

REFERENCA: DEDUPLIKIMI I TE DHENAVE

CFARE ESHTË DEDUPLIKIMI?

Deduplikimi është procesi i identifikimit dhe eliminimit të rekordeve të dyfishta në një databazë. Rekorde duplikate krijohen kur i njëjti entitet (person, biznes, adresë) futet me të dhëna paksa të ndryshme.

PSE ESHTË I RENDESISHEM?

- * Statistika të sakta - numri real i qytetareve/sherbimeve
- * Kursim kostosh - shmangje e sherbimeve të dyfishta
- * Cilësi e të dhënave - databazë e pastër dhe e besueshme
- * Perputhje ligjore - GDPR kërkon saktësi të dhënash

TEKNIKAT E DEDUPLIKIMIT

Teknika	Pershkrimi	Perdorimi
Exact Match	Perputhje e plote e fushave	Duplikate identike
Fuzzy Match	Perputhje e përafërt	Gabime shkrimi
Phonetic Match	Tingëllim i ngjashëm	Variante emrash
Token Match	Fjalë të njëjta në rend tjetër	Emra të invertuar

ALGORITMET E NGJASHMERISE

Algoritmi	Si funksionon
Levenshtein	Numeron ndryshimet minimale (insert/delete/replace)
Jaro-Winkler	Favorizon perputhjet në fillim të stringes
Soundex	Kodon tingëllimin fonetik të fjales
N-gram	Krahason sekuenca karakteresh

PRAKTIKAT ME TË MIRA

- Standardizoni të dhënat para krahasimit (kapitalizim, formate)
- Perdorni shumë fusha për krahasim, jo vetëm emrin
- Vendosni pragje të ndryshme për fusha të ndryshme
- Rishikoni manualisht përpara bashkimit
- Mbani log të të gjitha bashkimeve për auditim

6. Beni backup para cdo operacioni bashkimi

Referenca Deduplikimi - Workshop AI Biometrike