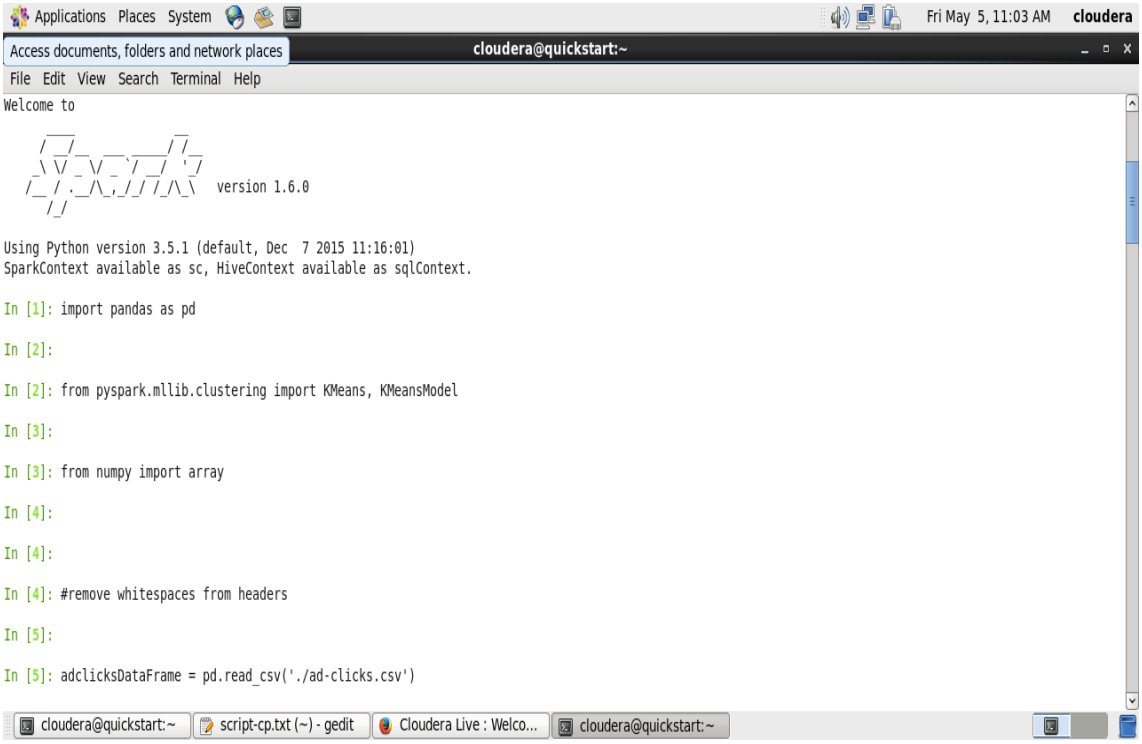
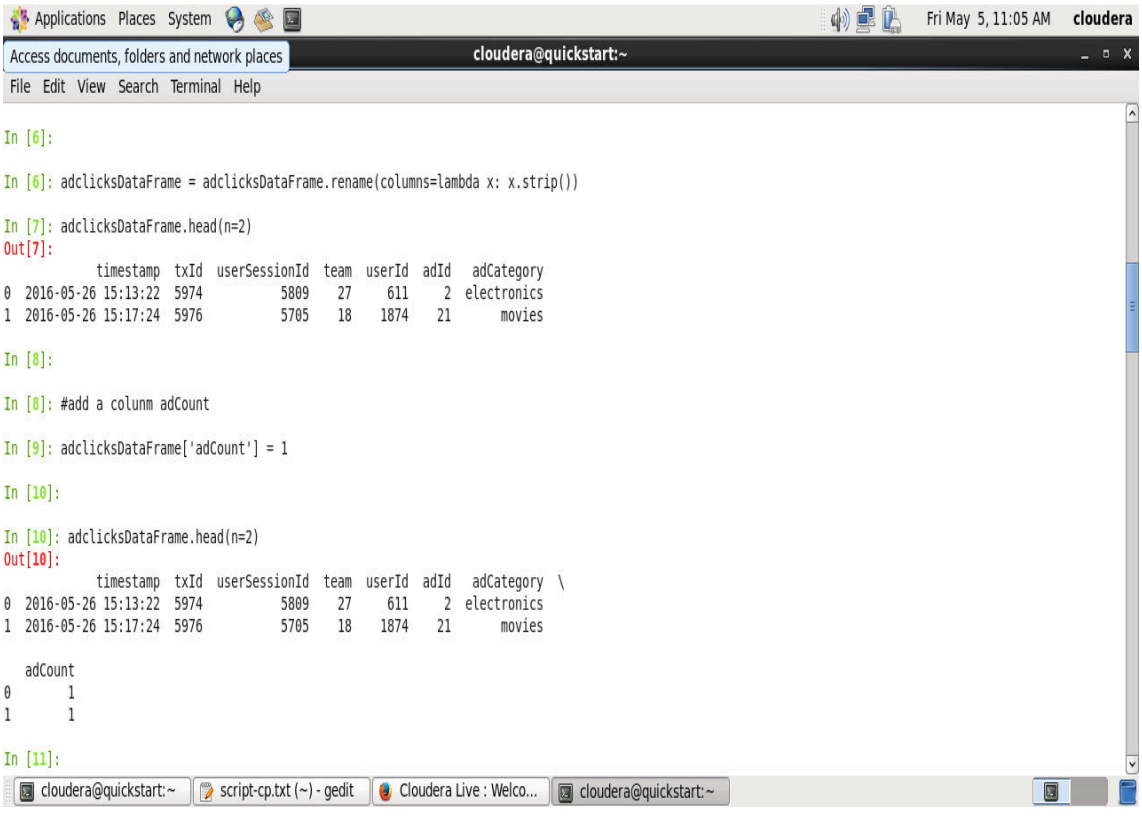
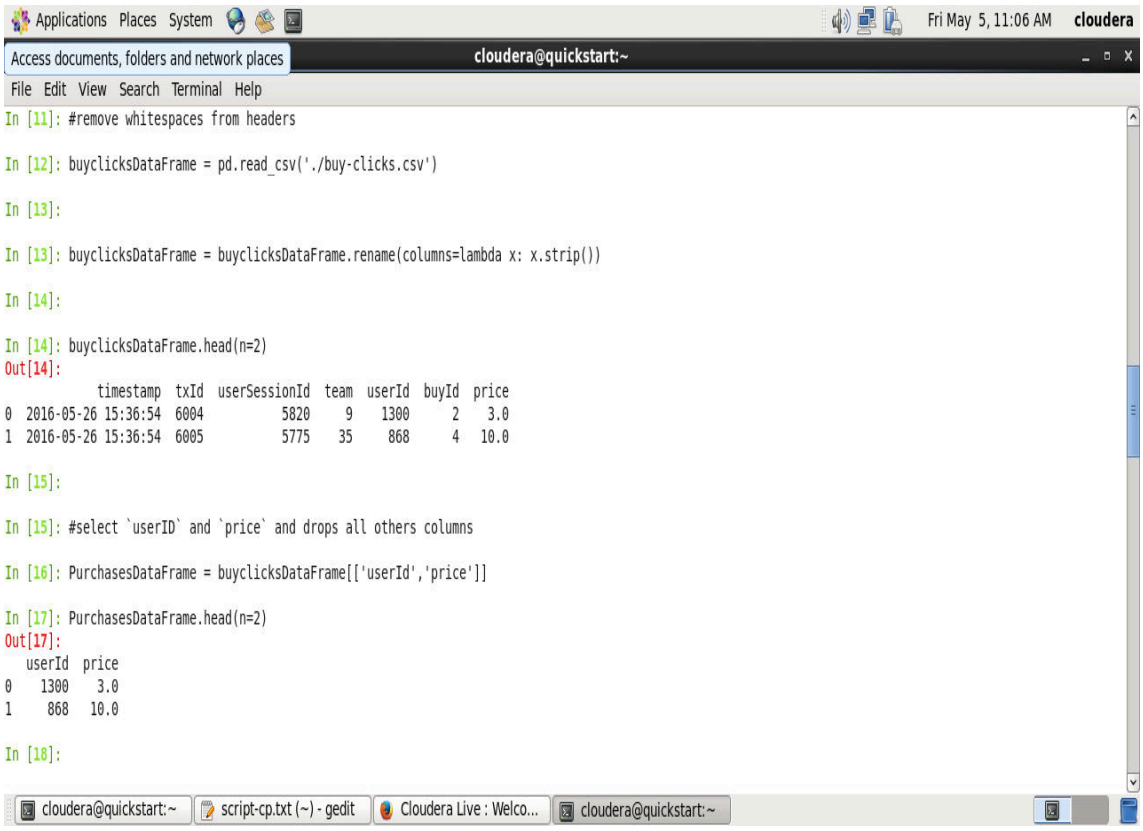
