**Name: Bharat Chandra**

AWS Big Data Project

**Problem statement:**

Imagine that you are working as an analyst in a famous taxi app company. Your organization

provides hassle-free travel to people all around the world. You are provided with an AWS

account to perform certain analytical tasks using AWS cloud services.

**Link for the dataset and resources:**

https://intellipaat-course-attachments.s3.ap-south-1.amazonaws.com/Hadoop/Hadoop+Datas

ets-20200609T120700Z-001.zip

**Dataset description:**

In here, you have a predefined dataset (yellow.csv), having more than 15 columns.

The dataset has different attributes as follows:

vendor\_id string,

pickup\_datetime string,

dropoff\_datetime string,

passenger\_count int,

trip\_distance DECIMAL(9,6),

pickup\_longitude DECIMAL(9,6),

pickup\_latitude DECIMAL(9,6),

rate\_code int,

store\_and\_fwd\_flag string,

dropoff\_longitude DECIMAL(9,6),

dropoff\_latitude DECIMAL(9,6),

payment\_type string,

fare\_amount DECIMAL(9,6),

extra DECIMAL(9,6),

mta\_tax DECIMAL(9,6),

tip\_amount DECIMAL(9,6),

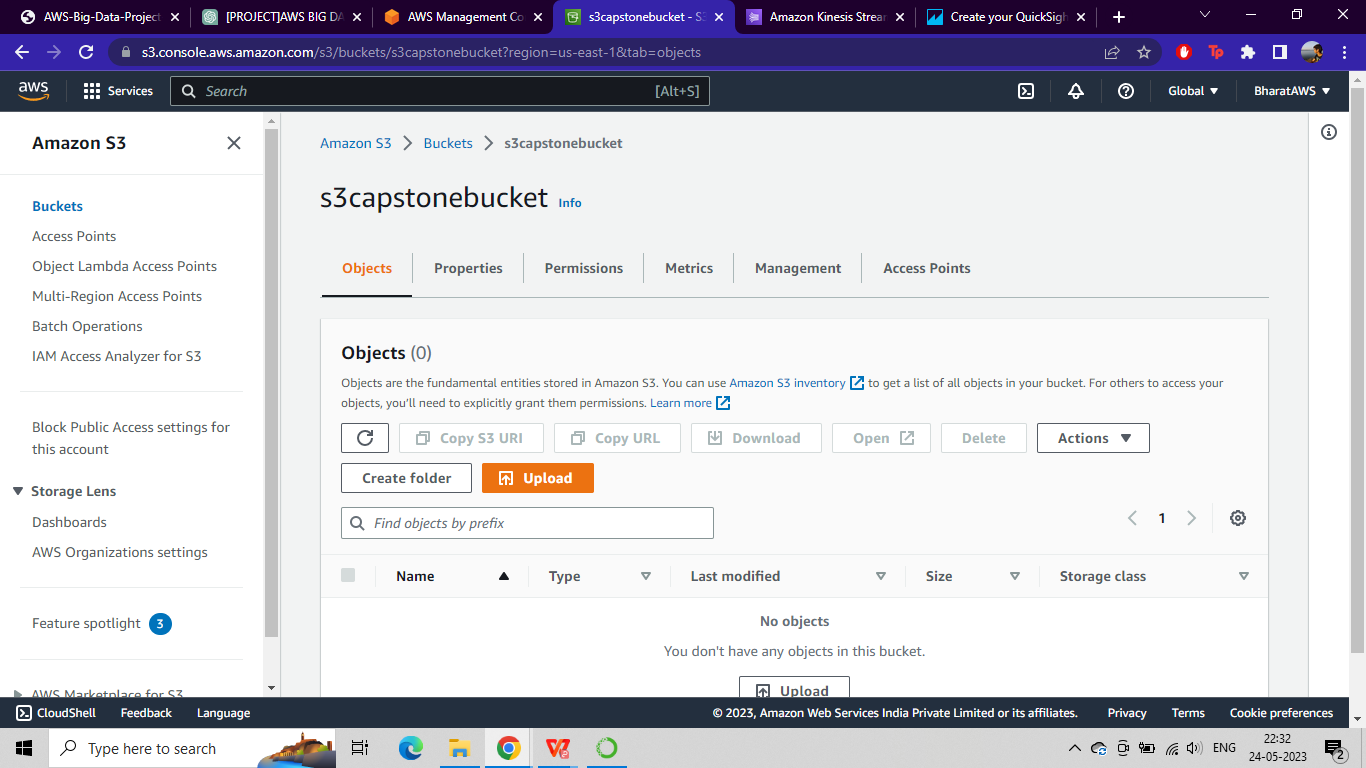
tolls\_amount DECIMAL(9,6),

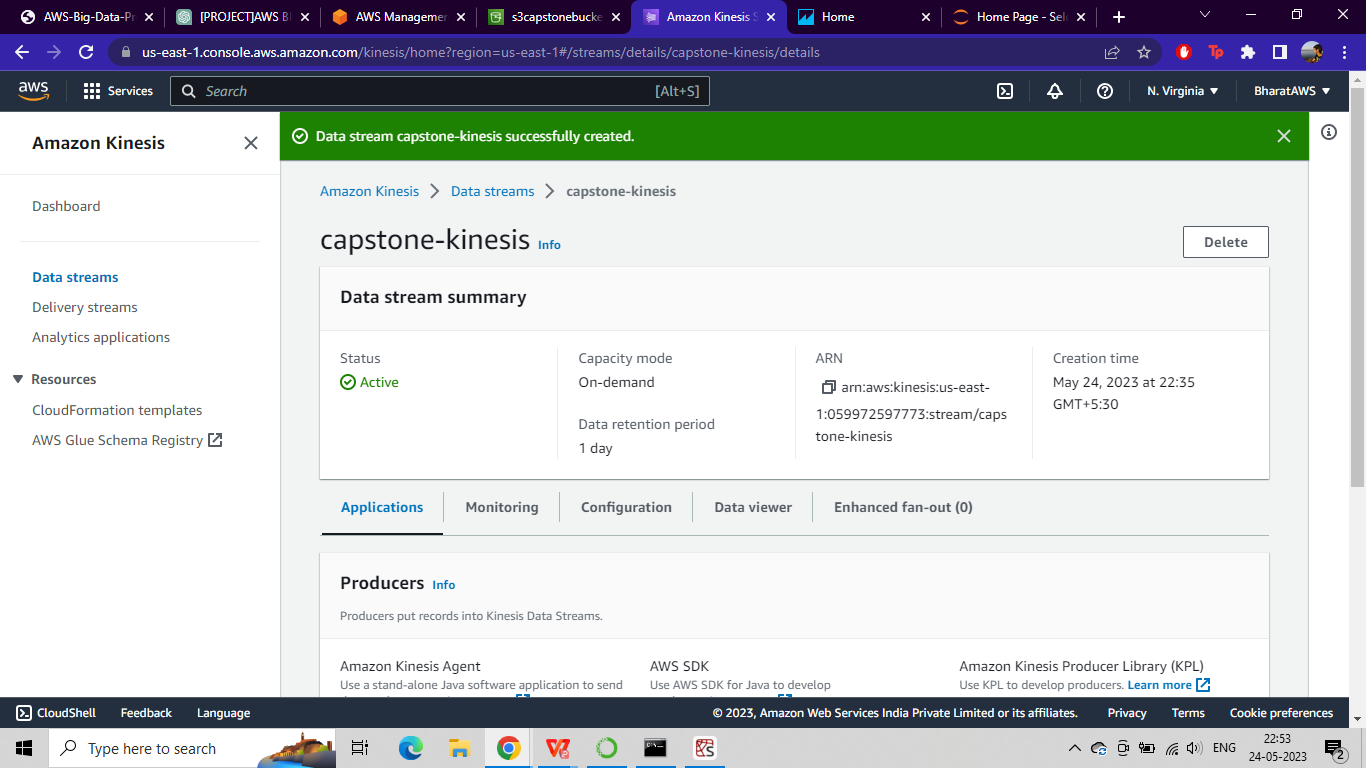
total\_amount DECIMAL(9,6),

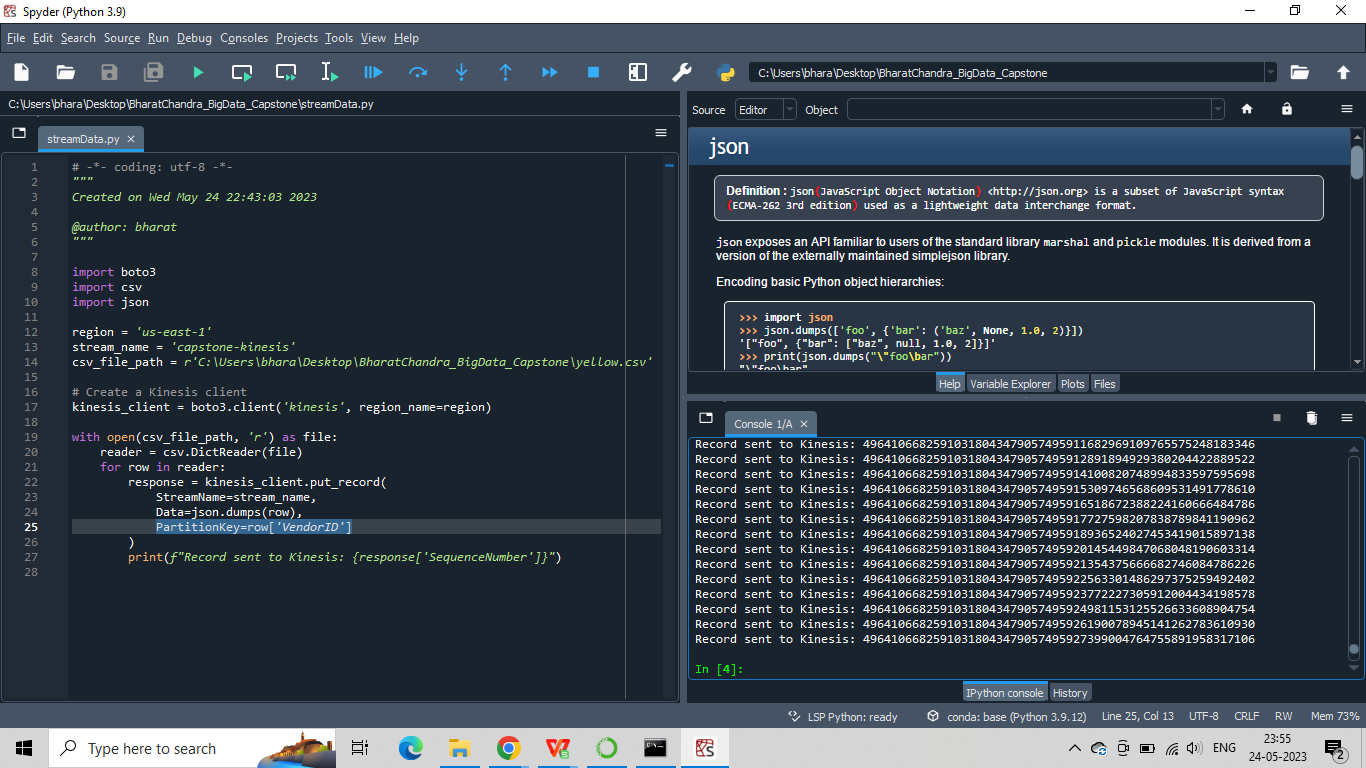
trip\_time\_in\_secs int

**Tasks:**

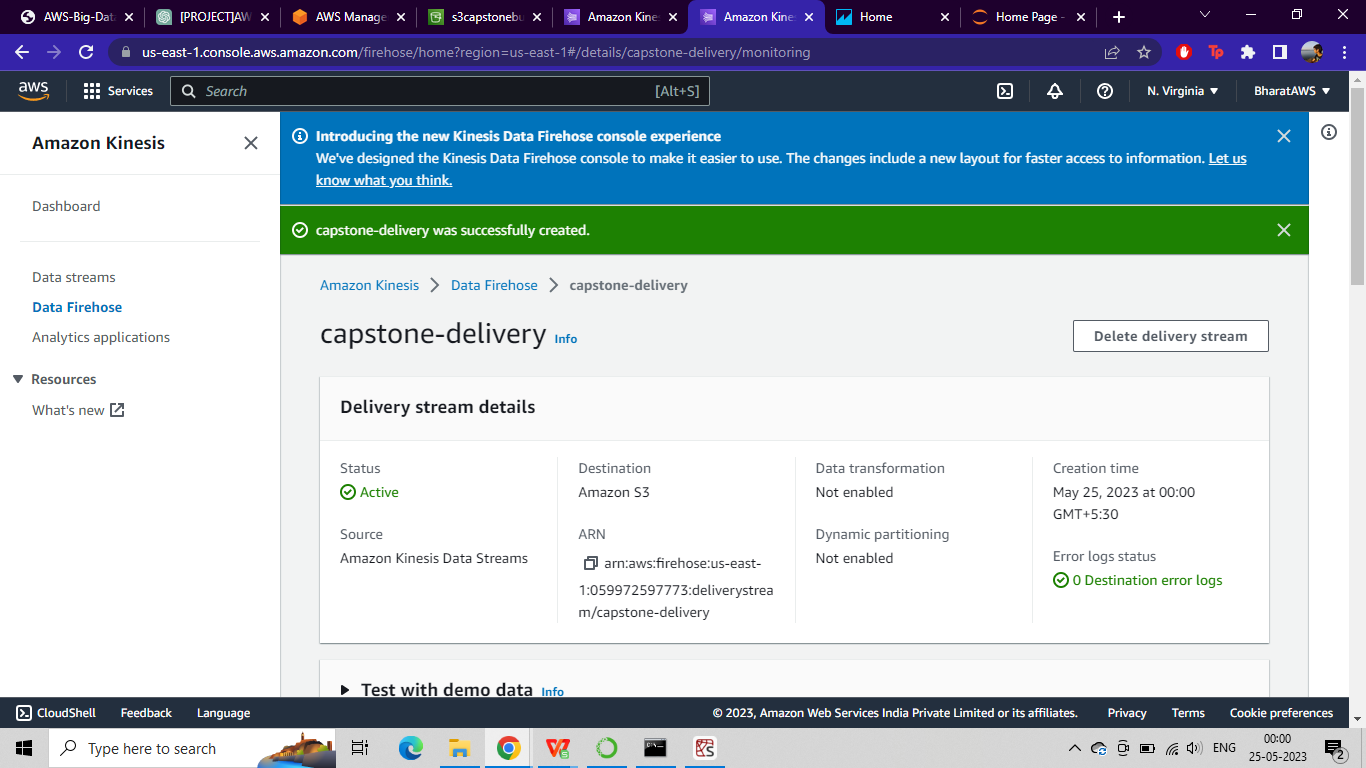
● Stream the taxi data to a Kinesis Stream



