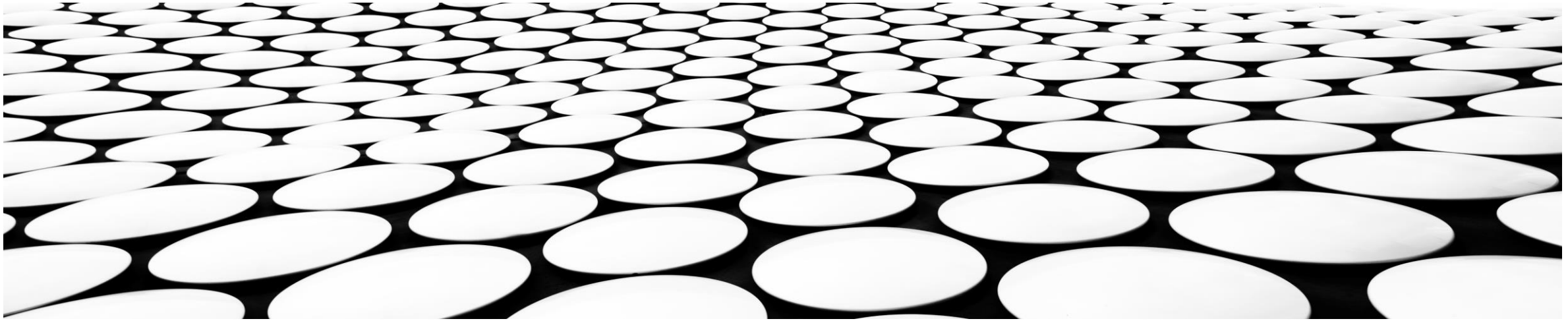
ΚΑΤΑΝΕΜΗΜΕΝΗ ΔΙΑΧΕΙΡΙΣΗ ΕΡΓΑΣΙΩΝ ΣΤΙΣ ΠΑΡΥΦΕΣ ΤΟΥ ΔΙΚΤΥΟΥ

ΔΗΜΗΤΡΗΣ ΦΛΟΥΡΗΣ

ΕΠΙΒΛΕΠΩΝ:

ΧΑΤΖΗΕΥΘΥΜΙΑΔΗΣ ΕΥΣΤΑΘΙΟΣ, ΚΑΘΗΓΗΤΗΣ ΕΚΠΑ

ΔΡ. ΚΟΛΟΜΒΑΤΣΟΣ ΚΩΝΣΤΑΝΤΙΝΟΣ

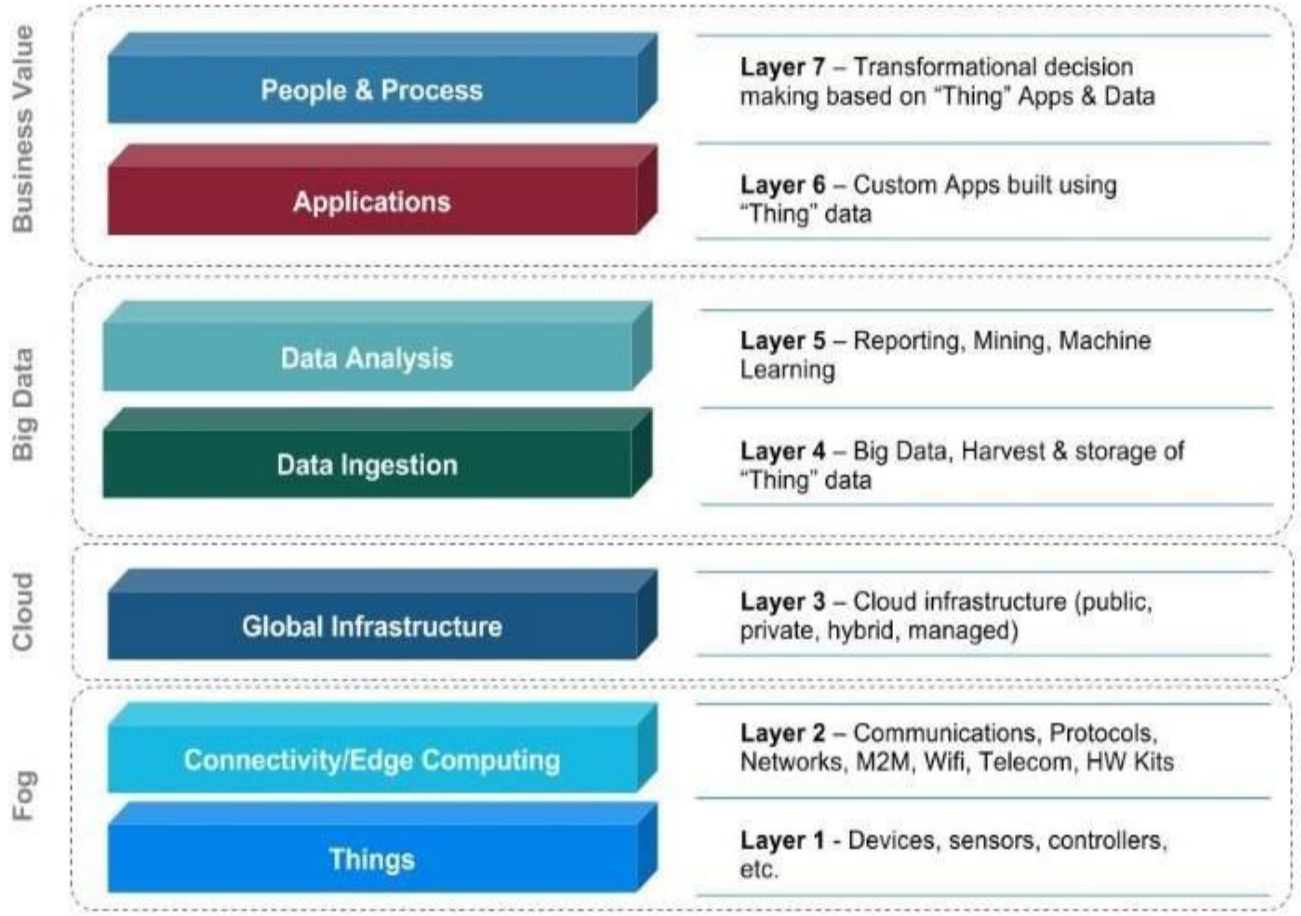


ΔΙΑΔΙΚΤΥΟ ΤΩΝ ΠΡΑΓΜΑΤΩΝ (INTERNET OF THINGS)

