Sportify Analysis problem:

#TO unzip file

Tar -zxvf data\_sample.tar

#To count the individual file rows

wc -l user\_data\_sample.csv

wc -l end\_song\_sample.csv

# Fetch top rows

head –n 10 user\_data\_sample.csv

head –n 10 end\_song\_sample.csv

|  |  |  |
| --- | --- | --- |
| File Name | # of Rows |  |
| user\_data\_sample.csv | 9566 |  |
| end\_song\_sample.csv | 1,342,892 |  |

Unique Track\_id = 314,985

Rows columns

404805 10

404804 – sample rows – 10% population data (33% overall)

Column Mean\_value

**time\_delta 32.609998**

Methods and Technique for correlation in pandas:

# This is to find correlation between two column/variables with specified methods

Df[‘a’].corr(df[‘b’],method = ‘spearman’)

# This is by default each column with all other columns

Df.corr()

Note that **non-numeric columns** will be automatically excluded from the correlation calculation

# Below are some methods for determine correlation

| **Method name** | **Description** |
| --- | --- |
| pearson (default) | Standard correlation coefficient |
| kendall | Kendall Tau correlation coefficient |
| spearman | Spearman rank correlation coefficient |

To install Pandas & numpy on AWS EC2 & EMR Cluster

sudo yum -y install numpy

sudo pip install pandas

pip install --upgrade pip

Hive EMR External Table

create external table user\_data(

Gender string

,age\_range string

,country string

,acct\_age INT

,user\_id string)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION 's3://spotifyc/user\_data/';

create external table song\_data(

ms\_played BigINT

,context string

,track\_id string

,product string

,end\_timestamp string

,user\_id string)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION 's3://spotifyc/end\_song\_data/';

create external table join\_data(

index int,

gender string,

age\_range string ,country string,

acct\_age\_weeks int,

user\_id string,

ms\_played bigint,

context string,

track\_id string,

product string,

end\_timestamp string,

date\_time timestamp)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION 's3://spotifyc/final\_data/';