Esperienze di Advanced Analytics nella statistica ufficiale: strumenti e progetti

Antonino Virgillito

Direzione Centrale per le tecnologie informatiche e della comunicazione

Introduzione

I Big Data nella statistica ufficiale

A partire dal 2013 la comunità statistica internazionale è impegnata nello studio di come sfruttare fonti dati alternative per la produzione di statistica ufficiale

Il percorso fatto finora ha evidenziato come l'impatto dei Big Data investa il processo di produzione a tutti i livelli

Innovazione tecnologica nei processi di produzione

Piattaforme Big Data

Visualizzazione

Machine learning

Advanced Analytics

A gennaio 2016 Istat ha completato il setup della piattaforma di produzione on-premise per la memorizzazione e l'elaborazione dei big data

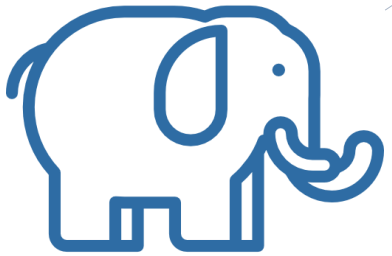
Perché non in cloud?
Non possibile per i vincoli di privacy sui dati

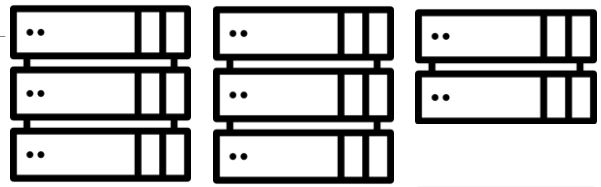
Non (solo) sperimentale
Acquisita e installata per rispondere a un requisito specifico (progetto Scanner Data)

Standard Hadoop
parallel storage/processing,
SQL, NoSQL, Spark…

Extensions
High-speed analytics engine
Administration console

Security
Advanced access control





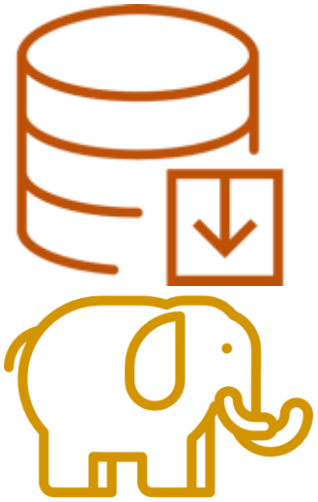
Specifiche tecniche
32/16 Core CPUs
128 Gb RAM per nodo
Connessione interna a 20Gbit
6 x 1.2Tb HD per nodo (60Tb in totale)

Cluster Hadoop da 8 nodi

La piattaforma è stata progettata con l'idea di ospitare dati per progetti diversi e ad alto livello di criticità dal punto di vista della privacy

E' stato implementato un meccanismo avanzato di sicurezza

- Integrazione con il back-end di autenticazione via Kerberos
- Definizione di permessi a livello dettagliato (tabella)

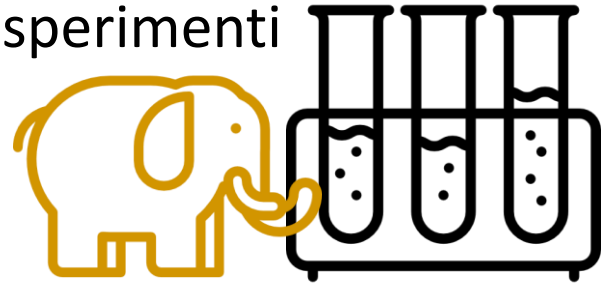


RDBMS
Offload

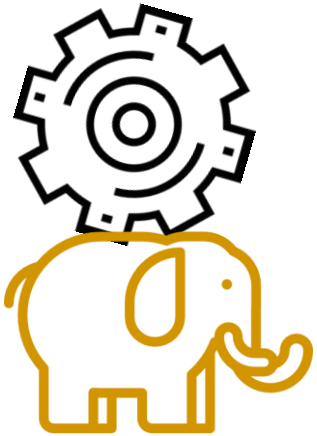
Big Data
staging

Scenari di utilizzo

Esperimenti



Elaborazioni
pesanti



Use Case 1
Scanner Data

Nuova sorgente dati per il calcolo dell'indice dei prezzi al consumo

Transazioni dei prodotti nei supermercati, registrate alle casse

Un record per prodotto ➔ quantità, fatturato (per settimana)