- Σύνολο συστημάτων δικτύου και επικοινωνίας
- Πρόσβαση σε βάσεις δεδομένων
- Παροχή ψηφιακών υπηρεσιών
- Τέλη 90’ από τον Άγγλο ερευνητή Kevin Ashton
- 1η απόπειρα σύνδεσης συσκευών μέσω RFID ετικέτας
- Μέχρι το 2021 θα υπάρχουν συνδεδεμένες 35 δισεκατομμύρια συσκευές παγκοσμίως

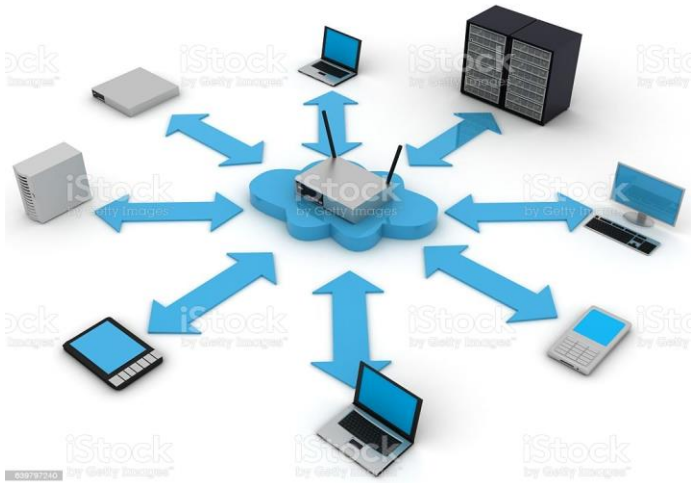
ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΔΙΑΔΙΚΤΥΟΥ ΤΩΝ ΠΡΑΓΜΑΤΩΝ

- Επίπεδο 1: Συσκευές
- Επίπεδο 2: Πρωτόκολλα επικοινωνίας
- Επίπεδο 3: Καθορισμός συνδεσιμότητας με τον έξω κόσμο του Διαδικτύου
- Επίπεδο 4: Καθαρισμός και αποθήκευση δεδομένων
- Επίπεδο 5: Ανάλυση δεδομένων - εξαγωγή συμπερασμάτων
- Επίπεδο 6: Λογισμικά που ελέγχουν τις συσκευές – πηγές δεδομένων
- Επίπεδο 7: Χρήστες – Επιχειρήσεις των συσκευών του ΔτΠ



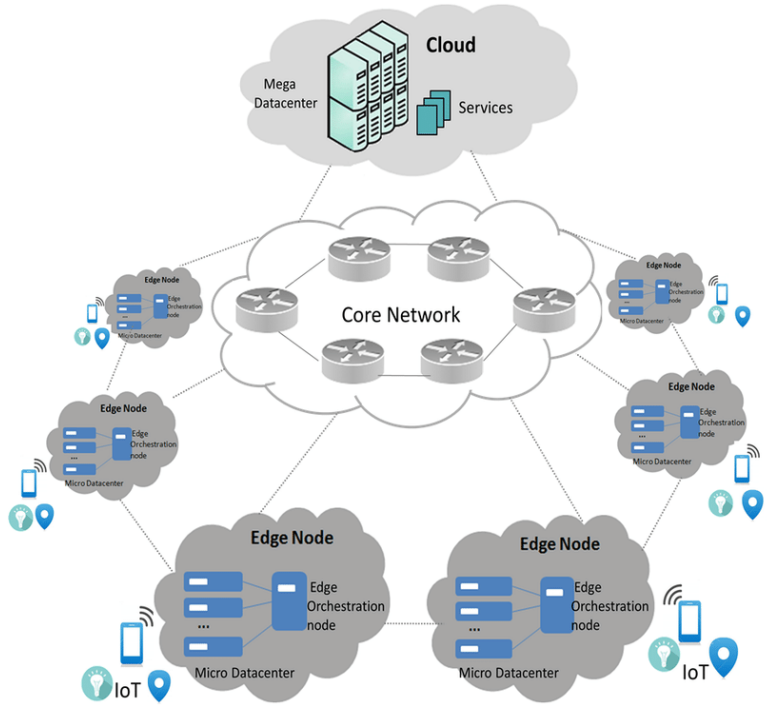
ΠΡΟΗΓΟΥΜΕΝΕΣ ΠΡΟΣΕΓΓΙΣΕΙΣ – CLOUD BASED

- Κεντροποιημένη Αρχιτεκτονική
- Συσκευές – Πηγές δεδομένων βρίσκονται στις άκρες του δικτύου.
- Κεντρικό υπολογιστικό νέφος όπου γίνεται η αποθήκευση και η επεξεργασία των δεδομένων.
- Μεγάλη κίνηση στο δίκτυο – καθυστέρηση στο χρόνο απόκριση του συστήματος
- Μεγάλο κόστος ανάπτυξης δικτύου
- Ευάλωτο σε κακόβουλες επιθέσεις ακόμα και σε διακοπή ρεύματος



EDGE COMPUTING

- Εμφανίστηκε με αφορμή των εφαρμογών που απαιτούσαν γρήγορες αποκρίσεις του συστήματος
- Κόμβοι που επικοινωνούν μεταξύ τους και αποτελούν το πυρήνα του δικτύου
- Οι κόμβοι βρίσκονται τοποθετημένοι κοντά στις πηγές δεδομένων
- Τα δεδομένα από τις πηγές αποθηκεύονται τοπικά στο πιο κοντινό κόμβο
- Στον κάθε κόμβο μπορούν να εκτελούνται διεργασίες για αιτήματα που αφορούν τα δεδομένα
- Σωστός διαχωρισμός των δεδομένων στους κόμβους
- Κόμβοι κοντά στις πηγές – λιγότερη κίνηση στο δίκτυο



ΠΛΕΟΝΕΚΤΗΜΑΤΑ EDGE COMPUTING

Ταχύτητα

- Τοποθετώντας του κόμβους κοντά στις πηγές δεδομένων μειώνεται η κίνηση που υπάρχει στο δίκτυο
- Διαχωρίζοντας τα δεδομένα στους κόμβους, παρέχουμε στο δίκτυο πρότερη γνώση με αποτέλεσμα να ξέρει που να κατευθύνει τα ερωτήματα που εισέρχονται σε αυτό.

