Flight Delay Analytics με Apache Spark

Ονοματεπώνυμο 1: *Συμεών Αλμανίδης (ics23083)*
Ονοματεπώνυμο 2: *Ευστράτιος Μαυρίγκος (ics23117)*

Εισαγωγή

Αυτή η ανάλυση εξετάζει καθυστερήσεις πτήσεων χρησιμοποιώντας Apache Spark, συγκρίνοντας δύο βασικά APIs — RDD και DataFrame — ως προς υλοποίηση, απόδοση και επεκτασιμότητα. Στόχοι του έργου:

- Να ποσοτικοποιηθούν οι μέσες καθυστερήσεις ανά αεροδρόμιο και ανά διαδρομή.
- Να συγκριθούν οι χρόνοι εκτέλεσης και οι απαιτήσεις κώδικα μεταξύ RDD και DataFrame υλοποιήσεων.
- Να αναδειχθούν βέλτιστες πρακτικές για παραγωγική χρήση μέσω του Catalyst engine.
- Να παρασχεθούν οπτικοποιήσεις και επιχειρησιακά συμπεράσματα για προτεραιοποίηση βελτιώσεων στις πιο προβληματικές διαδρομές/αεροδρόμια.

Υλικό (hardware) που χρησιμοποιήθηκε για την έρευνα:

Processor	Intel(R) Core(TM) i3-8100 CPU @ 3.60GHz 3.60 GHz
Installed RAM	8.00 GB
Storage	250GB
Graphics Card	NVIDIA GeForce GT 1030 (2 GB)

Θέμα 1: Spark RDD API

AIRPORT	AVERAGE DELAY
DFW	11.976923076923077
JFK	11.201680672268907
SEA	9.56115107913669
ORD	9.348837209302326
LAS	9.106382978723405
CLT	8.823529411764707
MCO	8.694656488549619
BOS	8.631205673758865
DEN	8.442857142857143
MIA	8.262295081967213

Total time	0.3009929	0.3888978	0.4449472	0.4480452	0.4527261
Action time	0.27897381	0.3705472	0.4236063	0.43217968	0.4332683

Code output:

```
------------------------------------------------------------
Μέσος συνολικός χρόνος εκτέλεσης (Total Time): 0.4273 δευτ.
Μέσος χρόνος υπολογισμού (Action Time): 0.4088 δευτ.
------------------------------------------------------------
```

Θέμα 2: Spark DataFrame API

ORIGIN_AIRPORT	DEST_AIRPORT	AVG_DELAY
DFW	JFK	23.3
JFK	LAS	22.8
MIA	JFK	20.5
MCO	ORD	19.818181818181817
MCO	LAX	19.142857142857142
BOS	SEA	18.857142857142858
DFW	ATL	18.625
BOS	PHX	18.181818181818183
JFK	MIA	18
DEN	SFO	17.857142857142858

Total time	0.30143809	0.3052673	0.342263	0.3828809	0.4364581
Action time	0.25328471	0.2668116	0.2940199	0.3280976	0.3757422

Code output:

```
------------------------------------------------------------
Μέσος συνολικός χρόνος εκτέλεσης (Total Time): 0.3435 δευτ.
Μέσος χρόνος υπολογισμού (Action Time): 0.2963 δευτ.
------------------------------------------------------------
```

Θέμα 3: Συγκριτική Ανάλυση RDD vs DataFrame

1)

Κριτήριο	RDD	DataFrame
Ευκολία υλοποίησης	Χαμηλότερη — χαμηλού επιπέδου API, χρειάζεται χειρισμός μετασχηματισμών με map/filter και parsing.	Υψηλότερη — δηλωτικό API (SQL-like), εύκολο φόρτωμα CSV/JSON με schema.
Χρόνος εκτέλεσης	Μεγαλύτερος κατά κανόνα — χωρίς λογικές βελτιστοποιήσεις, περισσότερα I/O και αντικειμενοποίηση.	Συνήθως μικρότερος — εκμετάλλευση Catalyst + Tungsten optimizations, vectorized execution.
Εκφραστικότητα κώδικα	Χαμηλή — λεπτομερής, απαιτεί περισσότερο boilerplate.	Υψηλή — σύντομη, περισσότερο δηλωτική και αναγνώσιμη.
Δυνατότητες βελτιστοποίησης (Catalyst)	Περιορισμένες — Catalyst δεν εφαρμόζεται σε RDDs (χειροκίνητες βελτιώσεις).	Πλήρεις — Catalyst optimizer, logical/physical plan, predicate pushdown, constant folding, join reordering κ.ά.