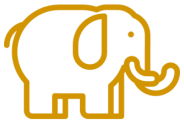
Fornitura dati settimanale

Campione di 2100 negozi che coprono 80 province

750 milioni di record all'anno

Use Case 1
Scanner Data

Architettura dati ibrida (database offload)



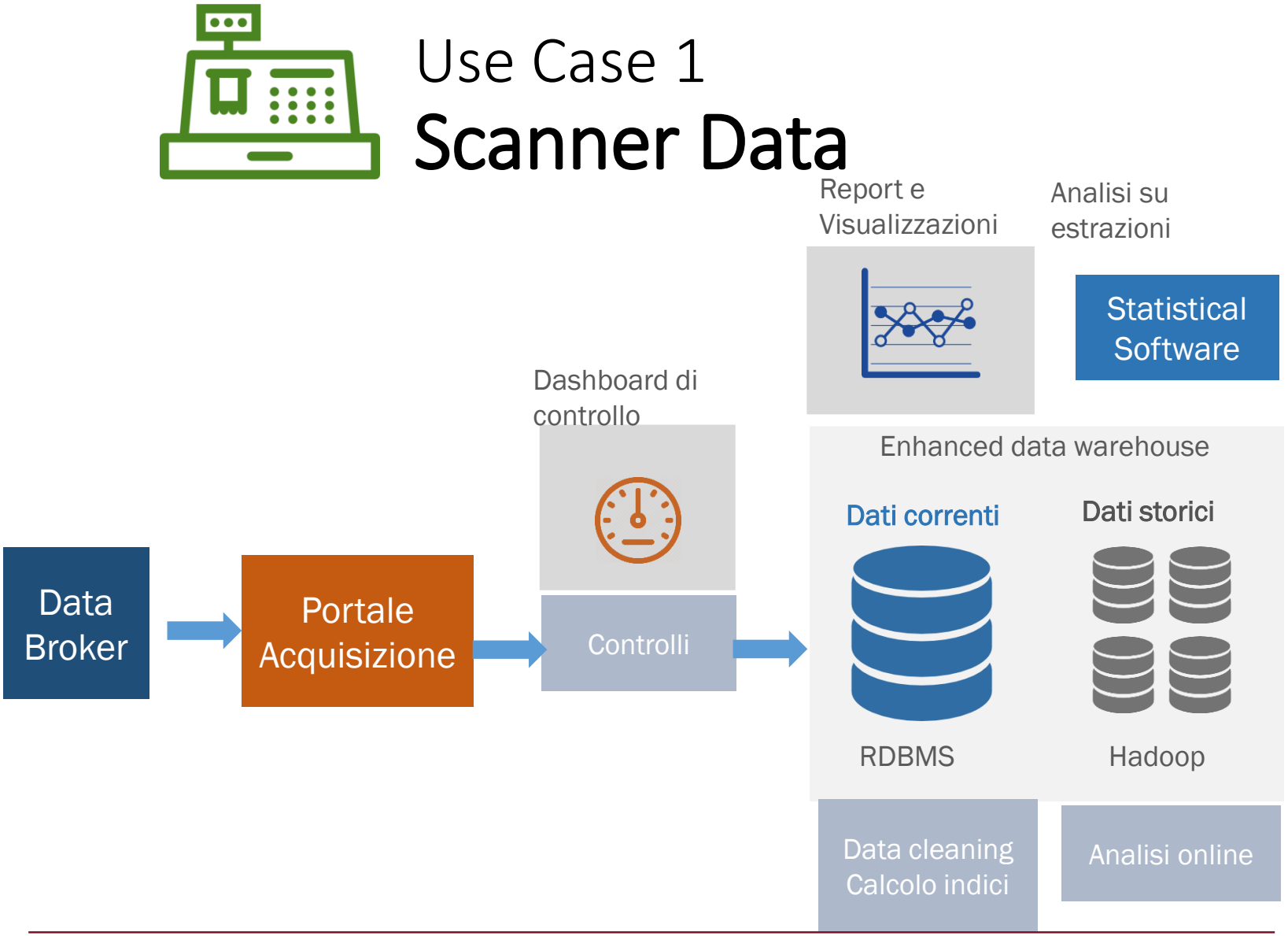
DBMS mantiene dati correnti

Procedure di data cleaning

Hadoop mantiene dati storici