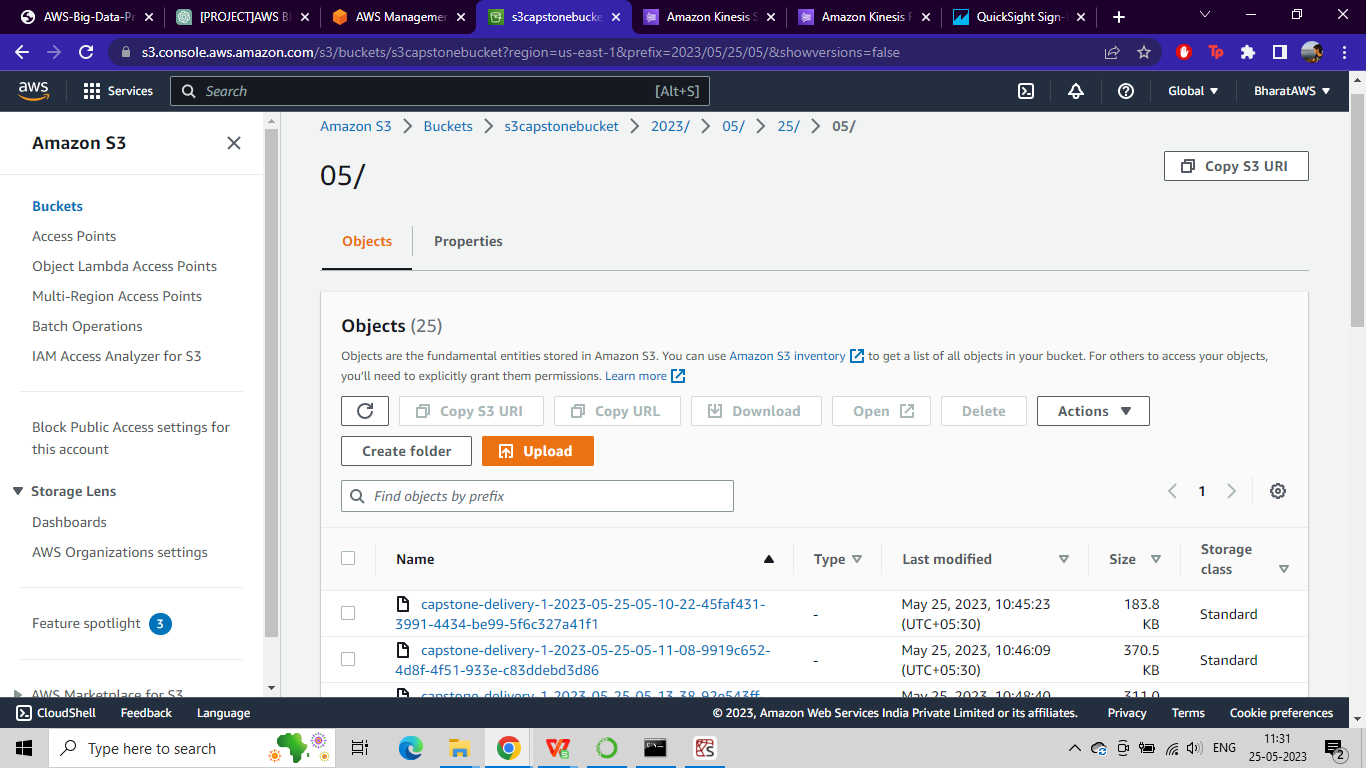


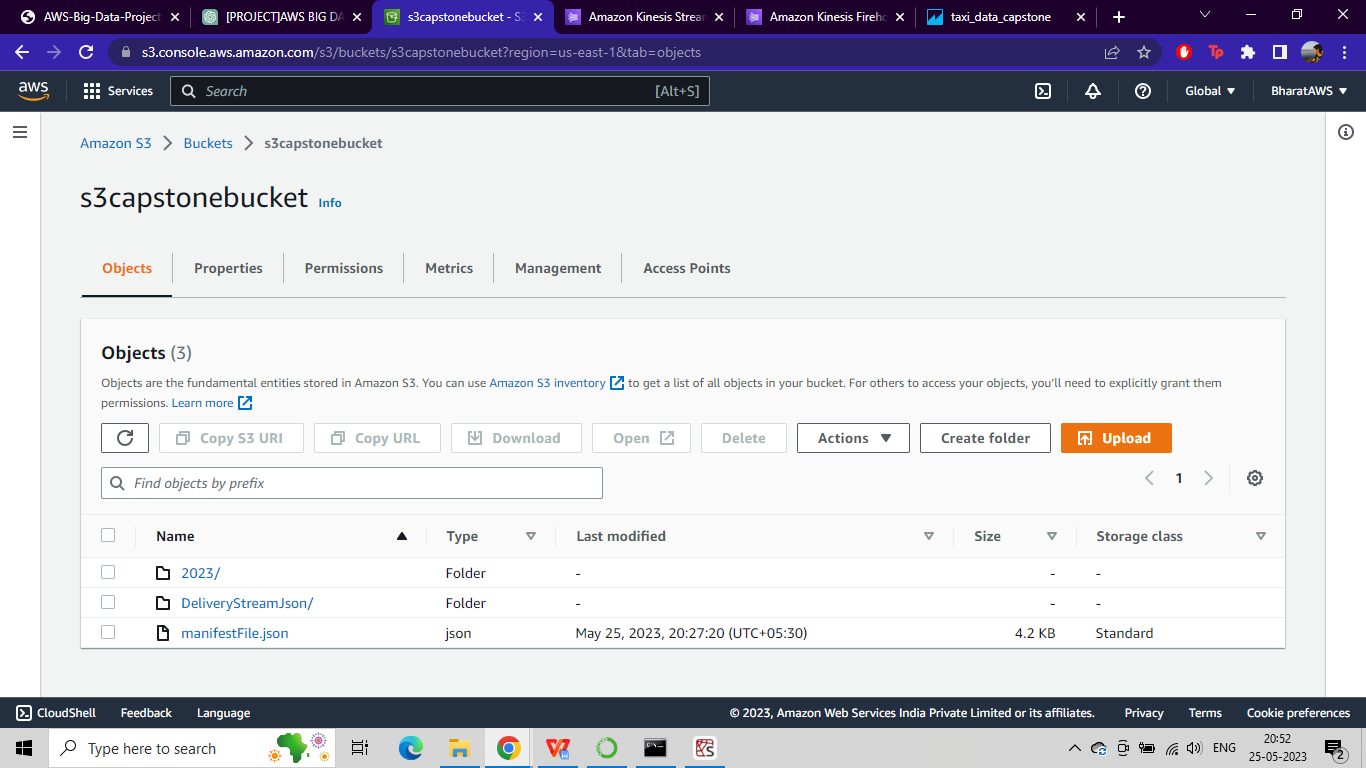