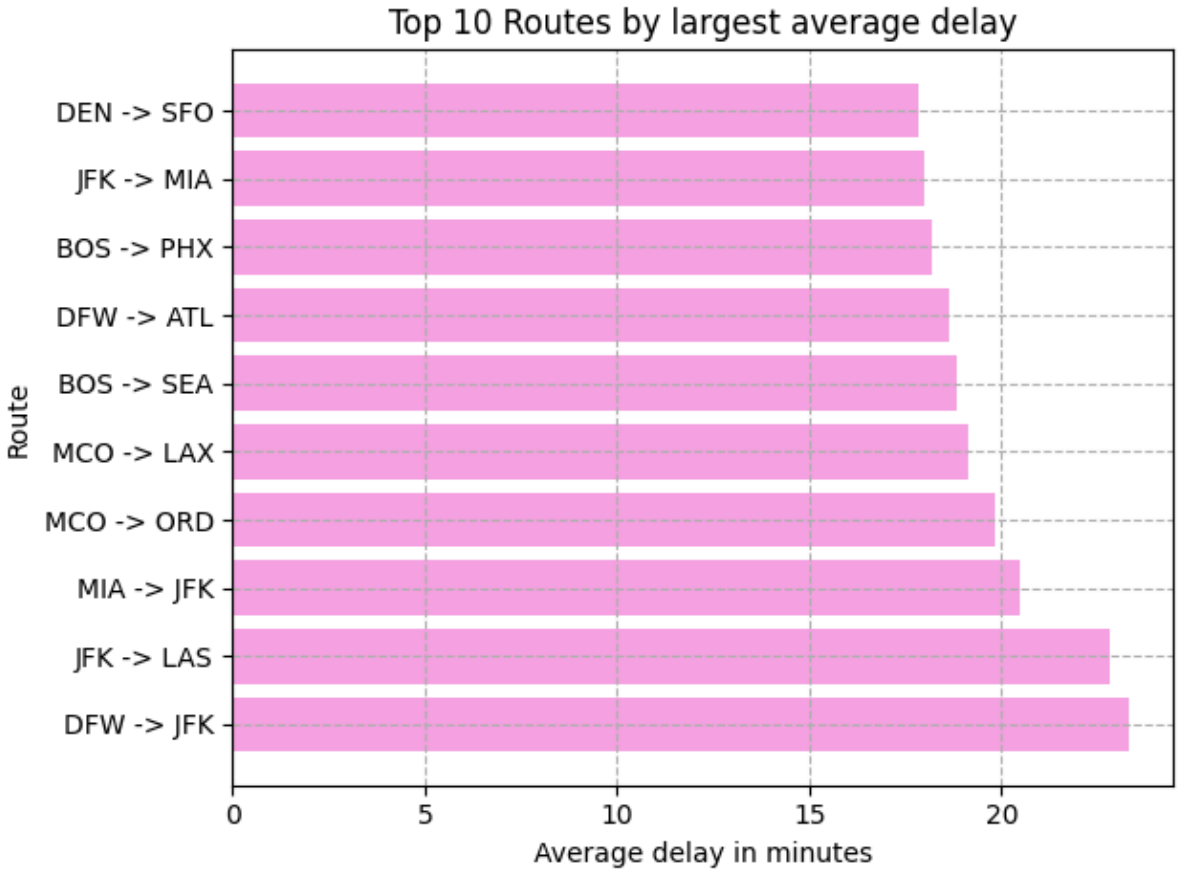
2) Τα DataFrames παρέχουν σαφώς ευκολία στην υλοποίηση και καλύτερη απόδοση για δομημένα/ημιδομημένα δεδομένα, λόγω του Catalyst optimizer και των βελτιώσεων εκτέλεσης (Tungsten, off-heap, vectorization). Η εκφραστικότητα και το declarative SQL-like API μειώνουν το χρόνο ανάπτυξης και τα λάθη. Τα RDDs είναι χρήσιμα όταν απαιτείται λεπτομερής έλεγχος, χειρισμός μη-δομημένων δεδομένων ή προσαρμοσμένων κατανεμημένων δομών, αλλά συνήθως οδηγούν σε υψηλότερο κόστος εκτέλεσης και

δυσκολότερη συντήρηση. Συστήνεται χρήση DataFrames για τις περισσότερες παραγωγικές εργασίες μεγάλης κλίμακας και μετατροπή σε RDD μόνο όταν μια αναγκαία χαμηλού-επιπέδου λειτουργία δεν υποστηρίζεται.

