ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΤΜΗΜΑ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

# Παραδοτέο 1ης υποχρεωτικής εργασίας

ΟΛΓΑ ΣΑΡΙΔΟΥ (ics23078) – ΑΠΟΣΤΟΛΟΣ ΣΙΝΙΟΡΗΣ(ics24123)

ΜΑΘΗΜΑ: BIG DATA
ΕΠΙΒΛΕΠΟΥΣΑ ΚΑΘΗΓΗΤΡΙΑ: ΔΡ. ΑΣΗΜΙΝΑ ΔΗΜΑΡΑ

## Περιβάλλον Εκτέλεσης

Για τους κώδικες χρησιμοποιήθηκαν:

Python 3, System RAM και Disk, και Spark version 3.5.1

## Μετρήσεις Θέματος 1 – RDD API

| **Εκτέλεση** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| **Χρόνοι εκτέλεσης προγράμματος** | 27 second | 11 second | 9 second | 8 second |
| **Χρόνοι εκτέλεσης 1st action** | 200 ms | 400 ms | 220 ms | 280 ms |

Στους χρόνους εκτέλεσης του προγράμματος έχουμε MAX = 27 seconds και MIN = 8 seconds.

Ο μέσος χρόνος εκτέλεσης του προγράμματος θα υπολογιστεί χωρίς τις εκτελέσεις MAX και MIN, και είναι AVG = (11 + 9 + 11)/3 = 10.333 seconds.

Το πρώτο action στον κώδικα είναι η συνάρτηση take(10), η οποία τυπώνει τα top 10 αεροδρόμια με τη μεγαλύτερη μέση καθυστέρηση πτήσεων.

Ο μέσος χρόνος εκτέλεσης του πρώτου action είναι AVG = (200 + 400 + 220 + 280 + 130)/5 = 246 ms.

| | |
|---|---|
| **Μέσος χρόνος εκτέλεσης προγράμματος** | 10.333 seconds |
| **Μέσος χρόνος εκτέλεσης 1st action** | 246 ms |

## Μετρήσεις Θέματος 2 - DATAFRAME API

| **Εκτέλεση** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| **Χρόνοι εκτέλεσης προγράμματος** | 36 second | 12 second | 11 second | 12 second |
| **Χρόνοι εκτέλεσης 1st action** | 4160 ms | 1235 ms | 1080 ms | 1090 ms |

Στους χρόνους εκτέλεσης του προγράμματος έχουμε MAX = 36 seconds και MIN = 10 seconds.

Ο μέσος χρόνος εκτέλεσης του προγράμματος θα υπολογιστεί χωρίς τις εκτελέσεις MAX και MIN, και είναι AVG = (12+11+12)/3 = 11.667 seconds.

Το πρώτο action στον κώδικα είναι η συνάρτηση write.csv, η οποία τυπώνει σε αρχείο csv(Comma-Separated Values) τα top 10 αεροδρόμια με τη μεγαλύτερη μέση καθυστέρηση πτήσεων.

Ο μέσος χρόνος εκτέλεσης του πρώτου action είναι AVG = ( 4160 + 1235 + 1080 + 1090 + 430 ) = 7995 ms = 7.9 seconds = 8 seconds.

| | |
|---|---|
| **Μέσος χρόνος εκτέλεσης προγράμματος** | 11.667 seconds |
| **Μέσος χρόνος εκτέλεσης 1st action** | 8 seconds |

## Θέμα 3ο - Συγκριτική Ανάλυση RDD vs DataFrame

Η προσέγγιση RDD χειρίζεται τα δεδομένα σε ένα χαμηλό και αδόμητο επίπεδο, κάτι το οποίο συνεπάγει τη δυσκολία υλοποίησής. Ο κώδικας RDD είναι συνήθως μακροσκελής και δύσκολος στην ανάγνωση, δηλαδή έχουμε χαμηλή εκφραστικότητα.

