Διαχείριση Μεγάλων Δεδομένων
1η Προγραμματιστική Εργασία

Διδάσκουσα:
Π. Ραυτοπούλου

Παράδοση μέχρι: **Κυριακή 06/12/2020 ώρα 23.59**
Εξέταση: εβδομάδα 25-29/01/2021
(η ακριβής ημερομηνία θα ανακοινωθεί έγκαιρα)

ΣΗΜΑΝΤΙΚΕΣ ΣΗΜΕΙΩΣΕΙΣ:

1. Σε όλα τα αρχεία που θα παραδώσετε θα πρέπει **ΟΠΩΣΔΗΠΟΤΕ** να βάλετε τα ονόματα, τους Α.Μ., και τα username/email των μελών της ομάδας (ομάδες 2 ατόμων).
2. Αφού έχετε ολοκληρώσει την εργασία που θέλετε να παραδώσετε την υποβάλετε στο eclass στο υποσύστημα «Εργασίες φοιτητών». Προσοχή: μόνο 1 άτομο από την ομάδα χρειάζεται να παραδώσει την εργασία μέσω του e-class! Η υποβολή πρέπει να γίνει **ΠΡΙΝ** την καταληκτική ημερομηνία παράδοσης. Παραδίδετε όλα τα αρχεία που σας ζητούνται μαζί σε ένα συμπιεσμένο αρχείο (το οποίο θα φέρει τα ονόματα της ομάδας π.χ., RaftopoulouPapadopoulos.zip).
3. Περιπτώσεις αντιγραφής θα μηδενίζονται κι οι εμπλεκόμενοι δε θα έχουν δικαίωμα παράδοσης άλλων εργασιών.
4. Η ημερομηνία παράδοσης είναι αυστηρή, και η παράδοση γίνεται μόνο μέσω του eclass και όχι με email στη διδάσκουσα. Ασκήσεις που παραδίδονται μετά τη λήξη της προθεσμίας δε γίνονται δεκτές.

120 χρόνια Ολυμπιακής ιστορίας – Πώς εξελίχθηκαν οι Ολυμπιακοί Αγώνες στο πέρασμα των χρόνων

Ο πρώτος καταγεγραμμένος εορτασμός των Ολυμπιακών Αγώνων στην αρχαιότητα ήταν στην Ολυμπία το 776π.Χ. Αν κι είναι σχεδόν σίγουρο ότι αυτή δεν ήταν η πρώτη φορά που γίνονταν οι Ολυμπιακοί Αγώνες, από εκείνη τη χρονιά κι έπειτα άρχισαν να είναι γνωστοί σε ολόκληρη την αρχαία Ελλάδα, φτάνοντας στο απόγειο της δόξας τους κατά τον 5ο και 6ο αιώνα π.Χ. Οι Ολυμπιακοί Αγώνες εκείνης της εποχής είχαν θρησκευτικό χαρακτήρα και ήταν αφιερωμένοι στο θεό Δία, το τεράστιο άγαλμα του οποίου είχε στηθεί στην Ολυμπία. Αρχικά η κούρσα στο στάδιο της Ολυμπίας ήταν το μόνο αγώνισμα που διεξαγόταν, στη συνέχεια όμως τα αγωνίσματα των Ολυμπιακών Αγώνων έγιναν είκοσι και ο εορτασμός προς τιμή του θεού διαρκούσε αρκετές ημέρες. Οι νικητές των αγώνων, οι οποίοι λάμβαναν ως έπαθλο τον κότινο, δηλαδή ένα στεφάνι από κλαδιά ελιάς, θαυμάζονταν και γίνονταν αθάνατοι μέσα από ποιήματα και αγάλματα.

Οι Ολυμπιακοί Αγώνες της αρχαιότητας έχασαν την αίγλη τους όταν οι Ρωμαίοι κατέλαβαν την Ελλάδα κι ο Χριστιανισμός έγινε η επίσημη θρησκεία της Ρωμαϊκής αυτοκρατορίας, οπότε και το 393 μ.Χ. ο αυτοκράτορας Θεοδόσιος απαγόρευσε την διεξαγωγή τους. Με αυτό τον τρόπο τελείωσε μια περίοδος χιλίων χρόνων.

