ambito delle attività di ricerca nelle quali è impegnato, l'Istat cura anche lo sviluppo di tecnologie innovative abilitanti alle nuove fonti di dati.

La disponibilità di grandi quantità di dati o big data (in seguito BD), con particolare riferimento a quelli della Grande Distribuzione Organizzata (in seguito GDO), è cresciuta considerevolmente e ha evidenziato la necessità di adeguare i sistemi IT per far fronte ai nuovi carichi elaborativi, abilitando contestualmente i ricercatori a nuovi scenari di indagine.

Il presente contributo descrive l'esperienza Istat nell'acquisizione di una Piattaforma Big Data (in seguito BDP), sulla quale sono stati implementati importanti processi di produzione statistica. La piattaforma ha inoltre reso possibile la sperimentazione di algoritmi complessi a livello computazionale, consentendo di sviluppare metodologie statistiche prima difficilmente implementabili.

1. I processi di produzione statistica con la BDP

1.1 Gli scanner data

Annualmente l'Istat rivede l'elenco dei prodotti che compongono il paniere di riferimento della rilevazione dell'Indice dei Prezzi al Consumo (in seguito IPC), anche per la misura dell'inflazione. Dal 2018 l'Istat utilizza i prezzi registrati alle casse della GDO (AA.VV. 2018) mediante scannerizzazione dei codici a barre o scanner data (in seguito SD).

Brescia, 29-31 maggio 2024 - ISBN: 978-88-946629-3-1

1.2 L'indice dei prezzi al consumo

Negli ultimi anni, il calcolo della stima dell'IPC è stato contraddistinto da un sostanziale aumento delle referenze prese in esame e del numero dei punti vendita da inserire nel campione. I database relazionali impiegati in precedenza hanno manifestato alcune criticità, fra le quali:

- la difficoltà di accogliere e scalare su una quantità di dati sempre crescente;
- la complessità nell'immagazzinare ed elaborare più annualità della stessa rilevazione;
- la difficoltà a raggiungere performance adeguate a svolgere in modo efficiente attività di controllo sulle forniture.

La BDP utilizzata in Istat ha consentito di risolvere tali problematiche, favorendo un'implementazione ottimale del processo di produzione statistica dell'IPC. Settimanalmente l'Istat acquisisce i dati di fatturato, quantità e tipologia dei prodotti dalla GDO per tutte le province italiane. Il campione dei punti vendita è rappresentativo dell'universo delle tipologie distributive della GDO e comprende circa 4.000 punti vendita distribuiti su tutto il territorio nazionale. Nel solo 2023 sono stati acquisiti circa 1 miliardo di dati provenienti dalla GDO.

Per ciascun prodotto e per ciascun punto vendita viene creato e gestito un micro-indice, per un totale di 35 milioni di micro-indici mensili e 421 milioni annuali. Sulla base dei micro-indici sono determinati, mensilmente, quasi 7 milioni di aggregazioni per livello territoriale (provinciale, regionale e nazionale), per raggruppamento di mercato e per tipologia di GDO, per un totale annuo di 80 milioni di aggregati (dati relativi al 2023).

