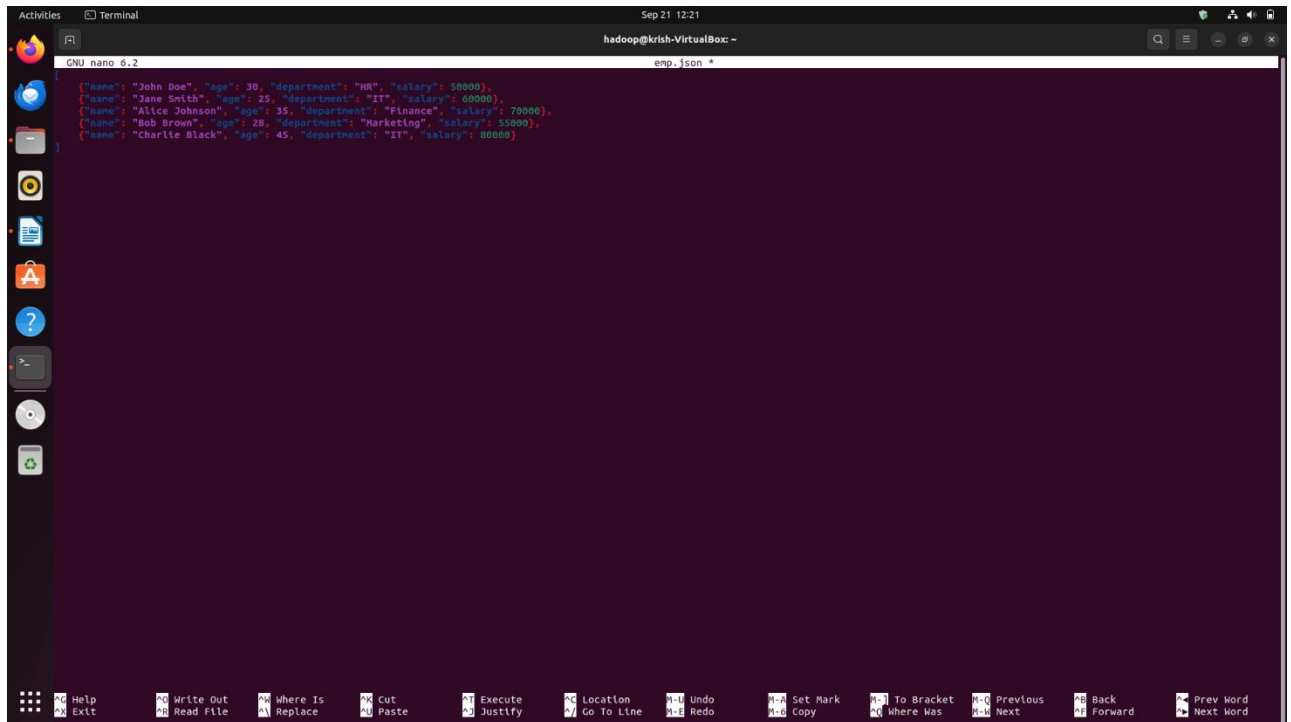


EXP NO: 6

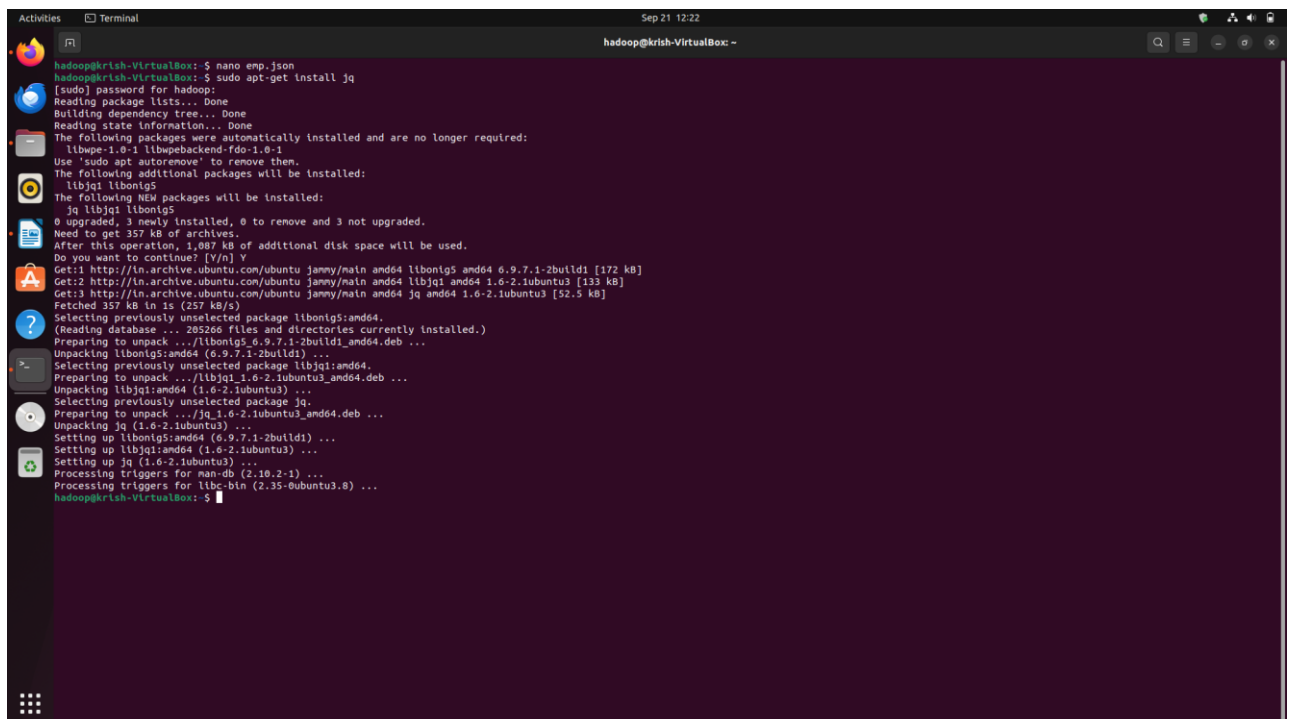
DATA MANAGEMENT WITH HDFS AND PANDAS

\$nano emp.json



```
hadoop@krish-VirtualBox: ~  
$ nano emp.json  
[{"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}]
```

sudo apt-get install jq



```
hadoop@krish-VirtualBox: ~  
$ sudo apt-get install jq  
[sudo] password for hadoop:  
Reading package lists... Done  
Building dependency tree... Done  
Reading state information... Done  
The following packages were automatically installed and are no longer required:  
  libwp6-1.0-1 libwp6backend-fdo-1.0-1  
Use 'sudo apt autoremove' to remove them.  
The following additional packages will be installed:  
  libjq1 libonig5  
The following NEW packages will be installed:  
  jq libjq1 libonig5  
0 upgraded, 3 newly installed, 0 to remove and 3 not upgraded.  
Need to get 357 kB of archives.  
After this operation, 1,087 kB of additional disk space will be used.  
Do you want to continue? [Y/n] Y  
Get:1 http://in.archive.ubuntu.com/ubuntu jammy/main amd64 libonig5 amd64 6.9.7.1-2build1 [172 kB]  
