



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Προχωρημένα Θέματα Βάσεων Δεδομένων

Αναφορά για την Εξαμηνιαία Εργασία:

Χρήση του Apache Spark στις Βάσεις Δεδομένων

Ον/μο : Μπερέτσος Θεόδωρος
Α.Μ.: : 03111612

Ον/μο : Ντόκος Χρήστος
Α.Μ.: : 03117xxx

Ον/μο : Στάβαρης Δημοσθένης
Α.Μ.: : 03117xxx

Ημερομηνία παράδοσης: 18/03/2022

Μέρος 1^ο: Υπολογισμός Αναλυτικών Ερωτημάτων με τα APIs του Apache Spark

Ζητούμενο 1

Έγινε λήψη του dataset `movie_data.tar.gz`. Αποσυμπίστηκε. Δημιουργήθηκαν τα directories **files** και **outputs** στο Hadoop file system. Τέλος, φορτώθηκαν τα 3 CSV αρχεία που μας δόθηκαν στο hdfs στο φάκελο **files** εκτελώντας τις παρακάτω εντολές στον **master** (βλ. Εικόνα 1).

```
user@master:~$ wget 'http://www.cslab.ntua.gr/courses/atds/movie_data.tar.gz'
--2022-03-11 19:41:19-- http://www.cslab.ntua.gr/courses/atds/movie_data.tar.gz
Resolving www.cslab.ntua.gr (www.cslab.ntua.gr)... 147.102.3.238
Connecting to www.cslab.ntua.gr (www.cslab.ntua.gr)|147.102.3.238|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 184259305 (176M) [application/x-gzip]
Saving to: 'movie_data.tar.gz'

movie_data.tar.gz      100%[=====] 175.72M  109MB/s   in 1.6s

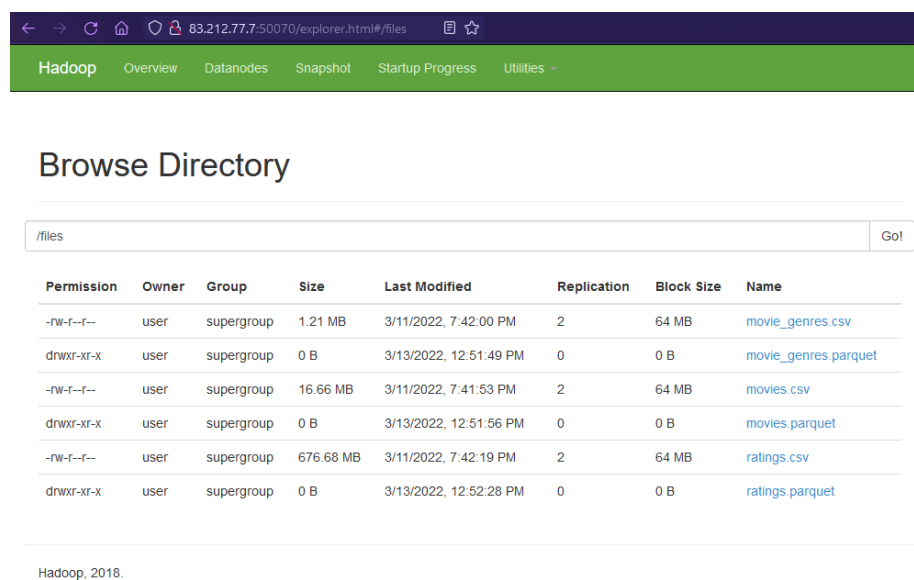
2022-03-11 19:41:21 (109 MB/s) - 'movie_data.tar.gz' saved [184259305/184259305]

user@master:~$ tar -xzf movie_data.tar.gz
user@master:~$ hadoop fs -mkdir hdfs://master:9000/files
user@master:~$ hadoop fs -put movies.csv hdfs://master:9000/files/.
user@master:~$ hadoop fs -put movie_genres.csv hdfs://master:9000/files/.
user@master:~$ hadoop fs -put ratings.csv hdfs://master:9000/files/.
user@master:~$ hadoop fs -mkdir hdfs://master:9000/outputs
user@master:~$ hadoop fs -ls hdfs://master:9000/
Found 2 items
drwxr-xr-x - user supergroup      0 2022-03-11 19:42 hdfs://master:9000/files
drwxr-xr-x - user supergroup      0 2022-03-11 19:42 hdfs://master:9000/outputs
user@master:~$ |
```

Εικόνα 1: Εντολές φόρτωσης .csv αρχείων στο hdfs

Ζητούμενο 2

Χρησιμοποιήθηκε το script με όνομα `csv2parquet.py` για την μετατροπή των αρχείων CSV σε Parquet. Εποπτικά η εικόνα των αρχείων στο directory files μέσα από το Web UI του Hadoop είναι η εξής (βλ. Εικόνα 2):



Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	user	supergroup	1.21 MB	3/11/2022, 7:42:00 PM	2	64 MB	movie_genres.csv
drwxr-xr-x	user	supergroup	0 B	3/13/2022, 12:51:49 PM	0	0 B	movie_genres.parquet
-rw-r--r--	user	supergroup	16.66 MB	3/11/2022, 7:41:53 PM	2	64 MB	movies.csv
drwxr-xr-x	user	supergroup	0 B	3/13/2022, 12:51:56 PM	0	0 B	movies.parquet
-rw-r--r--	user	supergroup	676.68 MB	3/11/2022, 7:42:19 PM	2	64 MB	ratings.csv
drwxr-xr-x	user	supergroup	0 B	3/13/2022, 12:52:28 PM	0	0 B	ratings.parquet

Εικόνα 2: Web UI Hadoop

Ζητούμενο 3

Lorem Ipsum

Ζητούμενο 4

Lorem Ipsum

Μέρος 2^ο: Υλοποίηση και μελέτη συνένωσης σε ερωτήματα και Μελέτη του βελτιστοποιητή του Spark

Ζητούμενο 1

Lorem Ipsum

Ζητούμενο 2

Lorem Ipsum

Ζητούμενο 3

Lorem Ipsum

Ζητούμενο 4

Lorem Ipsum