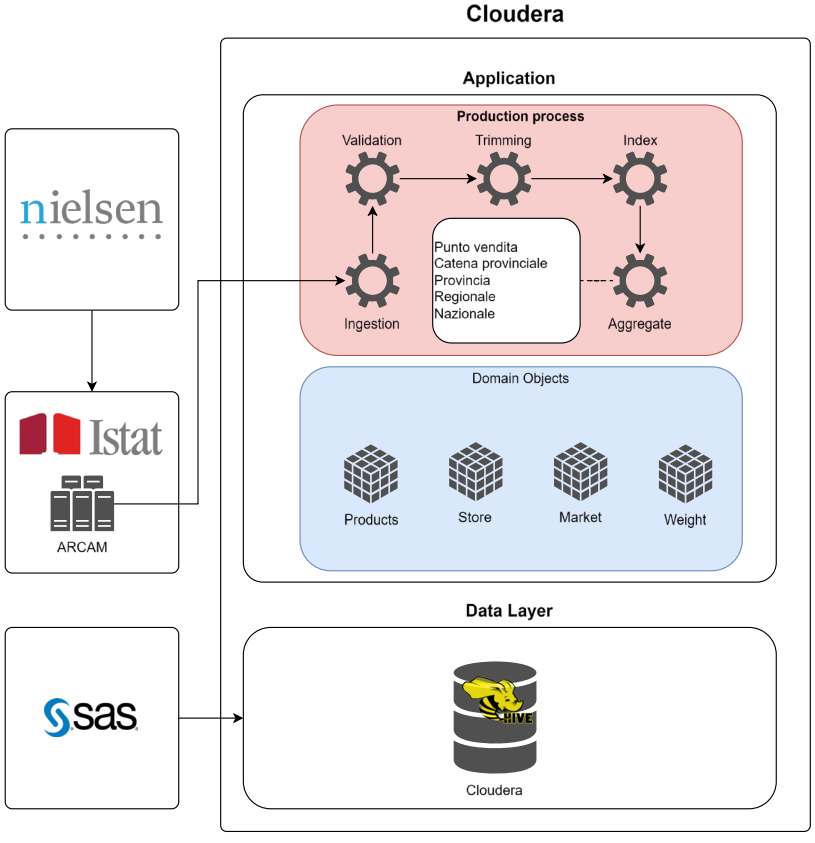


Fig. 1
Architettura del sistema per l'elaborazione degli scanner data

Nella Figura 1 è riportato lo schema del flusso di elaborazione dei dati necessario al processo SD per la produzione dell'IPC.

1.3 La stima della povertà e la parità regionale del potere d'acquisto

Le elaborazioni necessarie alle indagini sulla stima della povertà e della Parità Regionale del Potere d'Acquisto (in seguito PRPA), per la parte SD, sono realizzate utilizzando i dati elementari già acquisiti per l'indagine sugli IPC.
Fra i livelli considerati per la stima della povertà, oltre a quelli già utilizzati per l'IPC, viene incluso il formato del prodotto e sono presi in esame solo i formati di confezionamento più rappresentativi.
La BDP è utilizzata anche per il progetto sperimentale sulla PRPA, che fornisce una misura delle differenze nel livello medio dei prezzi di un paniere di prodotti fra una regione e l'altra e consente, quindi, di rilevare le diseguaglianze e le condizioni di vita delle famiglie nei diversi territori dovute alle differenze nel potere d'acquisto che le caratterizza (AA.VV. 2023).