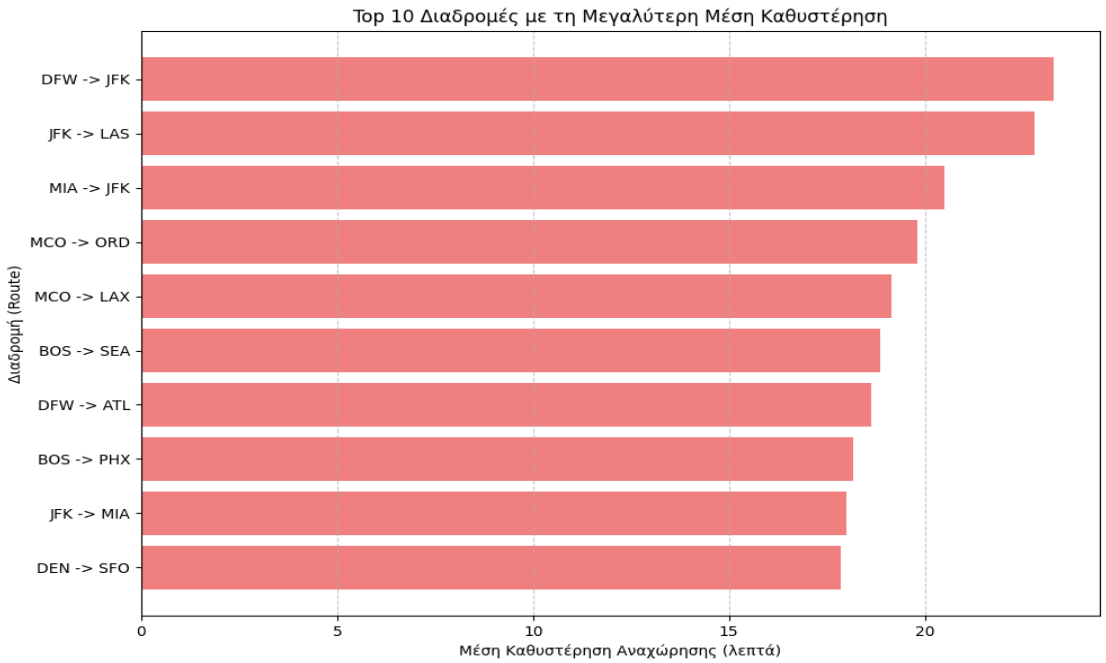
Εν αντιθέση με το DataFrame, το οποίο χρησιμοποιεί δομημένα δεδομένα και schema για τον χειρισμό των δεδομένων του, όπως και η SQL, αλλά και έναν πιο σύντομο και διαισθητικό κώδικα. Το DataFrame, λοιπόν, αποτελεί έναν ευκολότερο και πιο διαισθητικό τρόπο χειρισμού δεδομένων.

Είναι αντιληπτό πως το RDD υλοποιήθηκε ελαφρώς ταχύτερα από το DataFrame στον μέσο συνολικό χρόνο εκτέλεσης, ενώ για τον μέσο χρόνο του πρώτου action, το RDD είναι ασύγκριτα ταχύτερο από το DataFrame. Αυτή η διαφορά οφείλεται στη βελτιστοποίηση Catalyst που εφαρμόζει μόνο το DataFrame προτού εκτελέσει το πρώτο action, δηλαδή την εντολή write.csv.

| | **RDD** | **DATAFRAME** |
|---|---|---|
| **Μέσος χρόνος εκτέλεσης προγράμματος** | 10.333 seconds | 11.667 seconds |
| **Μέσος χρόνος εκτέλεσης 1st action** | 246 ms | 8 seconds |

Ο Catalyst αναλύει τον κώδικα ώστε να βρει τον πιο αποδοτικό τρόπο εκτέλεσης. Αυτό παρόλο που προκαλεί μεγάλο χρόνο στο πρώτο action, εξασφαλίζει επόμενες ταχύτερες εκτελέσεις σε μεγάλα σύνολα δεδομένων.