	Total Time (s)	Action Time (s)
RDD	0.4273	0.4088
DataFrame	0.3435	0.2963

3) Η καταλληλότερη επιλογή είναι το DataFrame, διότι εκμεταλλεύεται τον Catalyst optimizer για αυτόματες βελτιστοποιήσεις, προσφέρει βελτιωμένη εκτέλεση (Tungsten, διαχείριση μνήμης, vectorization), καλύτερη χρήση πόρων και απαιτεί λιγότερο κώδικα — όλα τα οποία οδηγούν σε μικρότερους χρόνους εκτέλεσης και καλύτερη επεκτασιμότητα. Τα RDD έχουν θέση μόνο σε ειδικές περιπτώσεις όπου το API των DataFrame δεν καλύπτει κάποια χαμηλού επιπέδου ανάγκη.

Θέμα 4: Οπτικοποίηση & Ανάλυση



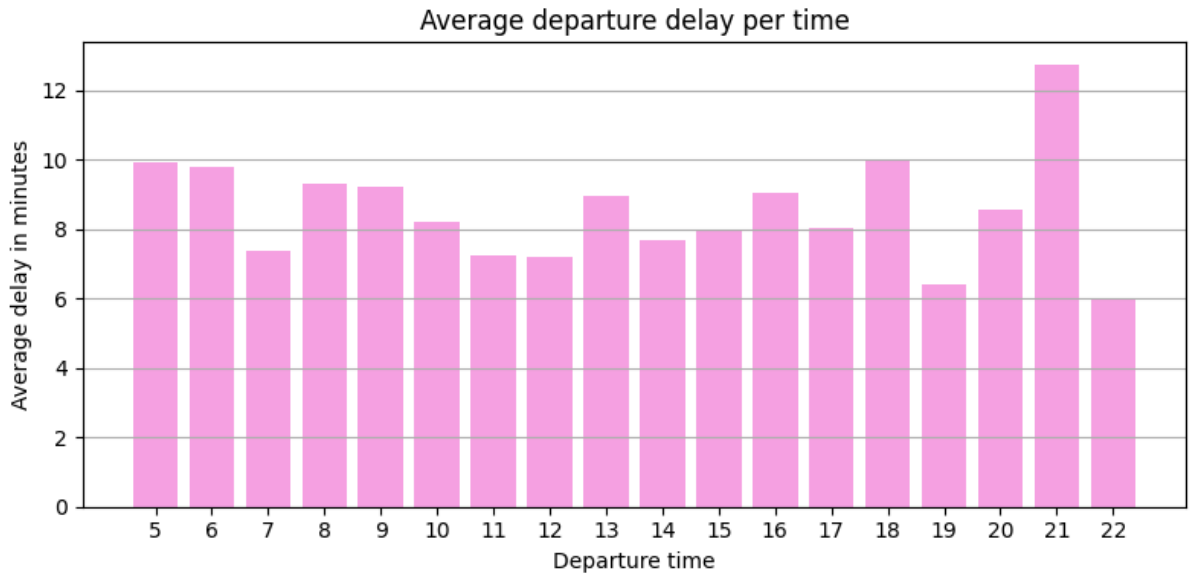
Το παραπάνω γράφημα, απεικονίζει τις 10 διαδρομές με τις υψηλότερες μέσες καθυστερήσεις σε λεπτά.

Βασικά Ευρήματα:

- Μεγαλύτερη καθυστέρηση: DFW -> JFK περίπου 23 λεπτά.
- Δεύτερη μεγαλύτερη: JFK -> LAS περίπου 22,5 λεπτά.
- Μικρότερη Καθυστέρηση: DEN -> SFO περίπου 18 λεπτά
- Αξιοσημείωτα Αεροδρόμια: Το αεροδρόμιο JFK εμπλέκεται σε τέσσερις από τις 10 κορυφαίες διαδρομές με τις χειρότερες καθυστερήσεις (DFW -> JFK, JFK -> LAS, MIA -> JFK και JFK -> MIA). Άλλα "αργά" αεροδρόμια που εμφανίζονται δύο φορές είναι το MCO, το BOS και το DFW.

