Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Προχωρημένα Θέματα Βάσεων Δεδομένων
9ο εξάμηνο, ακαδημαϊκό έτος 2025-2026

Εξαμηνιαία Εργασία

Ομάδα: group16

Τάτση Αικατερίνη Δανάη, ΑΜ: 03121146

Σπυρόπουλος Νικολας, ΑΜ: 03121202

Περιεχόμενα

1 Query 1

Το πρώτο ερώτημα μας ζητείται να υλοποιηθεί με 3 διαφορετικούς τρόπους - χρησιμοποιώντας το DataFrame API μία φορά χωρίς και μία φορά με User Defined Function (UDF) και χρησιμοποιώντας το RDD API. Από την θεωρία ξέρουμε ότι όταν χρησιμοποιούμε το RDD API, το φυσικό πλάνο των πράξεων το ορίζουμε εμείς επακριβώς. Δεν μπορεί, δηλαδή, να παρέμβει ο Catalyst Optimizer. Από την άλλη, όταν χρησιμοποιούμε το DataFrame API, ο Catalyst παρεμβαίνει και βελτιστοποιεί το query μας. Όμως, στην περίπτωση μιας συνάρτησης ορισμένης από τον χρήστη (UDF), ο Catalyst την "βλέπει" ως ένα μαύρο κουτί και το συγκεκριμένο κομμάτι δεν μπορεί να το αλλάξει. Επομένως, περιμένουμε η γρηγορότερη υλοποίηση να είναι το DataFrame χωρίς UDF, στην συνέχεια το DataFrame με UDF και τέλος το RDD.

Πράγματι, μετρώντας τους χρόνους εκτέλεσης των τριών παραπάνω υλοποιήσεων επαληθεύουμε αυτό που αναμέναμε. Οι χρόνοι φαίνονται στον πίνακα 1. Παρατηρούμε ότι η μη χρήση UDF είναι περίπου $1.3s$ γρηγορότερη από την χρήση μιας UDF, ενώ η υλοποίηση με RDD είναι περίπου 2 φορές πιο αργή από τις υλοποιήσεις με το DataFrame API.

Πίνακας 1: Χρόνοι εκτέλης των τριών υλοποιήσεων του Query 1

Method	Time (s)
DF-no-UDF	7.23
DF-UDF	8.53
RDD	16.9