Code for below implementation is uploaded into git repo.

<https://github.com/vignesh23git/clickstream_user_data.git>

**Project flow:**

Kafka -> FileStore -> Spark -> FileStore -> LogStash -> ElasticSearch (Kibana)

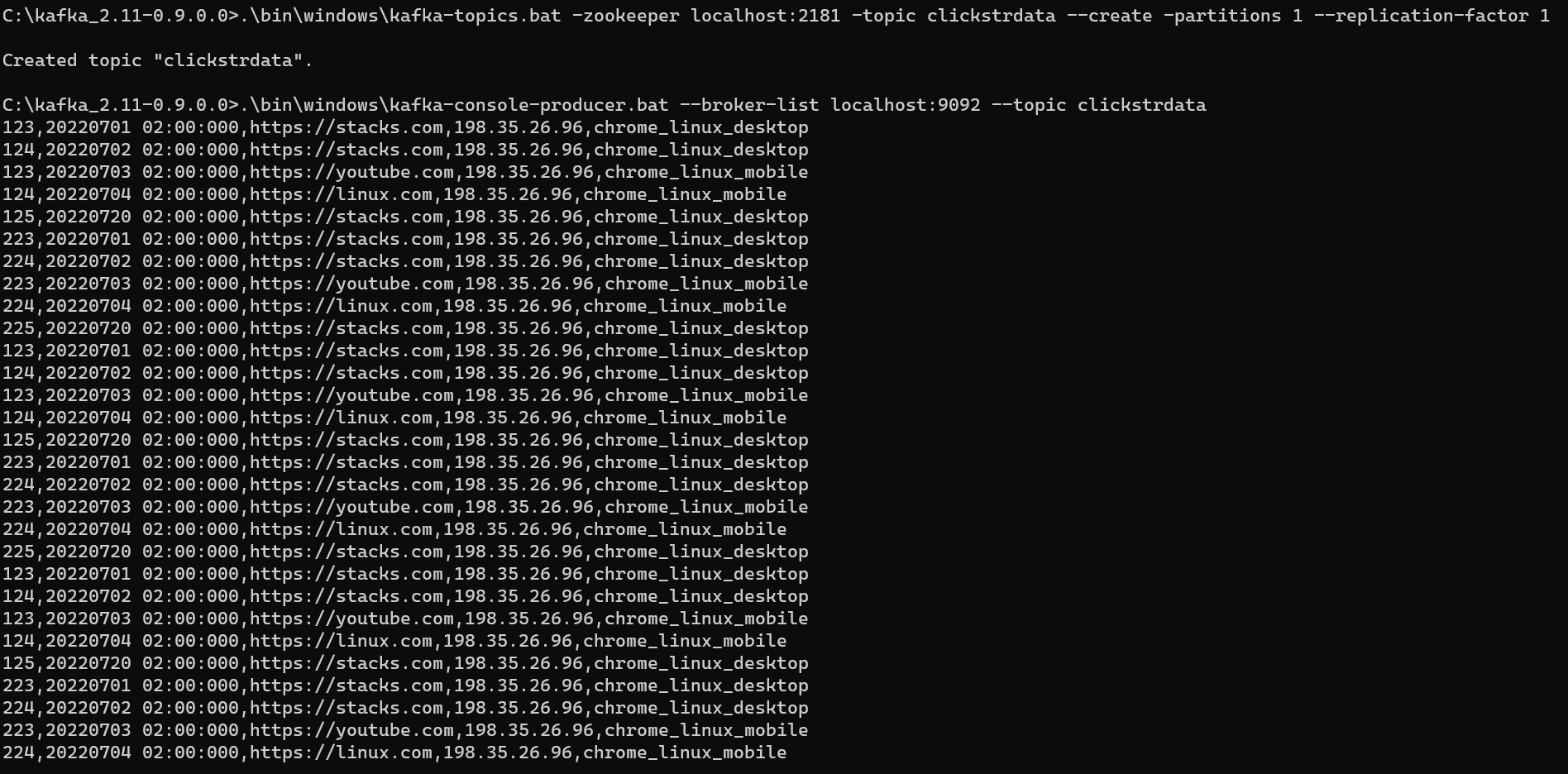
**Kafka:**

Started Zookeper server, Kafka broker.

Created a topic **clickstrdata** in Kafkawhere the user clickstream data is published and it is consumed into FileStore as a stream.

