

Name	Hatim Sawai
UID No.	2021300108

Experiment 7

HONOR PLEDGE	<p>I hereby declare that the documentation, code & output attached with this lab experiment has been completed by me in accordance with the highest standards of honesty. I confirm that I have not plagiarized OR used unauthorized materials OR given or received illegitimate help for completing this experiment. I will uphold equity & honesty in the evaluation of my work & if found guilty of plagiarism or dishonesty, will bear consequences as outlined in the 'integrity' section of the lab rubrics. I am doing so to maintain a community built around this code of honor.</p> <p>(Op) H.O + (Op) D.O = 4/4 Name: Hatim Sawai Sign:  AF + OD = 2/2 ASI = 2/2</p>
PROBLEM STATEMENT	<p>Title: Big data analysis using Hive/ Use of Graph database</p> <ol style="list-style-type: none"> Import the batch-specific data and store it as a Hive Table Perform a Hive query on your uploaded dataset Pull data into a spark dataframe and repeat the query using a spark dataframe Link neo4j to your underlying datastore and run a graph query
THEORY	<p>1. Hadoop Overview: Hadoop is an open-source framework designed for distributed storage and processing of large data sets across clusters of commodity hardware. It provides a distributed file system (HDFS) for storage and a framework (MapReduce) for processing and analyzing large datasets. Installation: Guide Followed: Hadoop-on-Ubuntu</p> <ol style="list-style-type: none"> Create New User for Hadoop: sudo adduser hdoop & su - hdoop Set Up SSH Keys: ssh-keygen -t rsa -P "" -f ~/.ssh/id_rsa Download the latest stable release of Hadoop from the Apache Hadoop Website. Configure Hadoop Environment: Set JAVA_HOME in /usr/local/hadoop/etc/hadoop/hadoop-env.sh

5. Update Hadoop's XML configuration files with default config:

6. Initialize the Hadoop filesystem namespace:
/usr/local/hadoop/bin/hdfs namenode -format

```
hadoop@DESKTOP-5UL8JKF:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [DESKTOP-5UL8JKF]
hadoop@DESKTOP-5UL8JKF:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@DESKTOP-5UL8JKF:~$
```

7. Start Hadoop, Launch HDFS and YARN services: start-all.sh, verify using: jps

```
hadoop@DESKTOP-5UL8JKF:~$ jps
64678 Jps
935 NameNode
1720 NodeManager
1242 SecondaryNameNode
1051 DataNode
1598 ResourceManager
hadoop@DESKTOP-5UL8JKF:~$
```

8. Check localhost:9870 for Hadoop UI and localhost:8088 for YARN UI:

Overview 'localhost:9000' (active)

Started:	Mon Mar 25 08:51:50 +0530 2024
Version:	3.3.6, r1bc78c3872bd9206a488195058f08fd12bf9c
Compiled:	Sun Jun 18 13:52:00 +0530 2023 by ubuntu from (HEAD detached at release-3.3.6-RC1)
Cluster ID:	CID-a02619bd-fcb8-4ea9-8b4d-dca159e61146
Block Pool ID:	BP-330841302-127.0.1.1-171259157079

Summary

Security is off.
Safemode is off.
14 files and directories, 4 blocks (4 replicated blocks, 0 ensure coded block groups) = 16 total filesystem object(s).
Heap Memory used 126.0 MB of 274 MB Heap Memory. Max Heap Memory is 846.5 MB.
Non Heap Memory used 49.7 MB of 51.36 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

All Applications

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[memory-mb (unit-MB), vcores]	<memory-1024, vCores-1>	<memory-8192, vCores-4>

Show 20 > entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU Vcores
No data available in table													

Showing 0 to 0 of 0 entries

2. Hive Overview:

Hive is a data warehouse infrastructure built on top of

Hadoop for providing data summarization, query, and analysis. It provides a SQL-like interface (HiveQL) to query and analyze data stored in Hadoop's distributed file system (HDFS).