Ασφάλεια - Αξιοπιστία

- Ο πυρήνας του δικτύου αποτελείται από πολλούς κόμβους - Κάποιος κόμβος τεθεί εκτός λειτουργίας, το δίκτυο μπορεί να συνεχίσει να λειτουργεί
- Τοπική αποθήκευση – λιγότερο εκτεθειμένα δεδομένα κατά την μεταφορά

Επεκτασιμότητα – Ευελιξία

- Αύξηση της εμβέλειας δικτύου με χαμηλό κόστος – προσθήκη νέων κόμβων
- Η προσθήκη νέων κόμβων δεν επηρεάζει τους υφιστάμενους

ΔΙΕΡΓΑΣΙΕΣ ΚΑΙ ΥΠΟΛΟΓΙΣΜΟΙ ΣΤΙΣ ΠΑΡΥΦΕΣ ΤΟΥ ΔΙΚΤΥΟΥ

Αλγόριθμοι Ταξινόμησης (Classification)

- Χρησιμοποιούνται κυρίως σε εφαρμογές μηχανικής μάθησης
- Δύο σύνολα δεδομένων: εκπαίδευσης αλγορίθμου και δοκιμής αλγορίθμου

Αλγόριθμοι Κατηγοριοποίησης (Clustering)

- Χωρίζουν τα δεδομένα σε κατηγορίες ανάλογα με τα δεδομένα που έχουν τη μεγαλύτερη βαρύτητα με βάση την έξοδο

Εντοπισμός Ανώμαλων Τιμών (Outlier Detection)

- Εντοπίζουν και αντικαθιστούν ελλιπής ή και ανώμαλες τιμές πιο συχνό τις αντικαθιστούν με το μέσο όρο της διάστασης

Μείωση Διαστάσεων σε Πολυδιάστατα Δεδομένα (Dimension Reduction)

- Εφαρμόζεται σε Big Data δεδομένα
- Λαμβάνουν υπόψιν μόνο τα σημαντικά χαρακτηριστικά του συνόλου των δεδομένων χωρίς να επηρεάζει το τελικό αποτέλεσμα

ΠΡΟΤΕΙΝΟΜΕΝΟΣ ΑΛΓΟΡΙΘΜΟΣ

- Προσομοίωση ενός Edge Computing δικτύου με πλήθος *n* Edge κόμβων.
- Μοιράζουμε στους κόμβους για αρχή από 100 εγγραφές δεδομένων στον κάθε ένα (σύνολο δεδομένων κόμβου)
- Εκτελούμε αλγόριθμους clustering πάνω στους κόμβους (K-Means) – Μείωση στο εύρος αναζήτησης
- Πηγές παράγουν και στέλνουν πολυδιάστατα δεδομένα σε τυχαίους κόμβους στη μορφή μονοδιάστατου πίνακα.
- Απόφαση Αποθήκευσης: P(Local Save) = 1 – *phi* και P(Remote Save) = *phi*
- Υπολογίζονται οι σημαντικές διαστάσεις και αναγνωρίζεται με βάση αυτών το cluster που ανήκουν τα νέα δεδομένα
- Οι κόμβοι που ανήκουν στο cluster που επιλέγηκε υπολογίζουν και στέλνουν τα στατιστικά τους μέσω μιας αναφοράς.
- Εκτελείται ένας μηχανισμός ανταμοιβής για το κάθε υποψήφιο κόμβο
- Κερδίζει ο κόμβος με το μεγαλύτερο σκορ

ΠΑΡΑΜΕΤΡΟΙ ΠΡΟΣΟΜΟΙΩΣΗΣ

- Ο αριθμός των edge κόμβων από τους οποίους θα αποτελείται το δίκτυο μας
- Το μέγεθος του συνόλου δεδομένων που θα χρησιμοποιήσουμε κατά την προσομοίωση
- Σε πόσα cluster θα χωριστούν οι κόμβοι του δικτύου
- Πόσες θα είναι οι σημαντικές διαστάσεις που θα λαμβάνονται υπόψιν κατά την επιλογή κόμβου
- Η πιθανότητα phi που επιλέγει τον τρόπο αποθήκευσης
- Κατώφλι πιθανότητας ο κόμβος να παρήγαγε τα νέα δεδομένα
- Κατώφλι κόστους μεταφοράς νέων δεδομένων

ΔΟΜΗ EDGE ΚΟΜΒΟΥ

ID: 0

Number of Rows: 1000

Number of Columns: 10

Dataset:

d0	d1	d2	d3	d4	d5	d6	d7	d8	d9
0.94	0.67	0.56	0.93	0.37	0.83	0.84	0.0	0.81	1
0.62	0.07	0.77	0.39	0.9	0.98	0.17	0.86	0.07	1
0.13	0.84	0.08	0.86	0.59	0.2	0.83	0.13	0.21	1
0.77	0.69	0.34	0.15	0.1	0.57	1.0	0.59	0.91	0
0.9	0.85	0.99	0.23	0.55	0.67	0.46	0.9	0.62	0
0.35	0.73	0.71	0.24	0.22	0.5	0.19	0.91	0.25	1

Average Value of each Dimension:

0.55	0.87	0.34	0.74	0.42	0.29	0.93	0.61	0.17	0.77

Most Important Dimension:

4	7	1	3	5	9	0	2	6	8

Report Time: 18:01:23

ΥΠΟΛΟΓΙΣΜΟΣ ΣΗΜΑΝΤΙΚΩΝ ΔΙΑΣΤΑΣΕΩΝ ΣΤΑ ΔΕΔΟΜΕΝΑ

- Χρησιμοποιώντας τη Chi-Square συνάρτηση $x^2 = \sum \frac{(<\tau\iota\mu\eta\ \delta\epsilon\iota\gamma\mu\alpha\tau o\varsigma> - <\alpha\nu\alpha\mu\epsilon\nu o\mu\epsilon\nu\eta\ \tau\iota\mu\eta>)^2}{\alpha\nu\alpha\mu\epsilon\nu\acute{o}\mu\epsilon\nu\eta\ \tau\iota\mu\eta}$
- Χρησιμοποιούμε την τελευταία στήλη των δεδομένων μας (ο η 1) ως το σημείο σύγκρισης για κάθε στήλη
- Βρίσκουμε μια τιμή για κάθε διάσταση και το πόσο σημαντική είναι με τα συγκεκριμένα δεδομένα
- Επιστρέφονται οι *k* πιο σημαντικές διαστάσεις