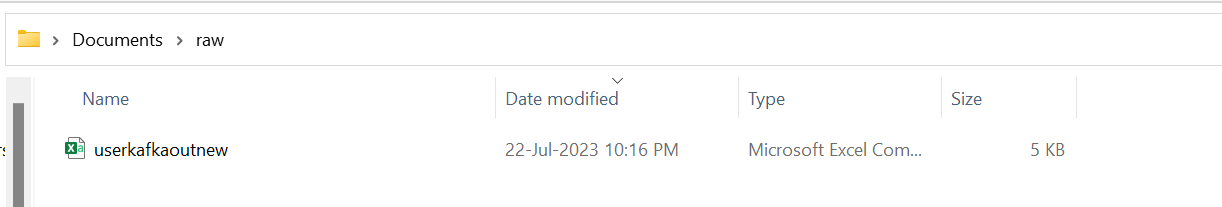
Below are commands which were executed to achieve this.

* Zookeeper start:
  + zkserver
* kafka server start:
  + C:\kafka\_2.11-0.9.0.0\>.\bin\windows\kafka-server-start.bat .\config\server.properties
* kafka create topic:
  + C:\kafka\_2.11-0.9.0.0>.\bin\windows\kafka-topics.bat -zookeeper localhost:2181 -topic clickstrdata --create -partitions 1 --replication-factor 1
* kafka producer console open:
  + C:\kafka\_2.11-0.9.0.0>.\bin\windows\kafka-console-producer.bat --broker-list localhost:9092 --topic clickstrdata
* kafka consumer console open:
  + C:\kafka\_2.11-0.9.0.0>.\bin\windows\kafka-console-consumer.bat --new-consumer --bootstrap-server localhost:9092 --topic clickstrdata --from-beginning > "C:\Users\SATHYA\Documents\raw\userkafkaoutnew.csv"



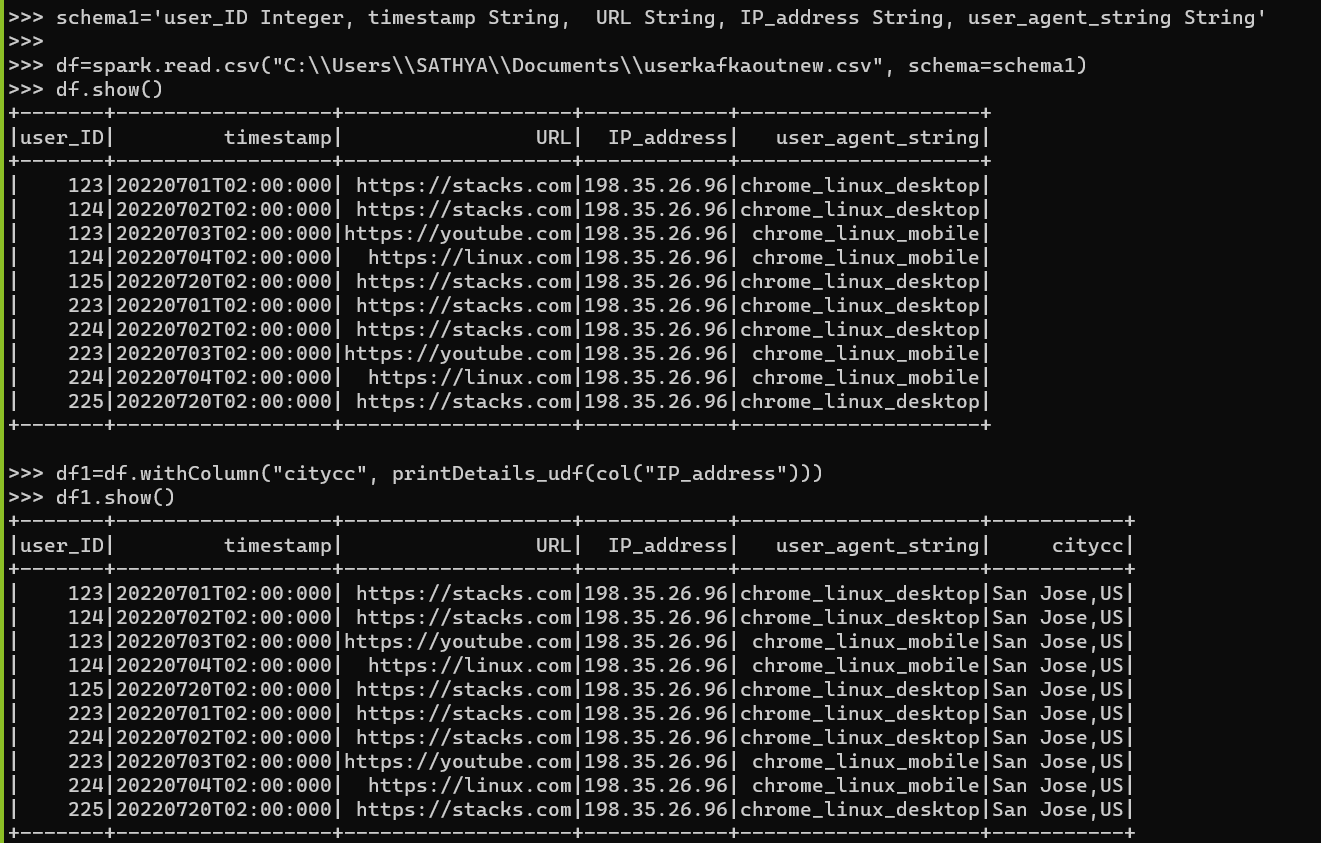
**FileStore:**

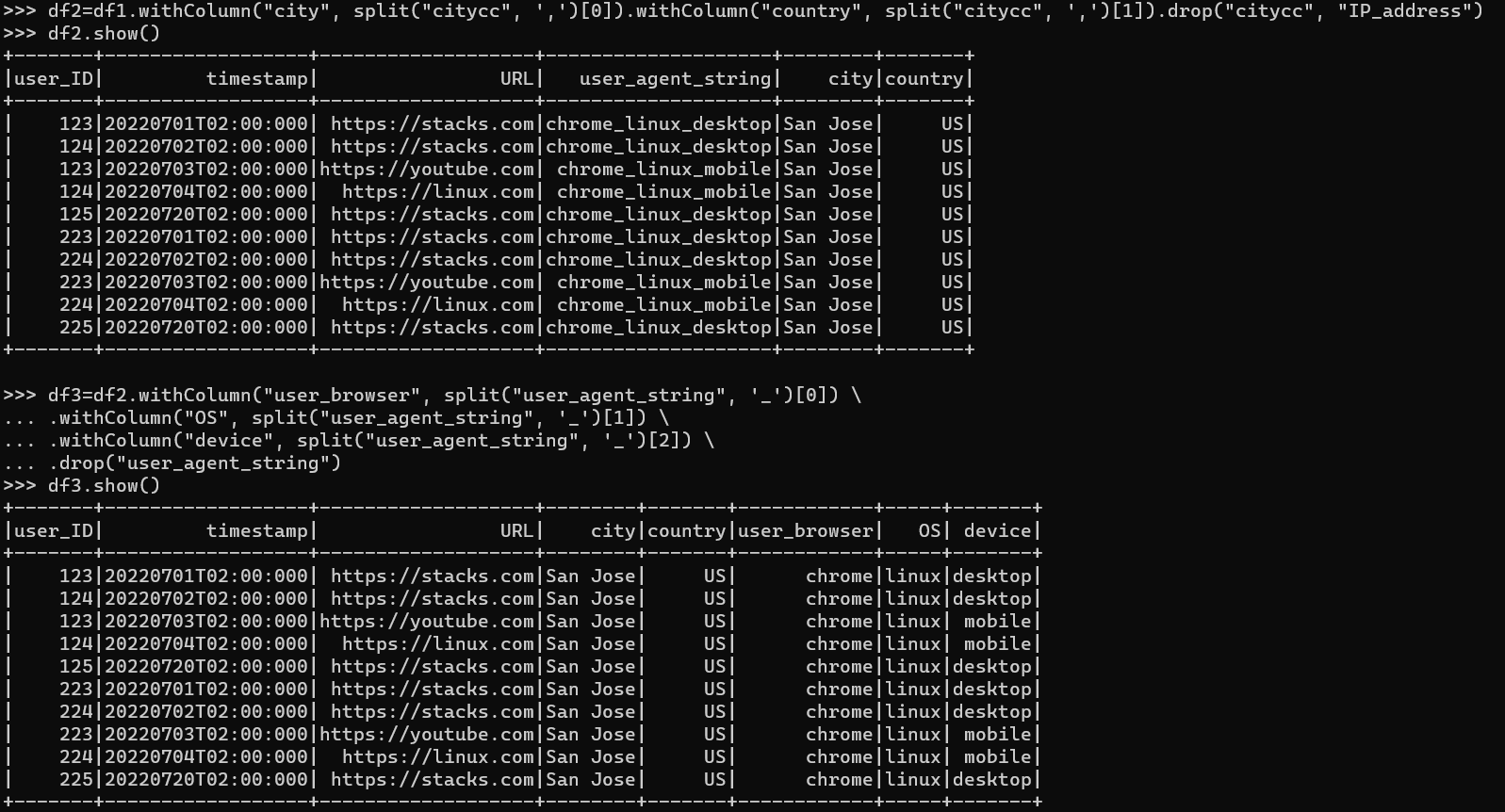
Data from kafka is saved as csv file in below path.

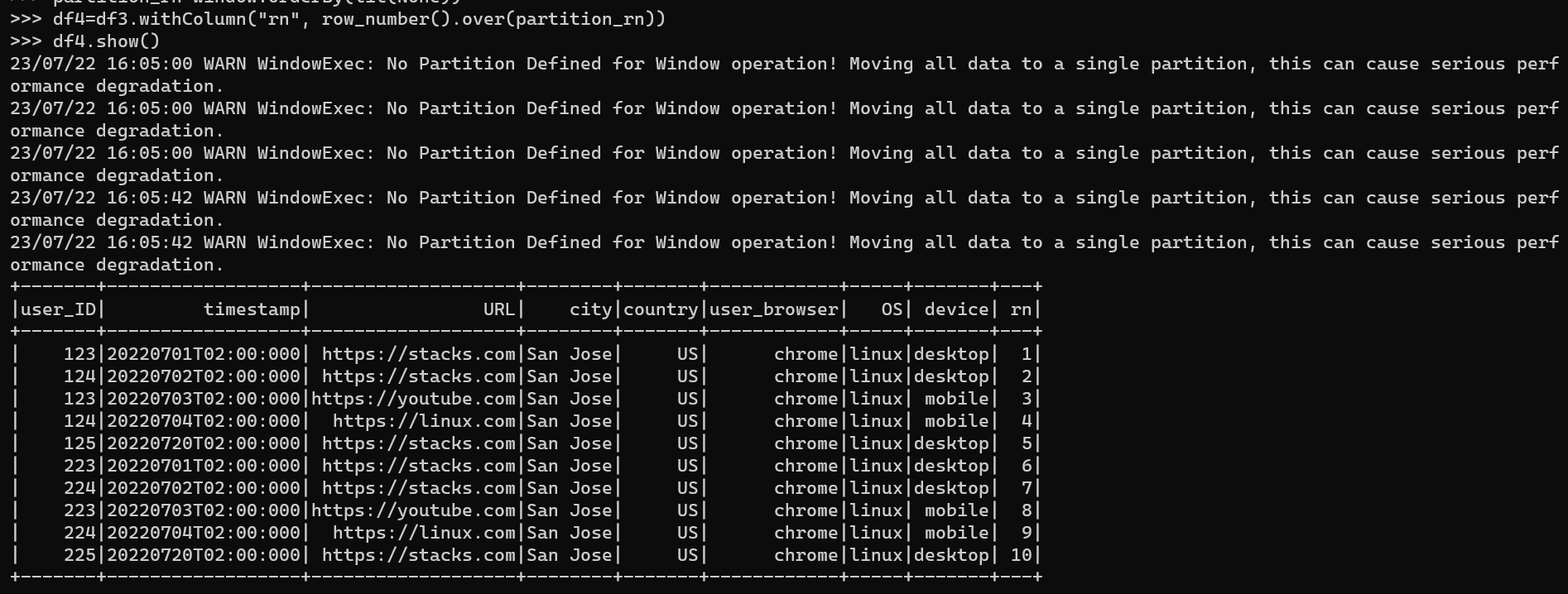


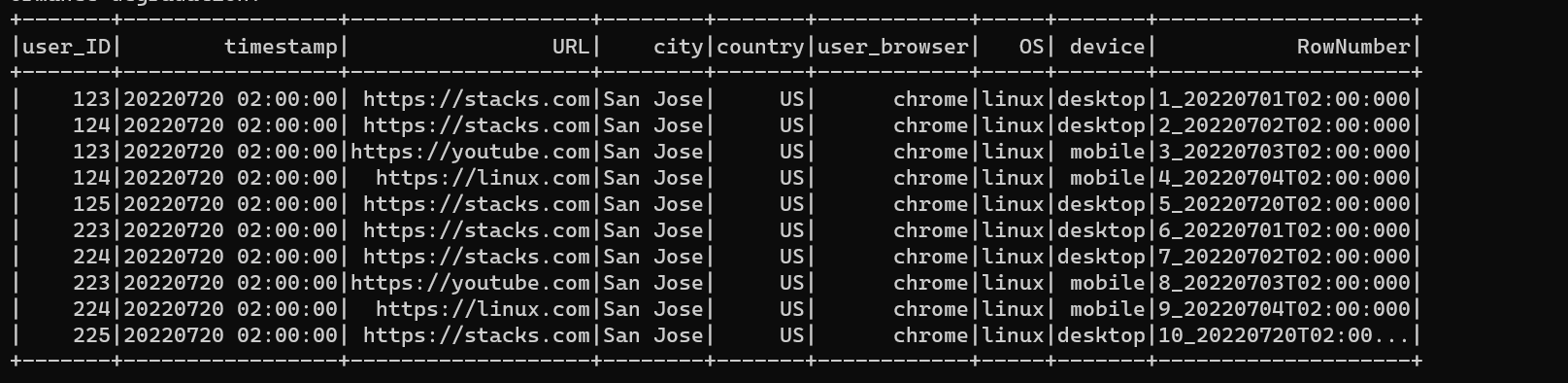
**Spark:**

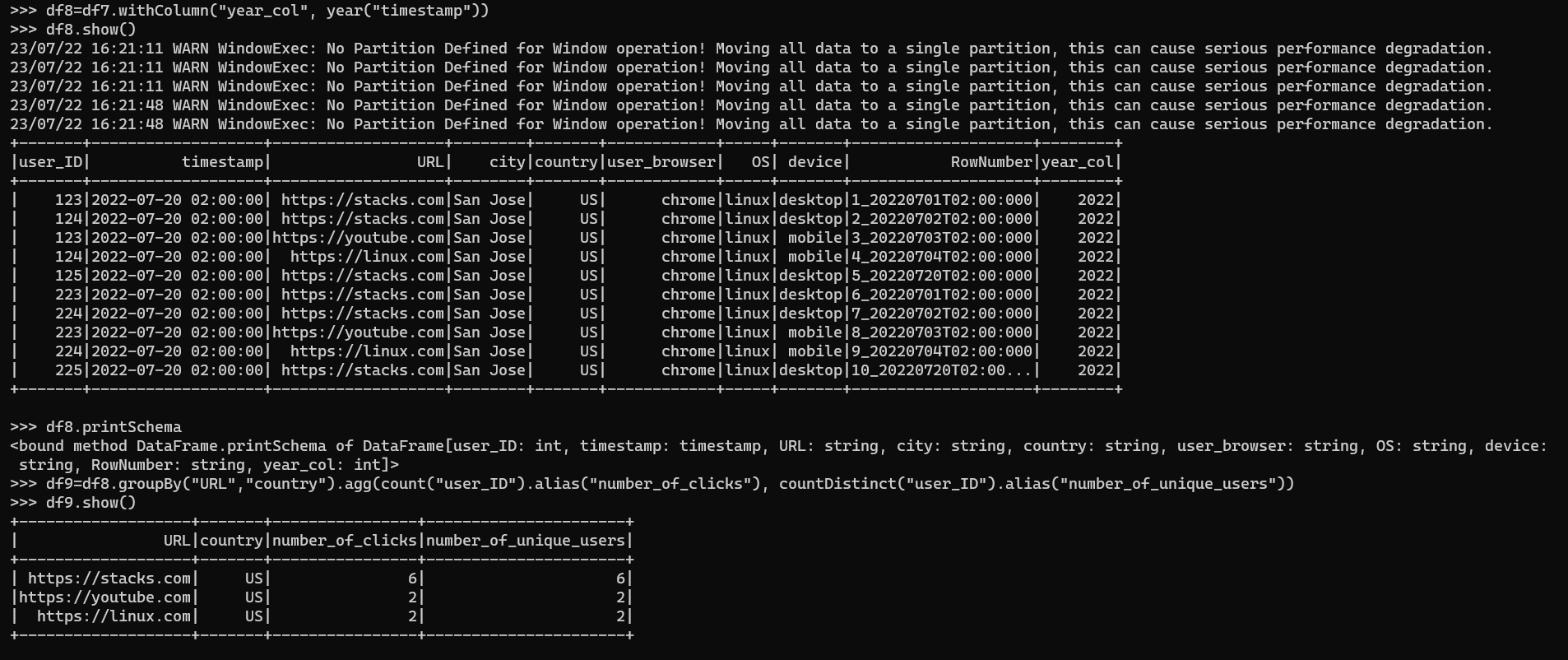
* Csv file above mentioned path is read from dataframe reader and required transformations and aggregations in Pyspark.
* While reading the data, only the incremental data load is filtered and processed further.
* While processing, the user country and city is derived from their IP address from the stream input using ip2geotools package.



