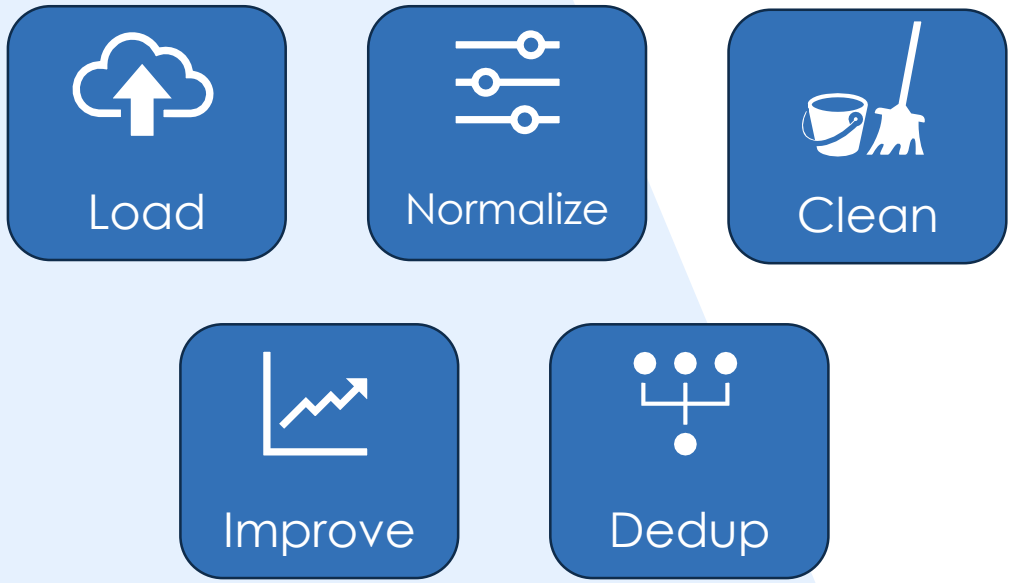
Gestione partecipata dei dati in IRIS… si può fare!

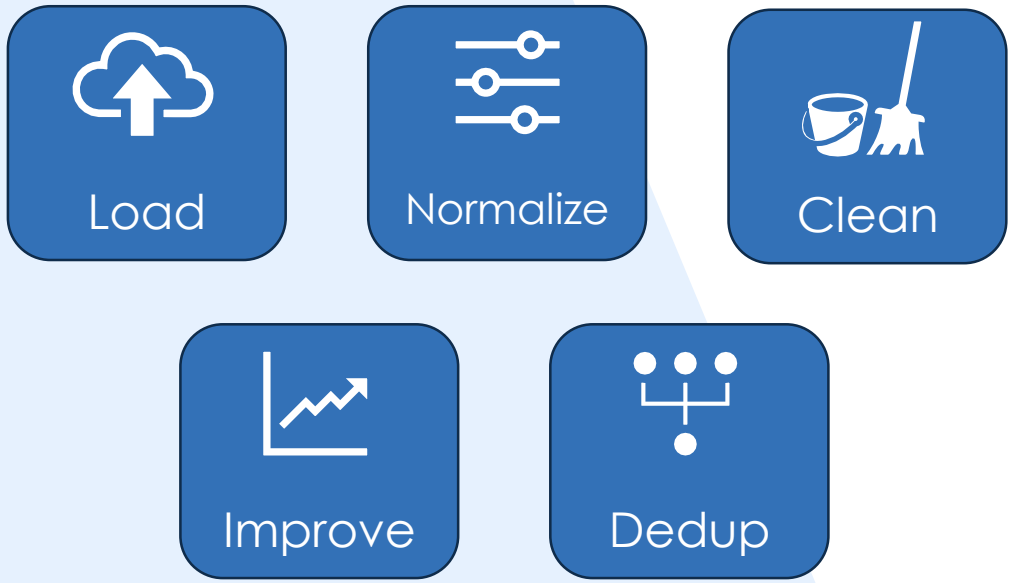
17 gennaio 2024

Fabrizio Luglio
f.luglio@cineca.it

Datalake: Le fasi di creazione



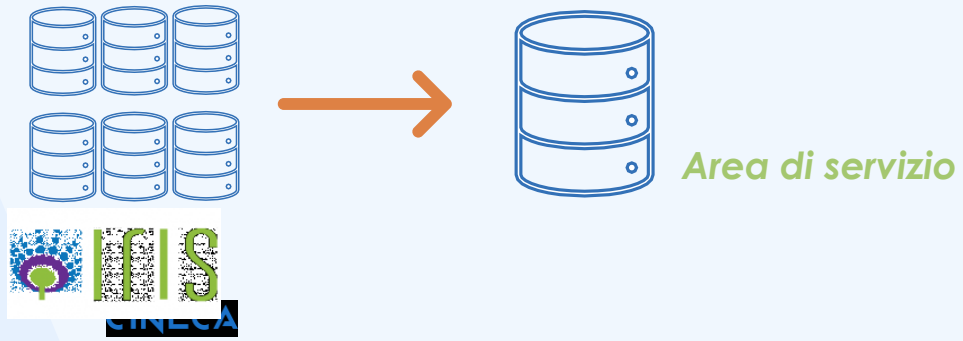
Datalake: Le fasi di creazione



Datalake: Le fasi di creazione

Load

I dati vengono prelevati dagli IRIS italiani….



