

Πρότυπο αναφοράς άσκησης
Συστήματα Διαχείρισης Δεδομένων Μεγάλου Όγκου
Εργαστηριακή Άσκηση 2023/24

Όνομα	Επώνυμο	ΑΜ
Πανταζή	Ηλιάνα	1072642
Πουργουρίδης	Ορφέας	1069664

Βεβαιώνω ότι είμαι συγγραφέας της παρούσας εργασίας και ότι έχω αναφέρει ή παραπέμψει σε αυτήν, ρητά και συγκεκριμένα, όλες τις πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, προτάσεων ή λέξεων, είτε αυτές μεταφέρονται επακριβώς (στο πρωτότυπο ή μεταφρασμένες) είτε παραφρασμένες. Επίσης βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά ειδικά για το συγκεκριμένο μάθημα/σεμινάριο/πρόγραμμα σπουδών.

Έχω ενημερωθεί ότι σύμφωνα με τον εσωτερικό κανονισμό λειτουργίας του Πανεπιστημίου Πατρών άρθρο 50§6, τυχόν προσπάθεια αντιγραφής ή εν γένει φαλκίδευσης της εξεταστικής και εκπαιδευτικής διαδικασίας από οιονδήποτε εξεταζόμενο, πέραν του μηδενισμού, συνιστά βαρύ πειθαρχικό παράπτωμα.

Υπογραφή

Πανταζή Ηλιάνα
20 / 09 / 2024

Υπογραφή

Πουργουρίδης Ορφέας
20 / 09 / 2024

Συνημμένα αρχεία κώδικα

Μαζί με την παρούσα αναφορά υποβάλλουμε τα παρακάτω αρχεία κώδικα

Αρχείο	Αφορά το ερώτημα	Περιγραφή/Σχόλιο
1. producere.py 2. consumer.py	1	<ul style="list-style-type: none">- producer: αφορά το υποερώτημα 4 (αποστολή δεδομένων)- consumer: αφορά το υποερώτημα 5 (παραλαβή δεδομένων)
1. producer.py 2. sparkjob.py	2	<ul style="list-style-type: none">- producer: από εκεί λαμβάνει η διεργασία μας τα δεδομένα- sparkjob: υλοποίηση διεργασίας για παραλαβή και επεξεργασία των raw data
1. sparkjob.py 2. mongoqueries.py	3	<ul style="list-style-type: none">- sparkjob: περιλαμβάνει και την δημιουργία κατάλληλων συλλογών για τα δεδομένα- mongoqueries: υλοποίηση queries

***** Επισυνάπτεται ένα επιπλέον αρχείο το οποίο αφορά το τι εντολή πρέπει να τρέξουμε στην terminal έτσι ώστε να κυλήσουν όλα ομαλά: `commands_for_servers.txt`**

Τεχνικά χαρακτηριστικά περιβάλλοντος λειτουργίας

[Τεχνικά χαρακτηριστικά φυσικού Η/Υ που χρησιμοποιήθηκε για την εργασία, αν χρησιμοποιήθηκε hosted υπηρεσία μπορείτε απλά να αναφέρετε αυτό αντί για τον πίνακα]

Χαρακτηριστικό	Τιμή
CPU model	Intel® Core™ i7-8550U
CPU clock speed	1.8GHz
Physical CPU cores	4
Logical CPU cores	8
RAM	16
Secondary Storage Type	SSD

Ερώτημα 1: Παραγωγή δεδομένων

1. Ο εξομοιωτής λειτουργεί κανονικά στο περιβάλλον μας.
2. Υλοποιήθηκε Kafka Broker ακολουθώντας τις οδηγίες.
3. Δημιουργήθηκε topic "vehicle_positions" στο οποίο θα στέλνονται τα δεδομένα το εξομοιωτή.
4. Από το αρχείο `producer.py`:

Λογική: μέσα στο αντικείμενο κλάσης `KafkaProducer` εισαγάγαμε το κομμάτι του εξομοιωτή το οποίο παράγει τα δεδομένα ου χρειαζόμαστε (περιλαμβάνουμε και την εκτύπωσή τους για επιβεβαίωση) στην συνέχεια δημιουργήσαμε την δομή με την οποία θέλουμε ο καταναλωτής να λαμβάνει τα δεδομένα μέσω `dataframe`. Επιπλέον αποθηκεύουμε τα ληφθέντα δεδομένα σε ένα `.json` αρχείο έτσι ώστε να βεβαιωθούμε πως η επεξεργασία έγινε σωστά. Στην συνέχεια στέλνουμε τα δεδομένα στον καταναλωτή.

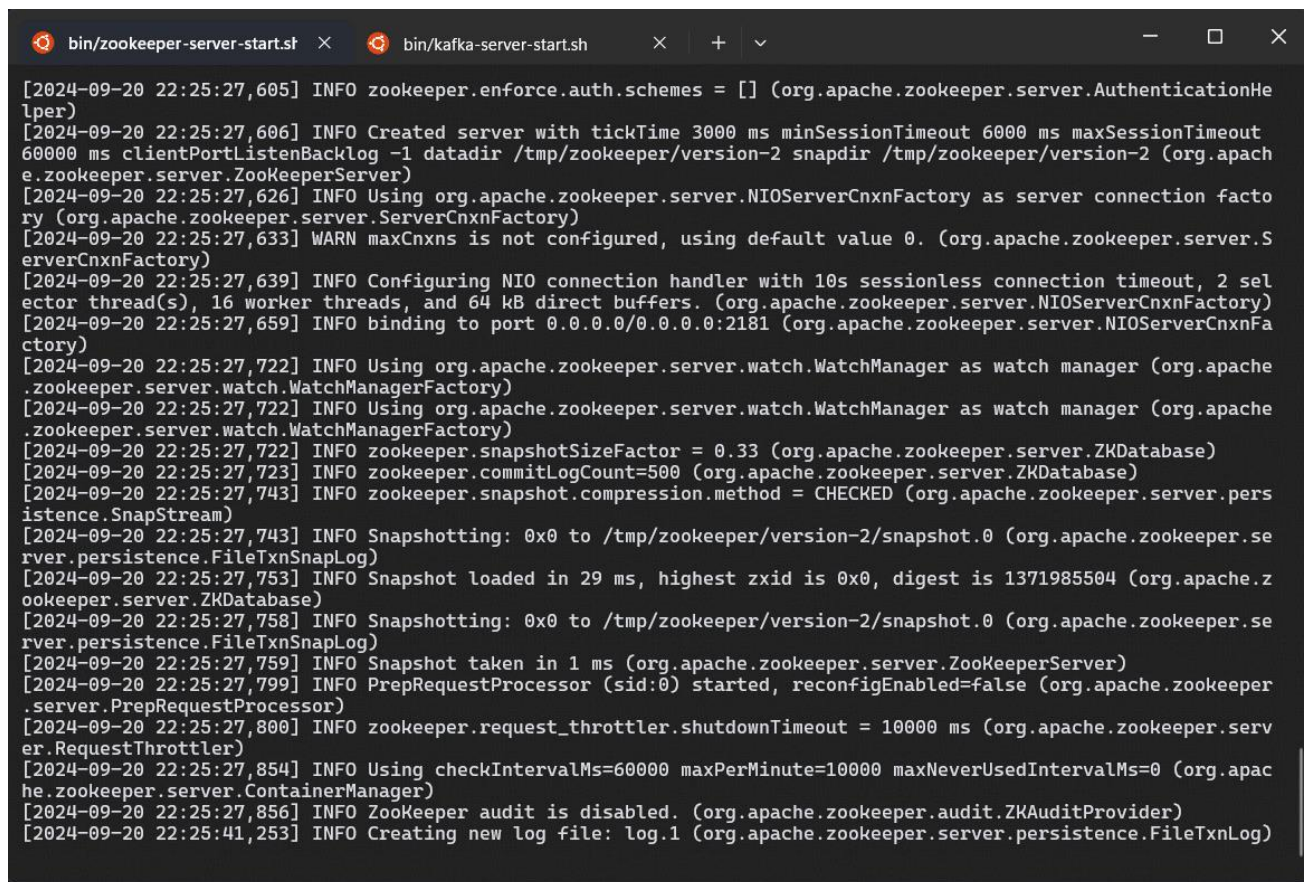
5. Από το αρχείο `consumer.py`:

Μέσω του τελευταίας εντολής `print` βεβαιωνόμαστε πως τα δεδομένα έχουν ληφθεί σωστά.

Screenshots:

Λειτουργία Kafka:

- Zookeeper:



```
[2024-09-20 22:25:27,605] INFO zookeeper.enforce.auth.schemes = [] (org.apache.zookeeper.server.AuthenticationHelper)
[2024-09-20 22:25:27,606] INFO Created server with tickTime 3000 ms minSessionTimeout 6000 ms maxSessionTimeout 60000 ms clientPortListenBacklog -1 datadir /tmp/zookeeper/version-2 snapdir /tmp/zookeeper/version-2 (org.apache.zookeeper.server.ZooKeeperServer)
[2024-09-20 22:25:27,626] INFO Using org.apache.zookeeper.server.NIOServerCnxnFactory as server connection factory (org.apache.zookeeper.server.ServerCnxnFactory)
[2024-09-20 22:25:27,633] WARN maxCnxns is not configured, using default value 0. (org.apache.zookeeper.server.ServerCnxnFactory)
[2024-09-20 22:25:27,639] INFO Configuring NIO connection handler with 10s sessionless connection timeout, 2 selector thread(s), 16 worker threads, and 64 kB direct buffers. (org.apache.zookeeper.server.NIOServerCnxnFactory)
[2024-09-20 22:25:27,659] INFO binding to port 0.0.0.0/0.0.0.0:2181 (org.apache.zookeeper.server.NIOServerCnxnFactory)
[2024-09-20 22:25:27,722] INFO Using org.apache.zookeeper.server.watch.WatchManager as watch manager (org.apache.zookeeper.server.watch.WatchManagerFactory)
[2024-09-20 22:25:27,722] INFO Using org.apache.zookeeper.server.watch.WatchManager as watch manager (org.apache.zookeeper.server.watch.WatchManagerFactory)
[2024-09-20 22:25:27,722] INFO zookeeper.snapshotSizeFactor = 0.33 (org.apache.zookeeper.server.ZKDatabase)
[2024-09-20 22:25:27,723] INFO zookeeper.commitLogCount=500 (org.apache.zookeeper.server.ZKDatabase)
[2024-09-20 22:25:27,743] INFO zookeeper.snapshot.compression.method = CHECKED (org.apache.zookeeper.server.persistence.SnapStream)
[2024-09-20 22:25:27,743] INFO Snapshotting: 0x0 to /tmp/zookeeper/version-2/snapshot.0 (org.apache.zookeeper.server.persistence.FileTxnSnapLog)
[2024-09-20 22:25:27,753] INFO Snapshot loaded in 29 ms, highest zxid is 0x0, digest is 1371985504 (org.apache.zookeeper.server.ZKDatabase)
[2024-09-20 22:25:27,758] INFO Snapshotting: 0x0 to /tmp/zookeeper/version-2/snapshot.0 (org.apache.zookeeper.server.persistence.FileTxnSnapLog)
[2024-09-20 22:25:27,759] INFO Snapshot taken in 1 ms (org.apache.zookeeper.server.ZooKeeperServer)
[2024-09-20 22:25:27,799] INFO PrepRequestProcessor (sid:0) started, reconfigEnabled=false (org.apache.zookeeper.server.PreRequestProcessor)
[2024-09-20 22:25:27,800] INFO zookeeper.request_throttler.shutdownTimeout = 10000 ms (org.apache.zookeeper.server.RequestThrottler)
[2024-09-20 22:25:27,854] INFO Using checkIntervalMs=60000 maxPerMinute=10000 maxNeverUsedIntervalMs=0 (org.apache.zookeeper.server.ContainerManager)
[2024-09-20 22:25:27,856] INFO ZooKeeper audit is disabled. (org.apache.zookeeper.audit.ZKAuditProvider)
[2024-09-20 22:25:41,253] INFO Creating new log file: log.1 (org.apache.zookeeper.server.persistence.FileTxnLog)
```