# Ημερών που	Θερμοκρασία Αέρα > 55	Επίπεδα Υγρασίας < 60	
Δεν Έβρεξε	42 45	33 30	75
Έβρεξε	18 15	7 10	25
	60	40	100

ΑΝΑΦΟΡΑ ΥΠΟΨΗΦΙΟΥ ΚΟΜΒΟΥ

- Την στέλνουν οι κόμβοι που βρίσκονται στο ίδιο cluster που επιλέγηκαν και τα νέα δεδομένα
- ID του κάθε κόμβου
- Ένας μονοδιάστατος πίνακας με τη μέση τιμή της κάθε διάστασης του συνόλου των δεδομένων του κόμβου
- Το κόστος που χρειάζεται για να σπαταλήσει ο τρέχων κόμβος για να στείλει τα δεδομένα στον υποψήφιο κόμβο
- Ο χρόνος που εκδόθηκαν τα πιο πάνω στατιστικά

ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ ΤΩΝ ΔΙΑΣΤΑΣΕΩΝ ΕΝΟΣ ΚΟΜΒΟΥ

- Ομοιότητα νέων δεδομένων με το σύνολο δεδομένων του κόμβου
- Κάθε κόμβος υπολογίζει τη μέση τιμή κάθε διάστασης από το σύνολο δεδομένων του
- Η τυπική απόκλιση δίνεται από τον τύπο

$$s = \sqrt{\frac{\left(d_0 - \overline{d_0}\right)^2 + \left(d_1 - \overline{d_1}\right)^2 + \ldots + \left(d_c - \overline{d_c}\right)^2}{c - 1}}$$

- Η τυπική απόκλιση του κάθε κόμβου αποθηκεύεται και στη συνέχεια χρησιμοποιείται για το μηχανισμό αμοιβής

ΜΗΧΑΝΙΣΜΟΣ ΑΝΤΑΜΟΙΒΗΣ ΥΠΟΨΗΦΙΩΝ ΚΟΜΒΩΝ

- Υπολογίζεται με βάση τη Gaussian πιθανότητα ο κόμβος να παρήγαγε τα νέα δεδομένα
- Υπολογίζεται το κόστος μεταφοράς από τον τρέχων κόμβο στο συγκεκριμένο υποψήφιο
- Υπολογίζεται η ομοιότητα των νέων δεδομένων με το σύνολο των δεδομένων του κόμβου
- Δίνονται οι ανάλογοι πόντοι στον υποψήφιο κόμβο

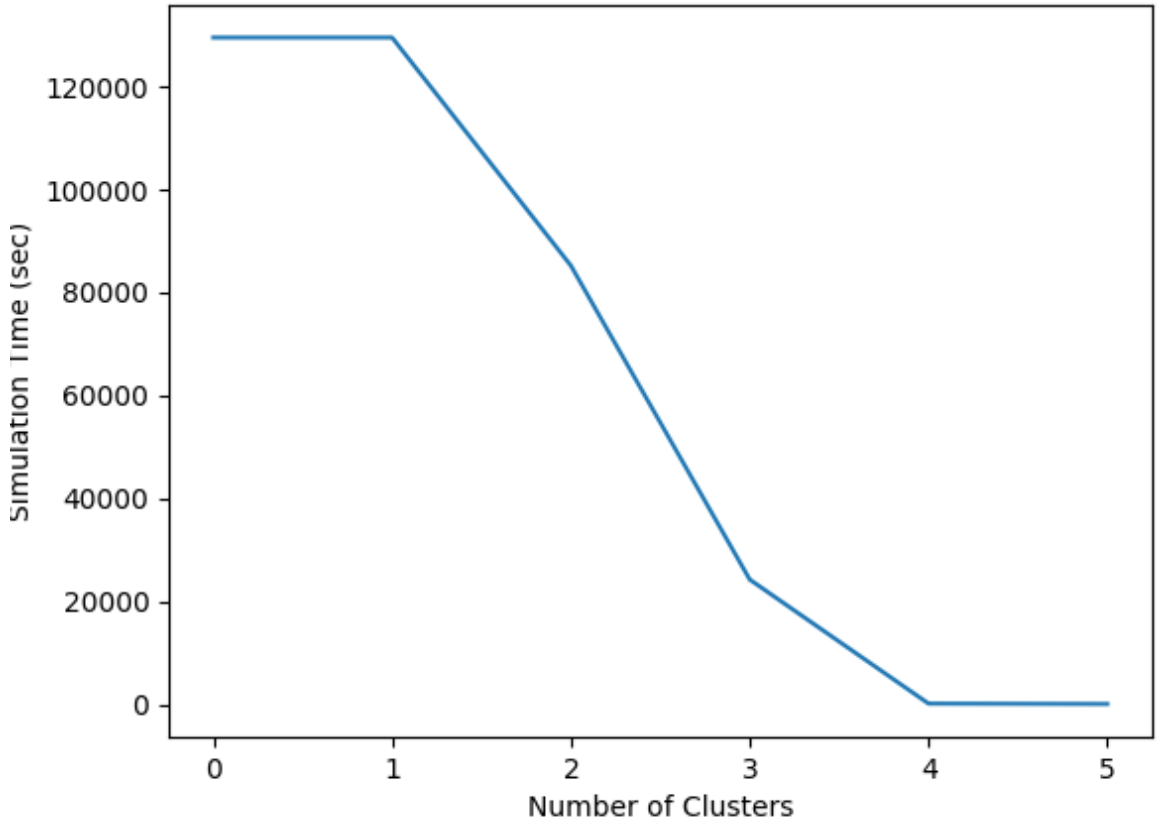
```
if Gauss_prob[node.id] > prob_thres:
    r += 0.5
if cost[node.id] > cost_thres:
    r += 0.15
if Similarity[node.id] < similar_thres:
    r += 0.35
```

