ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΤΜΗΜΑΤΟΣ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

# Ο ΡΟΛΟΣ ΤΩΝ BIG DATA ΣΤΗΝ ΥΓΕΙΟΝΟΜΙΚΗ ΠΕΡΙΘΑΛΨΗ ΜΕ ΕΜΦΑΣΗ ΣΤΗΝ ΟΓΚΟΛΟΓΙΑ

Διπλωματική Εργασία

της

Μαργαρίτη Μαρίας

Θεσσαλονίκη, Ιούνιος 2023

# Ο ΡΟΛΟΣ ΤΩΝ BIG DATA ΣΤΗΝ ΥΓΕΙΟΝΟΜΙΚΗ ΠΕΡΙΘΑΛΨΗ ΜΕ ΕΜΦΑΣΗ ΣΤΗΝ ΟΓΚΟΛΟΓΙΑ

Μαργαρίτη Μαρία

Πτυχίο Μαθηματικών, Πανεπιστήμιο Ιωαννίνων, 2021

Διπλωματική Εργασία

υποβαλλόμενη για τη μερική εκπλήρωση των απαιτήσεων του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΤΙΤΛΟΥ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΠΛΗΡΟΦΟΡΙΚΗ

Επιβλέπων Καθηγητής
Ψάννης Κωνσταντίνος

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 21/06/2023

| Ψάννης Κωνσταντίνος | Μαντάς Μιχαήλ | Ξυνόγαλος Στυλιανός |
|---|---|---|
| ..... | ..... | ..... |

Μαργαρίτη Μαρία

.....

# Περίληψη

Η παρούσα διπλωματική εργασία έχει ως στόχο τη διερεύνηση του ρόλου των Big Data και τις προοπτικές που προσφέρονται στον τομέα της υγείας και ιδιαίτερα στην ογκολογία. Αναμφίβολα ο όγκος των δεδομένων που παράγεται καθημερινά στον τομέα είναι τεράστιος και μη διαχειρίσιμος με παραδοσιακές μεθόδους. Αρχικά οι έγκυρες πηγές για τη συλλογή δεδομένων και στη συνέχεια η χρήση εξειδικευμένων τεχνικών για την ανάλυση τους αποτελούν βασικό κομμάτι της εκάστοτε έρευνας. Τα αποτελέσματα που εξάγονται προσφέρουν σημαντικά στην διαχείριση της υγειονομικής περίθαλψης, στη λήψη αποφάσεων, σε ακριβείς προβλέψεις, στην κατανόηση του μηχανισμού κάθε ασθένειας αλλά και στην μείωση του κόστους φροντίδας των ασθενών. Επιπλέον η ανάλυση ογκολογικών δεδομένων παρουσιάζει ιδιαίτερες προκλήσεις λόγω της μεγάλης ποικιλομορφίας τους με το 90% να αφορά σε μη δομημένα. Παράλληλα όμως καλείται να αντιμετωπίσει πρακτικές αλλά και ηθικές προκλήσεις, μεταξύ άλλων την αποθήκευση του τεράστιου όγκου των δεδομένων, την πρόσβαση σε αυτά, την οπτικοποίησή τους, την ασφάλεια και το απόρρητο των ασθενών. Η αναλυτική των μεγάλων δεδομένων αποτελεί ένα από τα σημαντικότερα όπλα των επιστημόνων για την αντιμετώπιση ασθενειών. Η ανάπτυξη προηγμένων αλγορίθμων είναι απαραίτητη ώστε να καταφέρουμε να μετατρέψουμε τον τεράστιο όγκο δεδομένων που λαμβάνουμε σε χρήσιμη πληροφορία. Στις μέρες μας οι πλατφόρμες που κυριαρχούν για την ανάλυση και την επεξεργασία των δεδομένων μεγάλης κλίμακας είναι το Apache Hadoop και το Apache Spark. Στη συγκεκριμένη εργασία έχει γίνει ανάλυση του πρώτου.

Σε ό,τι αφορά τη δομή της εργασίας, αρχικά γίνεται μια εκτενής βιβλιογραφική ανασκόπηση και μελέτη αντίστοιχων ερευνών, πραγματοποιώντας αναζήτηση σχετικών επιστημονικών άρθρων σε ηλεκτρονικές βάσεις δεδομένων. Στη συνέχεια γίνεται αναζήτηση και εύρεση ογκολογικών δεδομένων τα οποία αφορούν στον καρκίνο του πνεύμονα. Τα δεδομένα αυτά χρησιμοποιήθηκαν για εξαγωγή συμπερασμάτων με τη βοήθεια του Apache Hadoop.

**Λέξεις Κλειδιά:** Μεγάλα Δεδομένα, Αναλυτική Μεγάλων Δεδομένων, Υγειονομική Περίθαλψη, Ογκολογία, Apache Hadoop

# Abstract

This thesis aims to investigate the role of Big Data and the perspectives offered in the health sector and especially in oncology. Undoubtedly, the volume of data generated daily in the sector is massive and unmanageable with traditional methods. First of all, the valid sources for data collection, followed by the use of specialized techniques for data analysis, are essential part of any research. The extracted results contribute significantly to health care management, decision making, accurate predictions, understanding the mechanism of each disease and also reducing the cost of patient's medical care. Furthermore, the analysis of oncological data is featuring special challenges due to their large variety, with the 90% being unstructured data. However, at the same time, it is has to face practical and ethical challenges such as the storage of the enormous amount of data, the access to them, their visualization, the security and privacy of patients. The Big Data Analysis is one of the most important scientist's weapons to deal with diseases. The development of advanced algorithms is necessary to be able to transform the massive amount of data we receive into useful information. Nowadays the dominant platforms for large-scale data analysis and processing are Apache Hadoop and Apache Spark. In this project, has been analysed the first one platform.

As far as the structure of this paper is concerned, initially, was accomplished an extensive literature review and study of corresponding researches, by searching for relevant scientific articles in electronic databases. Then was conducted research for oncological data related to lung cancer. This dataset was used for conclusions with the support of Hadoop.



**Keywords:** Bid Data, Big Data Analytics, Healthcare, Oncology, Apache Hadoop

# Ευχαριστίες

Με την παρούσα διπλωματική εργασία ολοκληρώνεται ο κύκλος των μεταπτυχιακών σπουδών μου στην Επιστήμη και Τεχνολογία Η/Υ του Πανεπιστημίου Μακεδονίας. Θα ήθελα λοιπόν να ευχαριστήσω τους ανθρώπους που στάθηκαν υποστηρικτές μου όχι μόνο πρακτικά αλλά και ηθικά.

Η ολοκλήρωση της εργασίας θα ήταν αδύνατη χωρίς την πολύτιμη βοήθεια του επιβλέποντα καθηγητή μου κ. Ψάννη Κωνσταντίνου. Τον ευχαριστώ για την άψογη συνεργασία, την καθοδήγηση και την βοήθεια του σε όλη τη διάρκεια.

Τέλος ευχαριστώ την οικογένεια μου για την στήριξη τους σε κάθε μου προσπάθεια και στους οποίους οφείλω τις σπουδές μου, ως σήμερα.

# Περιεχόμενα

# Κατάλογος Εικόνων

# Κατάλογος Πινάκων

# Συμβολισμοί

EHR: Ηλεκτρονικά Μητρώα Υγείας

EMR: Ηλεκτρονικός Ιατρικός Φάκελος

# 1 Εισαγωγή

## 1.1 Σημαντικότητα του θέματος

Η υγεία αποτελεί αναμφίβολα το σημαντικότερο αγαθό του ανθρώπου. Στον τομέα της υγειονομικής περίθαλψης ο όγκος των δεδομένων που παράγεται καθημερινά είναι τεράστιος και προέρχεται από διάφορες πηγές όπως φαρμακοβιομηχανίες, νοσοκομεία αλλά και από τους ίδιους τους ασθενείς. Τα δεδομένα αυτά όμως για να αξιοποιηθούν είναι αναγκαίο να μετατραπούν σε χρήσιμη πληροφορία. Στόχος είναι η κατανόηση των ασθενειών, η αποτελεσματική θεραπεία, η πρόβλεψη ασθενειών και επιδημιών, η μείωση του κόστους περίθαλψης των ασθενών κ.ά. (Stergiou *et al.*, 2022).

Στις μέρες μας ο καρκίνος είναι ένα από τα σοβαρότερα προβλήματα υγείας που καλούμαστε να αντιμετωπίσουμε. Η ποικιλομορφία και ο μεγάλος αριθμός διαφορετικών τύπων καρκίνου καθιστούν δύσκολη την κατανόηση του μηχανισμού της ασθένειας, ιδιαίτερα αν αναλογιστούμε και το πλήθος των μεταλλάξεων. Σύμφωνα με έρευνες στη χώρα μας ο αριθμός των θανάτων ετησίως εξαιτίας της νόσου πλησιάζει τις 32.000.

Το μεγαλύτερο ποσοστό των ογκολογικών δεδομένων που παράγονται είναι μη δομημένα και συνεπώς ακόμα πιο δύσκολο να μελετηθούν. Τα Μεγάλα Δεδομένα και η Αναλυτική των Μεγάλων Δεδομένων είναι το όπλο για την αξιοποίηση τους και συνεπώς τη θεραπεία αλλά και την πρόληψη της ασθένειας. Φυσικά δεν μπορούμε να παραβλέψουμε και το κόστος των θεραπειών που σε αυτές τις περιπτώσεις είναι δυσβάσταχτο.

## 1.2 Βασική ορολογία Big Data

Τα Big Data, σύμφωνα και με το όνομά τους, αναφέρονται σε τεράστια ή πολύπλοκα σύνολα δεδομένων, τα οποία είναι αδύνατο να τα διαχειριστούμε με παραδοσιακές πλατφόρμες λογισμικού ή πλατφόρμες του διαδικτύου. Λόγω του όγκου τους προκύπτουν δυσκολίες στην αποθήκευση, την επεξεργασία αλλά και την ανάλυσή τους. Η πολυπλοκότητα και η ποικιλομορφία που παρουσιάζουν οδηγούν στην ανάγκη για νέους πόρους, τεχνικές και αλγορίθμους για την διαχείρισή τους και την εξαγωγή αξιόπιστων πληροφοριών.

Σύμφωνα με ερευνητές ο όρος Big Data διαδόθηκε από τον John Mashey τη δεκαετία του 1990 ενώ δεν υπάρχει ένας μοναδικά αποδεκτός ορισμός. Κατά καιρούς

έχουν δοθεί διαφορετικοί ορισμοί σχετικά με την έννοια των Big Data, με πιο δημοφιλή αυτόν του Douglas Laney.

*Troy Segal, March 2022*

"Τα Big Data αναφέρονται στα μεγάλα και διαφορετικά σύνολα πληροφοριών που αναπτύσσονται με ολοένα αυξανόμενους ρυθμούς. Ο όρος περιλαμβάνει τον όγκο των πληροφοριών, την ταχύτητα με την οποία δημιουργούνται και συλλέγονται, αλλά και την ποικιλία ή το εύρος των σημείων δεδομένων που καλύπτουν. Τα μεγάλα δεδομένα συχνά προέρχονται από εξόρυξη δεδομένων και φτάνουν σε πολλαπλές μορφές" ('Μεταδεδομένα', 2022).

*Bridget Botelho, Stephen J. Bigelow, January 2022*

"Τα μεγάλα δεδομένα είναι ένας συνδυασμός δομημένων, ημιδομημένων και μη δομημένων δεδομένων που συλλέγονται από οργανισμούς, μπορούν να εξορυχθούν για πληροφορίες και να χρησιμοποιηθούν σε έργα μηχανικής μάθησης, μοντελοποίηση προβλέψεων και άλλες προηγμένες εφαρμογές ανάλυσης" (*What is Big Data and Why is it Important?,* no date).