- *Kafka:*

```
bin/zookeeper-server-start.sh x bin/kafka-server-start.sh x + v
nt)
[2024-09-20 22:25:44,130] INFO [GroupCoordinator 0]: Starting up. (kafka.coordinator.group.GroupCoordinator)
[2024-09-20 22:25:44,131] INFO Feature ZK node created at path: /feature (kafka.server.FinalizedFeatureChangeListener)
[2024-09-20 22:25:44,143] INFO [GroupCoordinator 0]: Startup complete. (kafka.coordinator.group.GroupCoordinator)
[2024-09-20 22:25:44,184] INFO [TransactionCoordinator id=0] Starting up. (kafka.coordinator.transaction.TransactionCoordinator)
[2024-09-20 22:25:44,192] INFO [TxnMarkerSenderThread-0]: Starting (kafka.coordinator.transaction.TransactionMarkerChannelManager)
[2024-09-20 22:25:44,192] INFO [TransactionCoordinator id=0] Startup complete. (kafka.coordinator.transaction.TransactionCoordinator)
[2024-09-20 22:25:44,202] INFO [MetadataCache brokerId=0] Updated cache from existing None to latest Features(version=3.7-IV4, finalizedFeatures={}, finalizedFeaturesEpoch=0). (kafka.server.metadata.ZkMetadataCache)
[2024-09-20 22:25:44,296] INFO [ExpirationReaper-0-AlterAcls]: Starting (kafka.server.DelayedOperationPurgatory$ExpiredOperationReaper)
[2024-09-20 22:25:44,392] INFO [/config/changes-event-process-thread]: Starting (kafka.common.ZkNodeChangeNotificationListener$ChangeEventProcessThread)
[2024-09-20 22:25:44,421] INFO [Controller id=0, targetBrokerId=0] Node 0 disconnected. (org.apache.kafka.clients.NetworkClient)
[2024-09-20 22:25:44,433] WARN [Controller id=0, targetBrokerId=0] Connection to node 0 (DESKTOP-75P75MR./127.0.1.1:9092) could not be established. Node may not be available. (org.apache.kafka.clients.NetworkClient)
[2024-09-20 22:25:44,444] INFO [Controller id=0, targetBrokerId=0] Client requested connection close from node 0 (org.apache.kafka.clients.NetworkClient)
[2024-09-20 22:25:44,449] INFO [SocketServer listenerType=ZK_BROKER, nodeId=0] Enabling request processing. (kafka.network.SocketServer)
[2024-09-20 22:25:44,459] INFO Awaiting socket connections on 0.0.0.0:9092. (kafka.network.DataPlaneAcceptor)
[2024-09-20 22:25:44,484] INFO Kafka version: 3.7.0 (org.apache.kafka.common.utils.AppInfoParser)
[2024-09-20 22:25:44,484] INFO Kafka commitId: 2ae524ed625438c5 (org.apache.kafka.common.utils.AppInfoParser)
[2024-09-20 22:25:44,485] INFO Kafka startTimeMs: 1726860344470 (org.apache.kafka.common.utils.AppInfoParser)
[2024-09-20 22:25:44,488] INFO [KafkaServer id=0] started (kafka.server.KafkaServer)
[2024-09-20 22:25:44,722] INFO [zk-broker-0-to-controller-alter-partition-channel-manager]: Recorded new controller, from now on will use node DESKTOP-75P75MR.:9092 (id: 0 rack: null) (kafka.server.NodeToControllerRequestThread)
[2024-09-20 22:25:44,722] INFO [zk-broker-0-to-controller-forwarding-channel-manager]: Recorded new controller, from now on will use node DESKTOP-75P75MR.:9092 (id: 0 rack: null) (kafka.server.NodeToControllerRequestThread)
```

