# Ανάθεση Big Data Analytics (CN7031) για το 2024-2025

|  |  |
| --- | --- |
| ¶**Κωδικός μαθήματος ¶** | DS7031¶ |
| **Τίτλος μαθήματος ¶** | Big Data Analytics ¶ |
| **Υπεύθυνος μαθήματος ¶** | Δρ. Ιωάννης Γουναρίδης ¶ |
| **Καθηγητής ανάθεσης ¶** | Δρ. Ιωάννης Γουναρίδης ¶ |
| **Τίτλος ανάθεσης ¶** | Coursework ¶ |
| **Αριθμός ανάθεσης¶** |  |
| **Βαρύτητα ¶** | 100%¶ |
| **Ημερομηνία εξέτασης ¶** | 13η Εβδομάδα (Ιανουάριος 2025) |
| **Ώρα** |  |
| **Διάρκεια** |  |

Τίτλος εργασίας: Ανάλυσης μεγάλων δεδομένων που βασίζονται σε Python

Στάθμιση: 100%

Ημερομηνία φυλλαδίου: Εβδομάδα 6

Ημερομηνία υποβολής: Μέσα Δεκεμβρίου (προσαρμογή με βάση το ακαδημαϊκό σας ημερολόγιο)

## Επισκόπηση ανάθεσης

Η εργασία αυτή θα ολοκληρωθεί σε ομάδες των 3 ή 4 φοιτητών και περιλαμβάνει:

1. Αναφορά ανάλυσης μεγάλων δεδομένων χρησιμοποιώντας Python.

2. Ομαδική παρουσίαση.

## Λεπτομέρειες Ανάθεσης

1. Αναφορά ανάλυσης μεγάλων δεδομένων (60%)

- Επεξεργασία και ανάλυση δεδομένων:

- Φορτώστε και επεξεργαστείτε ένα αρκετά μεγάλο σύνολο δεδομένων καταγραφής ιστού χρησιμοποιώντας την Python.

- Μετατρέψτε το σύνολο δεδομένων σε κατάλληλη δομημένη μορφή χρησιμοποιώντας pandas και PySpark.

- Προηγμένη ανάλυση δεδομένων:

- Χρησιμοποιήστε Python και PySpark για να εκτελέσετε σύνθετη ανάλυση δεδομένων. Αυτό θα πρέπει να περιλαμβάνει σύνθετα ερωτήματα που χρησιμοποιούν τόσο PySpark SQL όσο και εγγενείς μετασχηματισμούς/ενέργειες PySpark σε RDD.

- Εργασία 1 (PySpark SQL): Κάθε μαθητής θα γράψει δύο προχωρημένα ερωτήματα PySpark SQL στο σύνολο δεδομένων. Αυτά θα πρέπει να επιδεικνύουν την ικανότητα χειρισμού της ανάλυσης δεδομένων πραγματικού κόσμου με το PySpark, εξάγοντας σημαντικές πληροφορίες.

- Δραστηριότητα 2 (PySpark RDD): Κάθε μαθητής θα πρέπει να αναπτύξει τρία ερωτήματα βασισμένα σε RDD. Τουλάχιστον δύο από αυτά θα πρέπει να προωθηθούν, αναδεικνύοντας την ικανότητα αξιοποίησης του πλήρους δυναμικού των RDD στο Spark για σύνθετους μετασχηματισμούς δεδομένων.

- Οπτικοποίηση δεδομένων:

- Δημιουργήστε απεικονίσεις χρησιμοποιώντας matplotlib ή seaborn για να απεικονίσετε βασικά ευρήματα από την ανάλυση. Κάθε μαθητής πρέπει να συμπεριλάβει τουλάχιστον δύο απεικονίσεις στην αναφορά.

- Εκτιμήσεις LSEP:

- Συζητήστε τις νομικές, κοινωνικές, ηθικές και επαγγελματικές (LSEP) επιπτώσεις των μεγάλων δεδομένων, όπως το απόρρητο των δεδομένων, η μεροληψία και η προστασία των δεδομένων. Κάθε φοιτητής θα πρέπει να συνεισφέρει σε αυτή την ενότητα, εξετάζοντας λεπτομερώς έναν από τους παράγοντες LSEP.

- Μορφή αναφοράς:

- Η αναφορά θα πρέπει να υποβληθεί ως Jupyter Notebook που μετατρέπεται σε HTML. Χρησιμοποιήστε τα κελιά εμπλουτισμένου κειμένου της Python για να τεκμηριώσετε τη διαδικασία ανάλυσης, να παρουσιάσετε αποτελέσματα και να συζητήσετε τις επιπτώσεις.

2. Παρουσίαση (40%)

-Περιεχόμενο:

- Κάθε ομάδα θα παρουσιάσει την ανάλυσή της, εστιάζοντας στην επεξεργασία δεδομένων, τα ερωτήματα SQL και RDD και τις εκτιμήσεις LSEP. Η παρουσίαση θα πρέπει να επισημαίνει μεμονωμένες συνεισφορές, ειδικά τα προηγμένα ερωτήματα και απεικονίσεις που δημιουργούνται από κάθε μαθητή.

-Μορφή:

- Η παρουσίαση θα πραγματοποιηθεί μέσω του Microsoft Teams και κάθε μέλος πρέπει να συνεισφέρει. Μια αποτυχία παρουσίασης θα οδηγήσει σε αποτυχία της λειτουργικής μονάδας.

- Κριτήρια αξιολόγησης:

- Σαφήνεια και οργάνωση της παρουσίασης.

- Κατανόηση της ανάλυσης μεγάλων δεδομένων χρησιμοποιώντας Python και Spark.

- Ικανότητα απάντησης σε τεχνικές ερωτήσεις που σχετίζονται με την επεξεργασία και ανάλυση δεδομένων.

- Ποιότητα οπτικοποιήσεων και συνολική αφήγηση.

## Οδηγίες υποβολής

- Υποβάλετε ένα μόνο αρχείο HTML (που εξάγεται από το Jupyter Notebook) ανά ομάδα στο Moodle.

- Ονομάστε το αρχείο ως "Your\_Group\_ID\_CN7031.html".

- Ισχύουν κανόνες λογοκλοπής, με μηδενικό βαθμό και περαιτέρω πειθαρχικά μέτρα για τυχόν εντοπισμένες περιπτώσεις.