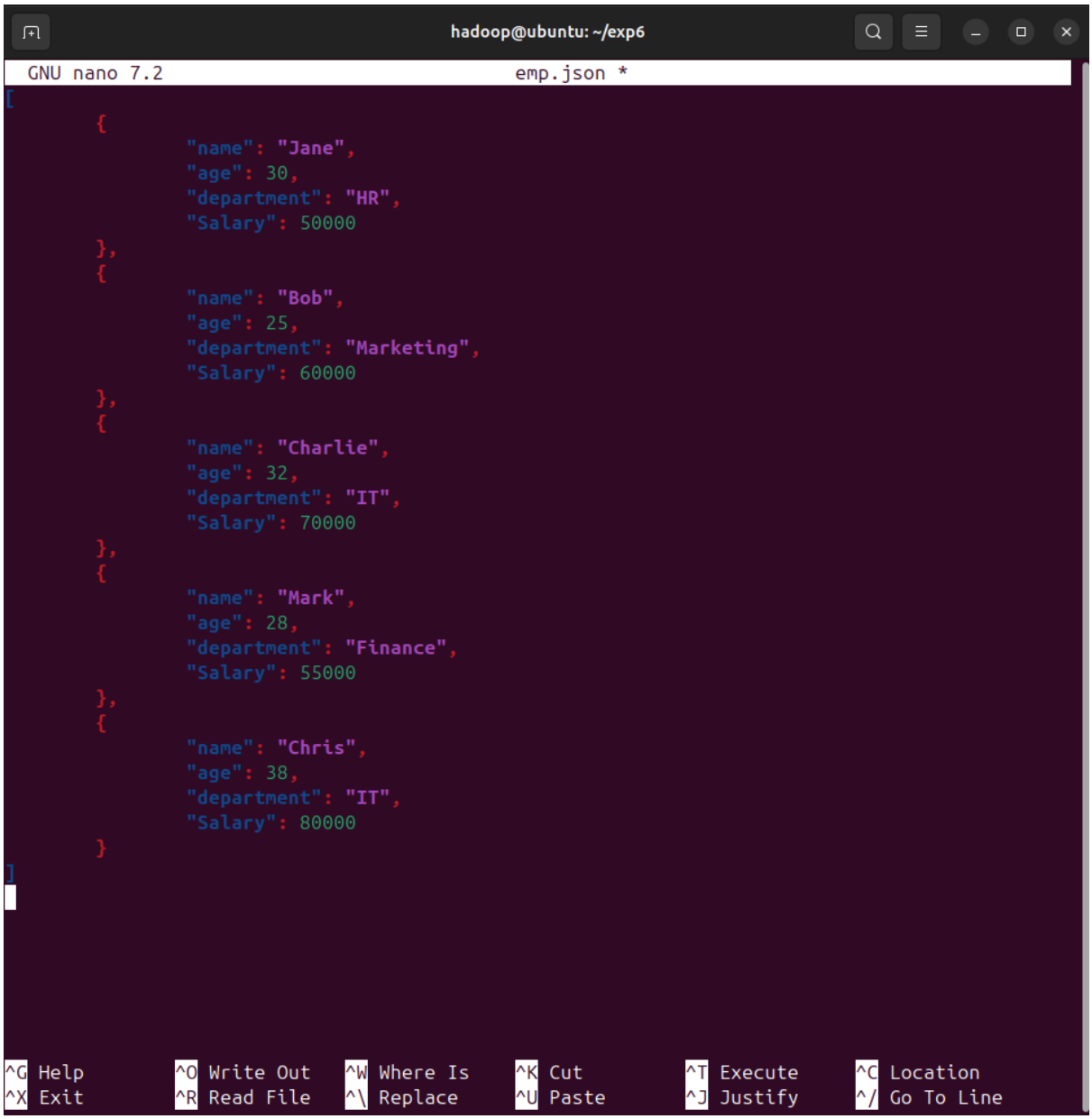
Exp. No : 6

Handling JSON data using HDFS and Python

1. Create emp.json file



2. Install jq package

```
hadoop@vikram: ~
hadoop@vikram:~$ sudo apt install jp
[sudo] password for hadoop:
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following NEW packages will be installed:
  jp
0 upgraded, 1 newly installed, 0 to remove and 7 not upgraded.
Need to get 1,292 kB of archives.
After this operation, 3,608 kB of additional disk space will be used.
Get:1 http://in.archive.ubuntu.com/ubuntu noble-updates/universe amd64 jp amd64 0.2.1+ds1-1ubu
ntu0.24.04.1 [1,292 kB]
Fetched 1,292 kB in 3s (422 kB/s)
Selecting previously unselected package jp.
(Reading database ... 204936 files and directories currently installed.)
Preparing to unpack .../jp_0.2.1+ds1-1ubuntu0.24.04.1_amd64.deb ...
Unpacking jp (0.2.1+ds1-1ubuntu0.24.04.1) ...
Setting up jp (0.2.1+ds1-1ubuntu0.24.04.1) ...
hadoop@vikram:~$
```

3. Execute jq . emp.json command

```
hadoop@ubuntu:~/exp6$ jq . 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
  }
]
hadoop@ubuntu:~/exp6$
```

4. pip install pandas

```
(exp6) hadoop@ubuntu:~/exp6$ pip install pandas
Collecting pandas
  Downloading pandas-2.2.3-cp312-cp312-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata
 (89 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 89.9/89.9 kB 4.4 MB/s eta 0:00:00
Collecting numpy>=1.26.0 (from pandas)
  Downloading numpy-2.1.2-cp312-cp312-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata
(60 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 60.9/60.9 kB 6.3 MB/s eta 0:00:00
Collecting python-dateutil>=2.8.2 (from pandas)
  Downloading python_dateutil-2.9.0.post0-py2.py3-none-any.whl.metadata (8.4 kB)
Collecting pytz>=2020.1 (from pandas)
  Downloading pytz-2024.2-py2.py3-none-any.whl.metadata (22 kB)
Collecting tzdata>=2022.7 (from pandas)
  Downloading tzdata-2024.2-py2.py3-none-any.whl.metadata (1.4 kB)
Requirement already satisfied: six>=1.5 in ./lib/python3.12/site-packages (from python-dateuti
l>=2.8.2->pandas) (1.16.0)
Downloading pandas-2.2.3-cp312-cp312-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (12.7 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 12.7/12.7 MB 1.5 MB/s eta 0:00:00
Downloading numpy-2.1.2-cp312-cp312-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (16.0 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 16.0/16.0 MB 1.2 MB/s eta 0:00:00
Downloading python_dateutil-2.9.0.post0-py2.py3-none-any.whl (229 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 229.9/229.9 kB 1.9 MB/s eta 0:00:00
Downloading pytz-2024.2-py2.py3-none-any.whl (508 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 508.0/508.0 kB 1.0 MB/s eta 0:00:00
Downloading tzdata-2024.2-py2.py3-none-any.whl (346 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 346.6/346.6 kB 2.2 MB/s eta 0:00:00
Installing collected packages: pytz, tzdata, python-dateutil, numpy, pandas
Successfully installed numpy-2.1.2 pandas-2.2.3 python-dateutil-2.9.0.post0 pytz-2024.2 tzdata
-2024.2
(exp6) hadoop@ubuntu:~/exp6$
```

