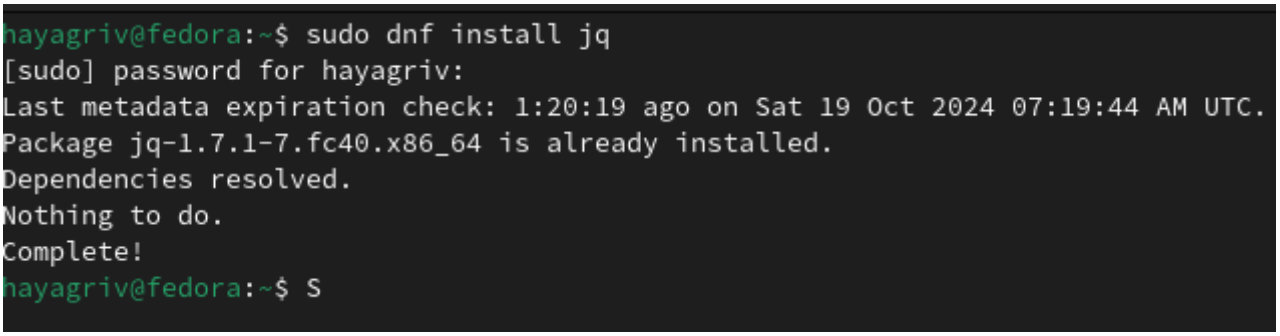
Exp. No : 6

Handling JSON data using HDFS and Python

1. Create emp.json file

```
hayagriv@fedora:~$ hadoop fs -cat /exp6/emp.json
[
        {
                "name": "Jane",
                "age": 30,
                "department": "HR",
                "Salary": 50000
        },
        {
                "name": "Bob",
                "age": 25,
                "department": "Marketing",
                "Salary": 60000
        },
        {
                "name": "Charlie",
                "age": 32,
                "department": "IT",
                "Salary": 70000
        },
        {
                "name": "Mark",
                "age": 28,
                "department": "Finance",
                "Salary": 55000
        },
        {
                "name": "Chris",
                "age": 38,
                "department": "IT",
                "Salary": 80000
        }
]
hayagriv@fedora:~$
```

2. Install jq package



3. Execute jq . emp.json command

```
hayagriv@fedora:~$ hadoop fs -cat /exp6/emp.json
[
        {
                "name": "Jane",
                "age": 30,
                "department": "HR",
                "Salary": 50000
        },
        {
                "name": "Bob",
                "age": 25,
                "department": "Marketing",
                "Salary": 60000
        },
        {
                "name": "Charlie",
                "age": 32,
                "department": "IT",
                "Salary": 70000
        },
        {
                "name": "Mark",
                "age": 28,
                "department": "Finance",
                "Salary": 55000
        },
        {
                "name": "Chris",
                "age": 38,
                "department": "IT",
                "Salary": 80000
        }
]
hayagriv@fedora:~$
```