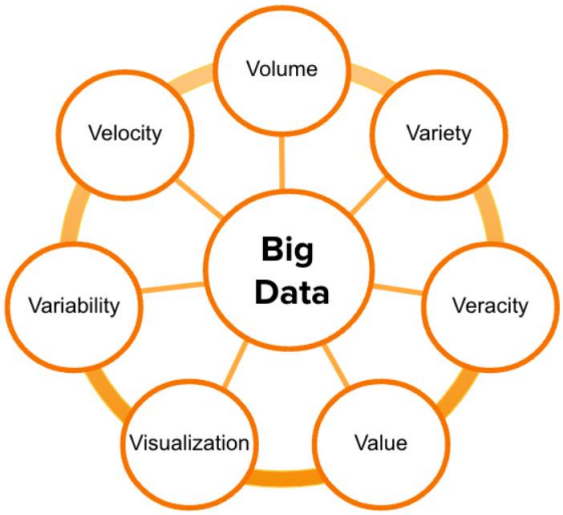
*Douglas Laney, 2001*

"Τα μεγάλα δεδομένα χαρακτηρίζονται από τον όγκο (volume), την ταχύτητα (velocity) και την ποικιλία (variety) και αυξάνονται και ως προς τις τρεις αυτές διαστάσεις. Οι τρεις διαστάσεις έχουν οριστεί ως τα "3 Vs" των Big Data και αποτελούν τον πιο αποδεκτό ορισμό τους" (*The V's of Big Data,* 2020).

Σε επόμενους ορισμούς στα τρία υπάρχοντα V's προστέθηκαν η εγκυρότητα (veracity), η αξία (value), η μεταβλητότητα (variability) και η οπτικοποίηση (visualization). Πιο αναλυτικά:

- **Volume**: Ο όγκος αναφέρεται στην ποσότητα των δεδομένων που παράγονται και αποθηκεύονται κάθε στιγμή. Η ποσότητα είναι αυτή που καθορίζει αν τα δεδομένα μπορούν να χαρακτηριστούν ως big data.
- **Velocity**: Ο όρος αφορά την ταχύτητα με την οποία παράγονται αλλά και την ταχύτητα με την οποία αναλύονται τα δεδομένα. Τις περισσότερες φορές τα δεδομένα διατίθενται σε πραγματικό χρόνο.

- **Variety**: Η ποικιλία υποδηλώνει την μορφή ή τον τύπο των δεδομένων. Τα δεδομένα ανάλογα με τη μορφή τους χωρίζονται σε δομημένα, ημί-δομημένα και μη δομημένα. Τα δομημένα έχουν αυστηρή οργάνωση και παρουσιάζουν ευκολία στη χρήση (π.χ. τηλεφωνικός κατάλογος). Τα ημί-δομημένα παρουσιάζουν μία δομή αλλά δεν είναι οργανωμένα σύμφωνα με ένα πρότυπο όπως ένας πίνακας. Τα μη δομημένα δεν ακολουθούν καμία προκαθορισμένη μορφή (π.χ. ήχος, εικόνα, βίντεο).
- **Veracity**: Η εγκυρότητα αναφέρεται στην ακεραιότητα και την ακρίβεια των δεδομένων. Η ποιότητα των πληροφοριών που εξάγονται εξαρτάται από την εγκυρότητά τους.
- **Value**: Αξία είναι η μέθοδος κατά την οποία εξάγονται χρήσιμες πληροφορίες από μεγάλα σύνολα δεδομένων. Συνήθως συναντάται ο όρος Big Data Analytics. Η αξία των δεδομένων είναι το κλειδί για την λήψη αποτελεσματικών αποφάσεων.
- **Variability**: Τα δεδομένα συνεχώς αλλάζουν και εκεί αναφέρεται ο όρος της μεταβλητότητας. Η συνεχής αλλαγή των δεδομένων μπορεί να επηρεάσει την ποιότητά τους.
- **Visualization**: Η οπτικοποίηση των δεδομένων εστιάζει στον τρόπο με τον οποίο παρουσιάζονται τα δεδομένα. Μία πλήρης και ορθά δομημένη παρουσίαση βοηθά στην κατανόηση και την ανάγνωση των δεδομένων και οδηγεί σε λήψη σωστών αποφάσεων.




**Εικόνα 1-1: 7 V's των Big Data (Manager, 2022)**

## 1.3 Χαρακτηριστικά των Big Data στην υγειονομική περίθαλψη

Η ολική σχεδόν ψηφιοποίηση των δεδομένων σε όλους τους τομείς έχει αντίκτυπο και στην υγειονομική περίθαλψη. Η υγεία αποτελεί ίσως τον τομέα με την μεγαλύτερη παραγωγή δεδομένων ανά δευτερόλεπτο. Αναμφίβολα η αξιοποίηση και η μετατροπή τους σε χρήσιμη πληροφορία είναι υψίστης σημασίας, καθώς ο όγκος τους είναι τεράστιος. Η αδιάκοπη αύξηση των δεδομένων και των πληροφοριών έχει στρέψει το ενδιαφέρον των επιστημόνων από τη θεραπεία των ασθενειών στην πρόληψη. Στόχο πλέον αποτελεί η ελαχιστοποίηση του κόστους της θεραπείας και η βελτίωση της φαρμακευτικής αγωγής που λαμβάνουν οι ασθενείς. Φυσικά προϋπόθεση των παραπάνω είναι ο ανασχηματισμός και η καλυτέρευση του συστήματος υγείας (Sarwar *et al.*, 2017).

Ενδιαφέρον παρουσιάζουν τα V's όπως αυτά προσαρμόζονται στον τομέα της υγείας:

- **Volume**: Σύμφωνα με έρευνα, το 2013 ο όγκος των δεδομένων υπολογιζόταν παγκοσμίως σε 4,4 zettabytes, ενώ υπερδιπλασιάζεται ετησίως. Αδιαμφισβήτητα, δεδομένα που αφορούν την υγεία όπως μελέτες για την αλληλουχία των γονιδίων, απεικονίσεις, στοιχεία αισθητήρων, αποτελέσματα εργαστηρίων, κ.λπ. φέρουν πληροφορίες οι οποίες απαιτούν καινοτόμους τρόπους για το συνδυασμό τους. Παράλληλα δεν πρέπει να αγνοούμε την ανάγκη επίλυσης θεμάτων αποθήκευσης είτε τοπικά είτε στο cloud (HealthITAnalytics, 2017).
- **Velocity**: Η ταχύτητα δεν αναφέρεται μόνο στον ρυθμό με τον οποίο παράγονται τα δεδομένα αλλά και στην επιτακτική ανάγκη να επεξεργάζονται έγκαιρα. Χαρακτηριστικό παράδειγμα αποτελούν οι περιπτώσεις ασθενών που βρίσκονται σε μηχανική υποστήριξη.
- **Variety**: Όπως σε κάθε τομέα έτσι και στην υγεία τα Big Data μπορεί να είναι δομημένα (εργαστηριακά αποτελέσματα, μετρήσεις αισθητήρων), ημι-δομημένα (δεδομένα αποθηκευμένα σε XML) ή μη δομημένα (μέτρηση καρδιακών παλμών χωρίς ιατρική συσκευή, χειρόγραφες σημειώσεις, ακτινογραφίες). Το μεγαλύτερο ποσοστό αφορά σε μη δομημένα και αγγίζει το 90%.
- **Veracity**: Λέγοντας εγκυρότητα αναφερόμαστε στην αξιοπιστία και την ακρίβεια που προσφέρουν τα δεδομένα. Η λήψη αποφάσεων για ζητήματα που αφορούν την ανθρώπινη ζωή προφανώς δεν μπορεί να βασίζεται σε

ανακρίβειες καθώς υπάρχουν περιπτώσεις λανθασμένων εκτιμήσεων, παραλείψεων, απατών ή ανακριβειών που επιβάλλεται να απορριφθούν (Raghupathi and Raghupathi, 2014) (HealthITAnalytics, 2017).

- **Value**: Η αξία συνεπάγεται ακριβή αποτελέσματα και ικανές στρατηγικές για τη λήψη αποφάσεων και προβλέψεων. Για να μπορούμε να κάνουμε λόγο για αξία είναι απαραίτητο να διαθέτουμε αποτελεσματικές πλατφόρμες για την επεξεργασία των δεδομένων, ικανές στρατηγικές ανάλυσης και τεχνικές αποθήκευσης. Η πρόβλεψη πανδημιών, καρκινικών όγκων και άλλων ασθενειών που μπορεί να πλήξουν ολόκληρους πληθυσμούς είναι το ιδανικό σενάριο (HealthITAnalytics, 2017).
- **Variability**: Τα υγειονομικά δεδομένα χαρακτηρίζονται από υψηλή μεταβλητότητα. Αυτό καθιστά δυσκολότερο το διαχωρισμό των δεδομένων που θα λάβουν μέρος σε μια ανάλυση και εγείρει ερωτήματα για τη διάρκεια διατήρησης αρχείου (HealthITAnalytics, 2017).
- **Visualization**: Η οπτικοποίηση των υγειονομικών δεδομένων παραπέμπει σε αναλυτικά αποτελέσματα με τη μορφή εικόνων, διαγραμμάτων, πινάκων κ.λπ.. Στόχος κάθε τέτοιας αναφοράς είναι η καλύτερη κατανόηση της πληροφορίας και της συσχέτισης των δεδομένων (HealthITAnalytics, 2017) (Raghupathi and Raghupathi, 2014).

## 1.4 Πηγές Big Data στην υγειονομική περίθαλψη

Τα Big Data στην υγειονομική περίθαλψη προέρχονται από πολλές και διαφορετικών ειδών πηγές. Μία σύγχρονη πηγή είναι τα κινητά τηλέφωνα από τα οποία μέσω των αισθητήρων λαμβάνονται πληροφορίες για την κινητικότητα των ατόμων (μέτρηση ημερήσιων βημάτων), τους καρδιακούς παλμούς και διάφορα βιομετρικά στοιχεία. Επίσης τα μέσα κοινωνικής δικτύωσης παρέχουν πληροφορίες για τον τρόπο ζωής, τη συμπεριφορά των ατόμων και δεδομένα που αφορούν στην μετάδοση και την εξάπλωση ασθενειών. Επιπλέον σημαντικές είναι οι πληροφορίες που προέρχονται από ιατρικές απεικονίσεις όπως μαγνητικές και αξονικές τομογραφίες (Alonso *et al.*, 2017). Εν συνεχεία από τις κλινικές δοκιμές παρέχονται δεδομένα για την ασφάλεια και την αποτελεσματικότητα των θεραπειών που εφαρμόζονται, ενώ από γονιδιακά δεδομένα εξακριβώνεται η γενετική σύνθεση των ασθενών και διερευνώνται υποκείμενες αιτίες

της ασθένειας. Εξίσου σημαντική πηγή αποτελούν τα Ηλεκτρονικά Μητρώα Υγείας (EHR) στα οποία αποθηκεύονται τα ιατρικά ιστορικά αλλά και θεραπείες ατομικά προσαρμοσμένες (Alonso *et al.*, 2017). Τέλος δεδομένα παρέχονται και από ασφαλιστικές που αφορούν στο πώς και το πόσο χρησιμοποιούνται οι υπηρεσίες υγειονομικής περίθαλψης και το κόστος.