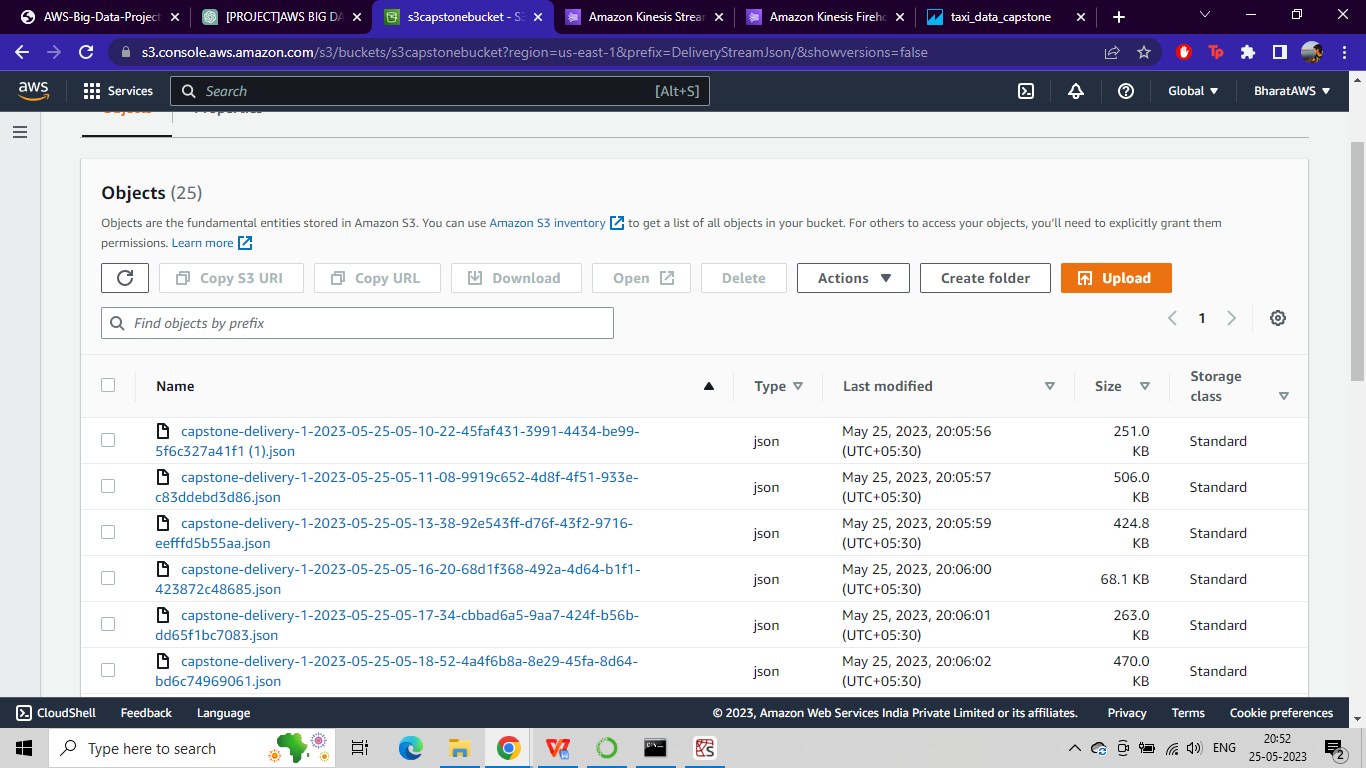
● Link the Kinesis Stream to a Kinesis Firehose delivery stream