- Τέλος υπολογίζεται ένα συνολικό σκορ ανταμοιβής υποψήφιου κόμβου χρησιμοποιώντας την εξίσωση:

$$total\ reward_{dn} = \frac{1}{1 + e^{(report_{time} + reward_r)}}$$

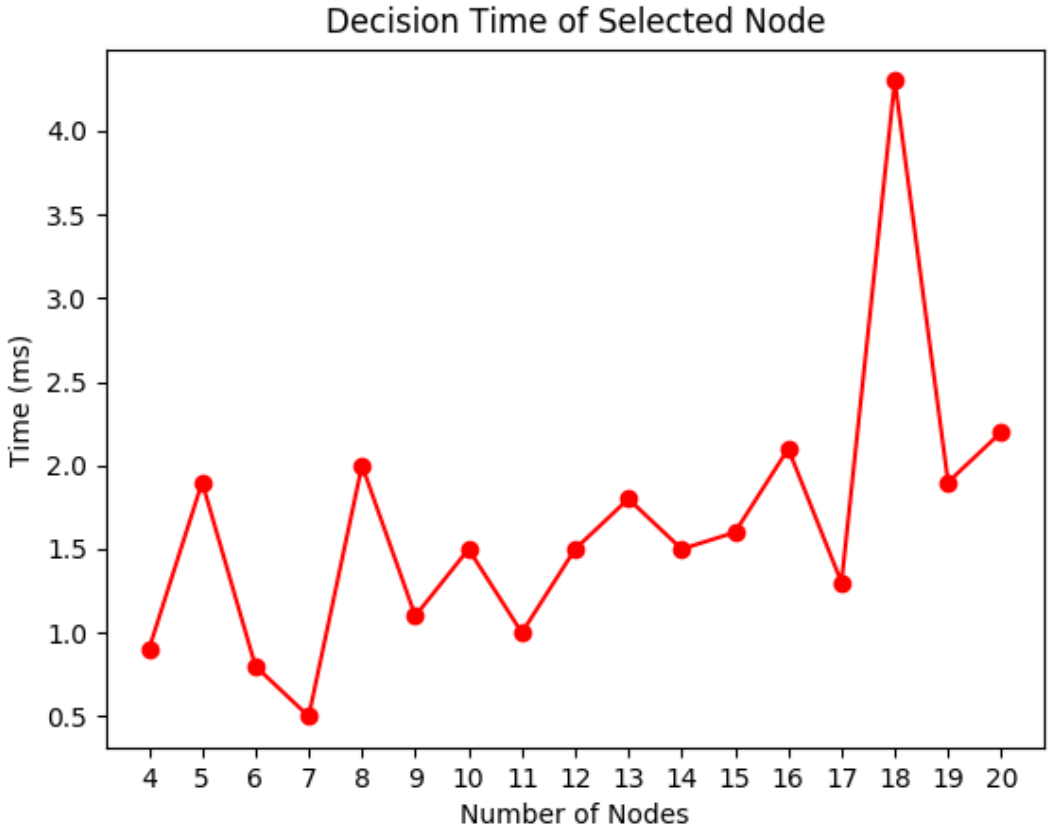
- Ο κόμβος με το μεγαλύτερο σκορ είναι κερδίζει.

ΠΕΙΡΑΜΑΤΙΚΗ ΑΠΟΤΙΜΗΣΗ
ΕΚΤΕΛΕΣΗ ΠΡΟΣΟΜΟΙΩΣΗΣ ΜΕ ΤΕΧΝΙΚΗ CLUSTERING ΚΟΜΒΩΝ



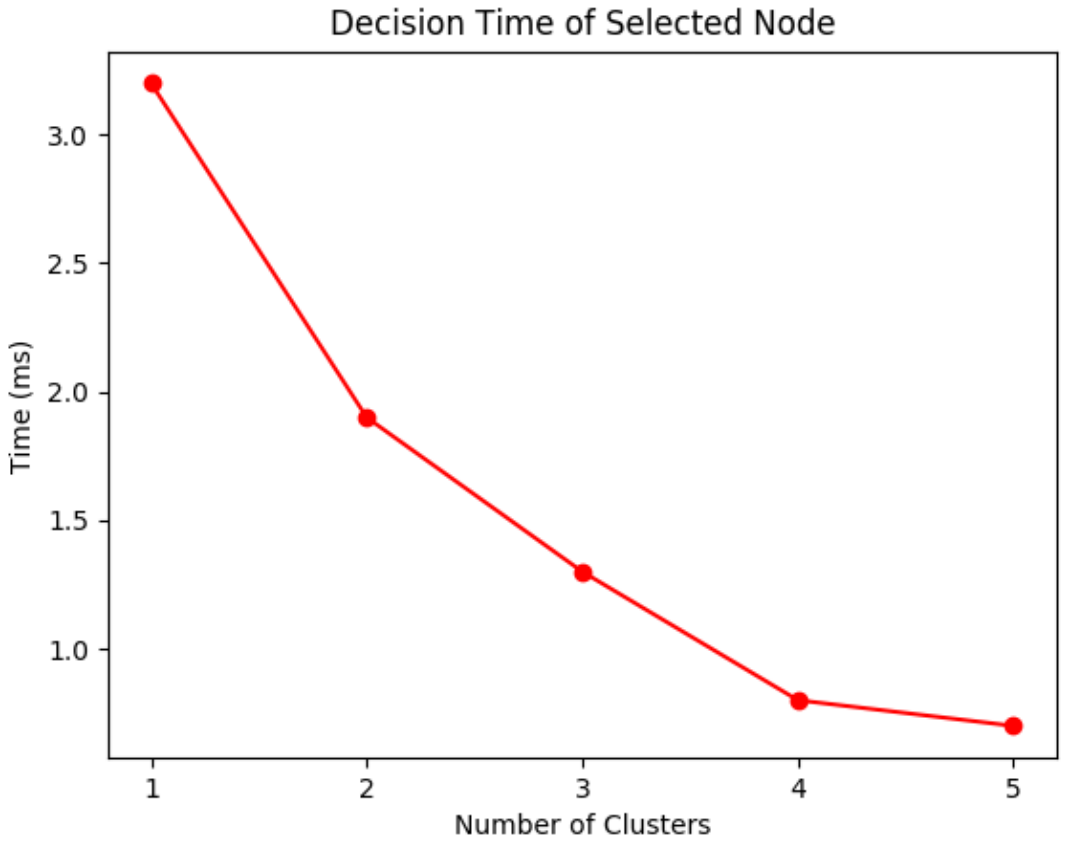
- Συνολικές Εγγραφές Δεδομένων: 25.000
- Συνολικές Διαστάσεις Δεδομένων: 10
- Πιθανότητα *phi = 0.75*
- Αριθμός Edge κόμβων δικτύου: 7
- Σημαντικές Διαστάσεις: 3