Get:2 http://in.archive.ubuntu.com/ubuntu jammy/main amd64 libjq1 amd64 1.6-2.1ubuntu3 [133 kB]  
Get:3 http://in.archive.ubuntu.com/ubuntu jammy/main amd64 jq amd64 1.6-2.1ubuntu3 [52.5 kB]  
Fetched 357 kB in 1s (257 kB/s)  
Selecting previously unselected package libonig5:amd64.  
(Reading database ... 265266 files and directories currently installed.)  
Preparing to unpack .../libonig5_6.9.7.1-2build1_amd64.deb ...  
Unpacking libonig5:amd64 (6.9.7.1-2build1) ...  
Selecting previously unselected package libjq1:amd64.  
Preparing to unpack .../libjq1_1.6-2.1ubuntu3_amd64.deb ...  
Unpacking libjq1:amd64 (1.6-2.1ubuntu3) ...  
Selecting previously unselected package jq.  
Preparing to unpack .../jq_1.6-2.1ubuntu3_amd64.deb ...  
Unpacking jq (1.6-2.1ubuntu3) ...  
Setting up libonig5:amd64 (6.9.7.1-2build1) ...  
Setting up libjq1:amd64 (1.6-2.1ubuntu3) ...  
Setting up jq (1.6-2.1ubuntu3) ...  
Processing triggers for man-db (2.10.2-1) ...  
Processing triggers for libc-bin (2.35-0ubuntu3.8) ...  
hadoop@krish-VirtualBox: ~  
$
```