## Θέμα 4ο - Οπτικοποίηση 10 routes με τη μεγαλύτερη μέση καθυστέρηση



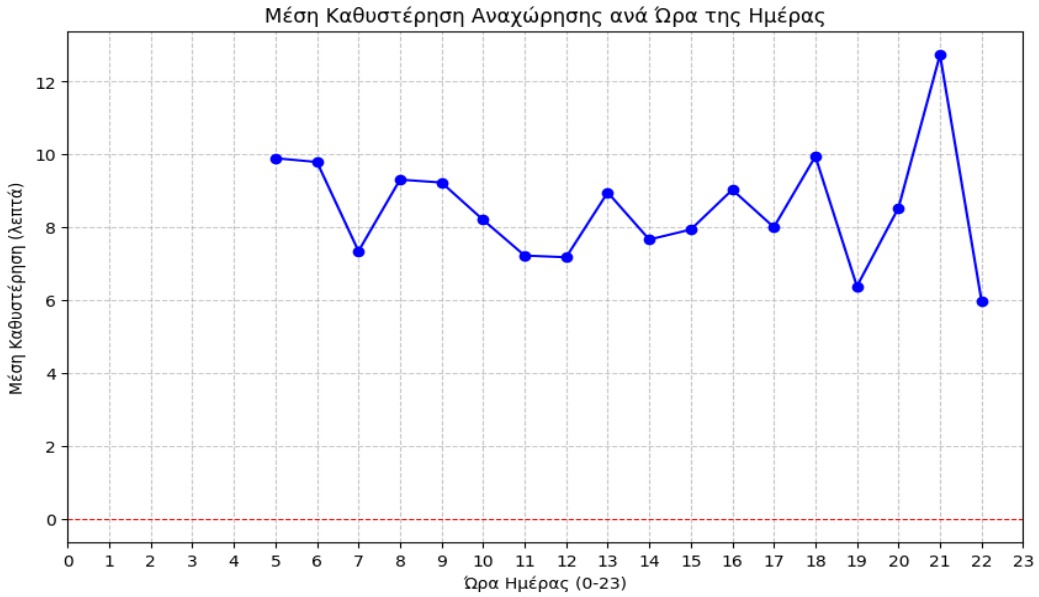

Παρατηρούμε πως οι περισσότερες καθυστερήσεις διαδρομών κυμαίνονται στα 17-20 λεπτά. Η διαδρομή που παρουσιάζει τη μεγαλύτερη μέση

καθυστέρηση είναι η DFW → JFK. Ακόμη, τα αεροδρόμια DFW και JFK εμφανίζονται σε πολλαπλές διαδρομές, υποδηλώνοντας ότι τα αεροδρόμια αυτά τείνουν να καθυστερούν συχνά τα routes τους.

## Θέμα 5ο – Εμπλουτισμός ανάλυσης

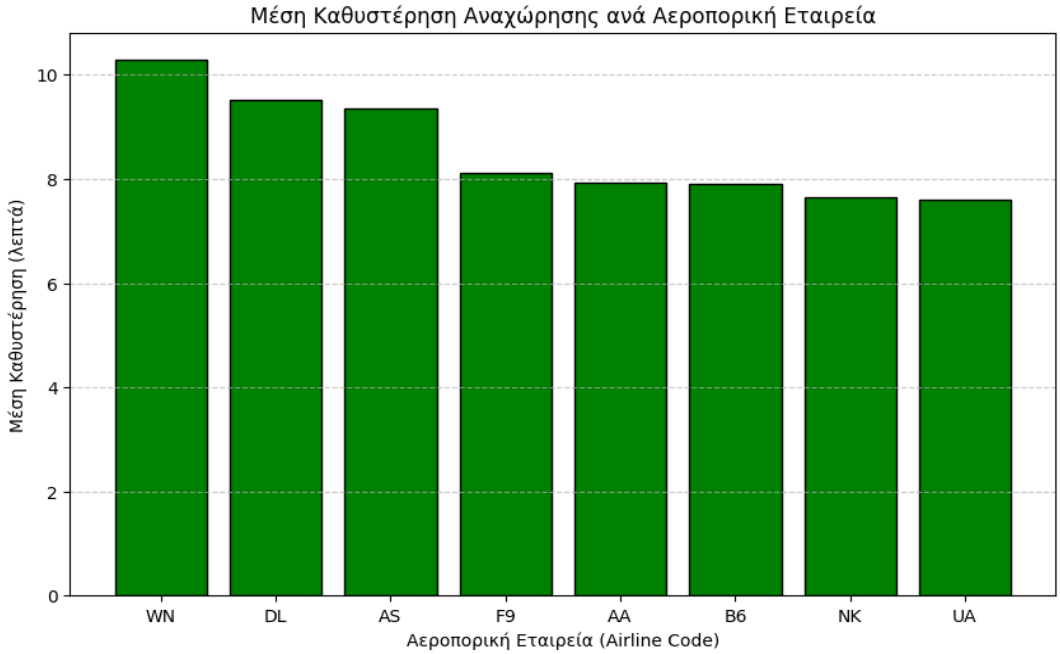
Δημιουργήθηκαν επιπλέον 3 γραφήματα επάνω στα αρχικά δεδομένα