Στις περιπτώσεις καρκίνου συλλέγονται κυρίως μη δομημένα δεδομένα τα οποία αναλύονται σε πραγματικό χρόνο για τη βέλτιστη αξιοποίησή τους (Plageras *et al.*, 2017). Τα δεδομένα που λαμβάνονται σχετίζονται με το φύλο, την ηλικία, την κληρονομικότητα, την αλληλουχία των DNA και RNA, την έκβαση και τη θεραπεία της ασθένειας, στοιχεία του γονιδιώματος, κ.ά.. Η κύρια πηγή είναι οι ίδιοι οι ασθενείς καθώς η παρακολούθηση της ασθένειας απαιτεί συνεχή χρήση ψηφιακών εργαλείων. Έτσι λαμβάνουμε εικόνα για τις συνέπειες του ίδιου του καρκίνου αλλά και της θεραπείας του στην ποιότητα ζωής των ασθενών. (Willems *et al.*, 2019)

Επίσης από τα μητρώα καρκίνου συλλέγονται δεδομένα για τα χαρακτηριστικά των όγκων και την αποτελεσματικότητα των θεραπειών καθώς και δημογραφικές πληροφορίες. Δεν είναι δύσκολο να αντιληφθούμε την ταχύτητα με την οποία επεκτείνονται τα στοιχεία συμπεριλαμβανομένων και των μεταστάσεων.

## 1.5 Προκλήσεις και προοπτικές

Η υιοθέτηση των μεγάλων δεδομένων από την υγειονομική περίθαλψη υπόσχεται βελτίωση των αποτελεσμάτων αλλά και έλεγχο του κόστους. Παρόλο όμως που ο κλάδος της πληροφορικής, ειδικότερα η τεχνολογία των Big Data, έχουν σημειώσει εντυπωσιακή πρόοδο θα μπορούσαμε να πούμε πως στον τομέα της υγείας και ιδιαίτερα της ογκολογίας δεν έχουμε φτάσει στο επιθυμητό επίπεδο. Οι συνεχείς έρευνες επιφέρουν μαζικές ποσότητες σύνθετων δεδομένων που έχουν ως επακόλουθο προκλήσεις που δεν μπορούν να μην αντιμετωπιστούν. Αναμφίβολα τα δεδομένα που αφορούν στην υγεία παρουσιάζουν μεγαλύτερες δυσκολίες τόσο στην ανάλυση όσο και στην ερμηνεία τους. Αυτό οφείλεται στην ετερογένεια του πληθυσμού, την μεταβλητότητα των ασθενειών στο χρόνο, τη διαφορετικότητα των χαρακτηριστικών τους αλλά κυρίως στο ότι αφορούν σε ανθρώπινες ζωές.

Η αναλυτική των μεγάλων δεδομένων με τη σειρά της στοχεύει στην μείωση του κόστους των θεραπειών, την πρόληψη ασθενειών και επιδημιών αλλά και γενικότερα στη βελτίωση της ποιότητας ζωής. Προκειμένου όμως να επιτευχθεί σωστά και με

ακρίβεια η ανάλυση των δεδομένων κρίνεται απαραίτητο τα δεδομένα να είναι οργανωμένα, ακριβή και κυρίως ασφαλή. Στην όλη διαδικασία εξακολουθούν να προκύπτουν ηθικές προκλήσεις που σχετίζονται με το προσωπικό απόρρητο (Garapati and Garapati, 2018).

Στη συνέχεια αναλύονται οι κυριότερες προκλήσεις που καλούμαστε να αντιμετωπίσουμε:

- **Αποθήκευση:** Δεδομένου του τεράστιου όγκου δεδομένων, η αποθήκευσή τους αποτελεί μια από τις κυριότερες προκλήσεις. Για τους οργανισμούς υγειονομικής περίθαλψης θα ήταν προτιμότερο η αποθήκευση να γίνεται στις εγκαταστάσεις τους. Αυτό τους προσφέρει μεγαλύτερο αίσθημα ασφάλειας αλλά και ευκολία πρόσβασης. Ωστόσο απαιτεί μεγάλο κόστος συντήρησης. Η επιλογή του cloud αποτελεί την πιο συμφέρουσα λύση ενώ προσφέρει και τη δυνατότητα ανάκτησης σε τυχόν φθορές (Dash *et al.*, 2019).
- **Ακρίβεια:** Σύμφωνα με μελέτες η ακρίβεια των δεδομένων συχνά αμφισβητείται. Η ελλιπής χρήση του ήδη υπάρχοντος ιατρικού ιστορικού και η δυσκολία πλήρους κατανόησης των δεδομένων λόγω της πολυπλοκότητάς τους δεν προσφέρουν πλήρη αξιοπιστία. Η εξέλιξη μέσω των Big Data Analytics απαιτεί και δεδομένα υψηλής ποιότητας (Dash *et al.*, 2019) (Raghupathi and Raghupathi, 2014).
- **Καθάρισμα:** Το καθάρισμα των δεδομένων είναι το σημαντικότερο μέρος κατά την ανάλυση. Με τον όρο καθαρισμός εννοούμε την απομάκρυνση μη σχετικών και διπλότυπων δεδομένων, το χειρισμό σφαλμάτων και μορφοποίησης κ.λπ.. Σύμφωνα με τους New York Times, οι επιστήμονες ξοδεύουν παραπάνω από το μισό χρόνο τους στον καθαρισμό των δεδομένων πριν τις αναλύσεις (HealthITAnalytics, 2017).
- **Ασφάλεια:** Δεν είναι λίγες οι φορές όπου έχουν παρατηρηθεί παραβιάσεις στην ασφάλεια και το απόρρητο των δεδομένων. Για το λόγο αυτό οι οργανισμοί υγείας έχουν αναπτύξει τους Κανόνες Προστασίας Προσωπικών Δεδομένων (HIPAA) (Rights (OCR), 2008) (Dash *et al.*, 2019). Από τη στιγμή που η αποθήκευση έχει μετακινηθεί στο cloud οι οργανισμοί υγείας επενδύουν περισσότερα χρήματα στην ασφάλεια καθώς οι παραδοσιακές μέθοδοι όπως το τείχος προστασίας ή η

κρυπτογράφηση δεν αρκούν (Memos *et al.*, 2021) (Stergiou *et al.*, 2018). Επιπλέον εκπαιδεύουν το ανθρώπινο δυναμικό τους ώστε να είναι σε θέση να διατηρήσει ιδιωτικά και ασφαλή τα δεδομένα (HealthITAnalytics, 2017).

- **Κοινή Χρήση Δεδομένων:** Κάθε ασθενής δεν λαμβάνει φροντίδα από ένα μέρος αποκλειστικά άρα οι πληροφορίες του ιστορικού του θα πρέπει να μεταφερθούν (Stergiou, Psannis and Gupta, 2020). Αυτό απαιτεί υψηλή διαλειτουργικότητα των δεδομένων ώστε η μετακίνηση να γίνεται ολοκληρωμένα και χωρίς εμπόδια (Dash *et al.*, 2019).
- **Μεταδεδομένα:** Τα μεταδεδομένα είναι σύνολα δεδομένων τα οποία περιγράφουν ένα άλλο σύνολο δεδομένων, με εφαρμογές φυσικά και στην υγεία (‘Μεταδεδομένα’, 2022). Πληροφορίες που αφορούν το άτομο το οποίο δημιούργησε τα δεδομένα, το χρόνο αλλά και το σκοπό ανήκουν στα μεταδεδομένα και είναι πολύ σημαντικό να είναι συνεχώς ενημερωμένα. Με τον τρόπο αυτό οι αναλυτές είναι σε θέση να διαχειρίζονται παλαιότερα ερωτήματα που θα εξελίξουν επόμενες έρευνες (Dash *et al.*, 2019).
- **Οπτικοποίηση:** Η οπτικοποίηση των δεδομένων και μάλιστα σε πραγματικό χρόνο είναι ικανή να βελτιώσει την κατάσταση της υγείας του ασθενούς (Minopoulos *et al.*, 2023). Μία καθαρή απεικόνιση (γράφημα, ιστόγραμμα, διάγραμμα πίτας, κ.λπ.) προσφέρει τη δυνατότητα στο ιατρικό προσωπικό να αντιληφθεί πιθανούς κινδύνους έγκαιρα και με μεγαλύτερη ευκολία (Dash *et al.*, 2019).

Αντιμετωπίζοντας τις προκλήσεις πλησιάζουμε ολοένα και περισσότερο στις προοπτικές που προσφέρουν τα Big Data στην υγεία. Ο συνδυασμός των δεδομένων του EHR και του EMR στην αναλυτική υπόσχεται εγκυρότερες προγνώσεις που θα συμβάλλουν σημαντικά στην πρόληψη. Στόχο αποτελεί επίσης η εξατομικευμένη θεραπεία που απαιτεί τη συνεργασία της βιοπληροφορικής, της ανάλυσης και της πληροφορικής υγείας (Dash *et al.*, 2019). Πλέον οι έρευνες είναι στραμμένες στην μελέτη του γονιδιώματος των οργανισμών με αποτέλεσμα ακόμη μεγαλύτερη αύξηση των δεδομένων που παράγονται. Τα δεδομένα αυτά βοηθούν στην έρευνα και οδηγούν σε καινοτόμες θεραπείες (Adibuzzaman *et al.*, 2018). Οι τεχνολογίες που διαθέτουμε και εξελίσσονται συνεχώς προσφέρουν παράλληλα και μάθηση στο ιατρικό προσωπικό και

πλήρη κατανόηση των καρκινοπαθών. Αυτό σαφώς θα συμβάλλει στην ακριβέστερη θεραπεία και φροντίδα των ασθενών με αποτελεσματικότητα (Tsai, Riaz and Gomez, 2019). Επιπλέον η συστηματική μελέτη των δεδομένων και η βελτίωση της υγειονομικής περίθαλψης υπόσχεται και σημαντική μείωση του κόστους της θεραπείας και της φροντίδας των ασθενών (Mehta and Pandit, 2018). Τα Big Data Analytics με τη σειρά τους θα αλλάξουν ριζικά τον τρόπο με τον οποίο οι ερευνητές λαμβάνουν τα δεδομένα από κλινικά ή άλλα αποθετήρια. Το ιδανικό σενάριο είναι η ευρεία χρήση της αναλυτικής των μεγάλων δεδομένων σε ολόκληρη την υγειονομική περίθαλψη και όπως φαίνεται αυτό θα επιτευχθεί (Raghupathi and Raghupathi, 2014).

## 1.6 Διάρθρωση της μελέτης

Στο Κεφάλαιο 1 γίνεται εισαγωγή στο πρόβλημα και τη σημαντικότητα του. Επίσης αναλύονται βασικές ορολογίες και παρουσιάζονται οι προκλήσεις και οι προοπτικές. Στο Κεφάλαιο 2 γίνεται αναφορά στα Big Data Analytics και τις εφαρμογές τους στον τομέα της υγείας. Στη συνέχεια στο Κεφάλαιο 3 αναπτύσσεται το απαραίτητο θεωρητικό υπόβαθρο για την υλοποίηση της εργασίας. Στο Κεφάλαιο 4 αναλύεται η μεθοδολογία για την υλοποίηση των εφαρμογών. Κλείνοντας στο Κεφάλαιο 5 αναπτύσσονται τα συμπεράσματα που προέκυψαν από τις εφαρμογές και μελλοντικές επεκτάσεις.

# 2 Big Data Analytics στην Υγειονομική Περίθαλψη

Για τη διαχείριση των Big Data επιλέγονται προηγμένες τεχνικές ανάλυσης, τα Big Data Analytics (Stergiou and Psannis, 2017). Με τον όρο Big Data Analytics αναφερόμαστε στην διαδικασία εξέτασης συνόλων δεδομένων που χαρακτηρίζονται από πολύ μεγάλο όγκο και πολυπλοκότητα με σκοπό την κατανόηση μοτίβων και πληροφοριών (Wang, Kung and Byrd, 2018).

Στον τομέα της υγείας τα Big Data Analytics είναι το μέσο των επιστημόνων για την κατανόηση της συμπεριφοράς των ασθενειών, την μελέτη για την αποτελεσματικότητα των θεραπειών και τη διαχείριση του κόστους περίθαλψης (Minopoulos *et al.*, 2022). Η αξιοποίηση των Big Data Analytics υπόσχεται βελτίωση στη θεραπεία των ασθενών και στη φροντίδα την οποία λαμβάνουν.

## 2.1 Εφαρμογές Big Data Analytics στην Υγειονομική Περίθαλψη

Κάθε εφαρμογή και κάθε τεχνική που χρησιμοποιείται από τους ειδικούς έχει στο επίκεντρό της τον ασθενή. Η έγκαιρη και έγκυρη διάγνωση, η εφαρμογή της αποτελεσματικότερης εξατομικευμένης θεραπείας και η ποιοτική περίθαλψη σε συνδυασμό με το ελάχιστο δυνατό κόστος είναι ο στόχος που οφείλει να επιτευχθεί.

### *2.1.1 Προγνωστική Αναλυτική (Predictive Analytics)*

Η προγνωστική αναλυτική έχοντας στο επίκεντρό της την πληροφορία που αντλεί και όχι τα δεδομένα από τα οποία αυτή προήλθε προσπαθεί να προβλέψει το μέλλον. Για την ανάλυση χρησιμοποιούνται ιστορικά δεδομένα, παρελθοντικοί δείκτες αλλά και δεδομένα πραγματικού χρόνου. Στόχος της είναι η πρόβλεψη μελλοντικών επιδόσεων όπως για παράδειγμα η πιθανότητα εμφάνισης επιπλοκών σε έναν ασθενή.

Μελετάει δεδομένα υγείας αναζητώντας μοτίβα και σχέσεις μεταξύ τους τα οποία θα οδηγήσουν σε έγκυρες προβλέψεις. Επιπλέον καθίσταται δυνατός ο εντοπισμός ασθενών που είναι περισσότερο επιρρεπείς σε μία συγκεκριμένη ασθένεια αλλά και η πρόβλεψη εξάπλωσης πιθανών επιδημιών. Τα παραπάνω συμβάλλουν στην οργάνωση των απαραίτητων υγειονομικών πόρων που απαιτούνται και στην χρήση εξατομικευμένων θεραπειών (Batko and Ślęzak, 2022).

### *2.1.2 Υποστήριξη Κλινικών Αποφάσεων (Clinical Decision Support)*

Η υποστήριξη κλινικών αποφάσεων συναντάται στο στάδιο της περίθαλψης των ασθενών. Χρησιμοποιεί την τεχνολογία και τη γνώση που έχει αποκτηθεί από τη μελέτη δεδομένων οδηγώντας τους ειδικούς στη λήψη ενημερωμένων αποφάσεων για την φροντίδα τους. Τα συστήματα που αφορούν την υγεία διακρίνονται από μεγάλη ακρίβεια ώστε να είναι οι αποφάσεις όσο το δυνατό εγκυρότερες. Αυτό σημαίνει ότι το ιατρικό προσωπικό πρέπει να έχει αρκετές γνώσεις για να είναι σίγουρο ότι η διάγνωση που κάνει είναι σωστή (Minopoulos *et al.*, 2022).

Τα εργαλεία υποστήριξης κλινικών αποφάσεων χρησιμοποιούν δεδομένα ιστορικού των ασθενών αλλά και πραγματικού χρόνου τα οποία συνδυάζονται με γνώσεις βασιζόμενες σε στοιχεία. Οι κλινικοί ιατροί ειδοποιούνται από το σύστημα για αλληλεπιδράσεις φαρμάκων, πιθανά λάθη στη χορήγηση φαρμακευτικής αγωγής κ.ά.. Συνεπώς μειώνεται η πιθανότητα ιατρικού λάθους βελτιώνοντας την αποτελεσματικότητα της θεραπείας και την ασφάλεια του εκάστοτε ασθενή.

### *2.1.3 Παρακολούθηση Ασθενούς (Patient Monitoring)*

Η παρακολούθηση των ασθενών στοχεύει στα καλύτερα δυνατά αποτελέσματα για τους ασθενείς και την φροντίδα τους. Οι ασθενείς παρακολουθούνται σε πραγματικό χρόνο και γίνεται προσπάθεια για τον εντοπισμό προβλημάτων σε πρώιμο στάδιο και πριν γίνουν σοβαρά.

Τα Analytics συμβάλλουν στην εξατομίκευση της περίθαλψης του κάθε ασθενούς και παρέχουν κατανόηση των αναγκών του ατομικά όπως η φαρμακευτική αγωγή. Με τον τρόπο αυτό παρατηρείται μείωση στις επανεισαγωγές στις δομές υγείας και βελτιώνεται ποιοτικά η περίθαλψη στο σύνολό της. Βασικός στόχος είναι η διαχείριση των ασθενειών προληπτικά (Harb *et al.*, 2021).

### *2.1.4 Διαχείριση Υγείας Πληθυσμού (Population Health Management)*

Η διαχείριση υγείας του πληθυσμού συμπεριλαμβάνει τη συλλογή και την ανάλυση δεδομένων που προέρχονται από καθορισμένους πληθυσμούς με σκοπό την ερμηνεία τους και τη βελτίωση των αποτελεσμάτων υγείας.

Τα PHM Analytics στοχεύουν στον εντοπισμό πληθυσμών που κινδυνεύουν από χρόνιες παθήσεις και αναζητούν τρόπους ώστε να μειωθεί η χρήση των υπηρεσιών υγείας. Επιπλέον παρακολουθώντας την πορεία της υγειονομικής περίθαλψης επισημαίνει τους τομείς στους οποίους μπορεί να υπάρξει καλύτερη κατανομή πόρων (Mehta and Pandit, 2018).

### *2.1.5 Ανίχνευση Απάτης (Fraud Detection)*

Η αναλυτική των μεγάλων δεδομένων μπορεί να εντοπίσει παράτυπες δραστηριότητες, ψευδείς ισχυρισμούς, παράνομες χρεώσεις ακόμη και μη συνιστώμενη συνταγογράφηση. Υπάρχουν δείκτες που επισημαίνουν τον κίνδυνο σε περίπτωση δόλιας δραστηριότητας.

Για την ανίχνευση πιθανής απάτης χρησιμοποιούνται διάφορες τεχνικές όπως μηχανική μάθηση, εξόρυξη δεδομένων ή στατιστική ανάλυση. Οι ασφαλιστικές εταιρείες και οι οργανισμοί υγειονομικής περίθαλψης αξιοποιούν αυτές τις τεχνικές ώστε να αποφύγουν τις δολιοφθορές και να εξασφαλίσουν ποιοτική περίθαλψη με ελάχιστο κόστος (Jha, Sivasankari and Venugopal, 2020).

# 3 Apache Hadoop

## 3.1 Τι είναι το Apache Hadoop

Το Apache Hadoop είναι ένα από τα πιο διαδεδομένα πλαίσια ανοιχτού κώδικα που επιτρέπει τη διαχείριση δομημένων, ημί-δομημένων και αδόμητων δεδομένων. Σύμφωνα με τους Doug Cutting και Mike Cafarella, συνιδρυτές του Hadoop, η ύπαρξή του οφείλεται σε ερευνητική εργασία για το σύστημα αρχείων της Google, δημοσιευμένη τον Οκτώβριο του 2003. Με αφορμή την εργασία αυτή πραγματοποιήθηκε η νέα έρευνα «MapReduce: Simplified Data Processing on Large Clusters». Το Apache Hadoop άρχισε να αναπτύσσεται στο Apache Nutch, όμως τον Ιανουάριο του 2006 μεταφέρεται σε νέο έργο με αντίστοιχη ονομασία Apache Hadoop. Ο αρχικός του κώδικας αποτελείται από 5.000 για το HDFS και 6.000 για το MapReduce. Το 2006 κυκλοφορεί το Hadoop 0.1.0 και τον Αύγουστο του 2022 η πιο πρόσφατη έκδοση 3.3.4 (‘Apache Hadoop’, 2023a).

Το Hadoop προσφέρει τη δυνατότητα διαχείρισης και αποθήκευσης τεράστιου όγκου δεδομένων εφαρμόζοντας παραλληλία και επεξεργασία σε πολλούς κόμβους. Λόγω του μεγέθους των δεδομένων για να λάβουμε αποτελέσματα σε λογικά χρονικά πλαίσια χρησιμοποιούνται χιλιάδες υπολογιστικές μηχανές. Η εργασία σε πολλούς κόμβους μας καλεί να χειριστούμε τον τρόπο της παραλληλίας και τη διαχείριση σφαλμάτων (Dash *et al.*, 2019).


**Εικόνα 3-1: Apache Hadoop (*Apache Hadoop*, 2023b)**

Το Hadoop αποτελείται από δύο τμήματα, ένα αποθήκευσης και ένα επεξεργασίας. Το πρώτο ονομάζεται Hadoop Distributed File System (HDFS) ενώ το δεύτερο είναι ο αλγόριθμος MapReduce. Τα δύο τμήματα θα αναλυθούν εκτενέστερα παρακάτω.

## 3.2 Κύριες ενότητες Apache Hadoop

Το Hadoop αποτελείται από τις παρακάτω τέσσερις ενότητες (*Apache Hadoop*, 2023b):

- **Hadoop Common:** Πρόκειται για ένα σύνολο βιβλιοθηκών και βοηθητικών προγραμμάτων τα οποία υποστηρίζουν τις υπόλοιπες ενότητες.
- **Hadoop Distributed File System (HDFS):** Το HDFS σημαίνει Κατανεμημένο Σύστημα Αρχείων Hadoop. Προσφέρει αξιοπιστία καθώς βασίζεται στην αποθήκευση αντιγράφων των δεδομένων σε διαφορετικούς κόμβους.
- **Hadoop YARN:** Είναι η τεχνολογία με την οποία προγραμματίζονται οι εργασίες και διαχειρίζονται οι πόροι.
- **Hadoop MapReduce:** Το MapReduce βασίζεται στο YARN και επεξεργάζεται με παραλληλία μεγάλα σύνολα δεδομένων.

## 3.3 Hadoop Distributed File System (HDFS)

Το Κατανεμημένο Σύστημα Αρχείων Hadoop έχει σχεδιαστεί με τέτοιο τρόπο ώστε να αποθηκεύει μεγάλο όγκο δεδομένων. Τα αρχεία αποθηκεύονται ταυτόχρονα σε ένα σύνολο μηχανημάτων, γεγονός που προσφέρει αξιοπιστία, ανοχή σε σφάλματα και παράλληλη χρήση. Το HDFS έχει σχεδιαστεί σύμφωνα με το σύστημα αρχείων της Google καθώς έχει υιοθετήσει πολλές από τις λειτουργίες του. Όπως αναφέραμε και παραπάνω το HDFS έχει ανοχή σε σφάλματα και έτσι εξασφαλίζει τη λειτουργικότητα του σε οποιεσδήποτε συνθήκες (‘Apache Hadoop’, 2023a). Αρχικά γίνεται διαχωρισμός των δεδομένων σε ομάδες και έπειτα δημιουργούνται πολλά αντίγραφα της κάθε μίας, τα οποία αποθηκεύονται σε διαφορετικά μηχανήματα. Σημαντικό χαρακτηριστικό της λειτουργικότητας του HDFS είναι ότι από τη στιγμή που θα αποθηκευτεί μία πληροφορία δεν επιτρέπεται ξανά η επεξεργασία της, εξασφαλίζοντας την ακεραιότητα της και αποτρέποντας τον κίνδυνο αλλοίωσης των πληροφοριών (*Hadoop Distributed File System (HDFS) for Big Data Projects*, 2016).