jq . emp.json

```

After this operation, 1,087 kB of additional disk space will be used.
Do you want to continue? [Y/n] Y
Get:1 http://in.archive.ubuntu.com/ubuntu jammy/main amd64 libbonig5 amd64 6.9.7-1.2build1 [172 kB]
Get:2 http://in.archive.ubuntu.com/ubuntu jammy/main amd64 libjq1 amd64 1.6-2.1ubuntu3 [133 kB]
Get:3 http://in.archive.ubuntu.com/ubuntu jammy/main amd64 jq amd64 1.6-2.1ubuntu3 [52.5 kB]
Fetched 357 kB in 1s (257 kB/s)
Selecting previously unselected package libbonig5:amd64.
(Reading database ... 205266 files and directories currently installed.)
Preparing to unpack .../libbonig5_6.9.7-1.2build1_amd64.deb ...
Unpacking libbonig5:amd64 (6.9.7-1.2build1) ...
Selecting previously unselected package libjq1:amd64.
Preparing to unpack .../libjq1_1.6-2.1ubuntu3_amd64.deb ...
Unpacking libjq1:amd64 (1.6-2.1ubuntu3) ...
Preparing to unpack .../jq_1.6-2.1ubuntu3_amd64.deb ...
Unpacking jq (1.6-2.1ubuntu3) ...
Setting up libbonig5:amd64 (6.9.7-1.2build1) ...
Setting up libjq1:amd64 (1.6-2.1ubuntu3) ...
Setting up jq (1.6-2.1ubuntu3) ...
Processing triggers for man-db (2.10.2-1) ...
Processing triggers for libc-bin (2.35-0ubuntu3.8) ...
hadoop@krish-VirtualBox:~$ jq . emp.json
{
  {
    "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
  }
]
hadoop@krish-VirtualBox:~$