ΠΕΙΡΑΜΑΤΙΚΗ ΑΠΟΤΙΜΗΣΗ – ΧΡΟΝΟΣ ΑΠΟΦΑΣΗΣ ΥΠΟΨΗΦΙΟΥ ΚΟΜΒΟΥ ΣΥΝΟΛΙΚΟΣ ΑΡΙΘΜΟΣ ΚΟΜΒΩΝ ΔΙΚΤΥΟΥ



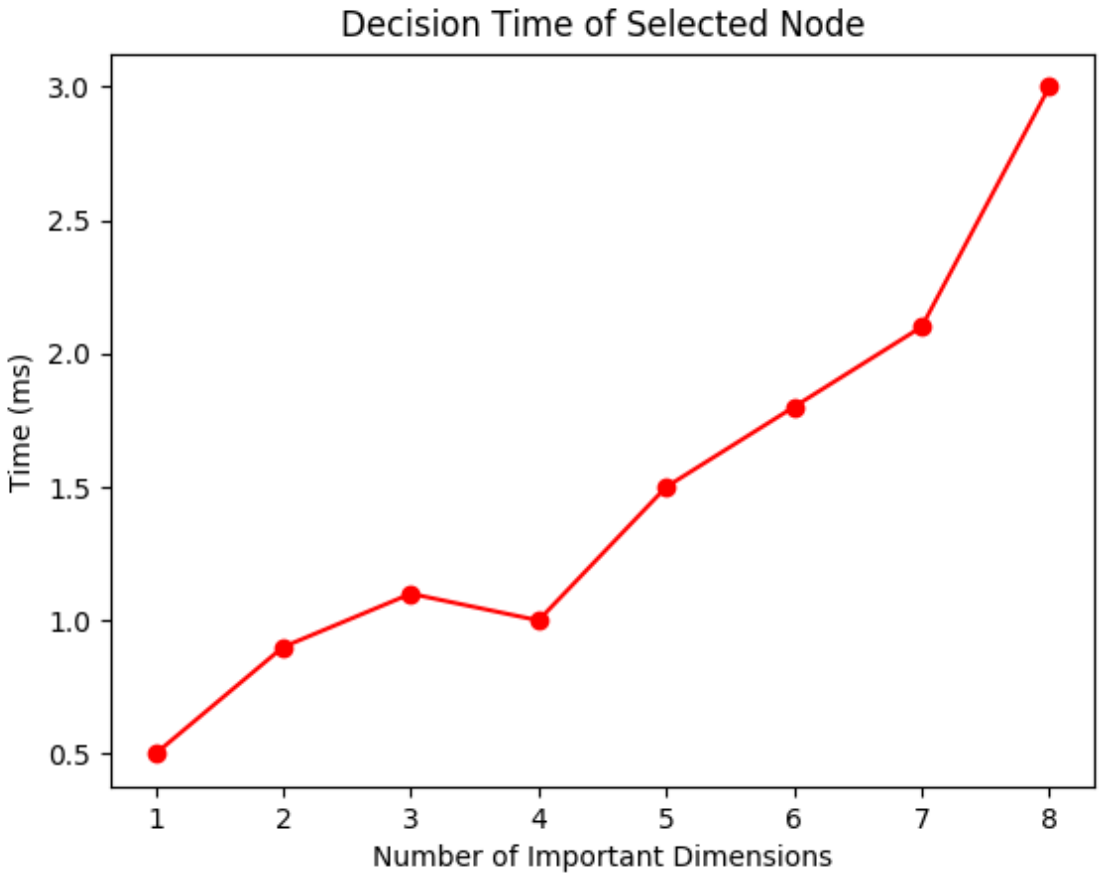
- Συνολικές Εγγραφές Δεδομένων: 25.000
- Συνολικές Διαστάσεις Δεδομένων: 10
- Πιθανότητα *phi = 0.75*
- Σημαντικές διαστάσεις: 3
- Αριθμός από Clusters: 3