Sempre disponibili per analisi via SQL o tool di BI/visualizzazione

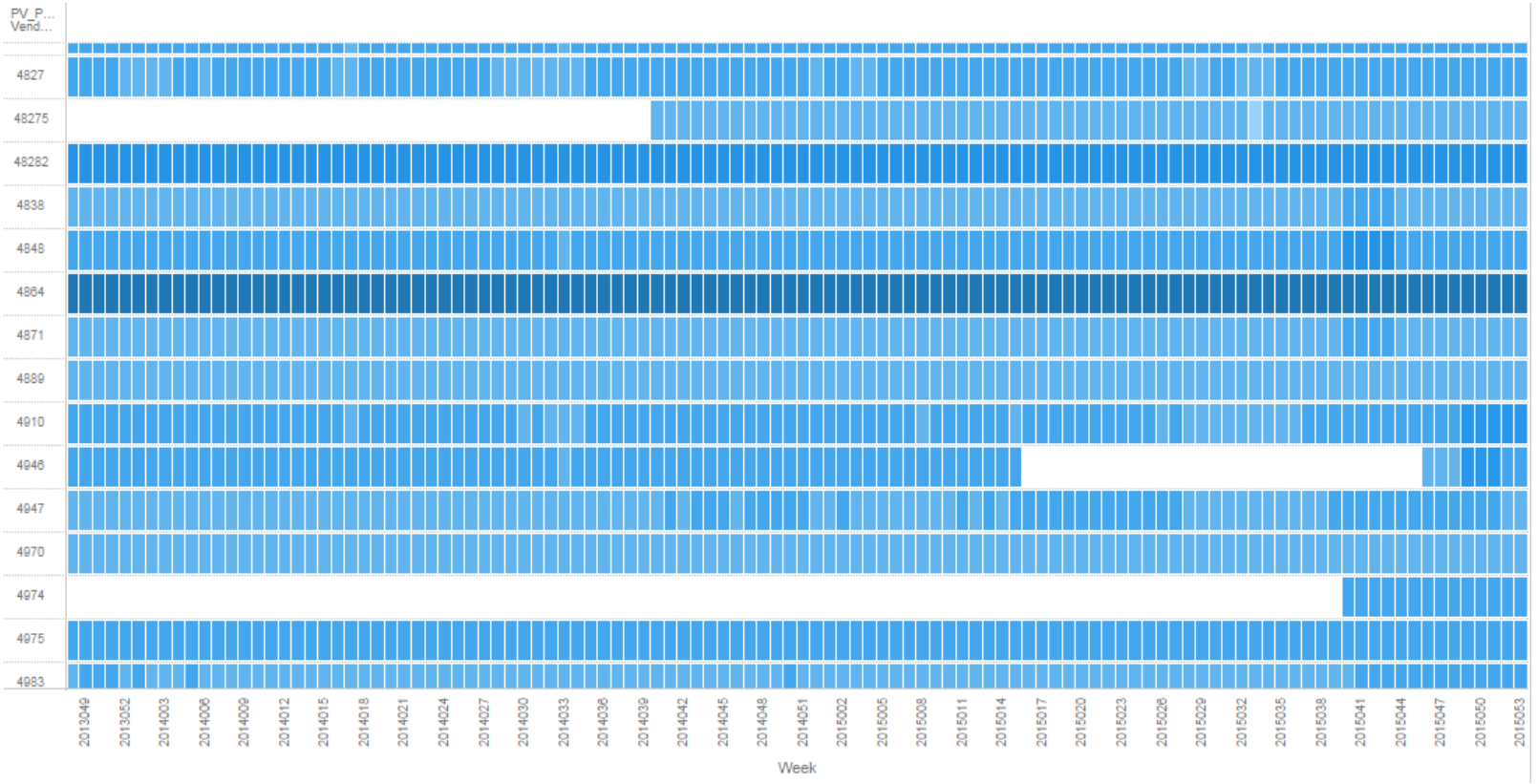
Use Case 1
Scanner Data



Use Case 1
Scanner Data

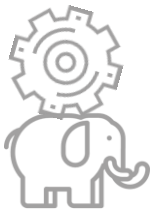
Timeline

2015 Inizio progetto e analisi preliminare dei dati
2016 Preparazione della data collection
2017 Calcolo degli indici in parallelo
2018 Produzione



