Documentation of the Global Terrorism Big Data Project:

Ekaterina Novgorodtseva
Lisa Hornung

Programs used:

Cloudera Quickstart CDH 5.12
Kafka package downloaded via Cloudera Parcels
Impala (preinstalled with Cloudera)
Spyder 3.2.4 included in Anaconda 5.0.1 version with Python 2.7 (downloaded from Anaconda website)

Procedure:

1. Download Cloudera Quickstart CDH 5.12 and install in VMWare.
2. Download Kafka via Cloudera Parcels
3. Download Anaconda
4. Install Anaconda packages „kafka-python", „numpy", „scikit-learn", „scipy" and „scikit-learn"
5. Process the data set file „globalterrorismdb_0617dist.csv" by inputting the command „tr -cd '\11\12\40-\176' < globalterrorismdb_0617dist.csv > TerrorDBinput.txt" into the command line. It removes all characters outside the ASCII range.
6. Now the data set can be streamed.
7. Execute the „TerrorConsumer.py" inside the first console window of Spyder.
8. Execute the „TerrorProducer.py" inside the second console window of Spyder.
9. Compare both windows for the output produced. Some data will not arrive as TerrorConsumer, probably due to formatting issues. To be more specific, years from 1998 to 2006, and 2008 to 2016.
10. The TerrorConsumer has written the received data into a text file called „TerrorDBoutput.txt" on the local system.
11. Impala can import this text file as a database. For this, we open Hue and go through follow steps:
12. Click on the „+" to add a database. Click on „Path" to choose the file to import from.

13. Choose „Upload file", since the created text file is stored on the local system. Choose the correct data path. Import.
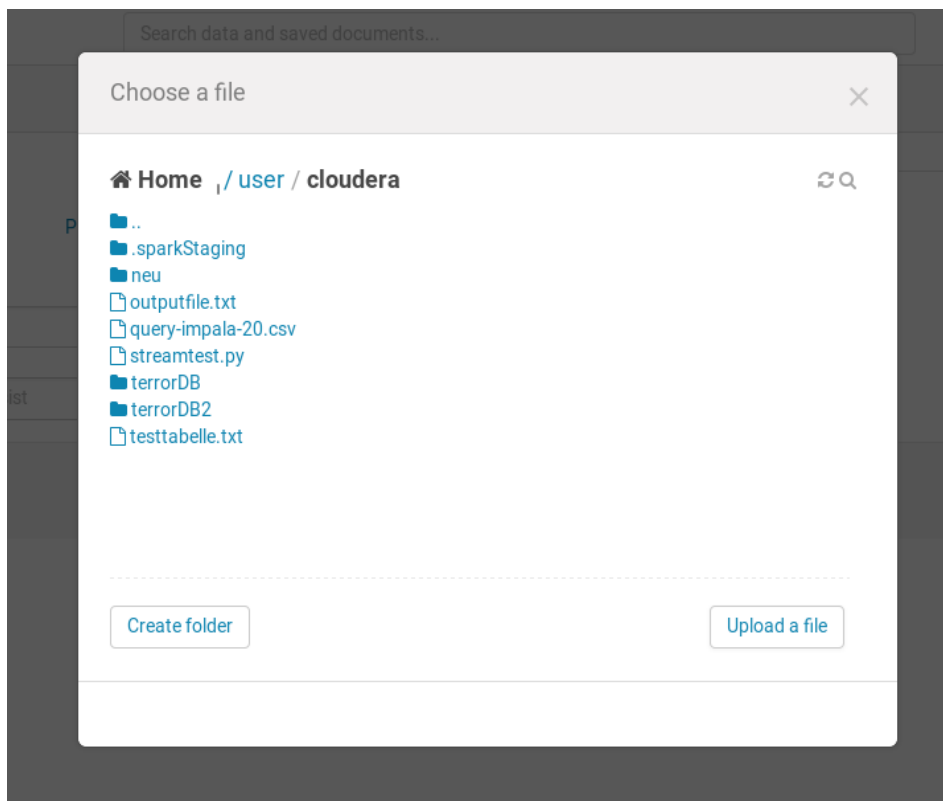
14. Proceed through the required steps, Rename database if needed.
15. Now open the Query window and input following query:

„SELECT eventid, iyear, imonth, iday, nkill FROM terrordb
WHERE iyear < 1990 AND nkill >–1 AND region = 2 AND weaptype1 = 6;"

16. It selects the region „Central America" and weapon type „Explosives", while also excluding all entries where the „nkill" value was NULL, and therefore an entry did not exist. The years chosen are from 1970 to 1989.
17. We export this data set as CSV file

```
1 SELECT eventid, iyear, imonth, iday, nkill FROM terrordb
2 WHERE iyear < 1990 AND nkill >-1 AND region = 2 AND weaptype1 = 6;
```

Query History Saved Queries Results (1,024+)

	eventid	iyear	imonth	iday	nkill
1	197200000001	1972	5	25	0
2	197211050001	1972	11	5	0
3	197211150001	1972	11	15	0
4	197304290001	1973	4	29	0
5	197403260002	1974	3	26	0

18. Import this file as database. We call the database camericaexp1990whole
19. Run the following query to extract the testing set:

SELECT eventid, iyear, imonth, iday, nkill
FROM camericaexp1990whole
ORDER BY RAND()
LIMIT 1200;

20. Export this as CSV file.
21. Run second query to extract the training set. The training set does not include the dataset in the testing set. Export the results as another csv file.

SELECT eventid, iyear, imonth, iday, nkill
FROM camericaexp1990whole
WHERE eventid NOT IN
(SELECT eventid FROM camericaexp1990bigtest)

22. With the extracted data, run the „centralamericaExpRidge.py", referencing the data path of the testing and training data set.
23. Review the graph.