Οι πρώτοι σύγχρονοι Ολυμπιακοί Αγώνες διεξήχθησαν το 1896 στην Αθήνα, μετά το τέλος ενός συνεδρίου στο Πανεπιστήμιο της Σορβόννης στο Παρίσι, οπότε και γεννήθηκε η Διεθνής Ολυμπιακή Επιτροπή (ΔΟΕ) με πρώτο πρόεδρο τον Μακεδόνα Δημήτριο Βικέλα, γενικό γραμματέα τον βαρόνο Πιέρ ντε Κουμπερντέν και μέλη προσωπικότητες από διάφορα κράτη.

Οι πρώτοι σύγχρονοι Ολυμπιακοί Αγώνες του 1896 έγιναν με μεγάλη επιτυχία· αν κι οι αθλητές που πήραν μέρος δεν ξεπερνούσαν τους 250, ήταν η μεγαλύτερη αθλητική διοργάνωση που είχε γίνει ποτέ ως τότε. Μετά όμως από την αρχική επιτυχία, οι Ολυμπιακοί Αγώνες, αν και διεξάγονταν κάθε 4 χρόνια, είχαν σοβαρά προβλήματα, οπότε και ακολούθησε μια αμφιλεγόμενη περίοδος μέχρι που τελικά το 1904 με το 80% των συμμετοχών να είναι Αμερικάνοι αθλητές σηματοδοτείται η αρχή της ανάπτυξης των αγώνων σε δημοσιότητα

και μέγεθος. Το 1925 η ΔΟΕ αποφασίζει μάλιστα να δημιουργήσει ξεχωριστή διοργάνωση για χειμερινά αθλήματα και η διοργάνωση το χειμώνα του 1924 χαρακτηρίστηκε ως οι πρώτοι Χειμερινοί Ολυμπιακοί Αγώνες. Οι χειμερινοί και οι θερινοί Ολυμπιακοί Αγώνες διεξάγονταν κάθε 4 χρόνια το ίδιο έτος μέχρι και το 1992. Έπειτα, αποφασίστηκε οι χειμερινοί αγώνες να ακολουθούν ένα δικό τους ξεχωριστό τετραετή κύκλο ξεκινώντας από το 1994. Οπότε οι θερινοί αγώνες διεξήχθησαν το καλοκαίρι του 1996, έπειτα οι χειμερινοί αγώνες το χειμώνα του 1998, και ούτω καθεξής.

Σκοπός της εργασίας

Σκοπός αυτής της εργασίας είναι να μελετήσετε ένα ιστορικό σύνολο δεδομένων που αφορά στους σύγχρονους Ολυμπιακούς Αγώνες **από την Αθήνα του 1896 ως το Ρίο του 2016** και να κάνετε ερωτήματα σχετικά με το πώς εξελίχθηκαν οι Ολυμπιακοί Αγώνες με την πάροδο του χρόνου, συμπεριλαμβανομένων ερωτήσεων για τους κορυφαίους αθλητές, τη συμμετοχή και την απόδοση των γυναικών, των διαφορετικών αθλημάτων και εκδηλώσεων, κοκ. Τα δεδομένα προέρχονται από το www.sports-reference.com, κι είναι διαθέσιμα (μετά από κατάλληλη επεξεργασία) μέσω του Kaggle[1].

Εκτός από τις διαφάνειες του μαθήματος, στις οποίες συζητήσαμε για το Hadoop και για το MapReduce, επιπλέον πληροφορίες για την ακριβή λειτουργία τους μπορείτε να βρείτε στα Έγγραφα και στους Συνδέσμους (όλα διαθέσιμα στο eclass του μαθήματος):

[1] Έγγραφα/Hadoop & MapReduce
[2] Σύνδεσμοι/Hadoop & MapReduce
[3] Σύνδεσμοι/Apache Pig

Η εργασία θα εκπονηθεί από ομάδες των **2 ατόμων**, θα υλοποιηθεί σε **Java**, και μπορεί να γίνει σε οποιοδήποτε λειτουργικό σύστημα (συστήνεται το GNU/Linux).

Για τις ανάγκες της εργασίας, θα κατεβάσετε και θα εγκαταστήσετε το Hadoop στο μηχάνημά σας κατά προτίμηση σε **Single-Node Local (Standalone) Mode**. Το mode αυτό προορίζεται για debugging και έχει ευκολότερη διαδικασία εγκατάστασης, οπότε το συστήνω μιας και δεν θα κάνετε την υλοποίηση σε κάποιον cluster αλλά στο μηχάνημά σας.