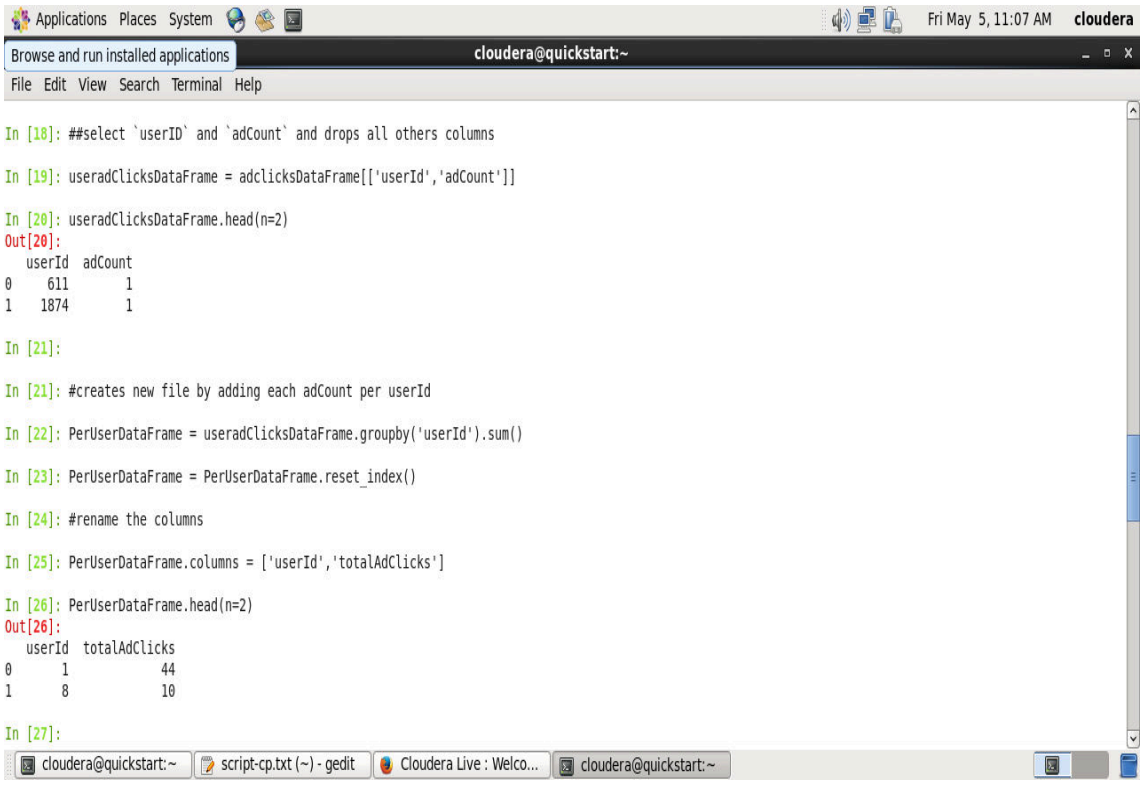
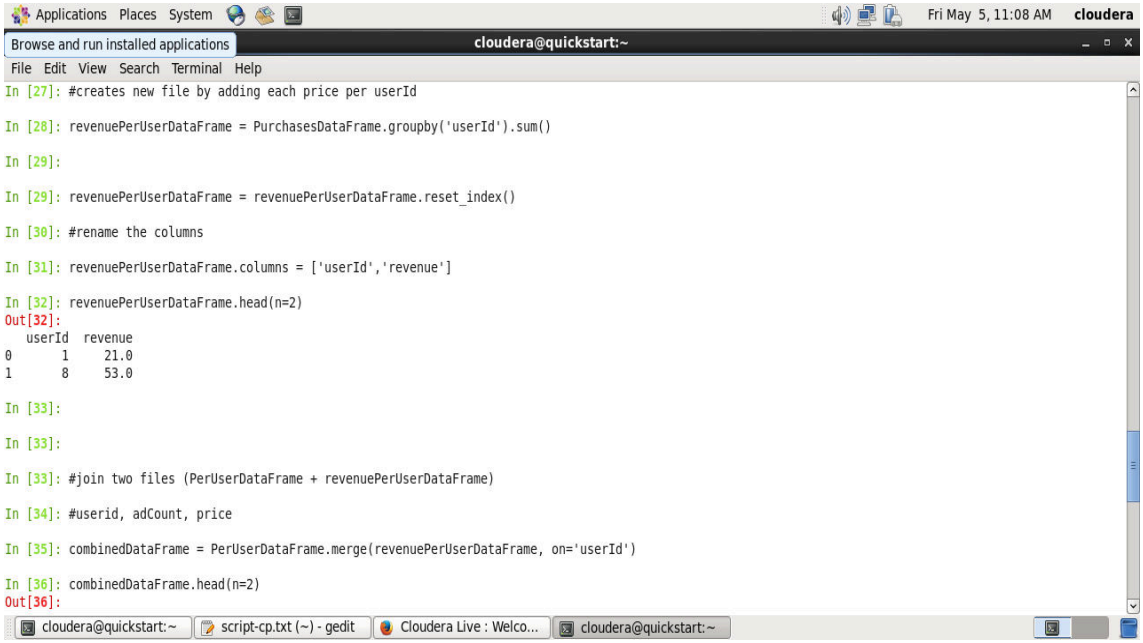
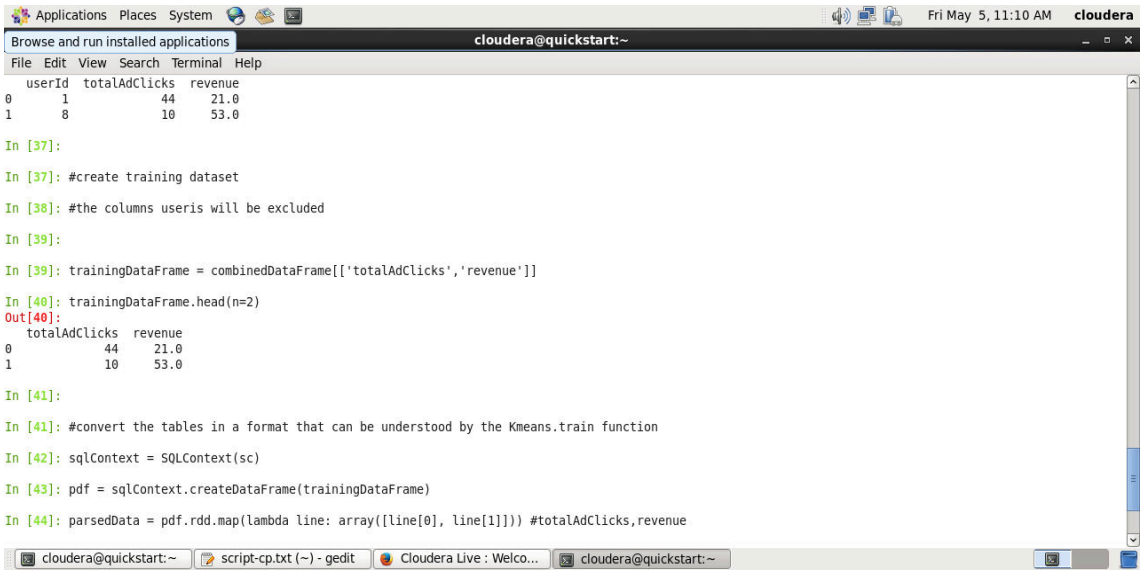

Applications Places System
Fri May 5, 11:03 AM cloudera
Access documents, folders and network places
cloudera@quickstart:~
File Edit View Search Terminal Help
Welcome to
version 1.6.0
Using Python version 3.5.1 (default, Dec 7 2015 11:16:01)
SparkContext available as sc, HiveContext available as sqlContext.