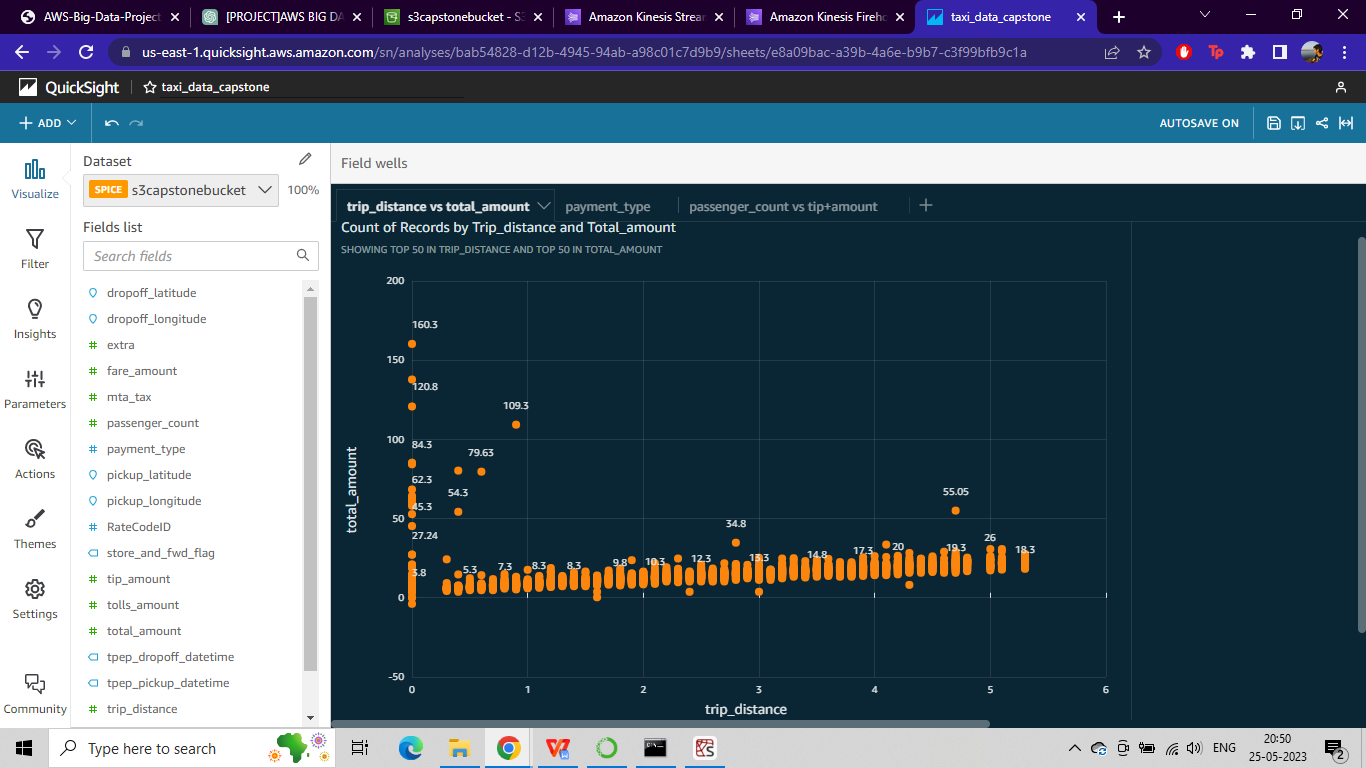
● Deliver the streamed data to Amazon S3