Installation:

Guide Followed: [Hive-on-Ubuntu](#)

1. Download the latest stable release of Hive from the [Apache Hive website](#).
2. Extract the downloaded archive to a directory on your WSL Ubuntu system.
3. Set up environment variables in your ` `.bashrc` file to specify the Hive home directory and add Hive's bin directory to your PATH.
4. Configure Hive's XML files (`hive-site.xml`, `hive-default.xml`, etc.) according to your Hadoop and metastore setup.
5. Create Hive Directories in HDFS, Create two separate directories to store data in the HDFS layer: tmp & warehouse
6. Initiate Derby Database:
\$HIVE_HOME/bin/schematool -dbType derby -initSchema
7. Start Hive services and the metastore using the provided scripts (`hive --service metastore`, etc.).

```
hadoop@DESKTOP-SUL8JKF:~/apache-hive-3.1.2-bin/bin$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hadoop/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.
        /impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-reload4j-1.7.
        4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = bc63537d-dd02-4292-806a-2baafc4bceee4f

Logging initialized using configuration in jar:file:/home/hadoop/apache-hive-3.1.2-bin/lib/hive-
r!/hive-log4j.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider u
        execution engine (i.e. spark, tez) or using Hive 1.X releases.
Hive Session ID = b7729195-232a-434e-ae7e-293192d9374d
hive> show tables;
OK
enron
temp_employee
Time taken: 1.668 seconds, Fetched: 2 row(s)
hive>
```

```
hadoop@DESKTOP-SUL8JKF: $ cd $HIVE_HOME/bin
hadoop@DESKTOP-SUL8JKF:~/apache-hive-3.1.2-bin/bin$ hive --service metastore
2024-03-25 08:57:18: Starting Hive Metastore Server
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hadoop/apache-hive-3.1.2-bin/lib/log4j-slf4
        /impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-
        4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
```

3. Spark Overview:

Apache Spark is an open-source, distributed computing system that provides an interface for programming entire clusters with implicit data parallelism and fault tolerance. It supports various programming languages such as Scala, Java, Python, and R.

Installation:

pip install spark (in python notebook)

localhost:4040/jobs/

Spark Jobs (3)

User: shubhar
Total Uptime: 44 s
Scheduling Mode: FIFO
Completed Jobs: 3

Event Timeline
Completed Jobs (3)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
2	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2024/03/25 09:14:03	0.5 s	1/1	1/1
1	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2024/03/25 09:13:58	4 s	1/1	4/4
0	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2024/03/25 09:13:54	2 s	1/1	1/1

Page: 1 / 1 Pages, Jump to: 1 Show 100 items in a page. Go

localhost:4040/SQL/execution/?id=1

Details for Query 1

Submitted Time: 2024/03/25 09:14:03
Duration: 1 s
Succeeded Jobs: 2

Show the Stage ID and Task ID that corresponds to the max metric

```

graph TD
    A[Scan csv] --> B[WholeStageCodegen (1)]
    B --> C[Project]
    C --> D[CollectLimit]
  
```

Scan csv
number of output rows: 21
number of files read: 1
metadata time: 0 ms
size of files read: 106.2 MiB

WholeStageCodegen (1)
duration: 0 ms

Project

CollectLimit

[Details](#)

4. Neo4j Overview:

Neo4j is a graph database management system that provides an efficient and expressive way to store, manage, and query highly connected data. It is optimized for handling graph data and supports powerful querying capabilities using the Cypher query language.

Installation:

Guide Followed: [Neo4j-on-Ubuntu](#)

1. Download the latest stable release of Neo4j from the [Neo4j Website](#)
2. Follow the installation instructions provided for Linux distributions.
3. After installation, start the Neo4j service using the provided startup scripts: sudo systemctl enable/start/edit/status/stop neo4j.

4. Configure username and password through cypher-shell:
:server change-password
5. Access the Neo4j browser interface using a web browser and connect to the Neo4j database.

```

hadoop@DESKTOP-5UL8JKF: $ sudo systemctl restart neo4j
hadoop@DESKTOP-5UL8JKF: $ cypher-shell -a 'neo4j://127.0.0.1:7687'
username: neo4j
password:
Password change required
new password:
confirm password:
A password must be at least 8 characters.
hadoop@DESKTOP-5UL8JKF: $ cypher-shell -a 'neo4j://127.0.0.1:7687'
username: neo4j
password:
Password change required
new password:
confirm password:
Connected to Neo4j using Bolt protocol version 5.4 at neo4j://127.0.0.1:7687
Type :help for a list of available commands or :exit to exit the shell.
Note that Cypher queries must end with a semicolon.
neo4j@neo4j> █

hadoop@DESKTOP-SUL8JKF: $ cypher-shell -a 'neo4j://127.0.0.1:7687'
username: neo4j
password:
Connected to Neo4j using Bolt protocol version 5.4 at neo4j://127.0.0.1:7687 as user neo4j.
Type :help for a list of available commands or :exit to exit the shell.
Note that Cypher queries must end with a semicolon.
neo4j@neo4j> CREATE (:Shark {name: 'Great White'});
0 rows
ready to start consuming query after 542 ms, results consumed after another 0 ms
Added 1 nodes, Set 1 properties, Added 1 labels
neo4j@neo4j> CREATE
  (:Shark {name: 'Hammerhead'})-[:FRIEND]→
  (:Shark {name: 'Sammy'})-[:FRIEND]→
  (:Shark {name: 'Megalodon'});
0 rows
ready to start consuming query after 250 ms, results consumed after another 0 ms
Added 3 nodes, Created 2 relationships, Set 3 properties, Added 3 labels
neo4j@neo4j> MATCH (a:Shark),(b:Shark)
  WHERE a.name = 'Sammy' AND b.name = 'Megalodon'
  CREATE (a)-[r:ORDER { name: 'Lamniformes' }]->(b)
  RETURN type(r), r.name;
+-----+
| type(r) | r.name      |
+-----+
| "ORDER" | "Lamniformes" |
+-----+

```

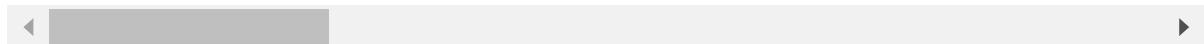
1. Importing Libraries & Dataset

```
In [ ]: import re
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from neo4j import GraphDatabase
from pyspark.sql import SparkSession
# read csv file
df = pd.read_csv('../Datasets/enron.csv', index_col=0, low_memory=False)
df.head()
```

Out[]:

		Message-ID	Date	From
0	<18782981.1075855378110.JavaMail.evans@thyme>		2001-05-14 23:39:00	frozenset({'phillip.allen@enron.cor
1	<15464986.1075855378456.JavaMail.evans@thyme>		2001-05-04 20:51:00	frozenset({'phillip.allen@enron.cor
2	<24216240.1075855687451.JavaMail.evans@thyme>		2000-10-18 10:00:00	frozenset({'phillip.allen@enron.cor
3	<13505866.1075863688222.JavaMail.evans@thyme>		2000-10-23 13:13:00	frozenset({'phillip.allen@enron.cor
4	<30922949.1075863688243.JavaMail.evans@thyme>		2000-08-31 12:07:00	frozenset({'phillip.allen@enron.cor

5 rows × 51 columns



2. Preprocessing Data

```
In [ ]: def sanitize_column_name(name):
    return re.sub(r'^[A-Za-z0-9_]', '_', name)

new_columns = [sanitize_column_name(col) for col in df.columns]
df.columns = new_columns
df.to_csv("../Datasets/cleaned_enron.csv", index=False)
```

```
In [ ]: df_new = pd.read_csv('../Datasets/cleaned_enron.csv', index_col=0, low_memory=False)
df_new.shape[0]
df_subset = df.iloc[:5000]
df_subset.to_csv("../Datasets/enron_chunk.csv", index=False)
```