5. pip install hdfs

```
(exp6) hadoop@ubuntu:~/exp6$ pip install hdfs
Collecting hdfs
  Downloading hdfs-2.7.3.tar.gz (43 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 43.5/43.5 kB 2.6 MB/s eta 0:00:00
  Installing build dependencies ... done
  Getting requirements to build wheel ... done
  Preparing metadata (pyproject.toml) ... done
Collecting docopt (from hdfs)
  Downloading docopt-0.6.2.tar.gz (25 kB)
  Installing build dependencies ... done
  Getting requirements to build wheel ... done
  Preparing metadata (pyproject.toml) ... done
Collecting requests>=2.7.0 (from hdfs)
  Downloading requests-2.32.3-py3-none-any.whl.metadata (4.6 kB)
Collecting six>=1.9.0 (from hdfs)
  Downloading six-1.16.0-py2.py3-none-any.whl.metadata (1.8 kB)
Collecting charset-normalizer<4,>=2 (from requests>=2.7.0->hdfs)
  Downloading charset_normalizer-3.4.0-cp312-cp312-manylinux_2_17_x86_64.manylinux2014_x86_64.
whl.metadata (34 kB)
Collecting idna<4,>=2.5 (from requests>=2.7.0->hdfs)
  Downloading idna-3.10-py3-none-any.whl.metadata (10 kB)
Collecting urllib3<3,>=1.21.1 (from requests>=2.7.0->hdfs)
  Downloading urllib3-2.2.3-py3-none-any.whl.metadata (6.5 kB)
Collecting certifi>=2017.4.17 (from requests>=2.7.0->hdfs)
  Downloading certifi-2024.8.30-py3-none-any.whl.metadata (2.2 kB)
Downloading requests-2.32.3-py3-none-any.whl (64 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 64.9/64.9 kB 15.6 MB/s eta 0:00:00
Downloading six-1.16.0-py2.py3-none-any.whl (11 kB)
Downloading certifi-2024.8.30-py3-none-any.whl (167 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 167.3/167.3 kB 12.2 MB/s eta 0:00:00
Downloading charset_normalizer-3.4.0-cp312-cp312-manylinux_2_17_x86_64.manylinux2014_x86_64.wh
l (143 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 143.8/143.8 kB 31.0 MB/s eta 0:00:00
Downloading idna-3.10-py3-none-any.whl (70 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 70.4/70.4 kB 22.2 MB/s eta 0:00:00
Downloading urllib3-2.2.3-py3-none-any.whl (126 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 126.3/126.3 kB 36.9 MB/s eta 0:00:00
Building wheels for collected packages: hdfs, docopt
  Building wheel for hdfs (pyproject.toml) ... done
  Created wheel for hdfs: filename=hdfs-2.7.3-py3-none-any.whl size=34323 sha256=f75adfa8348b1
7c6a762cf8cb20ca83640c67aa241affaa18c6beaa3828097f2
```

6. Create process_data.py

```
hadoop@ubuntu: ~/exp6
  GNU nano 7.2                    process_data.py
from hdfs import InsecureClient
import pandas as pd
import json

#Connect to HDFS
hdfs_client = InsecureClient("http://localhost:9870", user="hdfs")

#Read JSON data from HDFS
try:
        with hdfs_client.read("/home/hadoop/emp.json", encoding="utf-8") as reader:
                json_data = reader.read()
                if not json_data.strip()
                        raise ValueError("The JSON file is empty.")
                print(f"Raw JSON Data: {json_data[:1000]}")
                data = json.loads(json_data)
except json.JSONDecoderError as e:
        print(f"JSON Decode Error: {e}")
        exit(1)
except Exception as e:
        print(f"Error reading or parsing JSON data: {e}")
        exit(1)

#Convert JSON data to DataFrame
try:
        df = pd.DataFrame(data)
except ValueError as e:
        print(f"Error converting JSON data to DataFrame: {e}")
        exit(1)

#Projection: Select onlt 'name' and 'salary' columns
projected_df = df[['name','salary']]

#Aggregation: Calculation total salary
total_salary = df['salary'].sum()

                                [ Read 34 lines ]
^G Help      ^O Write Out   ^W Where Is   ^K Cut       ^T Execute   ^C Location
^X Exit      ^R Read File   ^\ Replace    ^U Paste     ^J Justify   ^/ Go To Line
```

Output:

```
hadoop@vikram: ~/exp6
hadoop@vikram:~/exp6$ python3 process_data.py
Raw JSON Data: [
        {
                "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
        }
]

hadoop@vikram:~/exp6$
```