# Data Engineer-Core

Time Duration [2, Day]

**Intro:**  
To be able to analyze our customer behavior and content we need to combine different data sets. The tasks are explained on a story level, which means the user and/or the backlog owner have created them. The ETL should be documented by having a readme which explains what you’re doing in the code and a dataset documentation explaining the new tables, datatypes and where the columns originate from. When you are done please zip your code and datasets in a file and send them to us. If any questions should arise then feel free to ask.

Some good to know nomenclature:  
  
**STARTED\_STREAMS** : One customer view of the title (or episode) per day, time, device, country, title, season, genre and product\_type

**SVOD** : Subscription based Video On Demand service

**TVOD** : Transaction based Video On Demand service (rent movie)

**EST** : Electronic Sell Through (buy movie)

**HOUSE\_NUMBER** : unique identifier for a title

**USER\_ID** : Users unique identifier

**WHATSON** : Source metadata manager for the programs and titles

**In this exercise you will use two input files:**

1. Data set started streams which contains the following fields (dt,"time", device\_name, house\_number,user\_id,country\_code,program\_title, season, season\_episode, genre,product\_type).
2. Data set whatson which contains the following fields (dt, house\_number, title,product\_category, broadcast\_right\_region, broadcast\_right\_vod\_type, broadcast\_right\_start\_date, broadcast\_right\_end\_date).

**Tasks:**

Using your own environment setup and your choice of etl language (Python/Java/Scala preferably) do the following ETL jobs

1. Sales and rentals broadcast rights:
   * We need to find the broadcast rights for a title to be able to expose it to the analytics team for further analysis. This should be for Product Types: TVOD and EST. Matching on most recent date for whatson data and joining based on the house\_number and country.
   * Input: whatson.csv and started\_streams.csv
   * Expected Output:

dt,

time,

device\_name,

house\_number,

user\_id,

country\_code,

program\_title,

season,

season\_episode,

genre,

product\_type,

broadcast\_right\_start\_date,

broadcast\_right\_end\_date

1. Product and user count:
   * We need to know how many watches a product is getting and how many unique users are watching the content, in what device, country and what product\_type .
   * Input: started\_streams.csv
   * Expected output:

dt,

program\_title,

device\_name,

country\_code,

product\_type,

unique\_users,

content\_count

1. Genre and time of day:
   * We need a list with the most popular Genre and what hours people watch?
   * Input: started\_streams.csv
   * Expected output:

watched\_hour,

genre,

unique\_users