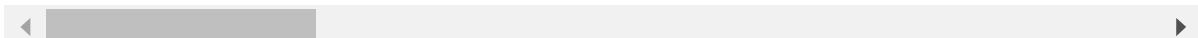
```
In [ ]: df_subset = pd.read_csv('../Datasets/enron_chunk.csv', index_col=0, low_memory=False)
print(df_subset.shape[0])
df_subset.head()
```

50000

Out[]:

	Date	Fr
	Message_ID	
<18782981.1075855378110.JavaMail.evans@thyme>	2001-05-14 23:39:00	frozenset({'phillip.allen@enron.co
<15464986.1075855378456.JavaMail.evans@thyme>	2001-05-04 20:51:00	frozenset({'phillip.allen@enron.co
<24216240.1075855687451.JavaMail.evans@thyme>	2000-10-18 10:00:00	frozenset({'phillip.allen@enron.co
<13505866.1075863688222.JavaMail.evans@thyme>	2000-10-23 13:13:00	frozenset({'phillip.allen@enron.co
<30922949.1075863688243.JavaMail.evans@thyme>	2000-08-31 12:07:00	frozenset({'phillip.allen@enron.co

5 rows × 50 columns



3. Loading Dataset to a Spark Dataframe

```
In [ ]: # Create SparkSession
spark = SparkSession.builder \
    .appName("CSV to Hive") \
    .config("hive.metastore.uris", "thrift://localhost:9083") \
    .enableHiveSupport() \
    .getOrCreate()

# Read CSV file into DataFrame
csv_file_path = "./eron2.csv"
df1 = spark.read.csv(csv_file_path, header=True, inferSchema=True)
# Define a function to sanitize column names
# def sanitize_column_name(name):
#     return re.sub(r'^[A-Za-z0-9_]', '_', name)

# new_columns = [sanitize_column_name(col) for col in df1.columns]
# df1 = df1.toDF(*new_columns)
df1.show()
```

```
your 131072x1 screen size is bogus. expect trouble
Warning: Ignoring non-Spark config property: hive.metastore.uris
24/03/24 11:55:07 WARN Utils: Your hostname, DESKTOP-5UL8JKF resolves to a loopback
address: 127.0.1.1; using 172.27.122.177 instead (on interface eth0)
24/03/24 11:55:07 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(ne
wLevel).
24/03/24 11:55:10 WARN NativeCodeLoader: Unable to load native-hadoop library for yo
ur platform... using builtin-java classes where applicable
24/03/24 11:55:30 WARN SparkStringUtils: Truncated the string representation of a pl
an since it was too large. This behavior can be adjusted by setting 'spark.sql.debu
g.maxToStringFields'.
```

	Message_ID	Date	From	To	
Subject	X_From	X_To	X_cc X_bcc		
X_Folder	X-Origin	X_FileName	content		
user	Cat_1_level_1	Cat_1_level_2	Cat_1_weight	Cat_2_level	
el_1	Cat_2_level_2	Cat_2_weight	Cat_3_level_1	Cat_3_level	
el_2	Cat_3_weight	Cat_4_level_1	Cat_4_level_2	Cat_4_weight	
ght	Cat_5_level_1	Cat_5_level_2	Cat_5_weight	Cat_6_level_1 Cat_6_level_2 Cat_6_weight Cat_7_level_1 Cat_7_level_2 Cat_7_weight Cat_8_level_1 Cat_8_level_2 Cat_8_weight Cat_9_level_1 Cat_9_level_2 Cat_9_weight Cat_10_level_1 Cat_10_level_2 Cat_10_weight Cat_11_level_1 Cat_11_level_2 Cat_11_weight Cat_12_level_1 Cat_12_level_2 Cat_12_weight labeled	
<18782981.1075855... 2001-05-14 23:39:00 frozenset({'phill... frozenset({'tim.b... NULL Phillip K Allen Tim Belden <Tim B... NULL NULL \\"Phillip_Allen_Ja... Allen-P pallen (Non-Privi... Here is our forecast al					
len-p	NULL	NULL	NULL	NULL	
NULL	NULL	NULL	NULL	NULL	
NULL	NULL	NULL	NULL	NULL	
ULL	NULL	NULL	NULL	NULL	
NULL	NULL	NULL	NULL	NULL	
LL	NULL	NULL	NULL	NULL	
LL	NULL	NULL	NULL	NULL	
NULL	NULL False				
<15464986.1075855... 2001-05-04 20:51:00 frozenset({'phill... frozenset({'john.... Re: Phillip K Allen John J Lavorato <... NULL NULL \\"Phillip_Allen_Ja... Allen-P pallen (Non-Privi... Traveling to have... all					
en-p	NULL	NULL	NULL	NULL	
NULL	NULL	NULL	NULL	NULL	
NULL	NULL	NULL	NULL	NULL	
ULL	NULL	NULL	NULL	NULL	
NULL	NULL	NULL	NULL	NULL	
LL	NULL	NULL	NULL	NULL	
LL	NULL	NULL	NULL	NULL	
NULL	NULL False				
<24216240.1075855... 2000-10-18 10:00:00 frozenset({'phill... frozenset({'leah....					