Για την εγκατάσταση ακολουθήστε τις αναλυτικές οδηγίες από εδώ:
https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html
Φυσικά ανάλογα με το setup του μηχανήματός σας μπορεί να χρειαστεί ελαφρώς διαφορετική διαδικασία.

Λειτουργικότητα του συστήματος

Σας δίνεται ένα αρχείο που αποτελείται από περισσότερες από 271.000 εγγραφές, όπου κάθε εγγραφή αντιστοιχεί σε έναν αθλητή (ID) που συμμετείχε σε κάποιον Ολυμπιακό Αγώνα (Games) με κάποιο συγκεκριμένο αγώνισμα (Event). Για κάθε εγγραφή έχουν αποθηκευτεί τα εξής 15 χαρακτηριστικά (attributes):

1. ID - Unique number for each athlete
2. Name - Athlete's name
3. Sex - M or F
4. Age – Integer
5. Height - In centimeters
6. Weight - In kilograms
7. Team - Team name
8. NOC - National Olympic Committee 3-letter code
9. Games - Year and season
10. Year – Integer
11. Season - Summer or Winter
12. City - Host city
13. Sport – Sport
14. Event – Event
15. Medal - Gold, Silver, Bronze, or NA

[1] https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results

Δείτε παρακάτω ένα παράδειγμα 4 εγγραφών:

1, A Dijiang, M, 24, 180, 80, China, CHN, 1992 Summer, 1992, Summer, Barcelona, Basketball, Basketball Men's Basketball, NA
2, A Lamusi, M, 23, 170, 60, China, CHN, 2012 Summer, 2012, Summer, London, Judo, Judo Men's Extra-Lightweight, NA
3, Gunnar Nielsen Aaby, M, 24, NA, NA, Denmark, DEN, 1920 Summer, 1920, Summer, Antwerpen, Football, Football Men's Football, NA
4, Edgar Lindenau Aabye, M, 34, NA, NA, Denmark/Sweden, DEN, 1900 Summer, 1900, Summer, Paris, Tug-Of-War, Tug-Of-War Men's Tug-Of-War, Gold

Για λόγους παρουσίασης, στο παραπάνω παράδειγμα έχουμε εισάγει κόμματα ανάμεσα στα χαρακτηριστικά κάθε εγγραφής.

Όλα τα δεδομένα είναι διαθέσιμα στο αρχείο athlete_events.csv που περιέχει 271.117 εγγραφές. Μελετήστε τα δεδομένα και τα χαρακτηριστικά των εγγραφών που εμφανίζονται σε αυτό το αρχείο, αν και για όσες εγγραφές το θεωρείτε σκόπιμο καθαρίστε τα δεδομένα σας (όχι με το χέρι, με τη χρήση ενός script!). Σημειώστε ότι ίσως χρειαστεί να διαχειριστείτε και τις τιμές NA όπου εμφανίζονται, π.χ. θεωρώντας τις ως μηδενικές. Έπειτα, χρησιμοποιείστε αυτό το αρχείο για να τρέξτε τον κώδικά σας και καταγράψτε σε αντίστοιχα αρχεία εξόδου τα αποτελέσματα που θα πάρετε για τα τρία παρακάτω ερωτήματα.

1. Μελέτη της επίδοσης κάθε αθλητή [15]

(α) Γράψτε τον **ψευδοκώδικα** για τη λειτουργία Map και για τη λειτουργία Reduce και βρείτε για κάθε αθλητή (ID, Name) το πλήθος των χρυσών μεταλλίων που έχει κερδίσει. Δώστε επίσης, ένα **παράδειγμα** και μια **σχηματική εκτέλεση** που να εξηγεί τη λογική πίσω από το ψευδοκώδικά σας.

(β) Υλοποιήστε στο Hadoop το σχεδιασμό που κάνατε παραπάνω. Το αρχείο εξόδου θα πρέπει να περιέχει τόσες γραμμές όσες και οι διαφορετικοί αθλητές, οι αθλητές να είναι ταξινομημένοι σε αύξουσα σειρά ως προς τον μοναδικό κωδικό τους, και σε κάθε γραμμή να εμφανίζονται (i) ο κωδικός του αθλητή (ID), (ii) το όνομά του (Name), (iii) το φύλλο του (Sex), και (iv) το πλήθος των χρυσών μεταλλίων που έχει κατακτήσει.