### *3.3.1 Αρχιτεκτονική HDFS*

Το HDFS ακολουθεί την αρχιτεκτονική τύπου master/slave σύμφωνα με την οποία ένας υπολογιστής (master) συντονίζει και κατανέμει τις εργασίες σε

περισσότερους διακομιστές (slaves) οι οποίοι τις εκτελούν. Ο master διακομιστής ονομάζεται NameNode ενώ οι slaves ονομάζονται DataNodes. Ο NameNode είναι υπεύθυνος για τη διαχείριση του namespace των αρχείων και ρυθμίζει την πρόσβαση των χρηστών. Επιπλέον ο NameNode χωρίζει τα δεδομένα σε blocks και περιέχει πληροφορίες για τα ονόματα των αρχείων, τα δικαιώματα και την τοποθεσία κάθε block των αρχείων. Οι DataNodes αποθηκεύουν και διαμοιράζουν τα δεδομένα ανάλογα με τα αιτήματα που λαμβάνουν, ενώ μπορούν να δημιουργούν και να διαγράφουν τα blocks. Σημαντικός είναι και ο ρόλος του Secondary NameNode, διακομιστής ο οποίος δημιουργεί αντίγραφα του NameNode. Σε περίπτωση αποτυχίας του NameNode γίνεται αντικατάσταση του με το πιο πρόσφατο αντίγραφο που έχει δημιουργήσει ο Secondary NameNode.

Μερος της αρχιτεκτονικής του HDFS είναι και η αποθήκευση των δεδομένων. Για κάθε block δεδομένων το HDFS δημιουργεί τρία αντίγραφα. Η αποθήκευση του πρώτου γίνεται τοπικά, ενώ τα άλλα δύο αποθηκεύονται το καθένα σε διαφορετικό DataNode. Το κάθε block έχει συνήθως μέγεθος 128MB (*Apache Hadoop*, 2023b).

## 3.4 MapReduce

Το MapReduce αποτελεί τον πυρήνα του Apache Hadoop. Είναι ένα μοντέλο προγραμματισμού σχεδιασμένο για την επεξεργασία δεδομένων μεγάλου όγκου κατανεμημένα σε clusters. Το Hadoop χρησιμοποιεί το μοντέλο MapReduce για την δημιουργία εφαρμογών επεξεργαζόμενο τεράστιες ποσότητες δεδομένων με τρόπο αξιόπιστο και ανθεκτικό σε λάθη.

Ο αλγόριθμος MapReduce αποτελείται από δύο φάσεις, την Map και την Reduce, με την Map να εκτελείται πρώτη. Τα παραγόμενα δεδομένα της πρώτης φάσης είναι της μορφής key – value και χρησιμοποιούνται στο στάδιο Reduce (‘MapReduce’, 2023).

Στη συνέχεια παραθέτουμε μια πιο λεπτομερή ανάλυση των φάσεων:

- **Map:** Προκειμένου να ξεκινήσει η φάση του Map τα δεδομένα χωρίζονται σε μικρότερα τμήματα και η επεξεργασία τους κατά τη διάρκεια της διαδικασίας καθορίζεται από τον προγραμματιστή. Όπως αναφέρθηκε και παραπάνω τα δεδομένα εισόδου είναι της μορφής key – value. Η συγκεκριμένη φάση στοχεύει στην ομαδοποίηση των τιμών με το ίδιο key.

- **Reduce:** Σε συνέχεια της πρώτης φάσης, η Reduce είναι σαν μια περίληψη των τιμών που έχουν παραχθεί. Όλα τα αποτελέσματα της Map συγκεντρώνονται και δίνεται στον χρήστη μια γενική απάντηση για το πρόβλημα που επιλύει το μοντέλο.

### *3.4.1 Αρχιτεκτονική MapReduce*

Όπως και στην περίπτωση του HDFS το MapReduce χρησιμοποιεί και αυτό την αρχιτεκτονική τύπου master/slave. Ο master διακομιστής ονομάζεται Job Tracker ενώ ο slave Task Tracker. Ο ρόλος του Job Tracker είναι να ελέγχει και να κατανέμει όλες τις διεργασίες που καλούνται να εκτελέσουν οι Task Trackers στη φάση Map. Βασικός σκοπός είναι η χρήση όσο το δυνατό λιγότερων πόρων.

Αρχικά τα αιτήματα των χρηστών αποστέλλονται στον Job Tracker ο οποίος εντοπίζει τη θέση των δεδομένων επικοινωνώντας με τον NameNode. Στη συνέχεια αναθέτει τις κατάλληλες εργασίες στους διαθέσιμους Task Trackers που έχει εντοπίσει. Σε περίπτωση που κάποιος Task Tracker δεν ανταποκριθεί ενημερώνεται ο Job Tracker ο οποίος μπορεί να τον χαρακτηρίσει ως αναξιόπιστο ή να αναθέσει την εργασία εκ νέου σε άλλο κόμβο. Με το πέρας των διεργασιών ενημερώνεται η κατάσταση του Job Tracker (*JobTracker - HADOOP2 - Apache Software Foundation*, 2019).

# 4 Μεθοδολογία

Στο παρόν κεφάλαιο θα παρουσιαστούν τα βήματα που ακολουθήθηκαν για την ανάλυση των ογκολογικών δεδομένων με στόχο την εξαγωγή συμπερασμάτων. Πιο συγκεκριμένα γίνεται αναφορά στα δεδομένα που χρησιμοποιήθηκαν, στη μορφή και την προετοιμασία τους. Επίσης παρουσιάζεται αναλυτικά κάθε εφαρμογή του MapReduce στα αρχεία jar που έχουν υλοποιηθεί.

## 4.1 Δεδομένα

Τα ογκολογικά δεδομένα που χρησιμοποιήθηκαν αφορούν στον καρκίνο του πνεύμονα και στο πως σχετίζεται με το κάπνισμα. Πρόκειται για ένα υποσύνολο δεδομένων το οποίο παρέχεται από το National Lung Screening Trial των ΗΠΑ και είναι διαθέσιμο στο The Cancer Imaging Archive (*Wiki - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki*, no date). Οι πληροφορίες που περιέχονται σχετίζονται με πρώην και νυν καπνιστές ενώ δεν υπάρχει δείγμα μη καπνιστών στην

έρευνα. Η εύρεση των δεδομένων αποτέλεσε το δυσκολότερο κομμάτι της εργασίας καθώς λόγω απορρήτου και προσωπικών δεδομένων δεν υπάρχουν διαθέσιμα προς ελεύθερη χρήση.

## 4.2 Μορφολογία δεδομένων

Τα δεδομένα περιέχουν πληροφορίες που λαμβάνονταν από καπνιστές οι οποίοι ελέγχονταν για καρκίνο κάθε χρόνο για επτά συνεχόμενα έτη. Επίσης δεν υπάρχουν δείγματα μη καπνιστών. Οι πληροφορίες αφορούν το φύλο, την ηλικία, το στάδιο του καρκίνου, την πρώτη επιβεβαιωμένη διάγνωση κ.ά.. Τα δεδομένα παρέχονται σε μορφή αρχείων .csv και η οριοθέτηση γίνεται με τον χαρακτήρα κόμμα(,). Κοινό σημείο των αρχείων είναι ο μοναδικός αριθμός id κάθε για κάθε άτομο που έλαβε μέρος.

| | A | B | C |
|---|---|---|---|
| 1 | 100012,61,2,1,1,3,1,110,1,3,110 | | |
| 2 | 100049,74,2,1,4,4,1,220,1,3, | | |
| 3 | 100055,67,1,1,1,1,0,400,1,3, | | |
| 4 | 100147,68,1,1,1,3,0,110,1,3,110 | | |
| 5 | 100158,65,1,1,1,4,0,110,1,3,110 | | |
| 6 | 100196,65,2,1,4,1,1,400,1,3, | | |
| 7 | 100242,60,2,1,1,3,0,110,1,3,999 | | |
| 8 | 100242,60,2,1,1,3,0,110,0,3,999 | | |
| 9 | 100264,69,1,1,3,5,1,400,1,3, | | |
| 10 | 100280,60,2,1,1,2,1,110,1,3,110 | | |
| 11 | 100292,68,1,1,1,4,1,110,1,3,110 | | |

**Εικόνα 4-1: Δεδομένα σε μορφή .csv**

## 4.3 Συλλογή και προετοιμασία δεδομένων

Αρχικά το σετ των δεδομένων αποτελείται από πέντε διαφορετικά αρχεία .csv τα οποία μορφοποιήθηκαν στο excel. Στη συνέχεια κάθε αρχείο εισήχθη στη βάση Oracle SQL Developer στην οποία δημιουργήθηκε ένα view αποτελούμενο από τα πεδία που χρησιμοποιήθηκαν για την εξαγωγή των συμπεράσματων. Στον παρακάτω πίνακα παρουσιάζονται τα τελικά πεδία που χρησιμοποιήθηκαν και η περιγραφή τους.

| | PID | AGE | GENDER | RACE | CAN_SCR | LC_GRADE | CIGSMOK | DE_STAG | FIRST_LC | LC_BEHAV |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 100012 | 61 | 2 | 1 | 1 | 2 | 1 | 110 | 1 | 3 |
| 2 | 100049 | 74 | 2 | 1 | 4 | 3 | 1 | 220 | 1 | 3 |
| 3 | 100055 | 67 | 1 | 1 | 1 | 9 | 0 | 400 | 1 | 3 |
| 4 | 100147 | 68 | 1 | 1 | 1 | 2 | 0 | 110 | 1 | 3 |
| 5 | 100158 | 65 | 1 | 1 | 1 | 3 | 0 | 110 | 1 | 3 |
| 6 | 100196 | 65 | 2 | 1 | 4 | 9 | 1 | 400 | 1 | 3 |
| 7 | 100242 | 60 | 2 | 1 | 1 | 2 | 0 | 110 | 1 | 3 |

**Εικόνα 4-2: Τελική μορφή δεδομένων προς επεξεργασία**

| ΠΕΔΙΟ | ΠΕΡΙΓΡΑΦΗ | ΤΙΜΕΣ |
|---|---|---|
| pid | Το μοναδικό αναγνωριστικό κάθε συμμετέχοντα | Αριθμητική τιμή |
| age | Ηλικία | 50 - 80 |
| gender | Φύλο | 1=Male<br>2=Female |
| race | Φυλή | 1=White<br>2=Black or African-American<br>3=Asian<br>4=American Indian or Alaskan Native<br>5=Native Hawaiian or Other Pacific Islander<br>6=More than one race<br>7=Participant refused to answer<br>95=Missing data form - form is not expected to ever be completed<br>99=Unknown/ decline to answer |
| can_scr | Πρώτη επιβεβαιωμένη διάγνωση καρκίνου. Υποδεικνύει εάν ο καρκίνος ακολούθησε θετικό/αρνητικό έλεγχο ή εμφανίστηκε μετά τα έτη προσυμπτωματικού ελέγχου. | 0=No Cancer<br>1=Positive Screen<br>2=Negative Screen<br>3=Missed Screen<br>4=Post Screening |
| lc_grade | Βαθμός καρκίνου | 1=Well Differentiated: Grade I<br>2=Moderately Differentiated: Grade II |

| | | |
|---|---|---|
| | | 3=Poorly Differentiated: Grade III<br>4=Undifferentiated: Grade IV<br>9=Unknown |
| cigsmok | Η κατάσταση του καπνιστή (πρώην ή νυν) τη χρονική στιγμή t. Πρώην καπνιστές θεωρούνται όσοι έχουν διακόψει εντός 15 ετών. | 0=Former<br>1=Current |
| de_stag | Στάδιο καρκίνου στην πρώτη επιβεβαιωμένη διάγνωση | 110=Stage IA<br>120=Stage IB<br>210=Stage IIA<br>220=Stage IIB<br>310=Stage IIIA<br>320=Stage IIIB<br>400=Stage IV<br>888=TNM not available<br>900=Occult Carcinoma<br>994=Carcinoid, cannot be assessed<br>999=Unknown, cannot be assessed |
| first_lc | Πρώτος καρκίνος του πνεύμονα | 0=No<br>1=Yes |
| lc_behav | Συμπεριφορά καρκίνου του πνεύμονα (π.χ. μεταστατικός) | 1=Borderline Malignancy<br>3=Invasive<br>6=Metastatic |

**Πίνακας 4-1: Περιγραφή Πεδίων**

Παραπάνω αναφέρεται η διαδικασία κατά την οποία ανακτήθηκαν και προετοιμάστηκαν τα προς ανάλυση δεδομένα. Επόμενο βήμα είναι η εισαγωγή τους στο σύστημα αρχείων του Hadoop, δηλαδή στο HDFS από το λειτουργικό σύστημα όπου βρίσκονται. Στη συνέχεια παρουσιάζονται αναλυτικά τα βήματα για τη διαδικασία.