ΠΕΙΡΑΜΑΤΙΚΗ ΑΠΟΤΙΜΗΣΗ – ΧΡΟΝΟΣ ΑΠΟΦΑΣΗΣ ΥΠΟΨΗΦΙΟΥ ΚΟΜΒΟΥ ΑΡΙΘΜΟΣ ΤΩΝ CLUSTERS



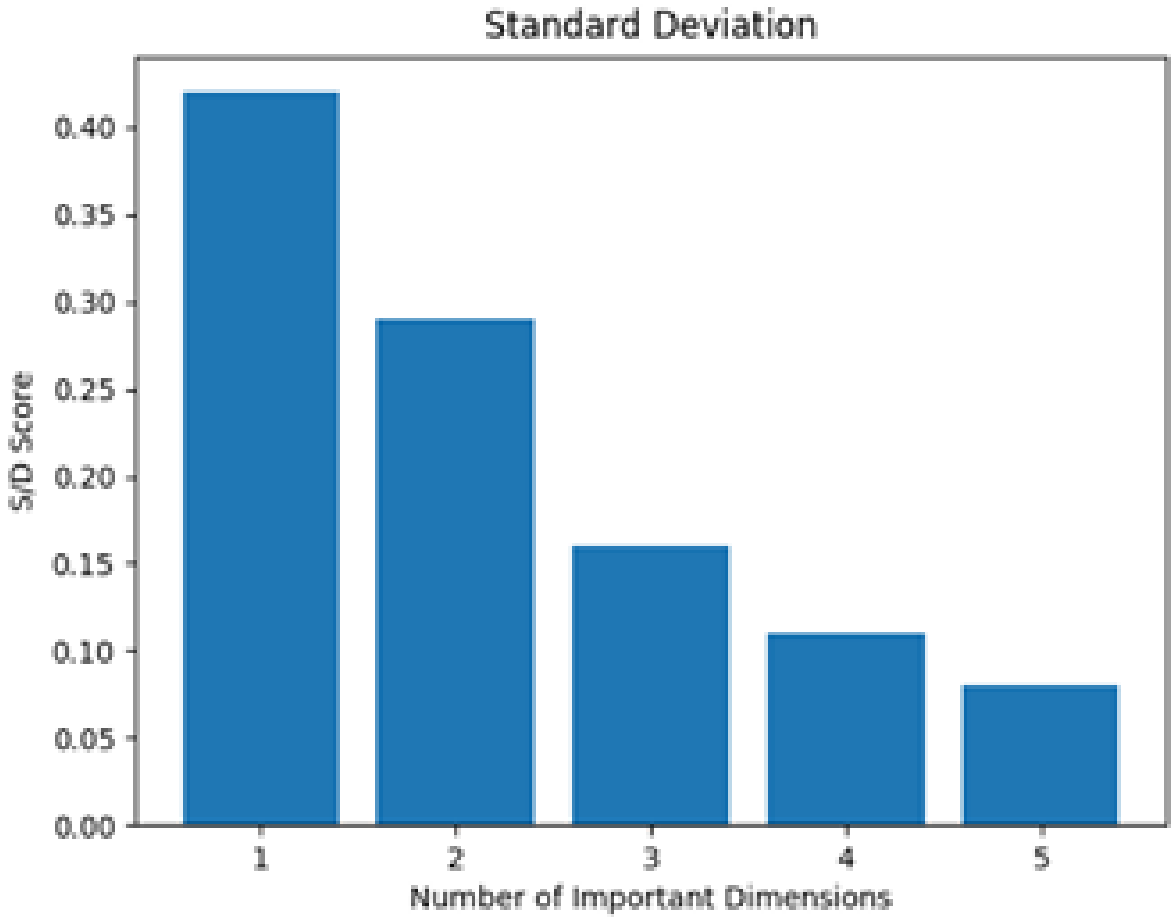
- Συνολικές Εγγραφές Δεδομένων: 25.000
- Συνολικές Διαστάσεις Δεδομένων: 10
- Πιθανότητα *phi = 0.75*
- Σημαντικές Διαστάσεις: 3
- *Αριθμός Edge Κόμβων δικτύου: 11*

ΠΕΙΡΑΜΑΤΙΚΗ ΑΠΟΤΙΜΗΣΗ – ΧΡΟΝΟΣ ΑΠΟΦΑΣΗΣ ΥΠΟΨΗΦΙΟΥ ΚΟΜΒΟΥ ΑΡΙΘΜΟΣ ΣΗΜΑΝΤΙΚΩΝ ΔΙΑΣΤΑΣΕΩΝ ΣΤΑ ΔΕΔΟΜΕΝΑ



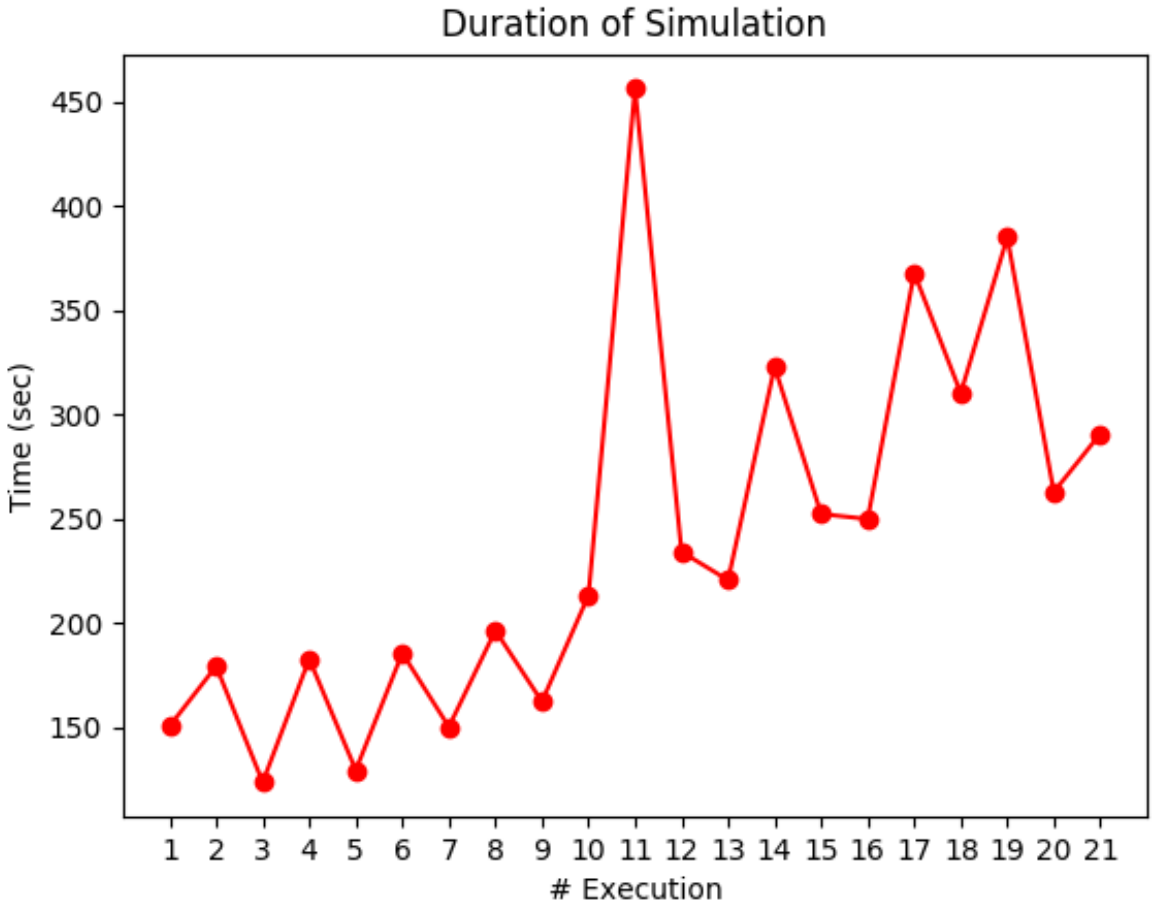
- Συνολικές Εγγραφές Δεδομένων: 25.000
- Συνολικές Διαστάσεις Δεδομένων: 10
- Πιθανότητα *phi = 0.75*
- Αριθμός Clusters: 5
- *Αριθμός Edge Κόμβων δικτύου: 11*