Δείτε παρακάτω ένα παράδειγμα:

4 Edgar Lindenau Aabye M 1
17 Paavo Johannes Aaltonen M 3
20 Kjetil Andr Aamodt M 4
21 Ragnhild Margrethe Aamodt F 1
...

2. Ανάδειξη των κορυφαίων αθλητών όλων των εποχών [35]

(α) Γράψτε τον **ψευδοκώδικα** για τη λειτουργία Map και για τη λειτουργία Reduce και βρείτε το πλήθος των χρυσών μεταλλίων, το πλήθος των ασημένιων μεταλλίων, το πλήθος των χάλκινων μεταλλίων, και τα συνολικά μετάλλια που έχει κερδίσει ένας αθλητής (Name) σε μια και μόνο διοργάνωση (Games). Δώστε ένα **παράδειγμα** και μια **σχηματική εκτέλεση** που να εξηγεί τη λογική πίσω από το ψευδοκώδικά σας.

(β) Υλοποιήστε στο Hadoop το σχεδιασμό που κάνατε παραπάνω. Το αρχείο εξόδου θα πρέπει να περιέχει τις δέκα κορυφαίες επιδόσεις όλων των εποχών, οι αθλητές να είναι ταξινομημένοι σε φθίνουσα σειρά ως προς το πλήθος των χρυσών μεταλλίων σε μία διοργάνωση και σε κάθε γραμμή να εμφανίζονται (i) η σειρά κατάταξης, (ii) το όνομα του αθλητή (Name), (iii) το φύλο του (Sex), (iv) η ηλικία του (Age), (v) το όνομα της ομάδας με την οποία συμμετείχε (Team), (vi) το άθλημα (Sport), (vii) η διοργάνωση (Games), και (viii) τα μετάλλια ως εξής: πλήθος χρυσών, πλήθος ασημένιων, πλήθος χάλκινων, συνολικά μετάλλια. Σε περίπτωση που υπάρχουν παραπάνω από ένας αθλητές με το ίδιο αριθμό χρυσών μεταλλίων σε μια διοργάνωση, τότε πιο ψηλά στην κατάταξη θεωρείστε τον αθλητή με τα περισσότερα συνολικά μετάλλια στη διοργάνωση. Σε περίπτωση που υπάρχουν παραπάνω από ένας αθλητές με το ίδιο αριθμό χρυσών και συνολικών μεταλλίων σε μια διοργάνωση, θεωρείστε ότι πρόκειται για την ίδια επίδοση και εμφανίστε όλους τους αθλητές ταξινομημένους σε αύξουσα σειρά (A-Z) ως προς το όνομά τους.

Δείτε παρακάτω ένα παράδειγμα:

1 Michael Fred Phelps, II M 23 United States Swimming 2008 Summer 8 0 0 8
2 Mark Andrew Spitz M 22 United States Swimming 1972 Summer 7 0 0 7
3 Michael Fred Phelps, II M 19 United States Swimming 2004 Summer 6 0 2 8
4 Kristin Otto F 22 East Germany Swimming 1972 Summer 6 0 0 6
4 Vitaly Venediktovich Shcherbo M 20 Unified Team Gymnastics 1992 Summer 6 0 0 6
...

3. Μελέτη των γυναικείων συμμετοχών ανά ομάδα στους Ολυμπιακούς αγώνες [50]

(α) Γράψτε τον **ψευδοκώδικα** για τη λειτουργία Map και για τη λειτουργία Reduce για να μελετήσετε τη συμμετοχή των γυναικών στους Ολυμπιακούς Αγώνες μέσα σε αυτά τα 120 χρόνια· θα πρέπει να βρείτε για κάθε διοργάνωση (Games) τις τρεις (3) πρώτες ομάδες (Team) από άποψη πλήθους γυναικείων συμμετοχών, καθώς και το πρώτης επιλογής άθλημα (Sport) των γυναικών αθλητριών σε κάθε μια από αυτές τις ομάδες. Σημειώστε ότι σε κάθε διοργάνωση προσμετράμε κάθε αθλήτρια (ID) μόνο μια φορά άσχετα με το πλήθος των αγωνισμάτων (Event) στα οποία έχει αυτή συμμετάσχει στη συγκεκριμένη διοργάνωση. Δώστε επίσης, ένα **παράδειγμα** και μια **σχηματική εκτέλεση** που να εξηγεί τη λογική πίσω από το ψευδοκώδικά σας.