1. Εκτέλεση της γραμμής εντολών με δικαιώματα διαχειριστή (σε Windows) και αλλαγή directory στον φάκελο \sbin.

   *cd C:\hadoop-3.3.0\sbin*

```
Microsoft Windows [Version 10.0.19045.2486]
(c) Microsoft Corporation. All rights reserved.

C:\windows\system32>cd C:\hadoop-3.3.0\sbin

C:\hadoop-3.3.0\sbin>
```

2. Format στο σύστημα των αρχείων ώστε να αρχικοποιηθεί το namenode. Αν εκτελεστεί ξανά χάνονται τα δεδομένα από το file system.

   *hdfs namenode -format*

```
C:\hadoop-3.3.0\sbin>hdfs namenode -format
2023-02-18 18:57:14,879 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
```

3. Εκκίνηση των dfs και yarn.

   *start-dfs*

   *start-yarn*

```
C:\hadoop-3.3.0\sbin>start-dfs

C:\hadoop-3.3.0\sbin>start-yarn
starting yarn daemons

C:\hadoop-3.3.0\sbin>
```

4. Δημιουργία directory στο HDFS που θα εισαχθούν τα δεδομένα.

   *hadoop fs -mkdir [input directory]*

```
C:\hadoop-3.3.0\sbin>hadoop fs -mkdir /input

C:\hadoop-3.3.0\sbin>
```

5. Εισαγωγή των δεδομένων από το λειτουργικό σύστημα στο directory που δημιουργήθηκε στο HDFS.

   *hadoop fs -put [source directory] [input directory]*

```
C:\hadoop-3.3.0\sbin>hadoop fs -put C:/v_lc_data.csv /input

C:\hadoop-3.3.0\sbin>
```

6. Προβολή περιεχομένου του HDFS και έλεγχος για την επιτυχή εισαγωγή των αρχείων στο directory.

   *hadoop fs -ls [input directory]/*

```
C:\hadoop-3.3.0\sbin>hadoop fs -ls /input/
Found 1 items
-rw-r--r--   1 maria supergroup      69071 2023-02-18 19:07 /input/v_lc_data.csv

C:\hadoop-3.3.0\sbin>
```

7. Προβολή των δεδομένων κάθε αρχείου.

   *hadoop dfs -cat [input directory]/[file name]*

```
C:\hadoop-3.3.0\sbin>hadoop dfs -cat /input/v_lc_data.csv
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
100012,61,2,1,1,3,1,110,1,3,110
100049,74,2,1,4,4,1,220,1,3,
100055,67,1,1,1,1,0,400,1,3,
100147,68,1,1,1,3,0,110,1,3,110
100158,65,1,1,1,4,0,110,1,3,110
100196,65,2,1,4,1,1,400,1,3,
100242,60,2,1,1,3,0,110,1,3,999
100242,60,2,1,1,3,0,110,0,3,999
```

## 4.4 Προγράμματα Hadoop

Για την εξαγωγή συμπερασμάτων σχετικά με τον καρκίνο του πνεύμονα υλοποιήθηκαν διαφορετικά προγράμματα τα οποία τρέχουν πάνω στο Hadoop. Τα προγράμματα υλοποιήθηκαν σε γλώσσα προγραμματισμού Java, χωρίς αυτό όμως να είναι αναγκαίο. Παρόλο που το ίδιο το Hadoop είναι γραμμένο σε Java επιτρέπει την υλοποίηση εφαρμογών και σε άλλες γλώσσες όπως Python ή C++. Όπως έχει αναφερθεί και παραπάνω στη φάση Map τα δεδομένα διαβάζονται και εισάγονται στη μορφή (key, value), ενώ στη φάση Reduce επιλύεται το πρόβλημα. Παρακάτω παρουσιάζονται συνοπτικά οι εφαρμογές και στο Παράρτημα Α βρίσκεται ολόκληρος ο κώδικας.

### *4.4.1 CountTotalRecords*

Με την πρώτη εφαρμογή βρίσκουμε το συνολικό αριθμό των εγγραφών του αρχείου .csv το οποίο μας είναι απαραίτητο σε επόμενους υπολογισμούς.

- **Φάση Map:** Διαβάζονται τα δεδομένα ανά id συμμετέχοντα και αποθηκεύονται στη μορφή (pid, 1).
- **Φάση Reduce:** Υλοποιείται η καταμέτρηση των pid.

### *4.4.2 AverageAgeCalculator*

Η συγκεκριμένη εφαρμογή υπολογίζει το μέσο όρο ηλικίας ανδρών και γυναικών που εμφανίζουν καρκίνο του πνεύμονα.

- **Φάση Map:** Διαβάζονται τα δεδομένα και αποθηκεύονται στη μορφή (gender, age).
- **Φάση Reduce:** Αφού έχουν ομαδοποιηθεί οι τιμές με βάση το φύλο στην προηγούμενη φάση, υπολογίζεται ο μέσος όρος.

### *4.4.3 PostScreeningCancer*

Η εφαρμογή αυτή υπολογίζει το ποσοστό των πρώην και νυν καπνιστών που εμφάνισαν καρκίνο μετά από έτη προσυμπτωματικού ελέγχου.

- **Φάση Map:** Διαβάζονται τα δεδομένα cigsmok και can_scr και αποθηκεύονται στη μορφή (smokerStatus, screenCancer)
- **Φάση Reduce:** Υπολογίζεται το ποσοστό ανά πρώην και νυν καπνιστή.

### *4.4.4 AgeRangeOfFirstLC*

Η υλοποίηση AgeRangeOfFirstLC υπολογίζει το εύρος ηλικίας με τα περισσότερα περιστατικά όπου εμφανίζεται για πρώτη φορά καρκίνος του πνεύμονα.

- **Φάση Map:** Διαβάζονται τα δεδομένα first_lc και age και αποθηκεύονται στη μορφή (first_lc, age)
- **Φάση Reduce:** Υπολογίζεται το εύρος ηλικίας και πιο συγκεκριμένα η δεκαετία κατά την οποία είναι πιο συχνή η πρώτη εμφάνιση του καρκίνου.

### *4.4.5 LCGradeByAge*

Η εφαρμογή αυτή υπολογίζει τον βαθμό (βλ. σελίδα 31) του καρκινικού όγκου ανάλογα με το εύρος ηλικίας.

- **Φάση Map:** Διαβάζονται τα δεδομένα lc_grade και age και αποθηκεύονται στη μορφή (lc_grade, age)
- **Φάση Reduce:** Υπολογίζεται ο συνολικός αριθμός των περιπτώσεων ανά βαθμό καρκίνου και εύρος ηλικίας.

## 4.5 Εκτέλεση Προγραμμάτων Hadoop

1. Εκτέλεση προγράμματος.

*hadoop jar [program directory] [input directory] [output direcrory]*

```
C:\hadoop-3.3.0\sbin>hadoop jar C:/Users/maria/AverageAgeCalculator/target/AverageAgeCalculator.jar
 org.example/AverageAgeCalculator /input /out
2023-02-18 19:15:41,277 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManag
er at /0.0.0.0:8032
2023-02-18 19:15:42,094 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not
performed. Implement the Tool interface and execute your application with ToolRunner to remedy this
.
```

2. Προβολή αποτελεσμάτων χωρίς αντιγραφή στο λειτουργικό σύστημα.

   *hadoop fs -cat [output directory]/**

```
C:\hadoop-3.3.0\sbin>hadoop fs -cat /out/*
The average age of Female is:   63
The average age of Male is:     64

C:\hadoop-3.3.0\sbin>
```

3. Εξαγωγή αποτελεσμάτων στο λειτουργικό σύστημα από το HDFS.

   *hadoop fs -copyToLocal [output directory] [destination directory]*

```
C:\hadoop-3.3.0\sbin>hadoop fs -copyToLocal /out /C:

C:\hadoop-3.3.0\sbin>
```

# 5 Επίλογος

## 5.1 Σύνοψη και συμπεράσματα

Σκοπός της παρούσας διπλωματικής ήταν να τονίσει τις εφαρμογές των Μεγάλων Δεδομένων στον τομέα της υγείας και κυρίως στην ογκολογία, αλλά και πως μπορούμε να εκμεταλλευτούμε τις δυνατότητες που προσφέρουν για συνολική βελτίωση της υγειονομικής περίθαλψης. Πιο συγκεκριμένα αναλύθηκε πλήρως η αξία και η προσφορά τους ώστε οι ειδικοί να μπορούν να επιτύχουν εξατομικευμένες θεραπείες για κάθε ασθενή, σε συνδυασμό πάντα με το ελάχιστο κόστος. Επιπλέον μελετήθηκε σε βάθος το πλαίσιο ανοιχτού κώδικα Apache Hadoop με στόχο την κατανόηση της λειτουργίας του και πώς συμβάλλει στην ανάλυση των δεδομένων.

Στη συνέχεια χρησιμοποιώντας ένα σύνολο δεδομένων σχετικών με τον καρκίνο του πνεύμονα υλοποιήθηκαν ορισμένες εφαρμογές με χρήση Hadoop από τις οποίες εξήχθησαν ορισμένα συμπεράσματα. Αρχικά δημιουργήθηκε μία εφαρμογή (CountTotalRecords.java) για τον υπολογισμό των συνολικών εγγραφών του δείγματος που ήταν απαραίτητος στη συνέχεια. Στην επόμενη υλοποίηση (AverageAgeCalculator.java) λαμβάνουμε το μέσο όρο ηλικίας ανδρών και γυναικών που εμφανίζουν καρκίνο του πνεύμονα. Διαπιστώνουμε ότι το όριο ηλικίας είναι πολύ

κοντά με τις γυναίκες να εμφανίζουν κατά μέσο όρο στην ηλικία των 63, ενώ οι άντρες των 64 ετών.

Ο όρος προσυμπτωματικός έλεγχος του καρκίνου του πνεύμονα αναφέρεται στην διαδικασία κατά την οποία το άτομο υποβάλλεται σε σχετικές εξετάσεις ώστε να γίνει διάγνωση της νόσου σε αρχικό στάδιο πριν εμφανιστούν συμπτώματα. Τα αποτελέσματα της τρίτης υλοποίησης (PostScreeningCancer.java) μας δίνουν τα ποσοστά των νυν και πρώην καπνιστών που εμφανίζουν καρκίνο μετά από έτη προσυμπτωματικού ελέγχου. Οι νυν καπνιστές φτάνουν το 25.95% ενώ οι πρώην το 15.63%, με τα δύο ποσοστά να μην είναι χαμηλά.