a) Μέση καθυστέρηση ανά ώρα αναχώρησης



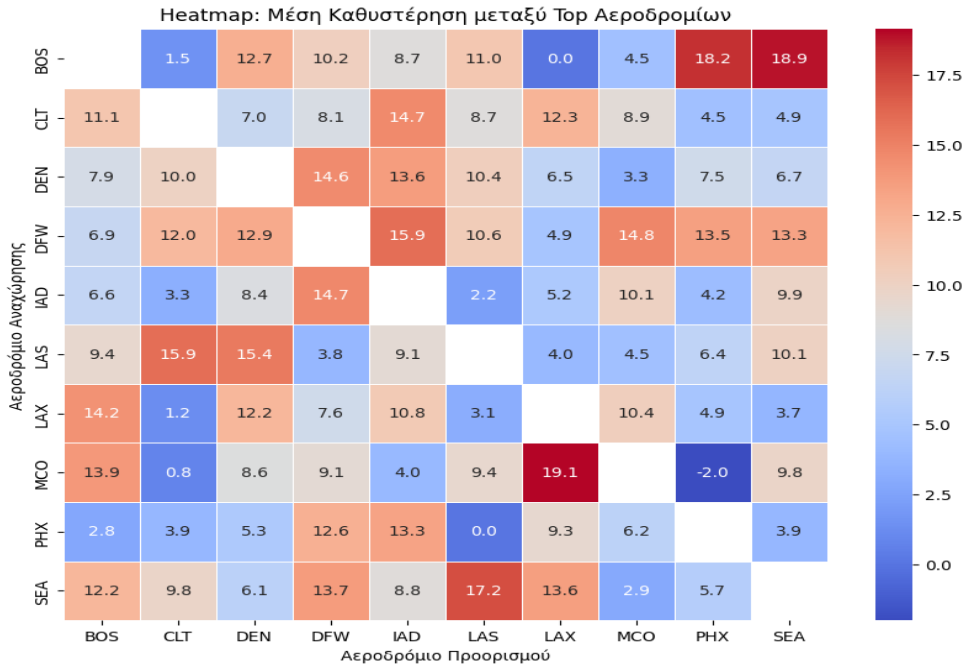

Σκοπός του γραφήματος, είναι να δείξει πως συσχετίζεται κάθε διαφορετική ώρα της ημέρας με την διάρκεια της καθυστέρησης της πτήσης. Παρατηρείται ότι, τις πρωινές ώρες οι καθυστερήσεις είναι λίγες και διαρκούν ελάχιστα (λιγότερο από 10 λεπτά), ενώ καθώς προχωράει η μέρα οι καθυστερήσεις συσσωρεύονται και κορυφώνονται τις βραδινές ώρες.

b) Κατανομή καθυστερήσεων ανά αεροπορική εταιρεία




Το ραβδόγραμμα αυτό, δείχνει την αξιοπιστία των αεροπορικών εταιρειών και επιτρέπει τον εντοπισμό αυτών με την μεγαλύτερη μέση καθυστέρηση. Με αυτές τις πληροφορίες κάθε επιβάτης ή/και κάθε αεροπορική εταιρεία θα γνωρίζει την ποιότητα των υπηρεσιών που του παρέχονται / που παρέχει.

c) Heatmap Routes vs delay



Μέσω της χρωματικής κλίμακας, το Heatmap υποδεικνύει συγκεκριμένα ζεύγη αεροδρομίων (Origin-Destination) που παρουσιάζουν «κόκκινα» σημεία (υψηλές καθυστερήσεις). Αυτό βοηθά να κατανοήσουμε αν οι καθυστερήσεις οφείλονται σε συγκεκριμένους πολυσύχναστους αεροδιαδρόμους ή κόμβους (hubs) που αδυνατούν να εξυπηρετήσουν τον όγκο των πτήσεων.