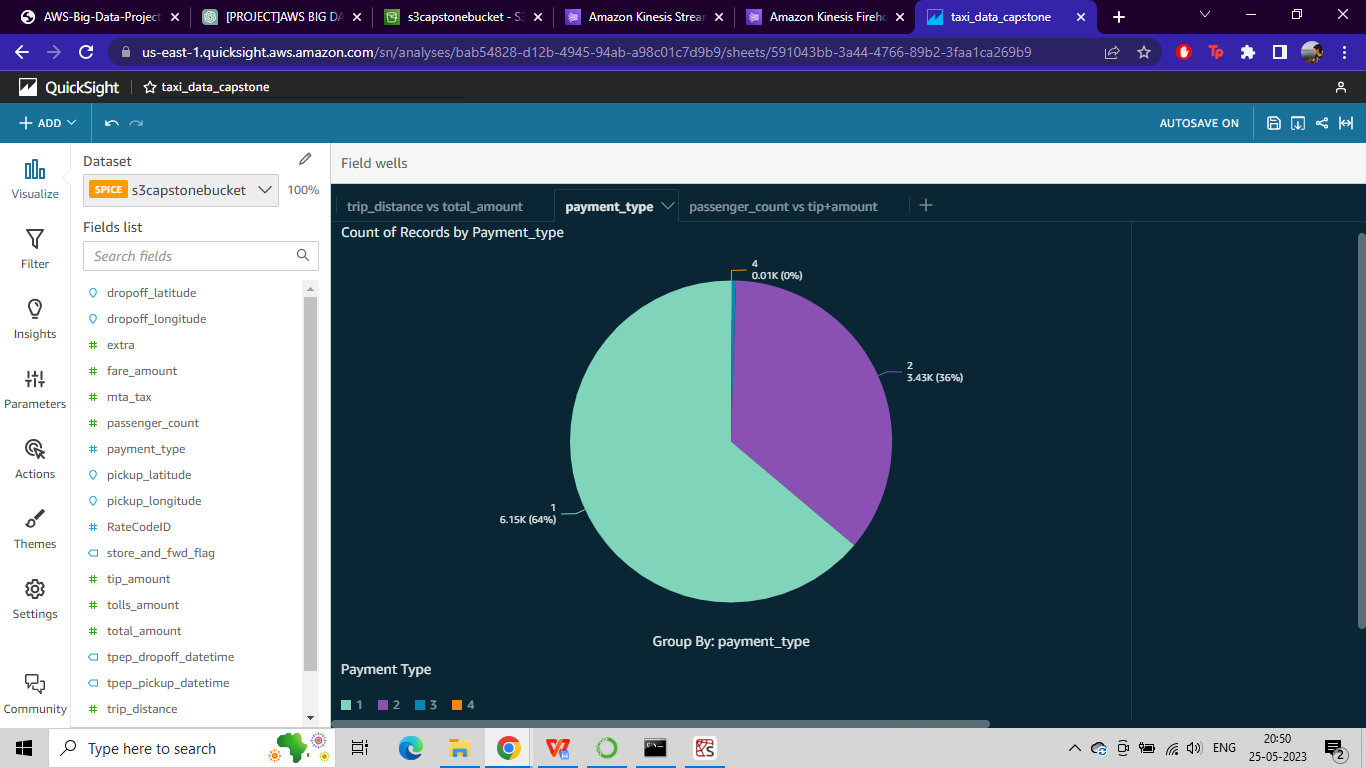




● Visualize the relationship between the total fare and the distance covered in trips(Amazon Quicksight)



● Visualize the percentage of payment types using a pie chart(Amazon Quicksight)



● Visualize the relationship between the passenger count and the tip amount(Amazon Quicksight)



**Use EMR to find out the following details:**

1. What is the total number of trips?

|  |
| --- |
| **total\_trips** |
| 10000 |

1. What is the total revenue generated by all the trips? (The fare is stored in the column

total\_amount)

|  |
| --- |
| **total\_revenue** |
| 160546.80999999488 |

1. What fraction of the total is paid for tolls? (Toll fares are stored in tolls\_amount)

|  |
| --- |
| **tolls\_fraction** |
| 0.015553034034124159 |

1. What fraction of the total is paid as driver tips? (Tips are stored in tip\_amount)

|  |
| --- |
| **tips\_fraction** |
| 0.10785203393328445 |

1. What is the average trip amount?

|  |
| --- |
| **average\_trip\_amount** |
| 16.054680999999487 |

1. What is the average distance of the trips? (Distances are stored in the column trip\_distance)

|  |
| --- |
| **average\_trip\_distance** |
| 3.253033000000003 |

1. How many different payment types are used?

|  |
| --- |
| **payment\_types\_count** |
| 4 |

1. For each payment type, display the following details:

● Average fare generated

● Average tip

● Average tax (Tax is stored in column mta\_tax)

|  |  |  |  |
| --- | --- | --- | --- |
| **payment\_type** | **average\_fare** | **average\_tip** | **average\_tax** |
| 2 | 11.393383098591547 | 0 | 0.4988732394366197 |
| 1 | 13.561018272684619 | 2.704248008745903 | 0.49711072934561923 |
| 3 | 13.21078947368421 | 0 | 0.42105263157894735 |
| 4 | 12.222222222222221 | 0 | 0.5 |

1. On average, which hour of the day generates the highest revenue?

|  |  |
| --- | --- |
| **hour** | **revenue** |
| 22 | 109750.60000000677 |