ΠΕΙΡΑΜΑΤΙΚΗ ΑΠΟΤΙΜΗΣΗ - ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ ΔΕΔΟΜΕΝΩΝ ΚΟΜΒΟΥ ΑΡΙΘΜΟΣ ΣΗΜΑΝΤΙΚΩΝ ΔΙΑΣΤΑΣΕΩΝ ΣΤΑ ΔΕΔΟΜΕΝΑ



- Ίδια εγγραφή δεδομένων για όλες τις περιπτώσεις
- Συνολικές Διαστάσεις Δεδομένων: 10
- Πιθανότητα *phi = 0.75*
- Αριθμός Clusters: 5
- *Αριθμός Edge Κόμβων δικτύου: 11*

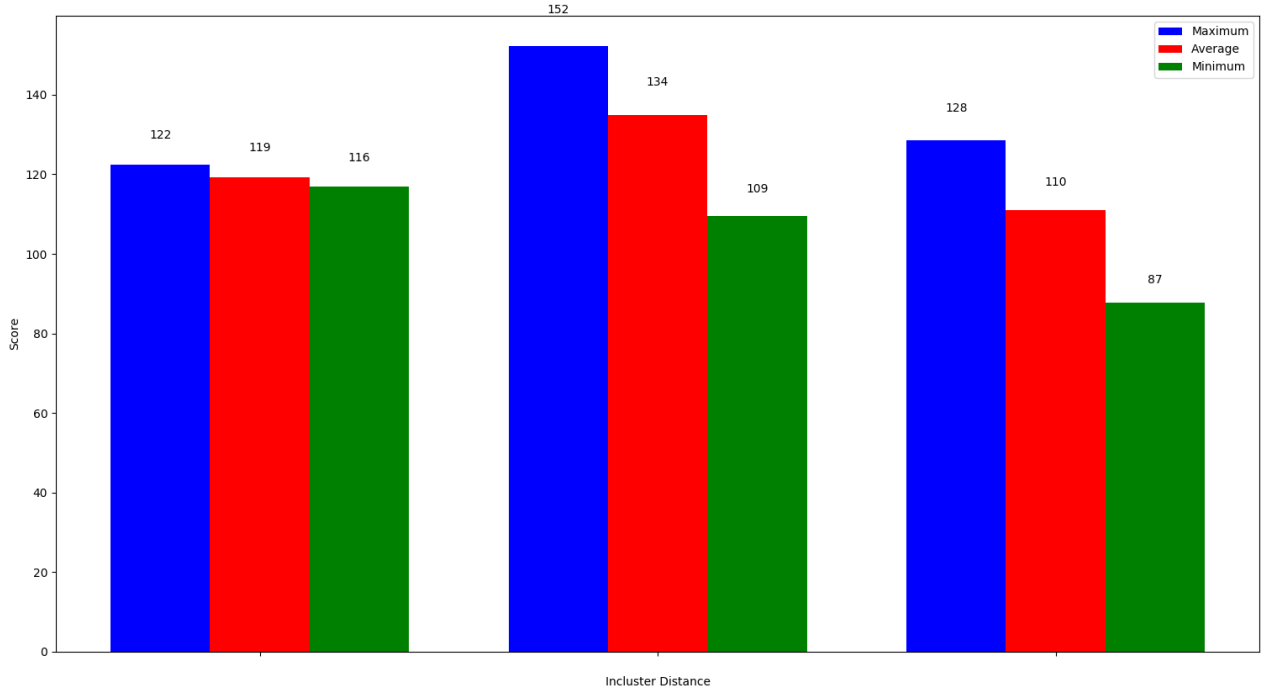
ΧΡΟΝΟΙ ΕΚΤΕΛΕΣΕΩΝ ΠΡΟΣΟΜΟΙΩΣΕΩΝ



- Συνολικές Εγγραφές Δεδομένων: 25.000
- Συνολικές Διαστάσεις Δεδομένων: 10
- Πιθανότητα *phi = 0.75*

A/A	Nodes	Clusters	Important Dimensions	Execution Time(s)
3	7	3	3	123.37
5	7	4	3	129.21
7	9	4	3	149.95
9	10	4	3	162.08
11	12	3	3	456.06
20	18	5	3	262.61

INTERCLUSTER ΑΠΟΣΤΑΣΗ



- *Ανά δύο cluster υπολογίζαμε την απόσταση μεταξύ τους κάθε φορά*
- *Υπολογισμός Incluster Απόστασης:*

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

- *Όσο πιο μακριά βρίσκονται τα clusters μεταξύ τους τόσο πιο όμοιο είναι οι κόμβοι*

A/A	Nodes	Clusters	Important Dimensions	Min	Max	Average
3	7	3	3	116	122	119
9	10	4	3	109	152	134
20	18	5	3	87	128	110

ΣΥΜΠΕΡΑΣΜΑΤΑ

- Ο αριθμός των κόμβων που βρίσκονται σε ένα δίκτυο επηρεάζει την απόκριση του συστήματος
- Ο αριθμός των clusters επηρεάζει την απόκριση του συστήματος
- Ο αριθμός των σημαντικών διαστάσεων επηρεάζει το solidity του συνόλου των δεδομένων του κάθε κόμβου είναι όμως αντιστρόφως ανάλογο με την απόκριση του συστήματος
- Όσο μικραίναμε τη πιθανότητα *phi τόσο χαλούσε το solidity των δεδομένων*
- *Τέλος να πούμε ότι ο αλγόριθμος προσομοίωσης δοκιμάστηκε μέχρι 1.5 εκατομμύριο εγγραφές και 15 διαστάσεις, με 150 κόμβους και 25 clusters, λαμβάνοντας υπόψιν 4 σημαντικές διαστάσεις και λειτούργησε χωρίς προβλήματα (Χρόνος εκτέλεσης ≈ 85 ώρες)*