Στο παρόν αρχείο υπάρχουν οι οδηγίες εκτέλεσης του κώδικα της εργασίας:

1) Επειδή η όλη υλοποίηση της εργασίας έγινε σε windows 10, τα βήματα για το σετάρισμα του spark και του hadoop έγιναν όπως αναλύονται στα παρακάτω sites:

α) Για το spark: https://phoenixnap.com/kb/install-spark-on-windows-10
(αλλά κατέβηκε η 2.4.7 αντί της 2.4.5, και επιπλέον ακολουθήθηκαν τα βήματα μέχρι το βήμα 7, Configure Environment Variables).

Για το hadoop: https://www.datasciencecentral.com/profiles/blogs/how-to-install-and-run-hadoop-on-windows-for-beginners

2) Φτιάχνουμε τα 2 maven projects στο Eclipse IDE
File -> New -> Project... -> Maven -> Maven Project κλπ
(Προσοχή πρέπει οι εκδόσεις jre της java στα maven projects να συμπίπτουν με αυτή που γνωρίζει το spark/hadoop)

3) Προσοχή: Αλλάζουμε τα paths, έτσι ώστε να ταιριάζουν με τα σημεία στα οποία υπάρχουν τα input files, και με τα σημεία στα οποία θέλουμε να πάνε τα output files. Τα paths δηλώνονται στη γραμμή 25 για το MoviesRDDOnly.java, και στην 20 για το BigDataDataframes.java. Αν προκύπτει error, ίσως πρέπει να προστεθεί και ο φάκελος input στο path (βλέπετε γραμμές 80-81 και 51-52 αντίστοιχα).

4) Προσθέτουμε τα dependencies των 2 poms που βρίσκονται στο zip αρχείο, στα dependencies των projects υπό δημιουργία.

5) Αφού κάνει build ο κώδικας παράγουμε το .jar από το .class αρχείο:

α) Δεξί κλικ πάνω στο projectName (από τον package explorer)

β) Export... -> Java -> JAR File -> Next

Προσοχή στο export destination του JAR File (και για το path και για το όνομα του jar, πρέπει να μετατραπεί αντιστοίχως και η εντολή στο βήμα 9.

γ) Next -> Next -> Browse main class (και επιλέγουμε ποια θέλουμε να είναι η main class). (Έτσι αποφεύγουμε το --class "classname")

δ) Finish (και λογικά τα errors που προκύπτουν είναι false positive, οπότε κλείνουμε το σχετικό dialog box και το jar έχει παραχθεί)

6) Μέσα στον κώδικα bin (του hadoop), φτιάχνουμε έναν άλλο φάκελο bin και βάζουμε μέσα το winutils.exe

7) Αλλάζουμε τα paths ώστε να μην έχουν ελληνικά (και κενά)

8) Μετονομάζουμε το αρχείο "log4j.properties.template" σε "log4j.properties". Το αρχείο αυτό βρίσκεται στο C:\Spark\spark-2.4.7-bin-hadoop2.7\conf

9) Τρέχουμε σε cmd:

α) cd C:\Spark\spark-2.4.7-bin-hadoop2.7\bin

β) spark-submit {{path}}\{{jarname}}.jar, δύο φορές, μία για κάθε project