….e caricati in **un'area di servizio** per subire elaborazioni prima di essere caricati nel datalake

85 Enti(ulteriori **7** in attivazione)

93,5% dei docenti/ricercatori italiani

Datalake: Le fasi di creazione

Alcune informazioni non sono standardizzate in tutti gli enti. Possono quindi

- Assumere valori differenti
- Possono non essere popolate

I dati devono quindi essere resi omogenei

Ruolo: Viene uniformato – PO, Professore ordinario, Professore di I° fascia, etc.

Sesso e anno nascita: recuperati dal CF se non presente

…..

Datalake: Le fasi di creazione

I Dati vengono selezionati e ripuliti:

- Non tutte i prodotti contribuiscono all'archivio finale

Tipologie: vengono scartate alcune tipologie non significative nelle valutazioni nazionali (VQR, ASN, etc.)

non ammesse "Recensione in rivista (263), Scheda bibliografica (264), Abstract in rivista (266), Abstract in Atti di convegno (274), Poster (275), Indice (278), Bibliografia (279), Manufatto (292), Recensione in volume (301)

Datalake: Le fasi di creazione

I Dati vengono selezionati e ripuliti:

- Non tutte i prodotti contribuiscono all'archivio finale
- Ove possibile vengono verificati e corretti

Identificativi

- Viene verificata la correttezza degli identificativi (WOS, Scopus, Medline, etc.) attraverso delle maschere- Lunghezza id, Presenza WOS: , 2-s2-, etc..
- Vengono corretti ove possibile o scartati

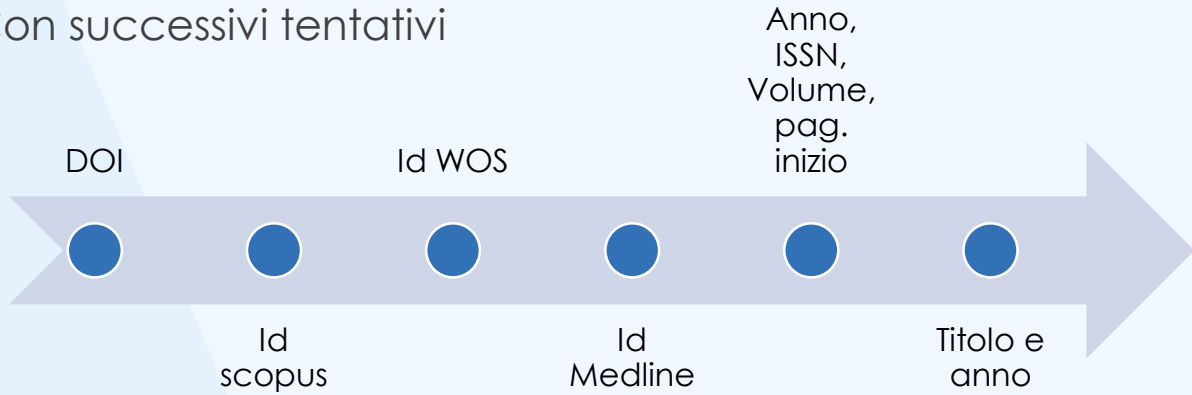
Datalake: Le fasi di creazione

Improve

La banca dati è ora pronta per l'identificazione dei duplicati ma prima di effettuare la deduplicazione vengono fatte delle attività di miglioramento dei prodotti.

Identifico gli stessi prodotti

Con successivi tentativi



Fusione dei medesimi prodotti

Creando un prodotto che è il meglio dei prodotti di partenza

Datalake: Le fasi di creazione

In caso di conflitto sui metadati come faccio a definire quale scegliere?

Soluzione attuale (subottimale)

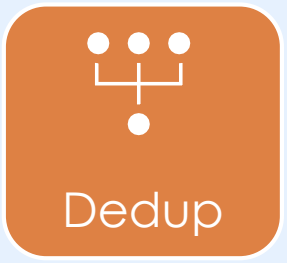
Il primo prodotto selezionato vince sui successivi

Soluzione da sviluppare (ottimale)

Viene redatto un **«Indice di qualità»** dell'archivio di provenienza basato sui seguenti parametri:

1. E' presente un processo di validazione
2. % di prodotti in attesa di validazione
3. % di prodotti con identificativi mancanti ma esistenti
4. % di prodotti con metadati discrepanti rispetto a banca dati scopus e wos

Datalake: Le fasi di creazione

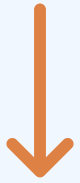


A questo punto l'archivio è pronto per essere deduplicato….

62.899

62.025

7.329.776



4.459.330

Citando quanto detto @UNIMORE… basare considerazioni sul datalake diventa interessante se…

1. **endorsement (atenei, MUR, ANVUR)**
2. **lo usano tutti o quasi**
3. **si potenziano gli interventi a supporto della qualità del dato negli Iris locali**

IRIS
INSTITUTIONAL RESEARCH INFORMATION SYSTEM

Grazie

FABRIZIO LUGLIO

IRIS PRODUCT OWNER