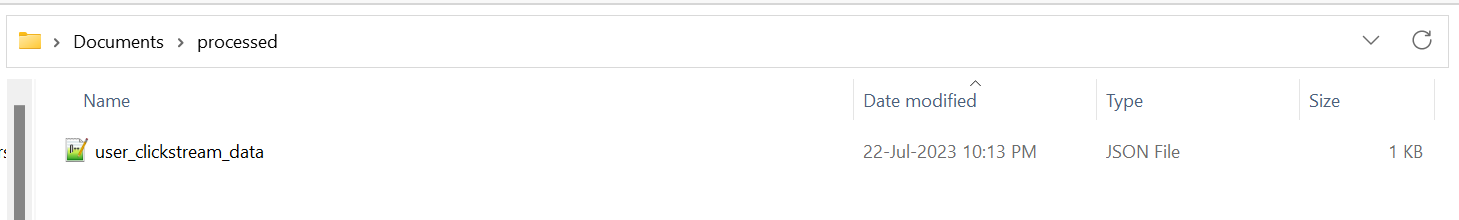






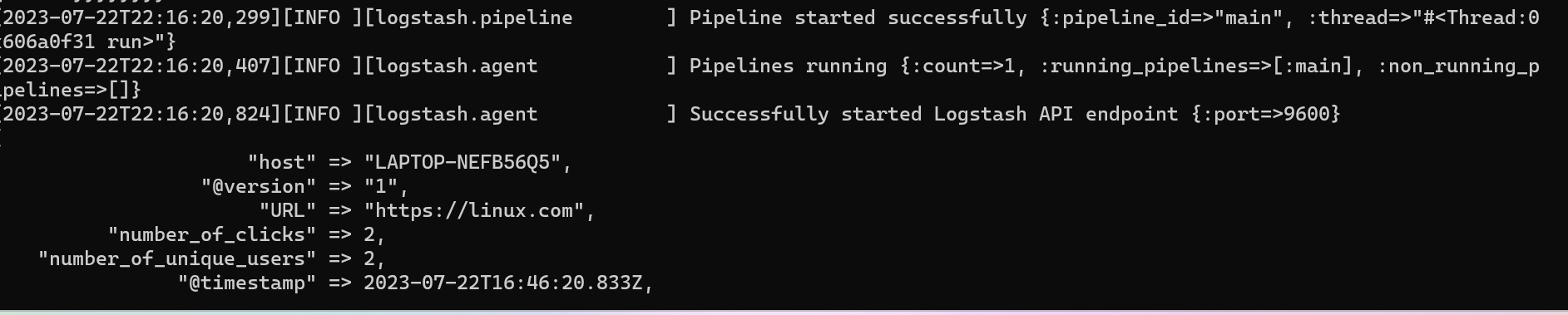


* Finally, the aggregated data is saved as json file in FileStore.



**Logstash:**

The processed stored in filestore is listened by logstash server for indexing into ElasticSearch.



Below are commands which were executed to achieve this.

* Starting LogStash server:
  + C:\Users\SATHYA\Desktop\elk\_couseudemy\logstash-6.3.2>.\bin\logstash.bat -f C:\Users\SATHYA\Desktop\elk\_couseudemy\data\userkafkaspark.conf
* Starting ElasticSearch server:
  + C:\Users\SATHYA\Desktop\elk\_couseudemy\elasticsearch-6.3.2>.bin\elasticsearch.bat
* Starting Kibana server:
  + C:\Users\SATHYA\Desktop\elk\_couseudemy\kibana-6.3.2-windows-x86\_64>.\bin\kibana.bat

**ElasticSearch/Kibana:**

Data indexed into elastic search is visualized using Kibana.