In [1]: import pandas as pd
In [2]:
In [2]: from pyspark.mllib.clustering import KMeans, KMeansModel
In [3]:
In [3]: from numpy import array
In [4]:
In [4]:
In [4]: #remove whitespaces from headers
In [5]:
In [5]: adclicksDataFrame = pd.read_csv('./ad-clicks.csv')
cloudera@quickstart:~
script-cp.txt (~) - gedit
Cloudera Live : Welco...
cloudera@quickstart:~


Applications Places System
Fri May 5, 11:05 AM cloudera
Access documents, folders and network places
cloudera@quickstart:~
File Edit View Search Terminal Help
In [6]:
In [6]: adclicksDataFrame = adclicksDataFrame.rename(columns=lambda x: x.strip())
In [7]: adclicksDataFrame.head(n=2)
Out[7]:
timestamp txId userSessionId team userId adId adCategory
0 2016-05-26 15:13:22 5974 5809 27 611 2 electronics
1 2016-05-26 15:17:24 5976 5705 18 1874 21 movies
In [8]:
In [8]: #add a columm adCount
In [9]: adclicksDataFrame['adCount'] = 1
In [10]:
In [10]: adclicksDataFrame.head(n=2)
Out[10]:
timestamp txId userSessionId team userId adId adCategory \
0 2016-05-26 15:13:22 5974 5809 27 611 2 electronics
1 2016-05-26 15:17:24 5976 5705 18 1874 21 movies
adCount
0 1
1 1
In [11]:
cloudera@quickstart:~
script-cp.txt (~) - gedit
Cloudera Live : Welco...
cloudera@quickstart:~


Applications Places System
Fri May 5, 11:06 AM cloudera
Access documents, folders and network places
cloudera@quickstart:~
File Edit View Search Terminal Help
In [11]: #remove whitespaces from headers
In [12]: buyclicksDataFrame = pd.read_csv('./buy-clicks.csv')
In [13]:
In [13]: buyclicksDataFrame = buyclicksDataFrame.rename(columns=lambda x: x.strip())
In [14]:
In [14]: buyclicksDataFrame.head(n=2)
Out[14]:
timestamp txId userSessionId team userId buyId price
0 2016-05-26 15:36:54 6004 5820 9 1300 2 3.0
1 2016-05-26 15:36:54 6005 5775 35 868 4 10.0
In [15]:
In [15]: #select `userID` and `price` and drops all others columns
In [16]: PurchasesDataFrame = buyclicksDataFrame[['userId','price']]
In [17]: PurchasesDataFrame.head(n=2)
Out[17]:
userId price
0 1300 3.0
1 868 10.0
In [18]:
cloudera@quickstart:~
script-cp.txt (~) - gedit
Cloudera Live : Welco...