Andamento della fornitura in termini di numero di record ricevuto per punto vendita per ogni settimana

Use Case 1
Scanner Data

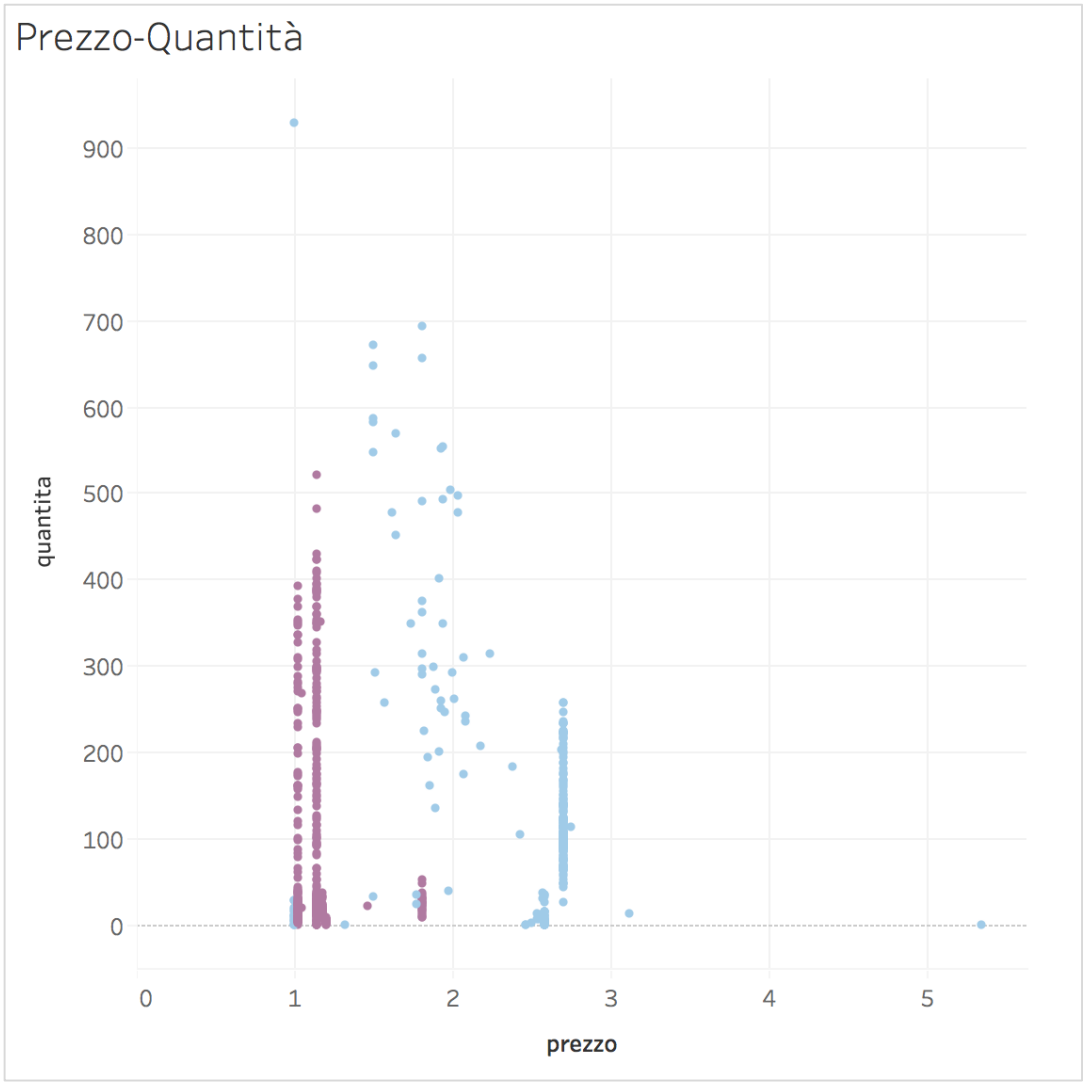


Utilizzo della piattaforma Big Data per il calcolo e l'analisi sull'intero dataset