```

pip install pandas

```

Setting up libcrypt-dev:amd64 (1:4.4.27-1) ...
Setting up libjs-jquery (3.6.0+dfsg-3.5.13-1) ...
Setting up libbinutils:amd64 (2.38-4ubuntu2.6) ...
Setting up libc-dev-bin (2.35-0ubuntu3.8) ...
Setting up libalgorithm-diff-xs-perl (0.04-6build3) ...
Setting up libcc1-0:amd64 (12.3.0-1ubuntu1-22.04) ...
Setting up liblsan0:amd64 (12.3.0-1ubuntu1-22.04) ...
Setting up libitm1:amd64 (12.3.0-1ubuntu1-22.04) ...
Setting up libc-devtools (2.35-0ubuntu3.8) ...
Setting up libjs-underscore (1.13.2+dfsg-2) ...
Setting up libalgorithm-merge-perl (0.08-3) ...
Setting up libtsan0:amd64 (11.4.0-1ubuntu1-22.04) ...
Setting up libctf0:amd64 (2.38-4ubuntu2.6) ...
Setting up libjs-sphinxdoc (4.3.2-1) ...
Setting up libgcc-11-dev:amd64 (11.4.0-1ubuntu1-22.04) ...
Setting up libc6-dev:amd64 (2.35-0ubuntu3.8) ...
Setting up binutils-x86_64-linux-gnu (2.38-4ubuntu2.6) ...
Setting up binutils (2.38-4ubuntu2.6) ...
Setting up dpkg-dev (1.21.1ubuntu2.3) ...
Setting up libxpat1-dev:amd64 (2.4.7-1ubuntu0.4) ...
Setting up libstdc++-11-dev:amd64 (11.4.0-1ubuntu1-22.04) ...
Setting up zlib1g-dev:amd64 (1:1.2.11.dfsg-2ubuntu9.2) ...
Setting up gcc-11 (11.4.0-1ubuntu1-22.04) ...
Setting up g++-11 (11.4.0-1ubuntu1-22.04) ...
Setting up gcc (4:11.2.0-1ubuntu1) ...
Setting up libpython3.10-dev:amd64 (3.10.12-1-22.04.6) ...
Setting up python3.10-dev (3.10.12-1-22.04.6) ...
Setting up g++ (4:11.2.0-1ubuntu1) ...
update-alternatives: using /usr/bin/g++ to provide /usr/bin/c++ (c++) in auto mode
Setting up build-essential (12.9ubuntu3) ...
Setting up libpython3-dev:amd64 (3.10.6-1-22.04.1) ...
Setting up python3-dev (3.10.6-1-22.04.1) ...
Processing triggers for man-db (2.10.2-1) ...
Processing triggers for libc-bin (2.35-0ubuntu3.8) ...
hadoop@krish-VirtualBox:~$ pip3 install pandas
Defaulting to user installation because normal site-packages is not writeable
Collecting pandas
  Downloading pandas-2.2.3-cp310-cp310-manylinux_2_17_x86_64_manylinux2014_x86_64.whl (13.1 MB)
    13.1/13.1 MB 3.2 MB/s eta 0:00:00
Collecting tzdata>=2022.7
  Downloading tzdata-2024.1-py2.py3-none-any.whl (345 kB)
    345.4/345.4 KB 2.7 MB/s eta 0:00:00
Collecting python-dateutil>=2.8.2
  Downloading python_dateutil-2.9.0.post0-py2.py3-none-any.whl (229 kB)
    229.9/229.9 KB 3.1 MB/s eta 0:00:00
Requirement already satisfied: pytz>=2020.1 in /usr/lib/python3/dist-packages (from pandas) (2022.1)
Collecting numpy>=1.22.4
  Downloading numpy-2.1.1-cp310-cp310-manylinux_2_17_x86_64_manylinux2014_x86_64.whl (16.3 MB)
    16.3/16.3 MB 4.7 MB/s eta 0:00:00
Requirement already satisfied: six>=1.5 in /usr/lib/python3/dist-packages (from python-dateutil>=2.8.2->pandas) (1.16.0)
Installing collected packages: tzdata, python-dateutil, numpy, pandas
WARNING: The scripts f2py and numpy-config are installed in /home/hadoop/.local/bin which is not on PATH.
Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
Successfully installed numpy-2.1.1 pandas-2.2.3 python-dateutil-2.9.0.post0 tzdata-2024.1
hadoop@krish-VirtualBox:~$

```

pip install hdf5

```

Activities Terminal Sep 21 12:30
hadoop@krish-VirtualBox: ~
Setting up g++-11 (11.4.0-1ubuntu1-22.04) ...
Setting up gcc (4:11.2.0-1ubuntu1) ...
Setting up libpython3.10-dev:amd64 (3.10.12-1-22.04.6) ...
Setting up python3.10-dev (3.10.12-1-22.04.6) ...
Setting up g++ (4:11.2.0-1ubuntu1) ...
update-alternatives: using /usr/bin/g++ to provide /usr/bin/c++ (c++) in auto mode
Setting up build-essential (12.9ubuntu3) ...
Setting up libpython3-dev:amd64 (3.10.6-1-22.04.1) ...
Setting up python3-dev (3.10.6-1-22.04.1) ...
Processing triggers for man-db (2.10.2-1) ...
Processing triggers for libc-bin (2.35-0ubuntu3.8) ...
hadoop@krish-VirtualBox: ~$ pip3 install pandas
Defaulting to user installation because normal site-packages is not writeable
Collecting pandas
  Downloading pandas-2.2.3-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (13.1 MB)
    11.1/13.1 MB 3.2 MB/s eta 0:00:00
Collecting tzdata>=2022.7
  Downloading tzdata-2024.1-py2.py3-none-any.whl (345 kB)
    345.4/345.4 KB 2.7 MB/s eta 0:00:00
Collecting python-dateutil>=2.8.2
  Downloading python_dateutil-2.9.0.post0-py2.py3-none-any.whl (229 kB)
    229.0/229.0 KB 1.7 MB/s eta 0:00:00
Requirement already satisfied: pytz>=2020.1 in /usr/lib/python3/dist-packages (from pandas) (2022.1)
Collecting numpy>=1.22.4
  Downloading numpy-2.1.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (16.3 MB)
    16.3/16.3 MB 3.2 MB/s eta 0:00:00
Requirement already satisfied: six>=1.5 in /usr/lib/python3/dist-packages (from python-dateutil>=2.8.2->pandas) (1.16.0)
Installing collected packages: tzdata, python-dateutil, numpy, pandas
WARNING: The scripts f2py and numpy-config are installed in '/home/hadoop/.local/bin' which is not on PATH.
Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
Successfully installed numpy-2.1.1 pandas-2.2.3 python-dateutil-2.9.0.post0 tzdata-2024.1
hadoop@krish-VirtualBox: ~$ pip install hdf5
Defaulting to user installation because normal site-packages is not writeable
Collecting hdf5
  Downloading hdf5-2.7.3.tar.gz (43 kB)
    43.5/43.5 KB 1.4 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Collecting docopt
  Downloading docopt-0.6.2.tar.gz (25 kB)
    25.0/25.0 KB 1.7 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: requests>=2.7.0 in /usr/lib/python3/dist-packages (from hdf5) (2.25.1)
Requirement already satisfied: six>=1.9.0 in /usr/lib/python3/dist-packages (from hdf5) (1.16.0)
Building wheels for collected packages: hdf5, docopt
  Building wheel for hdf5 (setup.py) ... done
  Created wheel for hdf5: filename=hdf5-2.7.3-py3-none-any.whl size=34347 sha256=70f6cd8058677699356d021f970f6214620418727d1511ca6d35cb0a85b13a40
  Stored in directory: /home/hadoop/.cache/pip/wheels/e5/8d/b0/99c1c8a3ac5788c86b08cd3f48b0134a5910e6ed26011808b
  Building wheel for docopt (setup.py) ... done
  Created wheel for docopt: filename=docopt-0.6.2-py2.py3-none-any.whl size=13723 sha256=89b1912e44563a48fa80b012f46646899f8571e9d56b96bd55074e043a76080b
  Stored in directory: /home/hadoop/.cache/pip/wheels/fc/ab/d4/5da2067ac95b36618c629a5f93f809425708506f72c9732fac
Successfully built hdf5 docopt
Installing collected packages: docopt, hdf5
WARNING: The scripts hdf5cli and hdf5cli-avro are installed in '/home/hadoop/.local/bin' which is not on PATH.
Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
Successfully installed docopt-0.6.2 hdf5-2.7.3
hadoop@krish-VirtualBox: ~$

```

nano process_data.py

```

Activities Terminal Sep 21 12:31
hadoop@krish-VirtualBox: ~
GNU nano 6.2 process_data.py
# Save the filtered result back to HDFS
filtered_json = filtered_df.to_json(orient='records')
try:
    with hdf5_client.write('/home/hadoop/filtered_employees.json', encoding='utf-8', overwrite=True) as writer:
        writer.write(filtered_json)
    print(f"Filtered JSON file saved successfully.")
except Exception as e:
    print(f"Error saving filtered JSON data: {e}")
    exit(1)

# Print results
print(f"Projection: Select only name and salary columns")
print(f"Projected DF:")

print(f"Aggregation: Calculate total salary")
print(f"Total Salary: {total_salary}\n")

print(f"# Count: Number of employees earning more than 50000")
print(f"Number of High Earners (>50000): {high_earners_count}\n")

print(f"Top 5 Earners:\n{top_5_earners}\n")

print(f"Skipped DataFrame (First 2 rows skipped): \n{skipped_df}\n")

print(f"Filtered DataFrame (IT department removed): \n{filtered_df}\n")

```

python3 process_data.py

```
Activities Terminal Sep 21 12:38
hadoop@krish-VirtualBox: ~
at java.security.AccessController.doPrivileged(Native Method)
at javax.security.auth.Subject.doAs(Subject.java:422)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1899)
at org.apache.hadoop ipc.ServerHandler.run(Server.java:3048)

hadoop@krish-VirtualBox: ~$ hdfs dfs -ls /home/
Found 1 items
drwxr-xr-x - hadoop supergroup 0 2024-09-21 12:35 /home/hadoop
hadoop@krish-VirtualBox: ~$ hdfs dfs -chown hdfs:hdfs /home/hadoop
hadoop@krish-VirtualBox: ~$ python3 process_data.py
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

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 (IT department removed):
  name age department salary
0 John Doe 30 HR 50000
2 Alice Johnson 35 Finance 70000
3 Bob Brown 28 Marketing 55000
hadoop@krish-VirtualBox: ~$
```