ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ, ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Προχωρημένα Θέματα Βάσεων Δεδομένων

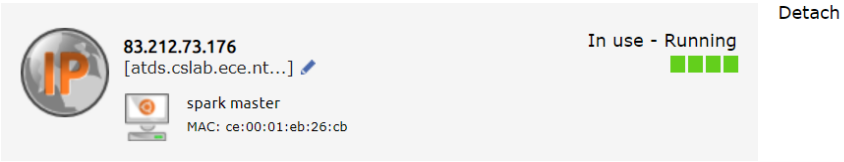
Προεργασία Δεύτερου Εργαστηρίου: *Οδηγίες ανεβάσματος αρχείων στο HDFS*

Στο δεύτερο εργαστήριο θα γίνει hands on επίδειξη στη δημιουργία και την εκτέλεση προγραμμάτων στο Apache Spark. Για να μπορείτε να εκτελείτε ταυτόχρονα τα προγράμματα, συνίσταται η εκ των προτέρων φόρτωση των αρχείων που θα χρησιμοποιήσουμε στα παραδείγματα στο hdfs, το οποίο εγκαταστάθηκε στο προηγούμενο εργαστήριο σε VM του Okeanos. Στη συνέχεια, παρατίθενται συνοπτικά τα βήματα για το ανέβασμα των αρχείων.

Βήμα 1: Σύνδεση ssh στο master vm που είχατε δημιουργήσει.

Αν δεν θυμάστε την public ip του master, μπορείτε να τη βρείτε το διαχειριστικό UI του Okeanos. Συνοπτικά υπενθυμίζουμε τη διαδικασία:

- https://accounts.okeanos.grnet.gr/ui/login : Προηγηθείτε εδώ και επιλέξτε Academic Login και συνδεθείτε με τα στοιχεία του ΕΜΠ.
- Αφού συνδεθείτε, πάνω αριστερά πατήστε στην επιλογή "Cyclades".
- Στη συνέχεια, επιλέξτε το tab "IP" και θα βρείτε την IP του master σε όπως φαίνεται στην ακόλουθη εικόνα (στο παράδειγμα είναι η 83.212.73.176).



Πραγματοποιείστε τη σύνδεση ssh σύμφωνα με την εντολή από terminal,

ssh user@<public_ip>

ή μέσω κάποιου Windows Client (π.χ. Putty).

Βήμα 2: Δημιουργία Φακέλου για τη λήψη των αρχείων και λήψη των αρχείων

Δημιουργήστε αρχικά έναν φάκελο για να βάλουμε τα δεδομένα μέσω της εντολής

```
mkdir data
```

και μεταβείτε σε αυτόν σύμφωνα με την επόμενη εντολή

```
cd data
```

Κατεβάστε τα δεδομένα στο master vm σύμφωνα με την εντολή

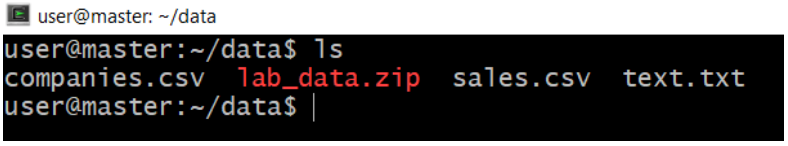
```
wget -q --no-check-certificate
'https://docs.google.com/uc?export=download&id=1FY5icf9RwwwDdYML9nrPA
6qR6CCMJbN2' -O lab_data.zip
```

Βήμα 3: Αποσυμπίεση των αρχείων

Στη συνέχεια, αποσυμπιέζουμε το lab_data.zip με χρήση της εντολής

```
unzip lab_data.zip
```

Κάντε ls για επιβεβαίωση ότι η εξαγωγή ήταν επιτυχής. Αν ναι, θα δείτε την ακόλουθη έξοδο



Βήμα 4: Φόρτωση των αρχείων στο hdfs

Εκτελέστε τις επόμενες τρεις εντολές για να φορτωθούν τα αρχεία στο hdfs και να μπορούμε να τα χρησιμοποιήσουμε στο εργαστήριο.

```
hadoop fs -put sales.csv hdfs://master:9000/.
```

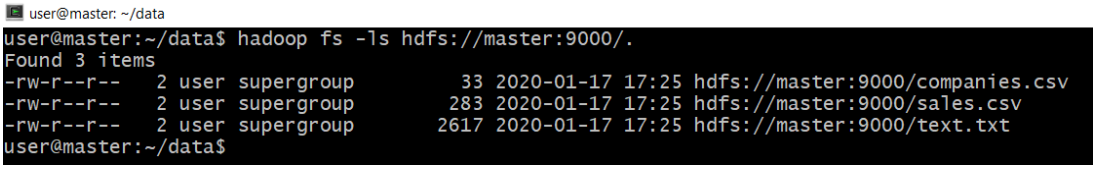
```
hadoop fs -put companies.csv hdfs://master:9000/.
```

```
hadoop fs -put text.txt hdfs://master:9000/.
```

Για να επιβεβαιώσετε ότι τα αρχεία έχουν φορτωθεί στο hdfs εκτελέστε

```
hadoop fs -ls hdfs://master:9000/.
```

και θα πρέπει να λάβετε το ακόλουθο αποτέλεσμα.



Βήμα 5: Βεβαιωθείτε ότι η python είναι εγκατεστημένη

Εκτελέστε

```
python –version
```

και θα πρέπει να λάβετε την έξοδο

```
Python 2.7.12
```