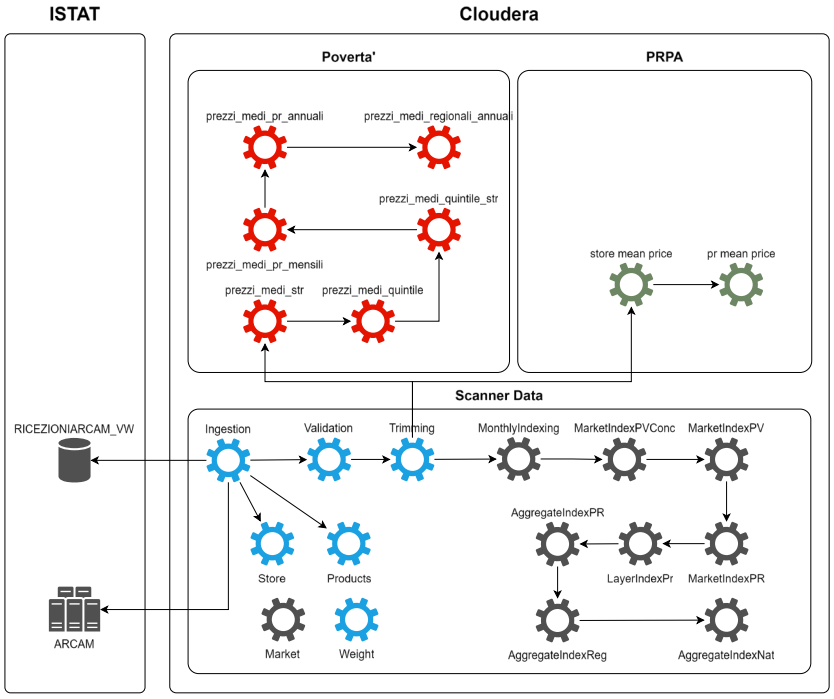


Fig. 2
Package per il trattamento dei dati del data-lake degli scanner data

La Figura 2 schematizza i moduli software utilizzati in comune per le tre indagini appena descritte.

1.4 L'infrastruttura tecnologica big data in Istat

L'Istat ha adottato una soluzione di BD innovativa, implementando un'architettura Cloudera su un ambiente virtuale, basato sulla tecnologia iper-convergente Nutanix. Tale tec-

nologia consente di sfruttare alcuni vantaggi caratteristici degli ambienti virtuali, quali la flessibilità, la scalabilità, la facilità di gestione e la possibilità di ottenere elevate prestazioni. La BDP risiede su cluster composto da 21 macchine virtuali, 40 CPU, 64 TB di spazio disco e 2.330 GB di RAM.

2. Le sperimentazioni per lo sviluppo di metodologie statistiche

2.1 Il record linkage probabilistico

Oltre ai processi di produzione statistica appena descritti, la BDP ha reso possibile lo sviluppo di alcune interessanti sperimentazioni.

La sperimentazione sul record linkage probabilistico ha avuto lo scopo di effettuare alcuni test di performance e di fattibilità rispetto all'attività di record linkage con il progetto Splink, sfruttando le potenzialità della BDP. Splink è un progetto libero che, utilizzando le API Spark Python, implementa il modello Fellegi-Sunter per il record linkage (Linacre, R., et al. 2022).

Nel corso dei test sono stati elaborati dataset di varie dimensioni, analizzati i tempi di esecuzione e individuata la configurazione migliore in base ai parametri utilizzati.

L'analisi dei risultati ottenuti si è rivelata utile per lo studio di fattibilità di altri progetti per i quali si prevede l'utilizzo della stessa tecnologia per elaborare una quantità di dati finora difficilmente gestibile.

2.2 Le basi di dati a grafo

Un'altra sperimentazione è stata condotta sul modello del registro base degli individui, che costituisce il riferimento unico per tutte le statistiche ufficiali riferite alla popolazione abitualmente dimorante.

La sperimentazione ha avuto lo scopo di indagare le potenzialità dei database a grafo nel collegare entità e caratteristiche degli individui in layer differenti e di valutare potenziali punti di forza rispetto ai database relazionali. Per tale sperimentazione è stata utilizzata la libreria GraphX (Gonzalez, J. E., et al. 2014) per il calcolo parallelo sulla BDP in modalità trasparente.

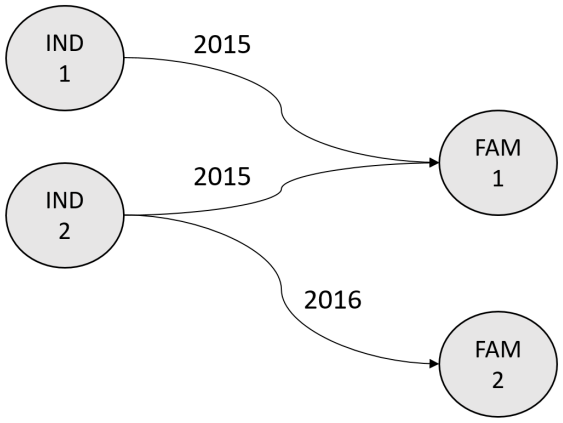
Il dataset di input era costituito da circa 3 milioni di record, dove ciascun record identificava un individuo e la relativa appartenenza a una famiglia durante un quinquennio.

