ΔΕΥΤΕΡΗ ΟΜΑΔΙΚΗ ΕΡΓΑΣΙΑ ΣΤΟ ΜΑΘΗΜΑ BIG DATA / ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕΓΑΛΟΥ ΟΓΚΟΥ

Μέλη ομάδας:

Ζερβας Μιχάλης (ics20015)

Δρίκος Χρήστος(ics20046)

Βήματα - Προβλήματα - Τρόποι επίλυσης

Τα βήματα που ακολουθήσαμε ήταν αρχικά η καλύτερη κατανόηση του τρόπου λειτουργίας του Spark και των εντολών που θα μας χρειαζόντουσαν. Έπειτα βασιζόμενοι κυρίως στο υλικό που μας δινόταν, προχωρήσαμε στη σύνθεση του κώδικα. Στον κώδικα τα προβλήματα που συναντήσαμε αφορούσαν τον τρόπο που θα υπολογίζεται το threshold και τον τρόπο αποθήκευσης των αποτελεσμάτων.

Όσων αφορά το threshold, στην αρχή τρέχαμε τον FP-growth με support=0 και μετά με την filter αφήναμε τα αποτελέσματα με frequency>=threshold. Όμως επειδή αυτή η υλοποίηση θα δημιουργούσε καθυστερήσεις λόγω επιπλέον υπολογισμών και επειδή δεν ήταν καλή προσέγγιση προγραμματιστικά, την αφήσαμε και βάλαμε το support να υπολογίζεται βάση του threshold και το σύνολο των συναλλαγών.

Σχετικά με τον τρόπο αποθήκευσης των αποτελεσμάτων, έπρεπε να γίνει μετασχηματισμός στο dataframe σε μία μονή στήλη string, έτσι ώστε να μπορέσει να γραφεί σε ένα αρχείο csv. Έπειτα, επειδή τα αποτελέσματα γραφόντουσαν σε πολλά μικρά parts, επιλέξαμε, για να είναι πιο ξεκάθαρα, να τα συγχωνεύουμε όλα αυτά σε ένα αρχείο csv. Επίσης όταν προσπαθούσαμε να το τρέξουμε σε 2 workers, εμφανιζόντουσαν κάποια errors στην έξοδο του spark, τα οποία όμως τελικά οφείλονταν στο ότι τα 2 vm είχαν διαφορετική έκδοση python.

Τέλος, αφού επιλύθηκαν αυτά τα προβλήματα και καταλήξαμε στην τελική μορφή του κώδικα, προχωρήσαμε στην συλλογή των αποτελεσμάτων και όλων των άλλων απαραίτητων πληροφοριών που ζητούνται για κάθε threshold.

Χρόνοι (http://83.212.80.243:8080/)

2 workers:

threshold: 5.000: **min_time** = 3.2 min, **max_time** = 3.6 min, **avg_time** = 3.4 min

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20230106170426-0008	FP-Growth approach	16	1024.0 MiB		2023/01/06 17:04:26	user	FINISHED	3.2 min
app-20230106170820-0009	FP-Growth approach	16	1024.0 MiB		2023/01/06 17:08:20	user	FINISHED	3.6 min
app-20230106171346-0010	FP-Growth approach	16	1024.0 MiB		2023/01/06 17:13:46	user	FINISHED	3.3 min

threshold: 10.000: **min_time** = 3.1 min, **max_time** = 3.5 min, **avg_time** = 3.3 min

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20230106164453-0004	FP-Growth approach	16	1024.0 MiB		2023/01/06 16:44:53	user	FINISHED	3.4 min
app-20230106165022-0005	FP-Growth approach	16	1024.0 MiB		2023/01/06 16:50:22	user	FINISHED	3.1 min
app-20230106165812-0007	FP-Growth approach	16	1024.0 MiB		2023/01/06 16:58:12	user	FINISHED	3.5 min

threshold: 50.000: **min_time** = 3.2 min, **max_time** = 3.3 min, **avg_time** = 3.2 min

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20230106152748-0000	FP-Growth approach	16	1024.0 MiB		2023/01/06 15:27:48	user	FINISHED	3.2 min
app-20230106154131-0001	FP-Growth approach	16	1024.0 MiB		2023/01/06 15:41:31	user	FINISHED	3.2 min
app-20230106163215-0002	FP-Growth approach	16	1024.0 MiB		2023/01/06 16:32:15	user	FINISHED	3.3 min

1 worker:

threshold: 5.000: **min_time** = 1.4 min, **max_time** = 1.5 min, **avg_time** = 1.4 min

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

app-20230106172016-0011	FP-Growth approach	8	1024.0 MiB		2023/01/06 17:20:16	user	FINISHED	1.4 min
app-20230106172300-0012	FP-Growth approach	8	1024.0 MiB		2023/01/06 17:23:00	user	FINISHED	1.4 min
app-20230106172454-0013	FP-Growth approach	8	1024.0 MiB		2023/01/06 17:24:54	user	FINISHED	1.5 min

threshold: 10.000: **min_time** = 1.4 min, **max_time** = 1.5 min, **avg_time** = 1.4 min

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20230106172728-0014	FP-Growth approach	8	1024.0 MiB		2023/01/06 17:27:28	user	FINISHED	1.4 min
app-20230106173114-0015	FP-Growth approach	8	1024.0 MiB		2023/01/06 17:31:14	user	FINISHED	1.5 min
app-20230106173309-0016	FP-Growth approach	8	1024.0 MiB		2023/01/06 17:33:09	user	FINISHED	1.4 min

threshold: 50.000: **min_time** = 1.3 min, **max_time** = 1.4 min, **avg_time** = 1.3 min

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20230106173624-0017	FP-Growth approach	8	1024.0 MiB		2023/01/06 17:36:24	user	FINISHED	1.3 min
app-20230106174117-0018	FP-Growth approach	8	1024.0 MiB		2023/01/06 17:41:17	user	FINISHED	1.4 min
app-20230106174318-0019	FP-Growth approach	8	1024.0 MiB		2023/01/06 17:43:18	user	FINISHED	1.4 min

Σχολιασμός χρόνων:

Παρατηρούμε ότι οι εκτελέσεις με 1 worker απαιτούν περίπου τον μισό χρόνο από αυτές με 2 workers. Αυτό δεν είναι κάτι αναμενόμενο, καθώς θα περιμέναμε με 2 workers να μοιράζεται ο υπολογιστικός φόρτος και να έχουμε πιο γρήγορες εκτελέσεις. Αυτό ίσως οφείλεται στο network overhead ή και στην εντολή coalesce(1) η οποία συγκεντρώνει όλα τα αποτελέσματα κάθε worker σε ένα αρχείο, οπότε όσο αυξάνονται οι workers πιθανόν να αυξάνεται και ο χρόνος για να τα συλλέξει.