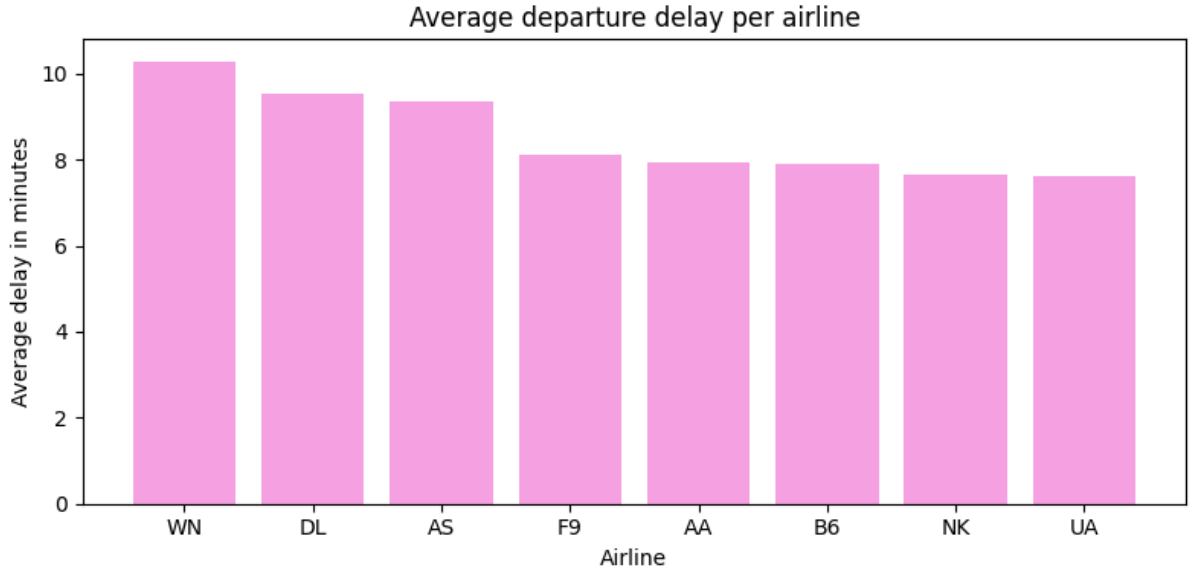
Θέμα 5: Επέκταση / Εμπλουτισμός Ανάλυσης

1. Μέση καθυστέρηση αναχώρησης ανά ώρα ημέρας



Σύμφωνα με τα αποτελέσματα της έρευνας όπως φαίνονται και παραπάνω,η χειρότερη ώρα για αναχώρηση είναι στις 21:00, φτάνοντας σχεδόν τα 13 λεπτά. Άλλες προβληματικές ώρες είναι νωρίς το πρωί, στις 5:00 και 6:00 (και οι δύο γύρω στα 10 λεπτά), καθώς και η βραδινή ώρα αιχμής στις 18:00 (επίσης 10 λεπτά). Οι ώρες με τις χαμηλότερες μέσες καθυστερήσεις είναι: 22:00 (περίπου 6 λεπτά) και 19:00 (περίπου 6,4 λεπτά). Επίσης, οι μεσημεριανές ώρες 11:00 και 12:00 είναι σχετικά καλές (περίπου 7,3 λεπτά).