Δυστυχώς η θεραπεία ενός καρκίνου του πνεύμονα δεν εγγυάται ότι δεν θα ξανά εμφανιστεί. Η επόμενη υλοποίηση (AgeRangeOfFirstLC.java ) όμως μας δίνει το εύρος ηλικίας στο οποίο τα άτομα εμφανίζουν για πρώτη φορά καρκίνο του πνεύμονα. Η δεκαετία από 58 έως 68 ετών είναι η πιο κρίσιμη και η πιο συχνή για τα άτομα που νοσούν πρώτη φορά. Συνεπώς ο προσυμπτωματικός έλεγχος σε αυτές τις ηλικίες θα πρέπει να είναι τακτικός.

Στην τελευταία υλοποίηση (LCGradeByAge.java) μελετήθηκε ο βαθμός του καρκίνου σχετικά με την ηλικία κάθε ατόμου. Ο βαθμός του όγκου διαφέρει από το στάδιο του. Το στάδιο περιγράφει το μέγεθος του όγκου και την ταχύτητα με την οποία εξαπλώνεται. Από την άλλη πλευρά ο βαθμός καθορίζεται από την εικόνα των κυττάρων στο μικροσκόπιο και κατά πόσο εμφανίζουν ομοιότητες με τα φυσιολογικά. Ο βαθμός ξεκινά από G1 για τα κύτταρα που έχουν πολλές ομοιότητες με τα φυσιολογικά και φτάνει έως G4 για αδιαφοροποίητα κύτταρα τα οποία είναι τα πιο επιθετικά. Από την μελέτη των δεδομένων διαπιστώνεται ότι σε όλες τις ηλικίες υπερισχύει ο βαθμός G3 ακολουθεί ο βαθμός G2. Όμως στις ηλικίες από 60 έως 70 ετών οι περιπτώσεις είναι σχεδόν διπλάσιες από τις ηλικίες έως 60 ετών για τους βαθμούς G3 και G4. Σε ηλικίες άνω των 70 ετών τα νούμερα είναι πιο χαμηλά.

Συμπερασματικά απαιτείται η άμεση και συνεχής συνεργασία του ιατρικού και ερευνητικού προσωπικού με τους επαγγελματίες της πληροφορικής. Με τον τρόπο αυτό όχι μόνο θα διασφαλιστεί η ακρίβεια και η αξιοπιστία των δεδομένων αλλά θα αξιοποιηθούν με το βέλτιστο δυνατό τρόπο.

## 5.2 Μελλοντικές επεκτάσεις

Όσον αφορά τις μελλοντικές επεκτάσεις είναι σημαντικό στην έρευνα να συμπεριλαμβάνονται ολοένα και περισσότερα γονιδιωματικά δεδομένα, τα οποία μπορούν να αναλυθούν αποτελεσματικά από λογισμικά όπως το Hadoop για την ανακάλυψη μοτίβων. Επίσης κρίνεται απαραίτητο να βελτιωθεί η κοινή χρήση δεδομένων υγειονομικής περίθαλψης σε ιδρύματα και ερευνητές ώστε να μπορούν να αναλυθούν αποτελεσματικά μεγαλύτερα σύνολα δεδομένων. Οι εφαρμογές που υλοποιήθηκαν μπορούν να βελτιωθούν εάν προστεθούν Combiner μέθοδοι συμβάλλοντας στην συνολική απόδοση. Τέλος οι έρευνες οφείλουν να επικεντρωθούν στην υλοποίηση εργαλείων που θα διαχειρίζονται με τη μέγιστη ασφάλεια τα δεδομένα υγειονομικής περίθαλψης.

# Παράρτημα Α

## 1. CountTotalRecords.java

```
package org.example;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

import java.io.IOException;

public class CountTotalRecords {
   public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {

      private Text id = new Text();
      private IntWritable one = new IntWritable(1);

      @Override
      public void map(LongWritable key, Text value, Context context)
            throws IOException, InterruptedException {

         String line = value.toString();
         String[] str = line.split(",");

         if (str.length > 0) {
            if (str[0].matches("[+-]?([0-9]*[.])?[0-9]+")) {
               context.write(id, one);
            }
         }
      }
   }

 public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {

    private IntWritable totalRecords = new IntWritable();

    @Override
    public void reduce(Text key, Iterable<IntWritable> values, Context context)
          throws IOException, InterruptedException {

       int counter = 0;

       for(IntWritable val : values){
          counter++;
       }

       totalRecords.set(counter);
       context.write(new Text("Total Records: "), totalRecords);
```

```
        }
    }

    public static void main(String[] args) throws Exception {

        if (args.length != 2) {
            System.err.println("Usage: CountTotalRecords <InPath> <OutPath>");
            System.exit(2);
        }

        Configuration conf = new Configuration();

        Job job = Job.getInstance(conf, "CountTotalRecords");

        job.setJarByClass(CountTotalRecords.class);
        job.setMapperClass(Map.class);
        job.setReducerClass(Reduce.class);
        job.setNumReduceTasks(1);

        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);

        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));

        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}
```

## 2. AverageAgeCalculator.java

```
package org.example;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

import java.io.IOException;

public class AverageAgeCalculator {
    public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {

        private Text gender = new Text();
        private IntWritable age = new IntWritable();

        @Override
        public void map(LongWritable key, Text value, Context context) throws IOException,
InterruptedException {
```

```
        String line = value.toString();
        String[] str = line.split(",");

        if (str.length > 0) {
           if(str[2].equals("1")){
              gender.set("Male");
           }
           else
              gender.set("Female");

           if (str[1].matches("[+-]?([0-9]*[.])?[0-9]+")) {
              float i = Float.parseFloat(str[1]);
              int j = Math.round(i);
              age.set(j);
           }
        }

        context.write(gender, age);
     }
  }

  public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {

     private IntWritable averageAge = new IntWritable();
     @Override
     public void reduce(Text key, Iterable<IntWritable> values, Context context)
           throws IOException, InterruptedException {

        int sumAges = 0;
        int counter = 0;

        for(IntWritable val : values){
           counter += 1;
           sumAges += val.get();
        }

        float average = sumAges/counter;

        averageAge.set(RoundAge(average));
        context.write(new Text("The average age of " + key + " is: "), averageAge);
     }
  }

  public static Integer RoundAge(Float f){
     int age = Math.round(f);
     return age;
  }

  public static void main(String[] args) throws Exception {

     if (args.length != 2) {
        System.err.println("Usage: AverageCalculator <InPath> <OutPath>");
        System.exit(2);
     }

     Configuration conf = new Configuration();

     Job job = Job.getInstance(conf, "AverageAgeCalculator");
```

```
        job.setJarByClass(AverageAgeCalculator.class);
        job.setMapperClass(Map.class);
        job.setReducerClass(Reduce.class);
        job.setNumReduceTasks(1);

        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);

        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));

        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}
```

### 3. PostScreeningCancer.java

```
package org.example;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import java.io.IOException;

public class PostScreeningCancer {
    public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {

        private Text smokerStatus = new Text();
        private IntWritable screenCancer = new IntWritable();

        @Override
        public void map(LongWritable key, Text value, Context context) throws IOException,
InterruptedException {

            String line = value.toString();
            String[] str = line.split(",");

            if (str.length > 0) {
                if (str[6].equals("1")) {
                    smokerStatus.set("Current");
                } else
                    smokerStatus.set("Former");

                if (str[4].matches("[+-]?([0-9]*[.])?[0-9]+")) {
                    screenCancer.set(Integer.parseInt(str[4]));
                }
            }

            context.write(smokerStatus, screenCancer);
        }
    }
```

```
    public static class Reduce extends Reducer<Text, IntWritable, Text, DoubleWritable> {

        private DoubleWritable finalPercent = new DoubleWritable();

        @Override
        public void reduce(Text key, Iterable<IntWritable> values, Context context)
                throws IOException, InterruptedException {

            int totalSmokers = 2150;
            int countPostScreening = 0;

            for(IntWritable val : values){
                if(val.get() == 4)
                    countPostScreening++;
            }

            double percent = ((double) countPostScreening/totalSmokers)*100.0;
            finalPercent.set(StringFormatter(percent));

            context.write(new Text("The percentage of " + key + " Smokers whose cancer occurred after the
screening years is: "), finalPercent);
        }
    }

    public static double StringFormatter(Double d){
        String str = String.format("%.2f", d);
        double doubleValue = Double.parseDouble(str.replaceAll(",","."));

        return doubleValue;
    }

    public static void main(String[] args) throws Exception {

        if (args.length != 2) {
            System.err.println("Usage: PostScreeningCancer <InPath> <OutPath>");
            System.exit(2);
        }

        Configuration conf = new Configuration();

        Job job = Job.getInstance(conf, "PostScreeningCancer");

        job.setJarByClass(PostScreeningCancer.class);
        job.setMapperClass(Map.class);
        job.setReducerClass(Reduce.class);
        job.setNumReduceTasks(1);

        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);

        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));

        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}
```

### 4. AgeRangeOfFirstLC.java

```
package org.example;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

import java.io.IOException;
import java.util.*;
import java.util.function.Function;
import java.util.stream.Collectors;

public class AgeRangeOfFirstLC {
   public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {

      private Text first_lc = new Text();
      private IntWritable age = new IntWritable();

      @Override
      public void map(LongWritable key, Text value, Context context) throws IOException,
InterruptedException {

         String line = value.toString();
         String[] str = line.split(",");

         if (str.length > 0) {
            if (str[8].equals("1")) {
               first_lc.set("First LC");
            } else
               first_lc.set("Not First LC");

            if (str[1].matches("[+-]?([0-9]*[.])?[0-9]+")) {
               age.set(Integer.parseInt(str[1]));
            }
         }

         context.write(first_lc, age);
      }
   }

   public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {

      private IntWritable ageRange = new IntWritable();

      @Override
      public void reduce(Text key, Iterable<IntWritable> values, Context context)
            throws IOException, InterruptedException {

         int counter = 0;
         List<Integer> ages = new ArrayList<>();
```

```
        for(IntWritable val : values){
            counter++;
            ages.add(val.get());
        }

        java.util.Map<Integer, Integer> counts = ages.stream()
                .collect(Collectors.groupingBy(Function.identity(), Collectors.summingInt(e -> 1)));

        List<java.util.Map.Entry<Integer, Integer>> list = new LinkedList<>(counts.entrySet());

        Collections.sort(list, (o1, o2) -> (o2.getValue()).compareTo(o1.getValue()));

        HashMap<Integer, Integer> countFirstLC = new LinkedHashMap<>();

        for (java.util.Map.Entry<Integer,Integer> l : list){
            countFirstLC.put(l.getKey(), l.getValue());
        }

        List<Integer> keys = new ArrayList<>();
        countFirstLC.keySet().forEach(k -> keys.add(k));

        List<Integer> firstKeys = keys.stream()
                .limit(10)
                .sorted(Comparator.reverseOrder())
                .collect(Collectors.toList());

        ageRange.set(firstKeys.get(firstKeys.size()-1));
        context.write(new Text(key + " from the age of"), ageRange);
        ageRange.set(firstKeys.get(0));
        context.write(new Text("to the age of"), ageRange);
    }
}

    public static void main(String[] args) throws Exception {

        if (args.length != 2) {
            System.err.println("Usage: AgeRangeOfFirstLC <InPath> <OutPath>");
            System.exit(2);
        }

        Configuration conf = new Configuration();

        Job job = Job.getInstance(conf, "AgeRangeOfFirstLC");

        job.setJarByClass(AgeRangeOfFirstLC.class);
        job.setMapperClass(Map.class);
        job.setReducerClass(Reduce.class);
        job.setNumReduceTasks(1);

        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);

        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));

        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}
```