Calcolo degli indici con diversi metodi e confronto dei risultati

Implementazione di diverse metodologie per l'eliminazione dei dati anomali e confronto dei risultati

Procedure implementate in Spark



Analisi della distribuzione dei dati per la valutazione delle performance delle procedure di identificazione dei dati anomali

Possibile sperimentazione di tecniche di machine learning

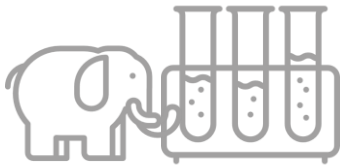
Use Case 1
Scanner Data

Prossimi passi

Consolidamento del processo di produzione e inizio del parallelo con la rilevazione tradizionale

Statistica sperimentale: implementazione di un modello per il calcolo di indici di parità del potere d'acquisto

Use Case 2
Dati telefonici



Attività sperimentale

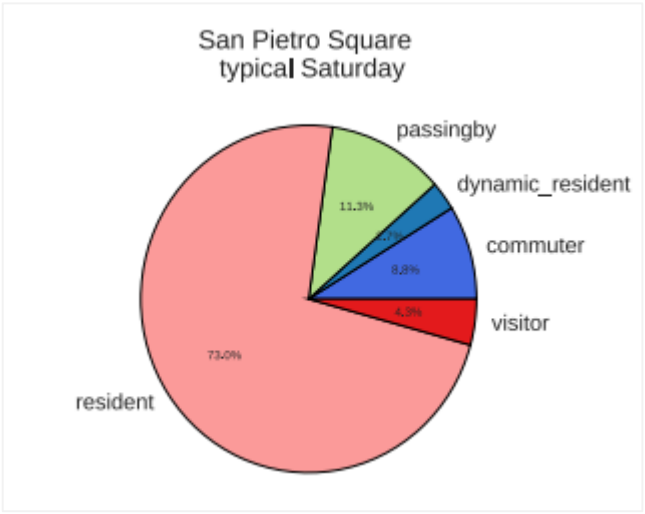
Analisi di dati telefonici per determinare pattern di movimento della popolazione

Campione di dati
Un mese di telefonate/SMS su Pisa e Roma

Call Detail Records ➔ ID chiamante, ora, durata, posizione antenna

Strumento: "Sociometer"
Software realizzato da Università di Pisa/CNR in Spark

Use Case 2
Dati telefonici



Esecuzione del software sul nostro cluster interno e analisi dei risultati per valutarne la possibilità di utilizzo per il calcolo di varie stime

- Presenze sui territori (flussi e stock)
- Mobilità/pendolarismo
- Domanda turistica domestica
- Densità di presenze in luoghi chiave

Conclusioni

Conclusioni:
Installazione e configurazione Hadoop

Processo **molto** complesso:

- Molti componenti interconnessi
- Non semplice capire gli errori

L'installazione è solo l'inizio…

- Aggiornamenti continui
- Gestione degli utenti
- Guasti, rallentamenti, etc.

Conclusioni: Costruzione delle competenze

- Profili e skill diversificati e molto specifici
- IT – Sistemisti
  - Hadoop coinvolge pesantemente il settore ed è necessaria una formazione mirata
- IT - analisti dati/sviluppatori
  - L'uso di SQL garantisce una transizione fluida ai nuovi strumenti per gli analisti DB
  - Spark trova ampio margine di applicazione ma ha una curva di apprendimento ed è più apprezzato dagli sviluppatori
- Statistici
  - Cambio di paradigma necessario per costruire la capacità di lavorare su dataset più grandi
  - Cooperazione più stretta con l'IT per sfruttare meglio il potenziale della tecnologia

Grazie!