(β) Υλοποιήστε στο Hadoop το σχεδιασμό που κάνατε παραπάνω. Το αρχείο εξόδου θα πρέπει να περιέχει τρεις (3) γραμμές[2] για κάθε διαφορετική διοργάνωση (Games), οι διοργανώσεις να είναι ταξινομημένες σε αύξουσα σειρά ως προς το χρόνο διεξαγωγής τους, οι ομάδες να είναι ταξινομημένες σε φθίνουσα σειρά ως προς το πλήθος γυναικείων συμμετοχών, και σε κάθε γραμμή να εμφανίζονται (i) η διοργάνωση (Games), (ii) η ομάδα (Team), (iii) ο μοναδικός κωδικός της ομάδας (NOC)[3], (iv) το πλήθος των γυναικείων συμμετοχών, και (v) το άθλημα προτίμησής τους (Sport). Αν μια διοργάνωση δεν έχει καθόλου γυναικείες συμμετοχές τότε δεν εμφανίζεται στο αρχείο εξόδου. Σε περίπτωση που ομάδες ισοψηφούν σε πλήθος γυναικείων συμμετοχών σε μια διοργάνωση εμφανίζονται όλες στο αρχείο εξόδου ταξινομημένες σε αύξουσα σειρά (A-Z). Δείτε παρακάτω ένα παράδειγμα:

```
1900 Summer France FRA 12 Croquet
1900 Summer United States USA 8 Golf Women's Individual
1900 Summer Great Britain GBR 1 Tennis
1900 Summer Italy ITA 1 Equestrianism
1900 Summer Lerina SUI 1 Sailing
1904 Summer United States USA 10 Archery
...
```

Bonus υλοποίηση 20%

Γράψτε τα ίδια ερωτήματα σε **Pig Latin**. Χρησιμοποιήστε Pig για να τα εκτελέσετε στο Hadoop και να συγκρίνετε τα αποτελέσματα με αυτά που πήρατε προηγουμένως.

Παραδοτέα και βαθμολόγηση

Πλήρης θεωρείται η εργασία η οποία **υλοποιεί σωστά τις βασικές απαιτήσεις** που περιγράφονται παραπάνω. Ασκήσεις που υλοποιούν μόνο ένα μέρος των βασικών απαιτήσεων λαμβάνουν και αντίστοιχο μέρος του βαθμού.

Τα παραδοτέα της εργασίας είναι:

- τα αρχεία πηγαίου κώδικα (μαζί με τυχόν scripts που χρησιμοποιήσατε)
- τα εκτελέσιμα αρχεία
- τα αρχεία με τα αποτελέσματα για κάθε ένα από τα παραπάνω ερωτήματα
- μία γραπτή αναφορά που θα περιέχει:
    - τις ενέργειες σας μαζί με κατάλληλη αιτιολόγηση σε σχέση με την (προ-)επεξεργασία των δεδομένων (π.χ., κάνατε κάποιου είδους καθαρισμό, πώς διαχειριστήκατε τις ομάδες με τον ίδιο κωδικό 3-γραμμάτων, κοκ.)
    - τον ψευδοκώδικα, ένα παράδειγμα και μια σχηματική εκτέλεση για κάθε ένα από τα παραπάνω ερωτήματα,
    - οδηγίες για την εκτέλεση του προγράμματος,
    - λεπτομέρειες της υλοποίησης που αξίζει να σημειωθούν,
    - τα αρχεία των αποτελεσμάτων για κάθε ένα από τα ερωτήματα
    - γραφικές απεικονίσεις (εσείς θα αποφασίσετε ποιες) των αποτελεσμάτων, και
    - τα σχόλιά σας σε σχέση με τα αποτελέσματα (π.χ., τι αναμένατε και τι πήρατε ως αποτέλεσμα, κα.) για κάθε ένα από τα ερωτήματα.

Καλή επιτυχία!

[2] Υπάρχει περίπτωση να είναι λιγότερες ή περισσότερες, δείτε παρακάτω γιατί.

[3] Κάποιες ομάδες που εμφανίζονται με διαφορετικό όνομα στο αρχείο δεδομένων έχουν τον ίδιο κωδικό 3-γραμμάτων· σκεφτείτε πώς θα διαχειριστείτε τέτοιες περιπτώσεις.