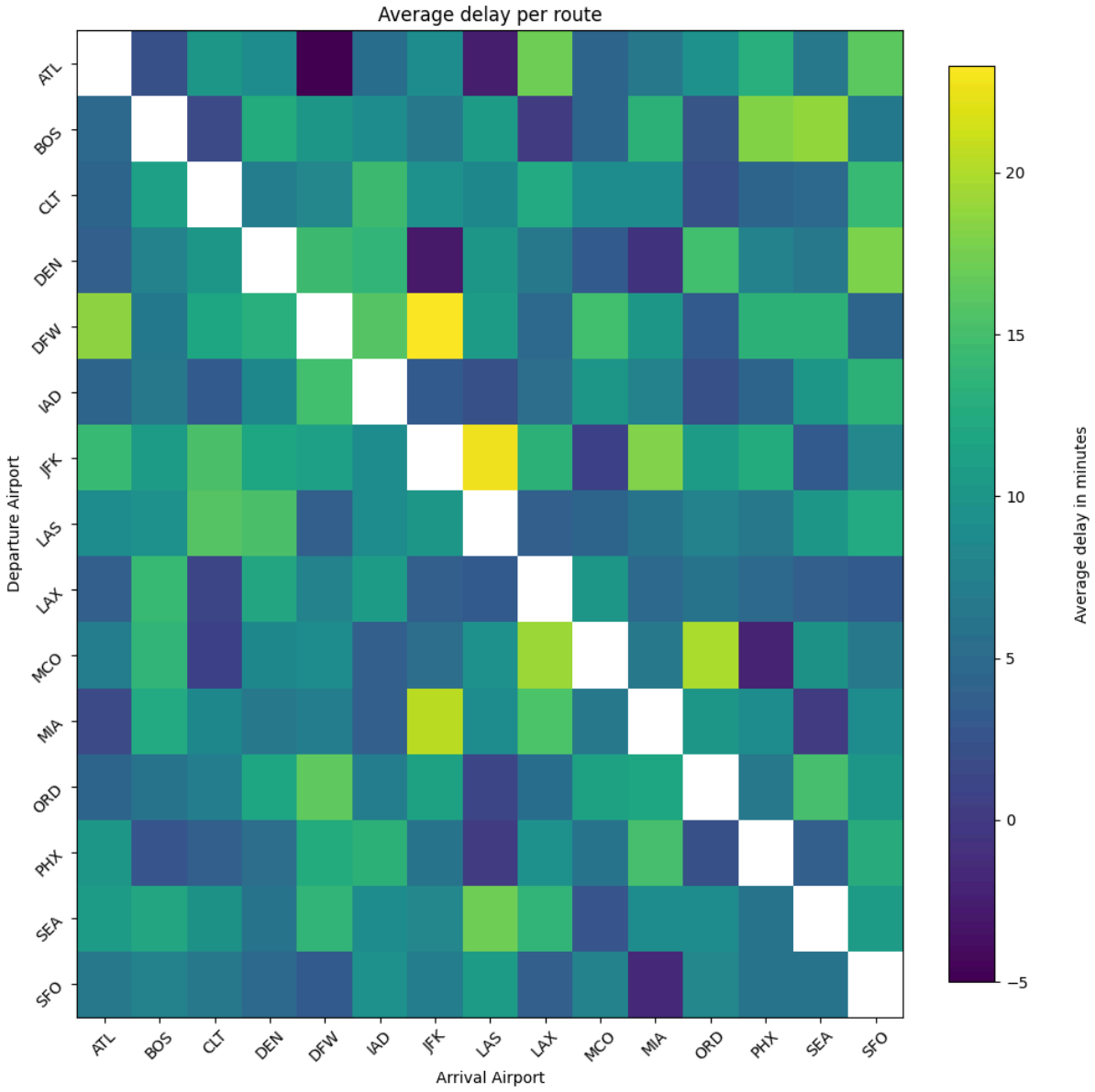
2. Μέση καθυστέρηση αναχώρησης ανά αεροπορική εταιρεία (Average departure delay per airline)



Η αεροπορική εταιρεία WN φαίνεται να έχει τη χειρότερη μέση καθυστέρηση αναχώρησης, με λίγο περισσότερα από 10 λεπτά. Οι DL (περίπου 9,6 λεπτά) και AS (περίπου 9,5 λεπτά) ακολουθούν αμέσως μετά με τις υψηλότερες καθυστερήσεις. Αντιθέτως, οι αεροπορικές εταιρείες NK και UA έχουν τις χαμηλότερες μέσες καθυστερήσεις αναχώρησης, με περίπου 7,6 λεπτά η καθεμία.

Γενικά, όλες οι αεροπορικές εταιρείες που εμφανίζονται έχουν, κατά μέσο όρο, καθυστέρηση στην αναχώρηση, με την διαφορά μεταξύ της καλύτερης και της χειρότερης σε αυτό το σύνολο να κυμαίνεται στα περίπου 2,5 λεπτά.

3. Μέση καθυστέρηση ανά διαδρομή (Average delay per route)



Ο θερμοχάρτης (heatmap) δείχνει την μέση καθυστέρηση πτήσης μεταξύ συγκεκριμένων αεροδρομίων. Διαδρομές με υψηλή καθυστέρηση (κίτρινο):

DFW -> JFK

JFK -> LAS

JFK -> SFO

Το αεροδρόμιο JFK φαίνεται να είναι ο πιο αργός κόμβος, καθώς εμφανίζει υψηλές καθυστερήσεις τόσο στις αναχωρήσεις όσο και στις αφίξεις