Python Scripts:

- *Producer Script:*

```
bin/zookeeper-se x bin/kafka-server- x python3 x python3 x + v
~/R/big_data | on main +3 !1 python3 producer.py at 22:27:40
simulation setting:
scenario name:
simulation duration: 3600 s
number of vehicles: 15315 veh
total road length: 6500 m
time discret. width: 5 s
platoon size: 5 veh
number of timesteps: 720
number of platoons: 3063
number of links: 13
number of nodes: 14
setup time: 1.09 s
simulating ...
  time | # of vehicles | ave speed | computation time
  0 s | 0 vehs | 0.0 m/s | 0.00 s
  600 s | 580 vehs | 5.0 m/s | 1.39 s
 1200 s | 580 vehs | 4.2 m/s | 2.60 s
 1800 s | 560 vehs | 4.9 m/s | 3.60 s
 2400 s | 555 vehs | 3.4 m/s | 4.54 s
 3000 s | 570 vehs | 3.8 m/s | 5.43 s
 3595 s | 585 vehs | 1.8 m/s | 6.31 s
simulation finished
results:
average speed: 7.1 m/s
number of completed trips: 6900 / 15315
average travel time of trips: 944.9 s
average delay of trips: 904.6 s
delay ratio: 0.957
name origin destination time link position spacing speed
0 0 N1 S1 20/09/2024 22:28:03 N1I1 0.0 125.0 30.0
name origin destination time link position spacing speed
1 0 N1 S1 20/09/2024 22:28:08 N1I1 100.0 175.0 20.0
name origin destination time link position spacing speed
2 0 N1 S1 20/09/2024 22:28:13 N1I1 250.0 175.0 30.0
name origin destination time link position spacing speed
3 0 N1 S1 20/09/2024 22:28:18 N1I1 400.0 -1.0 30.0
```

- Consumer Script:

```
bin/zookeeper-se  X bin/kafka-server-  X python3  X python3  X + - □ X
~ cd Repositories/big_data
~/R/big_data | on main +3 !1 python3 consumer.py
{'name': '0', 'origin': 'N1', 'destination': 'S1', 'time': '20/09/2024 22:28:03', 'link': 'N1I1', 'position': 0.0, 'spacing': 125.0, 'speed': 30.0}
{'name': '0', 'origin': 'N1', 'destination': 'S1', 'time': '20/09/2024 22:28:08', 'link': 'N1I1', 'position': 10.0, 'spacing': 175.0, 'speed': 20.0}
{'name': '0', 'origin': 'N1', 'destination': 'S1', 'time': '20/09/2024 22:28:13', 'link': 'N1I1', 'position': 25.0, 'spacing': 175.0, 'speed': 30.0}
{'name': '0', 'origin': 'N1', 'destination': 'S1', 'time': '20/09/2024 22:28:18', 'link': 'N1I1', 'position': 40.0, 'spacing': -1.0, 'speed': 30.0}
{'name': '0', 'origin': 'N1', 'destination': 'S1', 'time': '20/09/2024 22:28:23', 'link': 'I1S1', 'position': 50.0, 'spacing': -1.0, 'speed': 20.0}
{'name': '0', 'origin': 'N1', 'destination': 'S1', 'time': '20/09/2024 22:28:28', 'link': 'I1S1', 'position': 20.0, 'spacing': -1.0, 'speed': 30.0}
{'name': '0', 'origin': 'N1', 'destination': 'S1', 'time': '20/09/2024 22:28:33', 'link': 'I1S1', 'position': 35.0, 'spacing': -1.0, 'speed': 30.0}
{'name': '0', 'origin': 'N1', 'destination': 'S1', 'time': '20/09/2024 22:28:33', 'link': 'trip_end', 'position': -1.0, 'spacing': -1.0, 'speed': -1.0}
{'name': '1', 'origin': 'N1', 'destination': 'S1', 'time': '20/09/2024 22:28:28', 'link': 'N1I1', 'position': 0.0, 'spacing': 175.0, 'speed': 30.0}
{'name': '1', 'origin': 'N1', 'destination': 'S1', 'time': '20/09/2024 22:28:33', 'link': 'N1I1', 'position': 15.0, 'spacing': 175.0, 'speed': 30.0}
{'name': '1', 'origin': 'N1', 'destination': 'S1', 'time': '20/09/2024 22:28:38', 'link': 'N1I1', 'position': 30.0, 'spacing': 175.0, 'speed': 30.0}
{'name': '1', 'origin': 'N1', 'destination': 'S1', 'time': '20/09/2024 22:28:43', 'link': 'N1I1', 'position': 45.0, 'spacing': 50.0, 'speed': 30.0}
```