cloudera@quickstart:~


Applications Places System
Fri May 5, 11:07 AM cloudera
Browse and run installed applications
cloudera@quickstart:~
File Edit View Search Terminal Help
In [18]: ##select `userID` and `adCount` and drops all others columns
In [19]: useradClicksDataFrame = adclicksDataFrame[['userId','adCount']]
In [20]: useradClicksDataFrame.head(n=2)
Out[20]:
userId adCount
0 611 1
1 1874 1
In [21]:
In [21]: #creates new file by adding each adCount per userId
In [22]: PerUserDataFrame = useradClicksDataFrame.groupby('userId').sum()
In [23]: PerUserDataFrame = PerUserDataFrame.reset_index()
In [24]: #rename the columns
In [25]: PerUserDataFrame.columns = ['userId','totalAdClicks']
In [26]: PerUserDataFrame.head(n=2)
Out[26]:
userId totalAdClicks
0 1 44
1 8 10
In [27]:
cloudera@quickstart:~
script-cp.txt (~) - gedit
Cloudera Live : Welco...
cloudera@quickstart:~


Applications Places System
Fri May 5, 11:08 AM cloudera
Browse and run installed applications
cloudera@quickstart:~
File Edit View Search Terminal Help
In [27]: #creates new file by adding each price per userId
In [28]: revenuePerUserDataFrame = PurchasesDataFrame.groupby('userId').sum()
In [29]:
In [29]: revenuePerUserDataFrame = revenuePerUserDataFrame.reset_index()
In [30]: #rename the columns
In [31]: revenuePerUserDataFrame.columns = ['userId','revenue']
In [32]: revenuePerUserDataFrame.head(n=2)
Out[32]:
userId revenue
0 1 21.0
1 8 53.0
In [33]:
In [33]:
In [33]: #join two files (PerUserDataFrame + revenuePerUserDataFrame)
In [34]: #userid, adCount, price
In [35]: combinedDataFrame = PerUserDataFrame.merge(revenuePerUserDataFrame, on='userId')
In [36]: combinedDataFrame.head(n=2)
Out[36]:
cloudera@quickstart:~
script-cp.txt (~) - gedit
Cloudera Live : Welco...
cloudera@quickstart:~


Applications Places System
Fri May 5, 11:10 AM cloudera
Browse and run installed applications
cloudera@quickstart:~
File Edit View Search Terminal Help
userId totalAdClicks revenue
0 1 44 21.0
1 8 10 53.0
In [37]:
In [37]: #create training dataset
In [38]: #the columns useris will be excluded
In [39]:
In [39]: trainingDataFrame = combinedDataFrame[['totalAdClicks','revenue']]
In [40]: trainingDataFrame.head(n=2)
Out[40]:
totalAdClicks revenue
0 44 21.0
1 10 53.0
In [41]:
In [41]: #convert the tables in a format that can be understood by the Kmeans.train function
In [42]: sqlContext = SQLContext(sc)
In [43]: pdf = sqlContext.createDataFrame(trainingDataFrame)
In [44]: parsedData = pdf.rdd.map(lambda line: array([line[0], line[1]])) #totalAdClicks,revenue
cloudera@quickstart:~
script-cp.txt (~) - gedit
Cloudera Live : Welco...
cloudera@quickstart:~


Applications Places System
Fri May 5, 11:13 AM cloudera
Browse and run installed applications
cloudera@quickstart:~
File Edit View Search Terminal Help
In [45]: my_kmmodel = KMeans.train(parsedData, 2, maxIterations=10, runs=10, initializationMode="random")
/usr/lib/spark/python/pyspark/mllib/clustering.py:176: UserWarning: Support for runs is deprecated in 1.6.0. This param will have no effect in 1.7.0.
"Support for runs is deprecated in 1.6.0. This param will have no effect in 1.7.0.")
17/05/05 11:02:54 WARN netlib.BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeSystemBLAS
17/05/05 11:02:54 WARN netlib.BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeRefBLAS
In [46]: print(my_kmmodel.centers)
[array([27.39467849, 23.86474501]), array([39.07608696, 115.26086957])]
In [47]:
In [47]:
In [47]:
In [47]:
In [47]:
In [47]:
In [47]:
In [47]:
In [47]:
In [47]:
cloudera@quickstart:~
script-cp.txt (~) - gedit
Cloudera Live : Welco...
cloudera@quickstart:~