Data engineer screening questions

Query #1 Write a query to return a table of genres with the oldest and newest movie in that genre.

```
select m1.genre,
(
  select title from movies m2
  where m1.genre = m2.genre
  and m2.year = (
  select max(year) from movies m3
  where m2.genre = m3.genre
)

) newest,
(
  select title from movies m3
  where m1.genre = m3.genre
  and m3.year = (
  select min(year) from movies m4
  where m3.genre = m4.genre
)

) oldest

from movies m1
group by m1.genre;
```

genre	newest	oldest
Adventure	Jurassic Park	King Kong
Animated	Spirited Away	Snow White and the Seven Dwarfs
Biopic	Lincoln	Lawrence of Arabia
Action	Casino Royale	The Road Warrior

Query #2 Write a query to get the number of movies broken down by decade

```
select year/10*10 decade, count(*) movies_count
from movies
group by year/10*10
order by decade desc;
```

decade	movies_count
2010	1
2000	2
1990	5
1980	5
1970	1
1960	1
1940	2
1930	3

Query #3 How would you extract the rest of the 100 movies from that page and transform into the SQL DML like the above?

One way to get information from webpages is to use a webscripting technique. Webscripting allows you to get information off an html page. Python has severial packages that can be used, one is the Beautifulsoup package. With Beautifulsoup, I would be able to get the needed movie information (if webscripting is allowed on that page), and then once I have all what I need, I can then transform the information to SQL DML or anything else. Another way is to use the source API along with any authentication provided. This is the most effecient way of ingesting data from source to destination.