## 5. LCGradeByAge.java

```
package org.example;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

import java.io.IOException;
import java.util.*;
import java.util.function.Function;
import java.util.stream.Collectors;
public class LCGradeByAge {
   public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {

      private Text lc_grade = new Text();
      private IntWritable age = new IntWritable();

      @Override
      public void map(LongWritable key, Text value, Context context)
            throws IOException, InterruptedException {

         String line = value.toString();
         String[] str = line.split(",");

         if (str.length > 0) {
            if (str[5].equals("1")) {
               lc_grade.set("Well Differentiated (G1)");
            }
            if (str[5].equals("2")) {
               lc_grade.set("Moderately Differentiated (G2)");
            }
            if (str[5].equals("3")) {
               lc_grade.set("Poorly Differentiated (G3)");
            }
            if (str[5].equals("4")) {
               lc_grade.set("Undifferentiated (G4)");
            }
            if (str[5].equals("9")) {
               lc_grade.set("Missing or Undefined");
            }

            if (str[1].matches("[+-]?([0-9]*[.])?[0-9]+")) {
               float i = Float.parseFloat(str[1]);
               int j = Math.round(i);
               age.set(j);
            }
         }
         context.write(lc_grade, age);
      }
```

```
}

public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {

   @Override
   public void reduce(Text key, Iterable<IntWritable> values, Context context)
         throws IOException, InterruptedException {

      List<String> keySet1 = new ArrayList<>();
      List<String> keySet2 = new ArrayList<>();
      List<String> keySet3 = new ArrayList<>();

      for (IntWritable val : values) {
         if(val.get() <= 60) {
            keySet1.add(key.toString());
         }
         if(val.get() > 60 && val.get() <= 70) {
            keySet2.add(key.toString());
         }
         if(val.get() > 70) {
            keySet3.add(key.toString());
         }
      }

      context.write(new Text("----------"), new IntWritable());

      java.util.Map<String, Integer> sumGradesKeySet1 =
            keySet1.stream()
                  .collect(Collectors.groupingBy(Function.identity(), Collectors.summingInt(e -> 1)));

      List<java.util.Map.Entry<String, Integer>> listKeySet1
            = new LinkedList<>(sumGradesKeySet1.entrySet());

         for(java.util.Map.Entry<String, Integer> l : listKeySet1){
            context.write(new Text(l.getKey() + " for ages up to 60: "), new IntWritable(l.getValue()));
         }

         java.util.Map<String, Integer> sumGradesKeySet2 =
               keySet2.stream()
                     .collect(Collectors.groupingBy(Function.identity(), Collectors.summingInt(e -> 1)));

         List<java.util.Map.Entry<String, Integer>> listKeySet2
               = new LinkedList<>(sumGradesKeySet2.entrySet());

         for(java.util.Map.Entry<String, Integer> l : listKeySet2){
            context.write(new Text(l.getKey() + " for ages over 60 and up to 70: "), new
IntWritable(l.getValue()));
         }

         java.util.Map<String, Integer> sumGradesKeySet3 =
               keySet3.stream()
                     .collect(Collectors.groupingBy(Function.identity(), Collectors.summingInt(e -> 1)));

         List<java.util.Map.Entry<String, Integer>> listKeySet3 =
               new LinkedList<>(sumGradesKeySet3.entrySet());

         for(java.util.Map.Entry<String, Integer> l : listKeySet3){
            context.write(new Text(l.getKey() + " for ages over to 70: "), new IntWritable(l.getValue()));
         }
```

```
    }
}

    public static void main(String[] args) throws Exception {

        if (args.length != 2) {
            System.err.println("Usage: LCGradeByAge <InPath> <OutPath>");
            System.exit(2);
        }

        Configuration conf = new Configuration();

        Job job = Job.getInstance(conf, "LCGradeByAge");

        job.setJarByClass(LCGradeByAge.class);
        job.setMapperClass(Map.class);
        job.setReducerClass(Reduce.class);
        job.setNumReduceTasks(1);

        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);

        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));

        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}
```

# Βιβλιογραφία


Adibuzzaman, M. *et al.* (2018) 'Big data in healthcare – the promises, challenges and opportunities from a research perspective: A case study with a model database', *AMIA Annual Symposium Proceedings*, 2017, pp. 384–392.

Alonso, S.G. *et al.* (2017) 'A Systematic Review of Techniques and Sources of Big Data in the Healthcare Sector', *Journal of Medical Systems*, 41(11), p. 183. Available at: https://doi.org/10.1007/s10916-017-0832-2.

'Apache Hadoop' (2023a) *Wikipedia*. Available at: https://en.wikipedia.org/w/index.php?title=Apache_Hadoop&oldid=1137256045#History (Accessed: 12 February 2023).

*Apache Hadoop* (2023b). Available at: https://hadoop.apache.org/ (Accessed: 12 February 2023).

Batko, K. and Ślęzak, A. (2022) 'The use of Big Data Analytics in healthcare', *Journal of Big Data*, 9(1), p. 3. Available at: https://doi.org/10.1186/s40537-021-00553-4.

Dash, S. *et al.* (2019) 'Big data in healthcare: management, analysis and future prospects', *Journal of Big Data*, 6(1), p. 54. Available at: https://doi.org/10.1186/s40537-019-0217-0.

Garapati, S.L. and Garapati, D.S. (2018) 'Application of Big Data Analytics: An Innovation in Health Care'.

*Hadoop Distributed File System (HDFS) for Big Data Projects* (2016) *dummies*. Available at: https://www.dummies.com/article/technology/information-technology/data-science/big-data/hadoop-distributed-file-system-hdfs-for-big-data-projects-167408/ (Accessed: 12 February 2023).

Harb, H. *et al.* (2021) 'A Sensor-Based Data Analytics for Patient Monitoring in Connected Healthcare Applications', *IEEE Sensors Journal*, 21(2), pp. 974–984. Available at: https://doi.org/10.1109/JSEN.2020.2977352.

HealthITAnalytics (2017) *Understanding the Many V's of Healthcare Big Data Analytics*, *HealthITAnalytics*. Available at: https://healthitanalytics.com/news/understanding-the-many-vs-of-healthcare-big-data-analytics (Accessed: 12 February 2023).

Jha, B.K., Sivasankari, G.G. and Venugopal, K.R. (2020) 'Fraud Detection and Prevention by using Big Data Analytics', in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*. *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 267–274. Available at: https://doi.org/10.1109/ICCMC48092.2020.ICCMC-00050.

*JobTracker - HADOOP2 - Apache Software Foundation* (2019). Available at: https://cwiki.apache.org/confluence/display/HADOOP2/JobTracker (Accessed: 12 February 2023).

Manager, M.M., Product Marketing (2022) *The 7 V's of Big Data, impact.com*. Available at: https://impact.com/marketing-intelligence/7-vs-big-data/ (Accessed: 19 March 2023).

'MapReduce' (2023) *Wikipedia*. Available at: https://en.wikipedia.org/w/index.php?title=MapReduce&oldid=1138885793#Map_function (Accessed: 12 February 2023).

Mehta, N. and Pandit, A. (2018) 'Concurrence of big data analytics and healthcare: A systematic review', *International Journal of Medical Informatics*, 114, pp. 57–65. Available at: https://doi.org/10.1016/j.ijmedinf.2018.03.013.

Memos, V. *et al.* (2021) 'An Enhanced and Secure Cloud Infrastructure for e-Health Data Transmission', *Wireless Personal Communications*, 117. Available at: https://doi.org/10.1007/s11277-019-06874-1.

Minopoulos, G. *et al.* (2022) 'Exploitation of Emerging Technologies and Advanced Networks for a Smart Healthcare System', *Applied Sciences*, 12, p. 5859. Available at: https://doi.org/10.3390/app12125859.

Minopoulos, G. *et al.* (2023) 'A Medical Image Visualization Technique Assisted with AI-Based Haptic Feedback for Robotic Surgery and Healthcare', *Applied Sciences*, 13, p. 3592. Available at: https://doi.org/10.3390/app13063592.

Plageras, A.P. *et al.* (2017) 'Efficient Large-scale Medical Data (eHealth Big Data) Analytics in Internet of Things', in *2017 IEEE 19th Conference on Business Informatics (CBI). 2017 IEEE 19th Conference on Business Informatics (CBI)*, pp. 21–27. Available at: https://doi.org/10.1109/CBI.2017.3.

Raghupathi, W. and Raghupathi, V. (2014) 'Big data analytics in healthcare: Promise and potential', *Health Information Science and Systems*, 2, p. 3. Available at: https://doi.org/10.1186/2047-2501-2-3.

Rights (OCR), O. for C. (2008) *The HIPAA Privacy Rule, HHS.gov*. Available at: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html (Accessed: 12 February 2023).

Sarwar, M.U. *et al.* (2017) 'A Survey of Big Data Analytics in Healthcare', *International Journal of Advanced Computer Science and Applications*, 8(6).

Stergiou, C. *et al.* (2018) 'Security, privacy & efficiency of sustainable Cloud Computing for Big Data & IoT', *Sustainable Computing: Informatics and Systems*, 19, pp. 174–184. Available at: https://doi.org/10.1016/j.suscom.2018.06.003.

Stergiou, C., Psannis, K. and Gupta, B.B. (2020) 'IoT-Based Big Data Secure Management in the Fog Over a 6G Wireless Network', *IEEE Internet of Things Journal*, PP. Available at: https://doi.org/10.1109/JIOT.2020.3033131.

Stergiou, C. and Psannis, K.E. (2017) 'Efficient and secure BIG data delivery in Cloud Computing', *Multimedia Tools and Applications*, 76(21), pp. 22803–22822. Available at: https://doi.org/10.1007/s11042-017-4590-4.

Stergiou, K. *et al.* (2022) 'A Machine Learning-Based Model for Epidemic Forecasting and Faster Drug Discovery', *Applied Sciences*, 12, p. 10766. Available at: https://doi.org/10.3390/app122110766.

*The V's of Big Data* (2020) *Marbella International University Centre*. Available at: https://miuc.org/vs-big-data/ (Accessed: 12 February 2023).

Tsai, C.J., Riaz, N. and Gomez, S.L. (2019) 'Big Data in Cancer Research: Real-World Resources for Precision Oncology to Improve Cancer Care Delivery', *Seminars in Radiation Oncology*, 29(4), pp. 306–310. Available at: https://doi.org/10.1016/j.semradonc.2019.05.002.

Wang, Y., Kung, L. and Byrd, T.A. (2018) 'Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations', *Technological Forecasting and Social Change*, 126, pp. 3–13. Available at: https://doi.org/10.1016/j.techfore.2015.12.019.

*What is Big Data and Why is it Important?* (no date) *Data Management*. Available at: https://www.techtarget.com/searchdatamanagement/definition/big-data (Accessed: 12 February 2023).

*Wiki - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki* (no date). Available at: https://wiki.cancerimagingarchive.net/ (Accessed: 27 March 2023).

Willems, S.M. *et al.* (2019) 'The potential use of big data in oncology', *Oral Oncology*, 98, pp. 8–12. Available at: https://doi.org/10.1016/j.oraloncology.2019.09.003.

'Μεταδεδομένα' (2022) *Βικιπαίδεια*. Available at: https://el.wikipedia.org/w/index.php?title=%CE%9C%CE%B5%CF%84%CE%B1%CE%B4%CE%B5%CE%B4%CE%BF%CE%BC%CE%AD%CE%BD%CE%B1&oldid=9799324 (Accessed: 12 February 2023).