Ερώτημα 2: Κατανάλωση και επεξεργασία με Spark

Όπως και την περίπτωση του Kafka με στόχο να παραλάβουμε τα δεδομένα μας με την σωστή δομή δημιουργήσαμε ένα “template”. Στην αρχή το template αφορά τα δεδομένα όπως στέλνονται απευθείας από το Kafka ενώ στην συνέχεια με τα κατάλληλα processing και aggregations καταλήξαμε στο τελικό αποτέλεσμα που φαίνεται παρακάτω:

*****Σχόλιο: Για την σωστή ροή της υλοποίησης πρώτα τρέχουμε το script του producer (producer.py) και στη συνέχεια τρέχουμε το script της διεργασίας (sparkjob.py) ως εξής:**

Screenshots:

link	vcount	time	vspeed
I1S1	9	2024-09-20 22:55: ...	22.22
I1W1	1	2024-09-20 22:55: ...	0.0
E1I4	5	2024-09-20 22:55: ...	30.0
N1I1	33	2024-09-20 22:55: ...	20.91
trip_end	3	2024-09-20 22:55: ...	1.0
I3I2	2	2024-09-20 22:55: ...	50.0
I2I1	7	2024-09-20 22:55: ...	7.14
I4I3	1	2024-09-20 22:55: ...	0.0