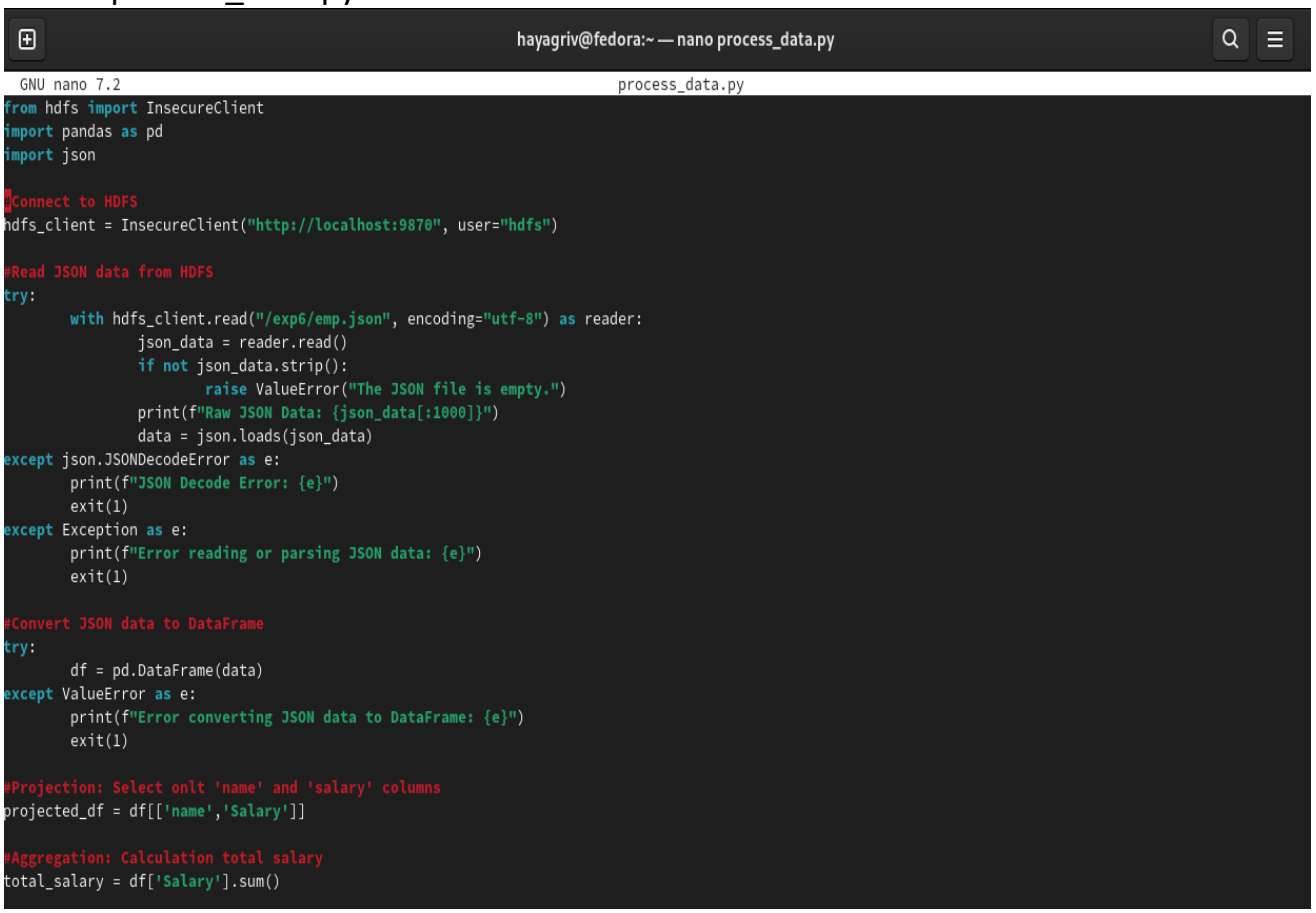
4. pip install pandas

```
hayagriv@fedora:~$ pip install pandas
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: pandas in ./.local/lib/python3.12/site-packages (
2.2.3)
Requirement already satisfied: numpy>=1.26.0 in ./.local/lib/python3.12/site-pac
kages (from pandas) (2.1.2)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/lib/python3.12/sit
e-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in ./.local/lib/python3.12/site-pack
ages (from pandas) (2024.2)
Requirement already satisfied: tzdata>=2022.7 in ./.local/lib/python3.12/site-pa
ckages (from pandas) (2024.2)
Requirement already satisfied: six>=1.5 in /usr/lib/python3.12/site-packages (fr
om python-dateutil>=2.8.2->pandas) (1.16.0)
```

5. pip install hdfs

```
ayagriv@fedora:~$ pip install hdfs
efaulting to user installation because normal site-packages is not writeable
equirement already satisfied: hdfs in ./.local/lib/python3.12/site-packages (2.
.3)
equirement already satisfied: docopt in ./.local/lib/python3.12/site-packages (
rom hdfs) (0.6.2)
equirement already satisfied: requests>=2.7.0 in /usr/lib/python3.12/site-packa
es (from hdfs) (2.31.0)
equirement already satisfied: six>=1.9.0 in /usr/lib/python3.12/site-packages (
rom hdfs) (1.16.0)
equirement already satisfied: charset-normalizer<4,>=2 in /usr/lib/python3.12/s
te-packages (from requests>=2.7.0->hdfs) (3.3.2)
equirement already satisfied: idna<4,>=2.5 in /usr/lib/python3.12/site-packages
(from requests>=2.7.0->hdfs) (3.7)
equirement already satisfied: urllib3<3,>=1.21.1 in /usr/lib/python3.12/site-pa
kages (from requests>=2.7.0->hdfs) (1.26.20)
ayagriv@fedora:~$
```

Create process_data.py



Output:

```
Raw JSON Data: [{"name": "John Doe", "age": 30, "department": "HR", "salary": 50000},
{"name": "Jane Smith", "age": 25, "department": "IT", "salary": 60000},
{"name": "Alice Johnson", "age": 35, "department": "Finance", "salary": 70000},
{"name": "Bob Brown", "age": 28, "department": "Marketing", "salary": 55000},
{"name": "Charlie Black", "age": 45, "department": "IT", "salary": 80000}]

Filtered JSON file saved successfully.
Projection: Select only name and salary columns
            name  salary
0       John Doe   50000
1     Jane Smith   60000
2  Alice Johnson   70000
3      Bob Brown   55000
4  Charlie Black   80000
Aggregation: Calculate total salary
Total Salary: 315000


# Count: Number of employees earning more than 50000
Number of High Earners (>50000): 4


Limit: Top 5 highest salary
Top 5 Earners:
            name  age department  salary
4  Charlie Black   45         IT   80000
2  Alice Johnson   35    Finance   70000
1     Jane Smith   25         IT   60000
3      Bob Brown   28  Marketing   55000
0       John Doe   30         HR   50000


Skipped DataFrame (First 2 rows skipped):
            name  age department  salary
2  Alice Johnson   35    Finance   70000
3      Bob Brown   28  Marketing   55000
4  Charlie Black   45         IT   80000


Filtered DataFrame (Sales department removed):
            name  age department  salary
0       John Doe   30         HR   50000
2  Alice Johnson   35    Finance   70000
3      Bob Brown   28  Marketing   55000
```