

## CS777 - Week 1 Homework Submission Template

### **Task 1 – Top-10 Active Taxis**

Many different taxis have had multiple drivers. Write and execute a Spark Python program that computes the top ten taxis that have had the largest number of drivers. Your output should be a set of (medallion, number of drivers) pairs.

*Note: You should consider that this is a real-world data set that might include wrongly formatted data lines. You should clean up the data before the main processing, a line might not include all of the fields. If a data line is not correctly formatted, you should drop that line and do not consider it.*

- Print a list of top 10 taxis having the largest number of drivers, and the amount of drivers (taxi ID and count)

#### Small Dataset

```
('65EFB7D02BAD12D5DE757CB5D350944E', 20)
('3C08296D0EB7ABE24FB7328DE9B62813', 20)
('55D311AD2752BC278BEF7386B25B28A9', 19)
('CD7B02776E69483397952DC5E1F44DFE', 19)
('7DEB25123AE57111F912C0EBF92F1F63', 19)
('F36564AB9C6EA3B6373EB0E1680A447A', 19)
('3B6AE3CF05F34ADC91DC68D20F2EB913', 19)
('799153A138F4E8334A1A95AE25040B83', 18)
('9FB7A7C1D7B960D8B17829145C6A1CF2', 18)
('F2A08960199BCDB7EE19411A8E7A4C5D', 18)
```

#### Large Dataset

```
('11DC93DD66D8A9A3DD9223122CF99EFD', 352)
('EE06BD8A621CAC3B608ACFDF0585A76A', 348)
('6C1132EF70BC0A7DB02174592F9A64A1', 341)
('A10A65AFD9F401BF3BDB79C84D3549E7', 340)
('23DB792D3F7EBA03004E470B684F2738', 339)
('7DA8DF1E4414F81EBD3A0140073B2630', 337)
('0318F7BBB8FF48688698F04016E67F49', 335)
('738A62EEE9EC371689751A864C5EF811', 333)
('B07944BF31699A169091D2B16597A4A9', 333)
('7D93E7FC4A7E4615A34B8286D92FF57F', 333)
```

### **Task 2 – Top-10 Best Drivers**

We would like to figure out who the top 10 best drivers are in terms of their average earned money per minute spent carrying a customer. The total amount field is the total money earned on a trip. In the end, we are interested in computing a set of (driver, money per minute) pairs.

- Print a list of top 10 best drivers based on earned money per minute carrying a customer (Driver ID and average earning)

#### Small Dataset

```
('32187D24B8C6D9DC9B1254330292F42E', 45.33626373626374)
('7930AF3429B99C59346390635AFDFFCF', 37.31236942250107)
('011AE79C7E609378068514E5C992B6D6', 31.470588235294116)
('583D58A6E31DBAF275DDFAD1857448D2', 26.619718309859156)
('C42ABFD8903E6A5B2E81EA6331B05700', 22.25260749613081)
('C742CFD86A6B2ABFB9CD7228286766CA', 17.313432835820894)
('62A757062319F29FA98D15C8DF8A6BF6', 13.844295302013425)
('7BDFF06419C23F667C5D69EEFBF091BE', 11.313253012048193)
('7F274F176A7BA8D55B5A7D0F2580634C', 7.301946902654868)
```

('BD50C5800362CB05615C52A0370E1A80', 7.247362012987013)

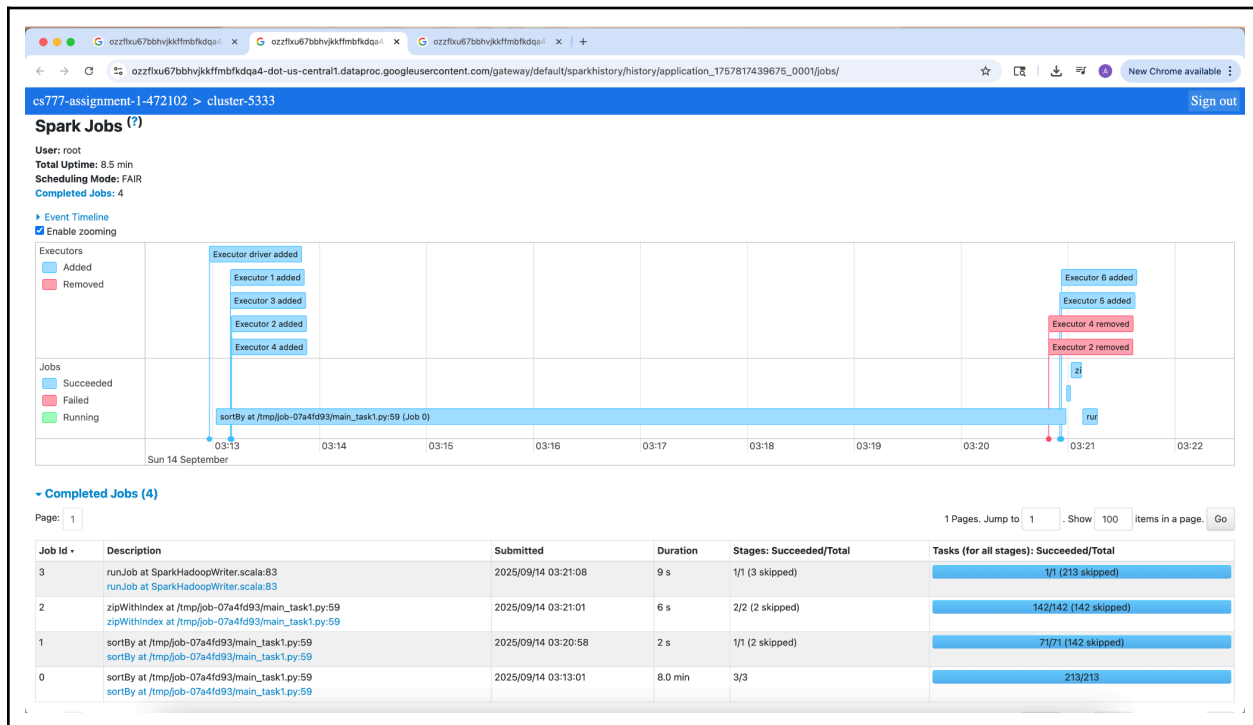
### Large Dataset

('E4F99C9ABE9861F18BCD38BC63D007A9', 40.88425676090866)  
('BA721F7DC14E1D7C93F30BB863E4B83C', 17.5)  
('AD4660069108F79079A23D5E05358565', 13.125)  
('664927CDE376A32789BA48BF55DFB7E3', 11.583838273060598)  
('32BB90E8976AAB5298D5DA10FE66F21D', 10.56700507614213)  
('021FF8A9BECC2EFF3B3EC40A10B397E6', 10.5)  
('19CA14E7EA6328A42E0EB13D585E4C22', 8.089887640449438)  
('BD8ED3630EB46807C78407285322257B', 7.82892732579303)  
('9C9540118D2725A4A63AF71BE096ABB4', 7.5)  
('094B6D6E876E9B6DF60B84B3F5567FC9', 7.166666666666667)

### Spark History Output:

To demonstrate that you did execute your code on the cloud it is important to include URLs in the screenshots. Otherwise, there is no way for us to verify if the code was executed in your cloud account.

### Task 1:



Task 2:

