**Project Report**

**Course Name: Software and Data Engineering (SDE)**

**Course Code: CSL7090**

# Exploring BigData: CDH Platform Architecture, Local Machine EDA

**Submitted To: -**

Dr. Sumit Kalra,

Assistant Professor,

Dept. of CSE,

IIT Jodhpur.

**Submitted By: -**

Bhawna Bhoria (M22MA003)

Jash Patel (M22CS061)

# Introduction

As part of this project , we will be observing and analyzing processing techniques across CDH, local machine platform to comprehend intricate systems and address data processing challenges in the context of expanding data volumes.

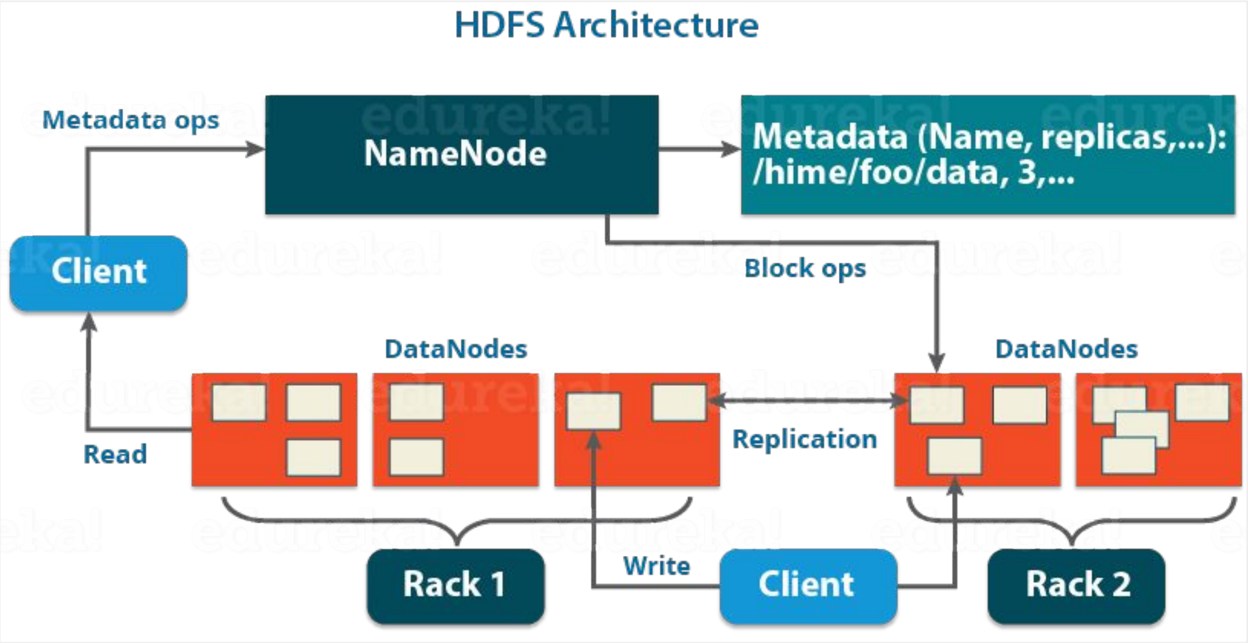
Platform 1:-

CDH - Architecture, Storage System and Job Scheduling.

Platform 2:-

EDA - Exploratory Data Analysis on Local Machine.

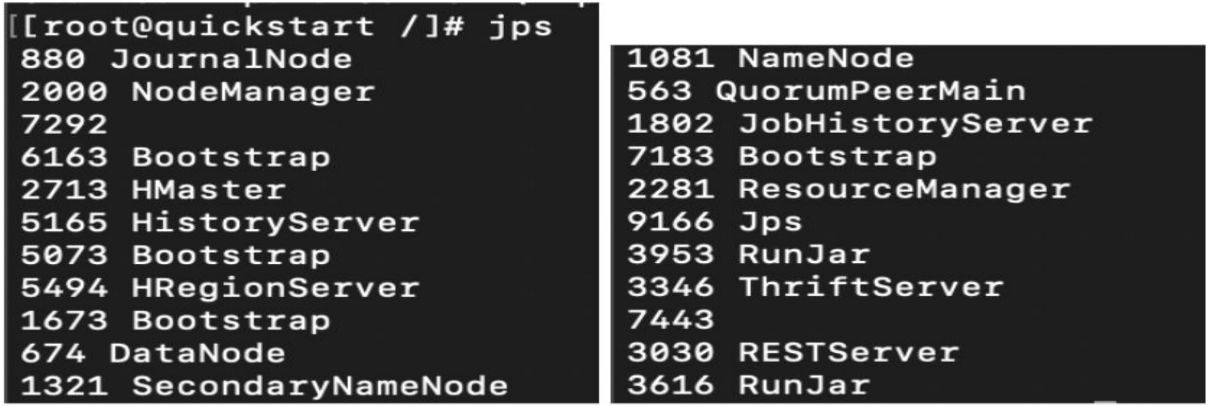
# Platform 1 : Cloudera Distribution of Hadoop



The customised distribution employs the Namenode and YARN Resource Manager to oversee and allocate resources in a Hadoop cluster. The functionalities of the Job Tracker and Task Tracker have been substituted by the ResourceManager and NodeManagers, providing a resource management architecture that is more adaptable and capable of handling larger workloads.

# Processes Running on CDH Architecture:-

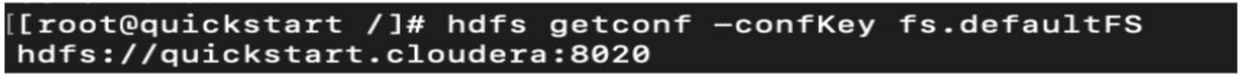
jps output :-



```
[[root@quickstart /]# jps
880 JournalNode
2000 NodeManager
7292
6163 Bootstrap
2713 HMaster
5165 HistoryServer
5073 Bootstrap
5494 HRegionServer
1673 Bootstrap
674 DataNode
1321 SecondaryNameNode
```

```
1081 NameNode
563 QuorumPeerMain
1802 JobHistoryServer
7183 Bootstrap
2281 ResourceManager
9166 Jps
3953 RunJar
3346 ThriftServer
7443
3030 RESTServer
3616 RunJar
```

Checking Namenode :-

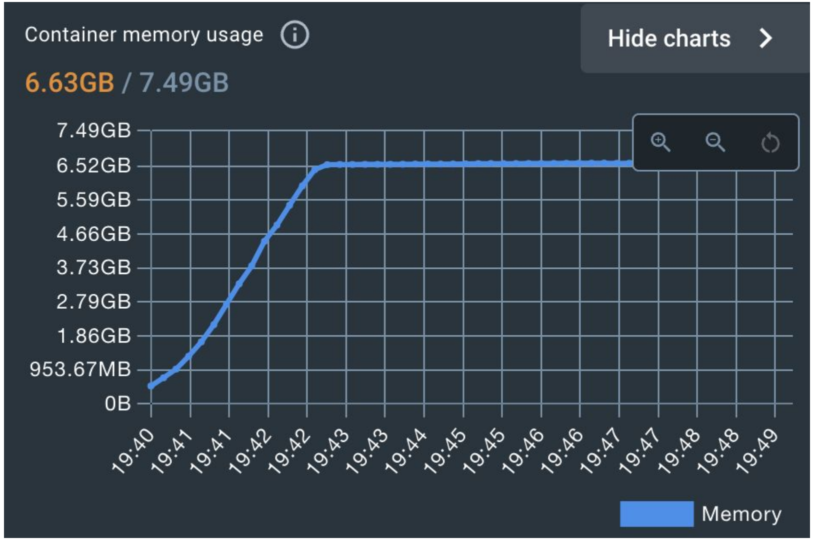


```
[[root@quickstart /]# hdfs getconf -confKey fs.defaultFS
hdfs://quickstart.cloudera:8020
```

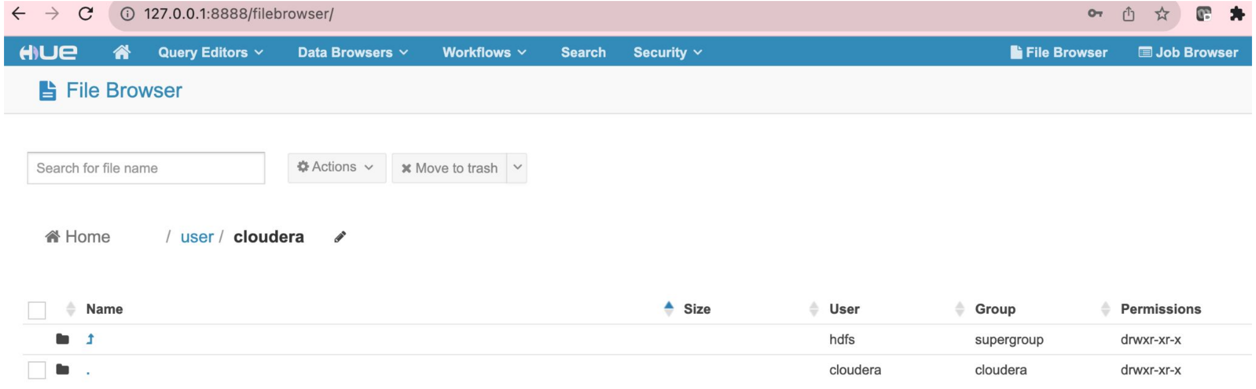
# Laptop configurations and CDH Image Size

Physical Memory on device: 16GB

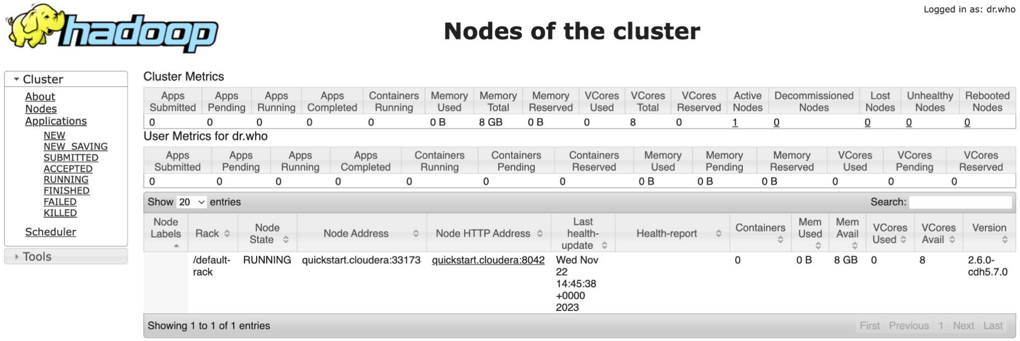
CDH Distribution Memory Occupance : 8 GB



# CDH Ecosystem : Cluster and HDFS Storage



| | Name | Size | User | Group | Permissions |
|---|---|---|---|---|---|
| | ↥ | | hdfs | supergroup | drwxr-xr-x |
| □ | . | | cloudera | cloudera | drwxr-xr-x |



Logged in as: dr.who

Nodes of the cluster

Cluster Metrics

| Apps Submitted | Apps Pending | Apps Running | Apps Completed | Containers Running | Memory Used | Memory Total | Memory Reserved | VCores Used | VCores Total | VCores Reserved | Active Nodes | Decommissioned Nodes | Lost Nodes | Unhealthy Nodes | Rebooted Nodes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 B | 8 GB | 0 B | 0 | 8 | 0 | 1 | 0 | 0 | 0 | 0 |

User Metrics for dr.who

| Apps Submitted | Apps Pending | Apps Running | Apps Completed | Containers Running | Containers Pending | Containers Reserved | Memory Used | Memory Pending | Memory Reserved | VCores Used | VCores Pending | VCores Reserved |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 B | 0 B | 0 B | 0 | 0 | 0 |

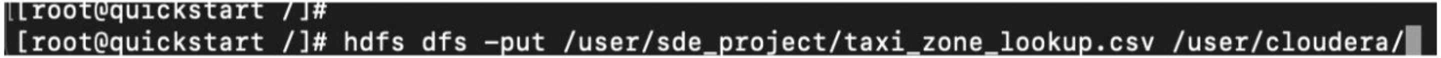
Show 20 entries Search:

| Node Labels | Rack | Node State | Node Address | Node HTTP Address | Last health-update | Health-report | Containers | Mem Used | Mem Avail | VCores Used | VCores Avail | Version |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | /default-rack | RUNNING | quickstart.cloudera:33173 | quickstart.cloudera:8042 | Wed Nov 22 14:45:38 +0000 2023 | | 0 | 0 B | 8 GB | 0 | 8 | 2.6.0-cdh5.7.0 |

Showing 1 to 1 of 1 entries

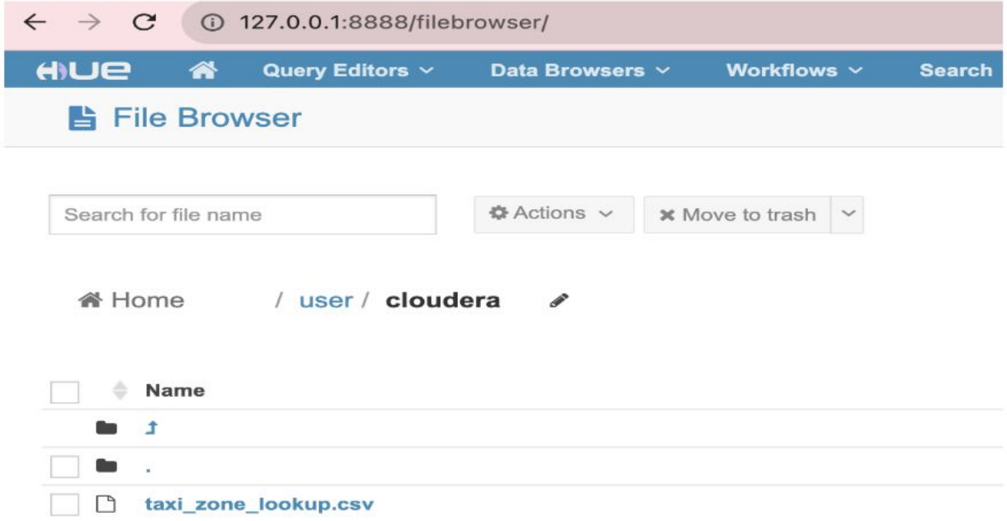
First Previous 1 Next Last

# Taking File to hdfs



```
[[root@quickstart /]#
 [root@quickstart /]# hdfs dfs -put /user/sde_project/taxi_zone_lookup.csv /user/cloudera/
```

The Hadoop Distributed File System (HDFS) is renowned for its fault-tolerant nature. It allows clients to upload files from NameNodes, which then direct them to DataNodes for storage information. The client then uploads the file to the specified DataNode, completing the write request. In a read operation, the client seeks metadata from NameNodes, which then identifies the DataNode server where the file resides in the cluster and randomly selects one.

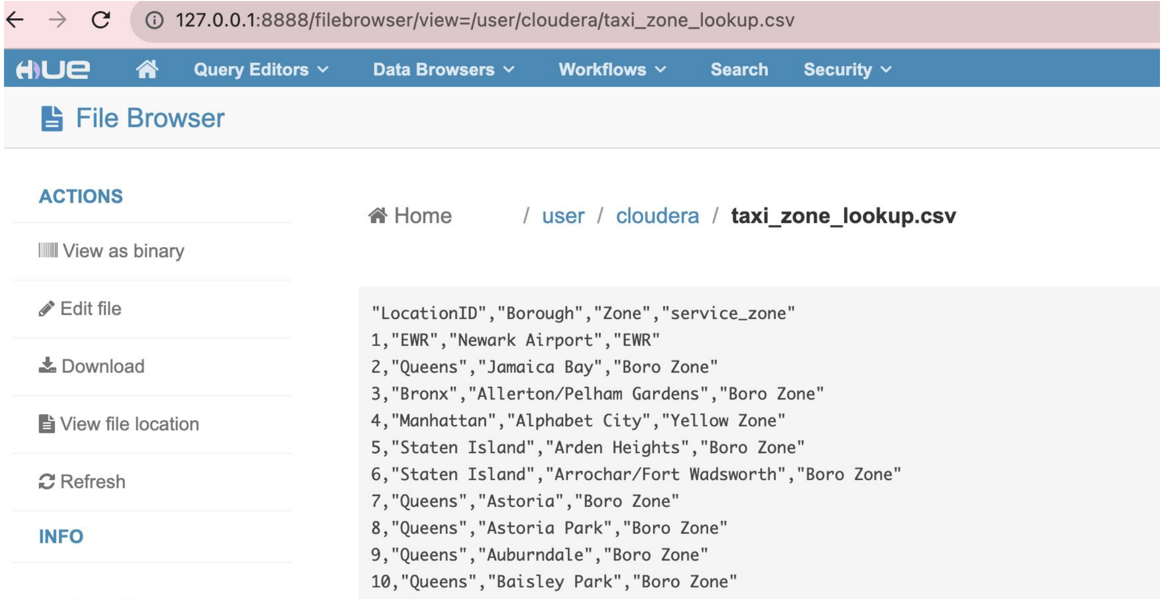


127.0.0.1:8888/filebrowser/

Home / user / cloudera

| Name |
|---|
| . |
| taxi_zone_lookup.csv |

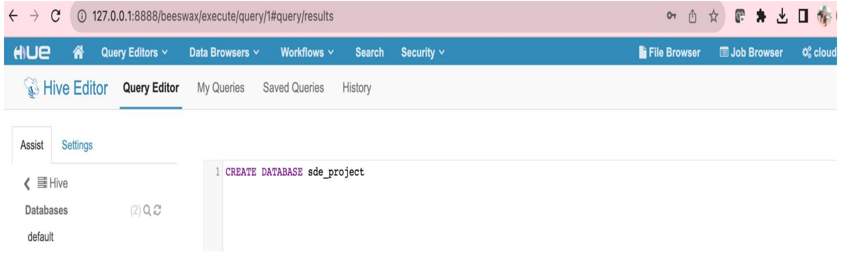
# File placed at HDFS storage system into distributed blocks of 64 bytes.



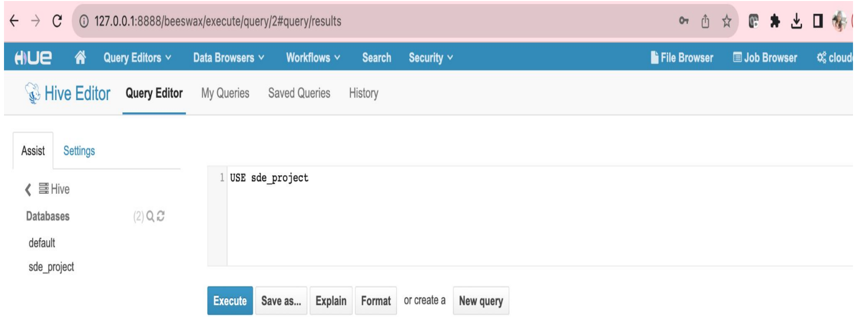
```
"LocationID","Borough","Zone","service_zone"
1,"EWR","Newark Airport","EWR"
2,"Queens","Jamaica Bay","Boro Zone"
3,"Bronx","Allerton/Pelham Gardens","Boro Zone"
4,"Manhattan","Alphabet City","Yellow Zone"
5,"Staten Island","Arden Heights","Boro Zone"
6,"Staten Island","Arrochar/Fort Wadsworth","Boro Zone"
7,"Queens","Astoria","Boro Zone"
8,"Queens","Astoria Park","Boro Zone"
9,"Queens","Auburndale","Boro Zone"
10,"Queens","Baisley Park","Boro Zone"
```

# HIVE

Hive, a Hadoop-based data warehouse tool, efficiently processes queries and generates results for extensive datasets. It executes statements, similar to MySQL, and takes charge of user tasks. The compiler retrieves metadata, compiles the task, selects the optimal strategy, and delivers results. The pseudocode outlines Hive's data transformation, loading, and extraction process.

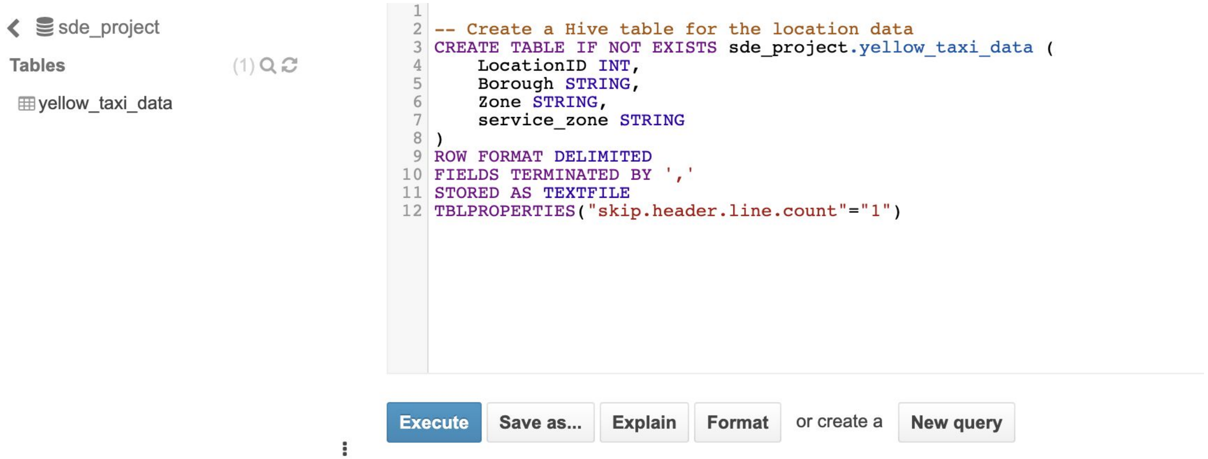


```
CREATE DATABASE sde_project
```

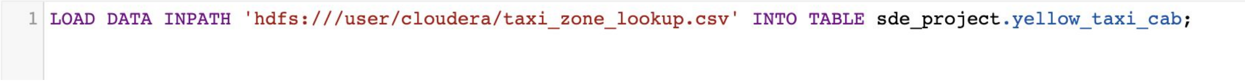


```
USE sde_project
```

# Querying on Hive:-

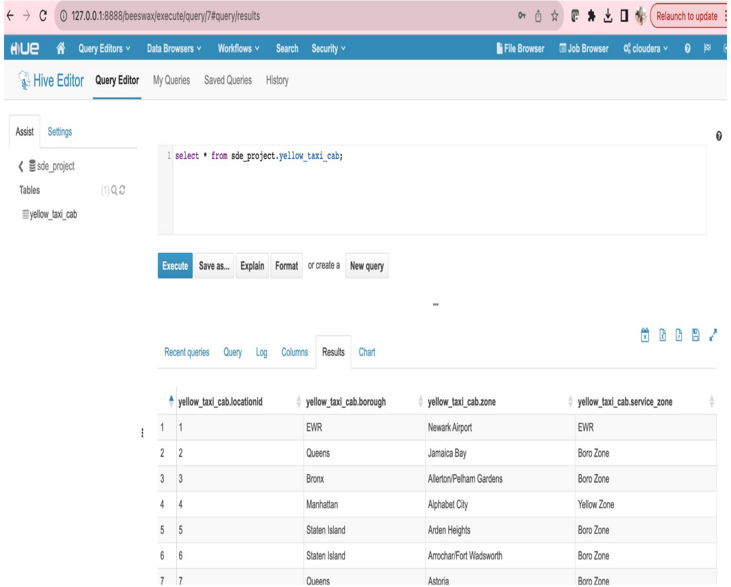


```

-- Create a Hive table for the location data
CREATE TABLE IF NOT EXISTS sde_project.yellow_taxi_data (
    LocationID INT,
    Borough STRING,
    Zone STRING,
    service_zone STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
TBLPROPERTIES("skip.header.line.count"="1")
```



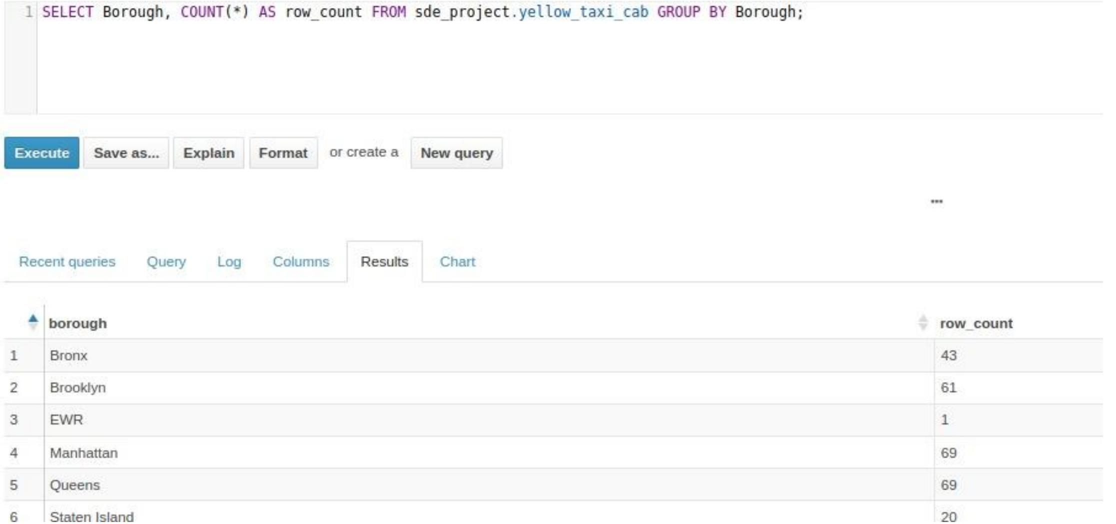
```
LOAD DATA INPATH 'hdfs:///user/cloudera/taxi_zone_lookup.csv' INTO TABLE sde_project.yellow_taxi_cab;
```

# Querying on Hive:-



```
select * from sde_project.yellow_taxi_cab;
```

| | yellow_taxi_cab.locationid | yellow_taxi_cab.borough | yellow_taxi_cab.zone | yellow_taxi_cab.service_zone |
|---|---|---|---|---|
| 1 | 1 | EWR | Newark Airport | EWR |
| 2 | 2 | Queens | Jamaica Bay | Boro Zone |
| 3 | 3 | Bronx | Allerton/Pelham Gardens | Boro Zone |
| 4 | 4 | Manhattan | Alphabet City | Yellow Zone |
| 5 | 5 | Staten Island | Arden Heights | Boro Zone |
| 6 | 6 | Staten Island | Arrochar/Fort Wadsworth | Boro Zone |
| 7 | 7 | Queens | Astoria | Boro Zone |



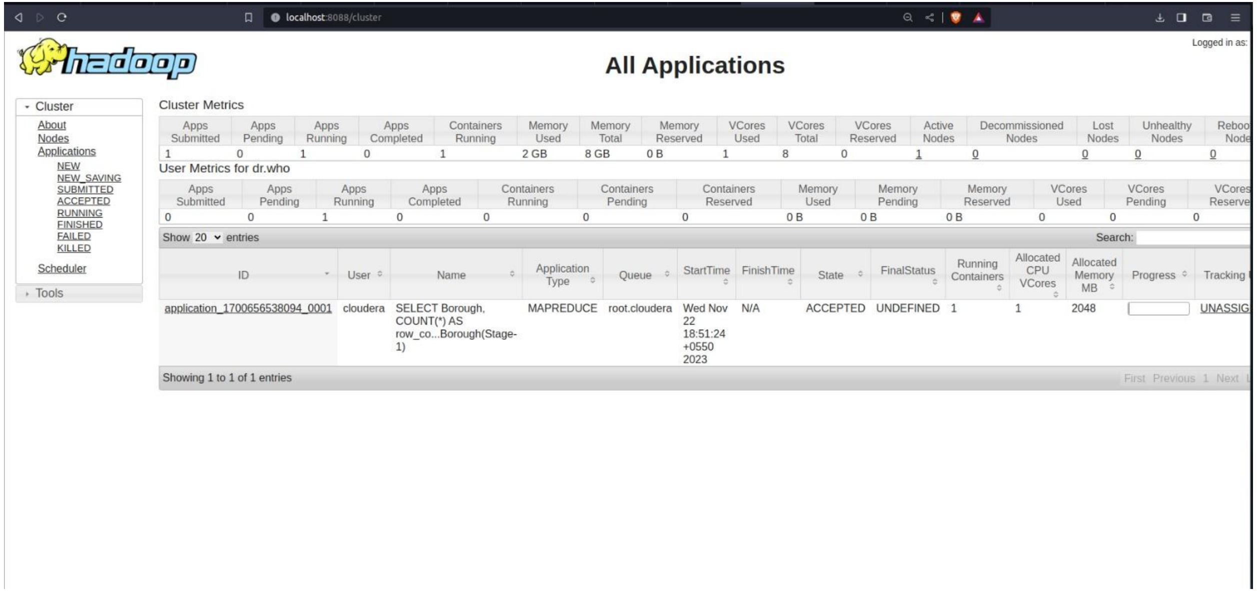
```
SELECT Borough, COUNT(*) AS row_count FROM sde_project.yellow_taxi_cab GROUP BY Borough;
```

| | borough | row_count |
|---|---|---|
| 1 | Bronx | 43 |
| 2 | Brooklyn | 61 |
| 3 | EWR | 1 |
| 4 | Manhattan | 69 |
| 5 | Queens | 69 |
| 6 | Staten Island | 20 |

Map Reduce job Runs in background on YARN.

# What happens behind the curtains?- YARN

Application Master Instance gets created and State is ACCEPTED.



localhost:8088/cluster

Logged in as:

## All Applications

- Cluster
  - About
  - Nodes
  - Applications
    - NEW
    - NEW_SAVING
    - SUBMITTED
    - ACCEPTED
    - RUNNING
    - FINISHED
    - FAILED
    - KILLED
  - Scheduler
- Tools

Cluster Metrics

| Apps Submitted | Apps Pending | Apps Running | Apps Completed | Containers Running | Memory Used | Memory Total | Memory Reserved | VCores Used | VCores Total | VCores Reserved | Active Nodes | Decommissioned Nodes | Lost Nodes | Unhealthy Nodes | Reboo Node |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 1 | 2 GB | 8 GB | 0 B | 1 | 8 | 0 | 1 | 0 | 0 | 0 | 0 |

User Metrics for dr.who

| Apps Submitted | Apps Pending | Apps Running | Apps Completed | Containers Running | Containers Pending | Containers Reserved | Memory Used | Memory Pending | Memory Reserved | VCores Used | VCores Pending | VCores Reserve |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 B | 0 B | 0 B | 0 | 0 | 0 |

Show 20 entries

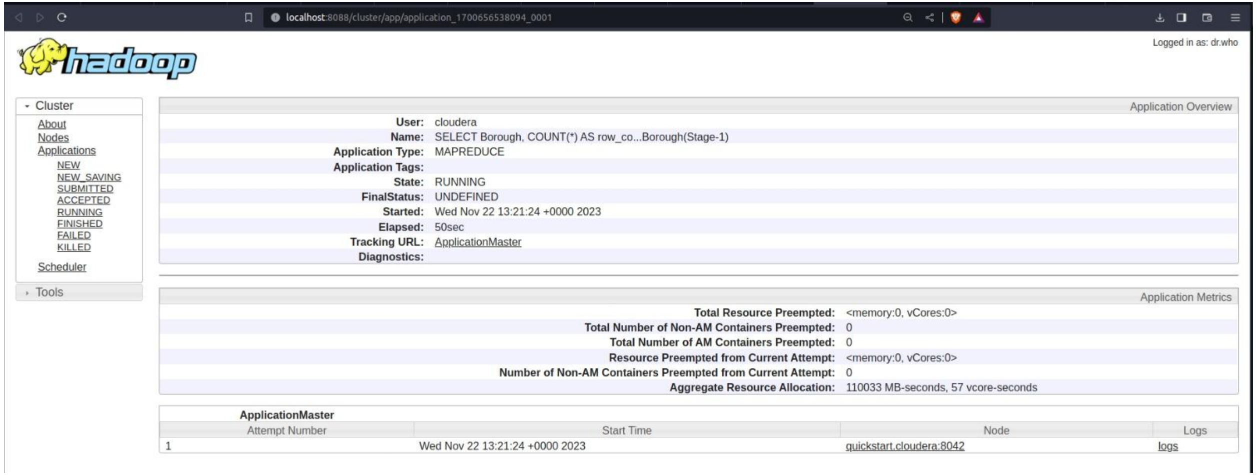
Search:

| ID | User | Name | Application Type | Queue | StartTime | FinishTime | State | FinalStatus | Running Containers | Allocated CPU VCores | Allocated Memory MB | Progress | Tracking |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| application_1700656538094_0001 | cloudera | SELECT Borough, COUNT(*) AS row_co...Borough(Stage-1) | MAPREDUCE | root.cloudera | Wed Nov 22 18:51:24 +0550 2023 | N/A | ACCEPTED | UNDEFINED | 1 | 1 | 2048 |  | UNASSIG |

Showing 1 to 1 of 1 entries

First Previous 1 Next

# Logs are stored on HistoryServer and Jobs is moved to FINISHED Applications.



Application Overview

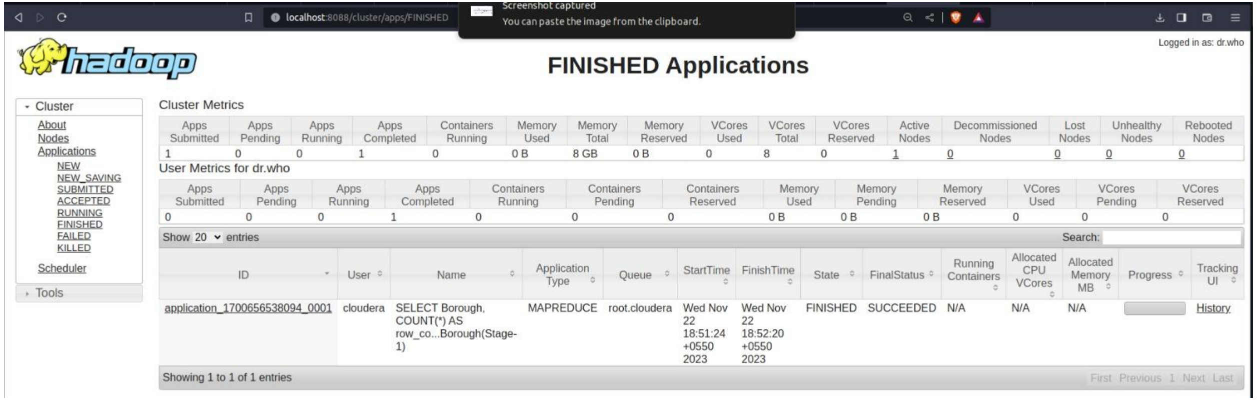
| | |
|---|---|
| User: | cloudera |
| Name: | SELECT Borough, COUNT(*) AS row_co...Borough(Stage-1) |
| Application Type: | MAPREDUCE |
| Application Tags: | |
| State: | RUNNING |
| FinalStatus: | UNDEFINED |
| Started: | Wed Nov 22 13:21:24 +0000 2023 |
| Elapsed: | 50sec |
| Tracking URL: | ApplicationMaster |
| Diagnostics: | |

Application Metrics

| | |
|---|---|
| Total Resource Preempted: | <memory:0, vCores:0> |
| Total Number of Non-AM Containers Preempted: | 0 |
| Total Number of AM Containers Preempted: | 0 |
| Resource Preempted from Current Attempt: | <memory:0, vCores:0> |
| Number of Non-AM Containers Preempted from Current Attempt: | 0 |
| Aggregate Resource Allocation: | 110033 MB-seconds, 57 vcore-seconds |

ApplicationMaster

| Attempt Number | Start Time | Node | Logs |
|---|---|---|---|
| 1 | Wed Nov 22 13:21:24 +0000 2023 | quickstart.cloudera:8042 | logs |



Screenshot captured
You can paste the image from the clipboard.

## FINISHED Applications

Cluster Metrics

| Apps Submitted | Apps Pending | Apps Running | Apps Completed | Containers Running | Memory Used | Memory Total | Memory Reserved | VCores Used | VCores Total | VCores Reserved | Active Nodes | Decommissioned Nodes | Lost Nodes | Unhealthy Nodes | Rebooted Nodes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 0 | 0 B | 8 GB | 0 B | 0 | 8 | 0 | 1 | 0 | 0 | 0 | 0 |

User Metrics for dr.who

| Apps Submitted | Apps Pending | Apps Running | Apps Completed | Containers Running | Containers Pending | Containers Reserved | Memory Used | Memory Pending | Memory Reserved | VCores Used | VCores Pending | VCores Reserved |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 B | 0 B | 0 B | 0 | 0 | 0 |

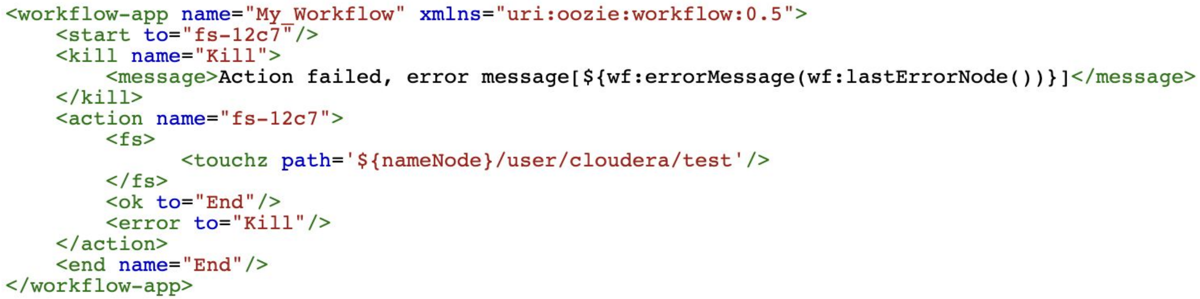
Show 20 entries Search:

| ID | User | Name | Application Type | Queue | StartTime | FinishTime | State | FinalStatus | Running Containers | Allocated CPU VCores | Allocated Memory MB | Progress | Tracking UI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| application_1700656538094_0001 | cloudera | SELECT Borough, COUNT(*) AS row_co...Borough(Stage-1) | MAPREDUCE | root.cloudera | Wed Nov 22 18:51:24 +0550 2023 | Wed Nov 22 18:52:20 +0550 2023 | FINISHED | SUCCEEDED | N/A | N/A | N/A | | History |

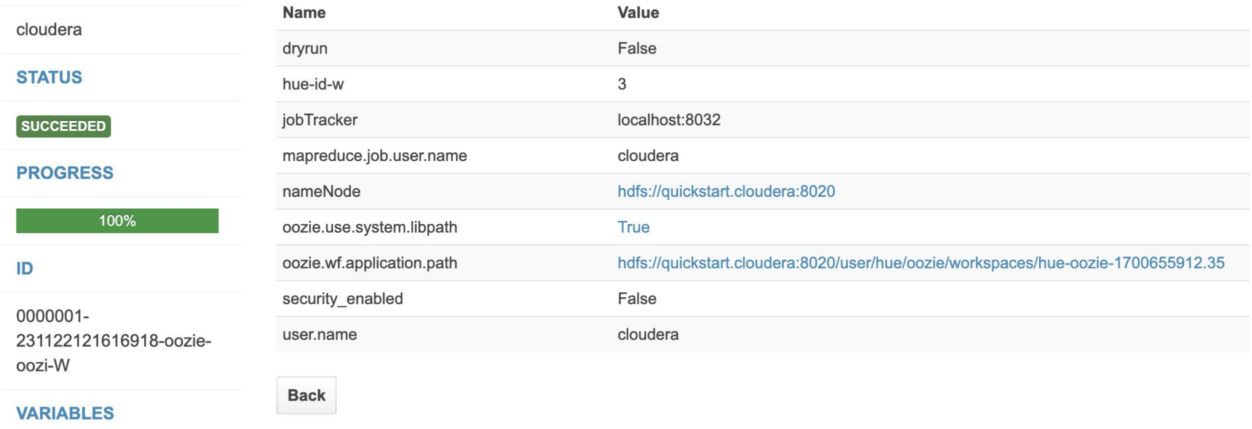
Showing 1 to 1 of 1 entries

# CDH Oozie Workflows

Workflow to create HDFS directories and files (action name : fs-12c7)



```
<workflow-app name="My_Workflow" xmlns="uri:oozie:workflow:0.5">
    <start to="fs-12c7"/>
    <kill name="Kill">
        <message>Action failed, error message[${wf:errorMessage(wf:lastErrorNode())}]</message>
    </kill>
    <action name="fs-12c7">
        <fs>
              <touchz path='${nameNode}/user/cloudera/test'/>
        </fs>
        <ok to="End"/>
        <error to="Kill"/>
    </action>
    <end name="End"/>
</workflow-app>
```



cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

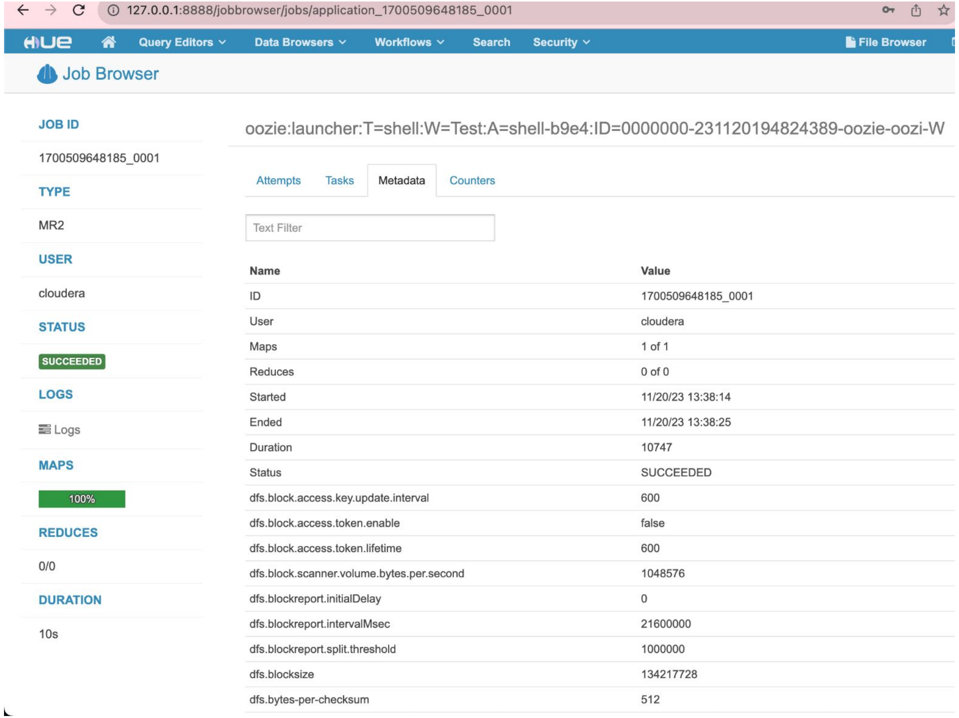
0000001-231122121616918-oozie-oozi-W

VARIABLES

| Name | Value |
|---|---|
| dryrun | False |
| hue-id-w | 3 |
| jobTracker | localhost:8032 |
| mapreduce.job.user.name | cloudera |
| nameNode | hdfs://quickstart.cloudera:8020 |
| oozie.use.system.libpath | True |
| oozie.wf.application.path | hdfs://quickstart.cloudera:8020/user/hue/oozie/workspaces/hue-oozie-1700655912.35 |
| security_enabled | False |
| user.name | cloudera |

# Workflow to run shell script checking availability of file on hdfs

```
<workflow-app name="Test" xmlns="uri:oozie:workflow:0.5">
    <start to="shell-869a"/>
    <kill name="Kill">
        <message>Action failed, error message[${wf:errorMessage(wf:lastErrorNode())}]</message>
    </kill>
    <action name="shell-869a">
        <shell xmlns="uri:oozie:shell-action:0.1">
            <job-tracker>${jobTracker}</job-tracker>
            <name-node>${nameNode}</name-node>
            <exec>/user/cloudera/file_avail_check.sh</exec>
              <capture-output/>
        </shell>
        <ok to="End"/>
        <error to="Kill"/>
    </action>
    <end name="End"/>
</workflow-app>
```



oozie:launcher:T=shell:W=Test:A=shell-b9e4:ID=0000000-231120194824389-oozie-oozi-W

JOB ID
1700509648185_0001

TYPE
MR2

USER
cloudera

STATUS
SUCCEEDED

LOGS
Logs

MAPS
100%

REDUCES
0/0

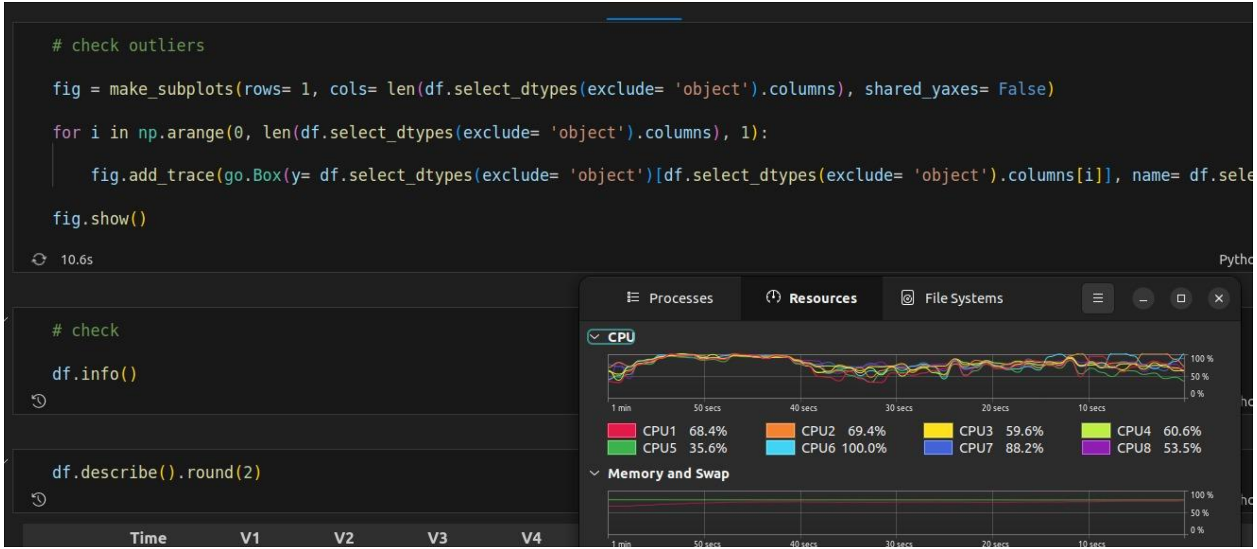
DURATION
10s

| Name | Value |
|---|---|
| ID | 1700509648185_0001 |
| User | cloudera |
| Maps | 1 of 1 |
| Reduces | 0 of 0 |
| Started | 11/20/23 13:38:14 |
| Ended | 11/20/23 13:38:25 |
| Duration | 10747 |
| Status | SUCCEEDED |
| dfs.block.access.key.update.interval | 600 |
| dfs.block.access.token.enable | false |
| dfs.block.access.token.lifetime | 600 |
| dfs.block.scanner.volume.bytes.per.second | 1048576 |
| dfs.blockreport.initialDelay | 0 |
| dfs.blockreport.intervalMsec | 21600000 |
| dfs.blockreport.split.threshold | 1000000 |
| dfs.blocksize | 134217728 |
| dfs.bytes-per-checksum | 512 |

# Platform 2 : EDA using standard libraries and dataprep

The main objective is to tune down the CPU Utilisation while performing Exploratory Data Analysis on Local Machine.

1. Traditional Method : CPU utilization of EDA using standard libraries



```
# check outliers

fig = make_subplots(rows= 1, cols= len(df.select_dtypes(exclude= 'object').columns), shared_yaxes= False)

for i in np.arange(0, len(df.select_dtypes(exclude= 'object').columns), 1):

    fig.add_trace(go.Box(y= df.select_dtypes(exclude= 'object')[df.select_dtypes(exclude= 'object').columns[i]], name= df.sele

fig.show()
```

10.6s

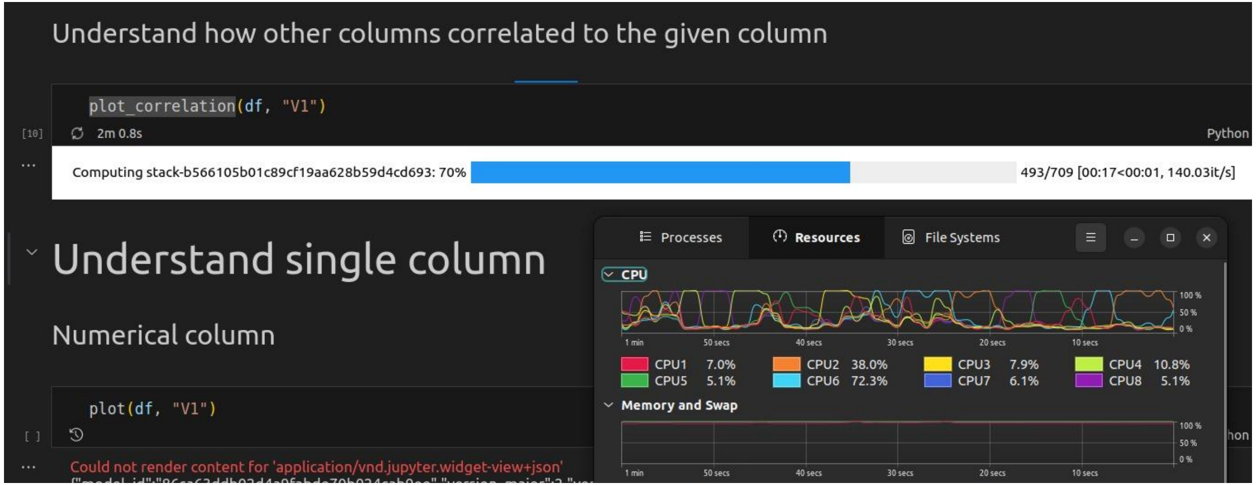
```
# check

df.info()
```

```
df.describe().round(2)
```

```
CPU1  68.4%    CPU2  69.4%    CPU3  59.6%    CPU4  60.6%
CPU5  35.6%    CPU6  100.0%   CPU7  88.2%    CPU8  53.5%
```

| Time | V1 | V2 | V3 | V4 |
|---|---|---|---|---|

## 2. EDA Library Method : CPU Utilisation after data profiling using “dataprep”



Understand how other columns correlated to the given column

```
plot_correlation(df, "V1")
```

[10] 2m 0.8s Python

```
Computing stack-b566105b01c89cf19aa628b59d4cd693: 70% 493/709 [00:17<00:01, 140.03it/s]
```

### Understand single column

Numerical column

```
plot(df, "V1")
```

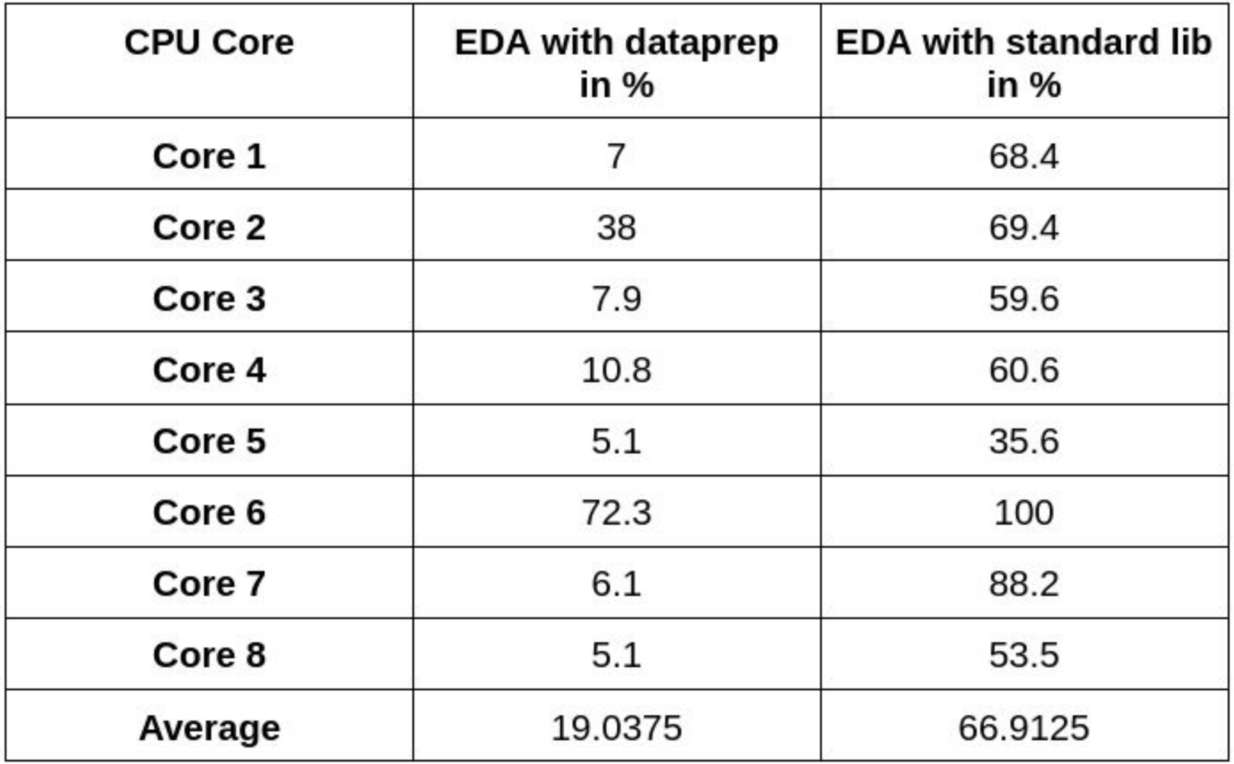
[ ]

Could not render content for 'application/vnd.jupyter.widget-view+json'

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 🟥 | CPU1 | 7.0% | 🟧 | CPU2 | 38.0% | 🟨 | CPU3 | 7.9% | 🟩 | CPU4 | 10.8% |
| 🟩 | CPU5 | 5.1% | 🟦 | CPU6 | 72.3% | 🟦 | CPU7 | 6.1% | 🟪 | CPU8 | 5.1% |

CPU Utilization has come down significantly after using the profiling method of “dataprep”.

# Performance comparison on EDA



| CPU Core | EDA with dataprep in % | EDA with standard lib in % |
|---|---|---|
| **Core 1** | 7 | 68.4 |
| **Core 2** | 38 | 69.4 |
| **Core 3** | 7.9 | 59.6 |
| **Core 4** | 10.8 | 60.6 |
| **Core 5** | 5.1 | 35.6 |
| **Core 6** | 72.3 | 100 |
| **Core 7** | 6.1 | 88.2 |
| **Core 8** | 5.1 | 53.5 |
| **Average** | 19.0375 | 66.9125 |

# Conclusion

The project explores big data using CDH platform architecture and local machine Exploratory Data Analysis (EDA). It highlights the need for robust big data ecosystems due to the surge in global data storage capacity. Practical insights into CDH, local machine EDA are presented, showcasing their processing techniques. The project provides a holistic view of big data tools and platforms, laying the foundation for future data engineering and analytics endeavors.

# References

https://blog.clairvoyantsoft.com/cloduera-quickstart-vm-using-docker-on-mac-2308acd196f2

https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9463522&tag=1

https://github.com/sfu-db/dataprep

# Thank You