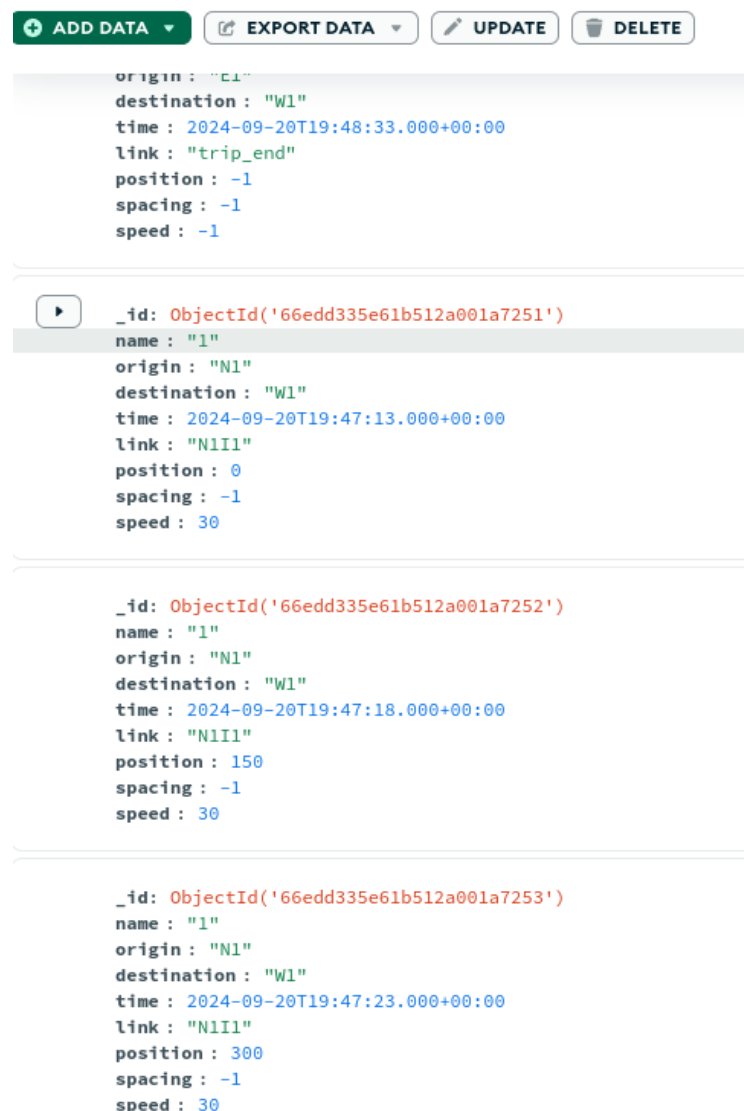
Ερώτημα 3: Αποθήκευση σε MongoDB

1. Εγκαταστάθηκε η MongoDB.
2. Ενσωματώθηκε ο driver.
3. Στο ίδιο αρχείο με την υλοποίηση του προηγούμενου ερωτήματος υλοποιήθηκε και η δημιουργία της Mongo Database με τα κατάλληλα collections για τα raw και τα processes data αντίστοιχα (collection: data = raw, collection: processed = processed)

Παρακάτω παρουσιάζονται Screenshots από το MongoDB Compass, το οποίο αποτελεί interface για ευκολότερη διαχείριση της βάσης μας, μετά την κλήση της sparkjob.py διεργασίας:

Screenshots:

- Collection data:



- Collection processed:

`_id: ObjectId('66edd339e61b512a001a7259')`
`link: "I1S1"`
`vcount: 9`
`time: 2024-09-20T19:55:36.738+00:00`
`vspeed: 22.22`

`_id: ObjectId('66edd339e61b512a001a725a')`
`link: "I1W1"`
`vcount: 1`
`time: 2024-09-20T19:55:36.738+00:00`
`vspeed: 0`

`_id: ObjectId('66edd339e61b512a001a725b')`
`link: "E1I4"`
`vcount: 5`
`time: 2024-09-20T19:55:36.738+00:00`
`vspeed: 30`

`_id: ObjectId('66edd339e61b512a001a725c')`
`link: "N1I1"`
`vcount: 34`
`time: 2024-09-20T19:55:36.738+00:00`
`vspeed: 20.29`

`_id: ObjectId('66edd339e61b512a001a725d')`
`link: "trip_end"`
`vcount: 3`
`time: 2024-09-20T19:55:36.738+00:00`
`vspeed: 1`

`_id: ObjectId('66edd339e61b512a001a725e')`
`link: "I3I2"`
`vcount: 2`

- Παραδείγματα υλοποίησης των Queries:

1.

link	vcount
I3I2	1

2.

link	vspeed
I3I2	50.0

3.

position
250.0

***Σημείωση: Όλη η εργασία υλοποιήθηκε σε περιβάλλον Linux Ubuntu μέσω WSL.

Σχολιασμός αποτελεσμάτων

Βιβλιογραφία

Πέραν των link που μας προτάθηκαν από την εκφώνηση για τυχόν errors που προκύπταν κατά την διαδικασία υλοποίησης οποιουδήποτε task βασιστήκαμε κυρίως στο Stack Overflow (<https://stackoverflow.com>) και στα official documentations των αντίστοιχων εργαλείων που χειριζόμασταν.