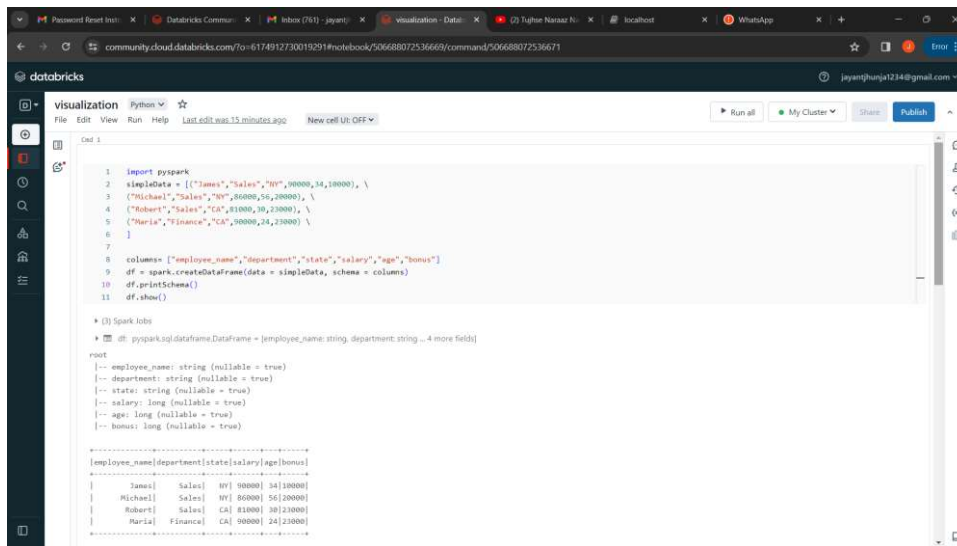


Name –Jayant Jhunja

Data Engineering Assignment



```
1 import pyspark
2 simpleData = [{"name": "James", "Sales", "NY", 90000, 34, 10000}, \
3 ("Michael", "Sales", "NY", 86000, 36, 20000), \
4 ("Robert", "Sales", "CA", 81000, 30, 20000), \
5 ("Maria", "Finance", "CA", 90000, 24, 20000) \
6 ]
7
8 columns = ["employee_name", "department", "state", "salary", "age", "bonus"]
9 df = spark.createDataFrame(data = simpleData, schema = columns)
10 df.printSchema()
11 df.show()
```

root

```
-- employee_name: string (nullable = true)
-- department: string (nullable = true)
-- state: string (nullable = true)
-- salary: long (nullable = true)
-- age: long (nullable = true)
-- bonus: long (nullable = true)
```

employee_name	department	state	salary	age	bonus
James	Sales	NY	90000	34	10000
Michael	Sales	NY	86000	36	20000
Robert	Sales	CA	81000	30	20000
Maria	Finance	CA	90000	24	20000