Sono stati quindi prodotti vertici per rappresentare individui e famiglie, mentre gli archi identificano l'appartenenza di un individuo ad una famiglia per ciascun anno (Figura 3).

L'esito della sperimentazione ha evidenziato la facilità e l'efficacia con cui è possibile estrarre la struttura, le relazioni, le particolarità del tessuto delle famiglie e gli spostamenti di individui da una famiglia ad un'altra.

Un ulteriore promettente sviluppo della soluzione prevede l'integrazione di ontologie per aggiungere una componente di reasoning al sistema.

Fig. 3
Esempio di modellazione di individui e di famiglie per la sperimentazione



3. Conclusioni

Partendo dai dati elementari del data-lake IPC, la BDP è stata impiegata per il parziale riuso di pacchetti, procedure e flussi di lavoro utili allo sviluppo di altre indagini (cfr. Figura 2). La capacità della BDP di definire strutture dati indipendentemente dal formato, dalle partizioni e dalle posizioni dei dati sottostanti ha agevolato notevolmente i lavori per integrare altri processi produttivi e di ricerca.
Abilitare i ricercatori e gli esperti di dominio all'uso di tali tecnologie per ottenere vantaggi, sia rispetto alla velocità dei risultati, sia sulla capacità di gestione di grandi quantità di dati, sta diventando sempre più un aspetto fondamentale nella produzione di ricerca e nell'innovazione dei processi.

Riferimenti bibliografici

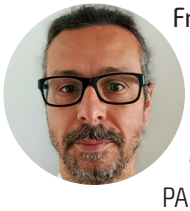
AA.VV. (2018) GLI INDICI DEI PREZZI AL CONSUMO. Aggiornamenti del paniere, della struttura di ponderazione e dell'indagine – Nota informativa ISTAT, https://www.istat.it/it/files//2018/02/Nota-informativa_-paniere2018_fp-1.pdf

AA.VV. (2023) Indici spaziali dei prezzi al consumo - Anno 2021 – ISTAT, Testo integrale e Nota metodologica https://www.istat.it/it/archivio/287297

Gonzalez, J. E., Xin, R. S., Dave, A., Crankshaw, D., Franklin, M. J., & Stoica, I. (2014). {GraphX}: Graph processing in a distributed dataflow framework. In 11th USENIX symposium on operating systems design and implementation (OSDI 14) (pp. 599-613)

Linacre, R., Lindsay, S., Manassis, T., Slade, Z., & Hepworth, T. (2022). ApplyiSplink: Free software for probabilistic record linkage at scale. International Journal of Population Data Science, 7(3), 1794. https://doi.org/10.23889/ijpds.v7i3.1794

Autori



Francesco Altarocca fraltaro@istat.it

Ricercatore presso l'Istat (Dir. Informata DCIT) da oltre 20 anni. È responsabile dell'iniziativa "Produzione sistemi IT per Big Data" ed è coinvolto in linee di attività, progetti strategici e comitati in ambito big data, nuove fonti, AI, strumenti innovativi e Trusted Smart Statistics. Coordina task force e gruppi di lavoro, ha effettuato docenze nell'ambito del corso "La PA nell'epoca dei big data" e collabora con altri autori per articoli e pubblicazioni.

Marco Polizzi polizzi@istat.it

Tecnologo presso la Direzione centrale per le tecnologie informatiche – Istat
Ha svolto attività di ricerca su remote sensing presso l'Università "La Sapienza" di Roma e l'Enea.
Come Istat è stato responsabile dell'iniziativa IT "Architetture dati, Big Data, LOD e BI", ora dell'iniziativa IT "Piattaforme per gestione e analisi dati", per il supporto informatico alle indagini statistiche con strumenti RDBMS e BI. È stato referente Istat per il progetto PNRR Catalogo Nazionale Dati.