Re: test|Phillip K Allen| Leah Van Arsdall| NULL| NULL|\Phillip_A1
len_De...| Allen-P| pallen.nsf|test successful. ...|
allen-p| NULL| NULL| NULL|
NULL| NULL| NULL| NULL|
NULL| NULL| NULL| NULL| N
ULL| NULL| NULL| NULL| NULL|
NULL| NULL| NULL| NULL| NULL| NU
LL| NULL| NULL| NULL| NULL| NU
LL| NULL| NULL| NULL| NULL| NU
NULL| NULL| False|
|<13505866.1075863...|2000-10-23 13:13:00|frozensest({'phill...|frozensest({'randa...|
NULL| Phillip K Allen| Randall L Gay| NULL| NULL|\Phillip_Alle...|
De...| Allen-P| pallen.nsf|Randy, Can you se...| al
len-p| NULL| NULL| NULL|
NULL| NULL| NULL| NULL|
NULL| NULL| NULL| NULL| N
ULL| NULL| NULL| NULL| NULL|
NULL| NULL| NULL| NULL| NULL| NU
LL| NULL| NULL| NULL| NULL| NU
LL| NULL| NULL| NULL| NULL| NU
NULL| NULL| False|
|<30922949.1075863...|2000-08-31 12:07:00|frozensest({'phill...|frozensest({'greg....|
Re: Hello|Phillip K Allen| Greg Piper| NULL| NULL|\Phillip_A
llen_De...| Allen-P| pallen.nsf|Let's shoot for T...|
allen-p| NULL| NULL|
NULL| NULL| NULL| NULL|
NULL| NULL| NULL| NULL| N
ULL| NULL| NULL| NULL| NULL|
NULL| NULL| NULL| NULL| NULL| NU
LL| NULL| NULL| NULL| NULL| NU
LL| NULL| NULL| NULL| NULL| NU
NULL| NULL| False|
|<30965995.1075863...|2000-08-31 11:17:00|frozensest({'phill...|frozensest({'greg....|
Re: Hello|Phillip K Allen| Greg Piper| NULL| NULL|\Phillip_A
llen_De...| Allen-P| pallen.nsf|Greg, How about e...|
allen-p| NULL| NULL|
NULL| NULL| NULL|
NULL| NULL| NULL| N
ULL| NULL| NULL| NULL| NULL|
NULL| NULL| NULL| NULL| NULL| NU
LL| NULL| NULL| NULL| NULL| NU
LL| NULL| NULL| NULL| NULL| NU
NULL| NULL| False|
|<16254169.1075863...|2000-08-22 14:44:00|frozensest({'phill...|frozensest({'david...|
NULL| Phillip K Allen|david.l.johnson@e...| NULL| NULL|\Phillip_Alle...|
De...| Allen-P| pallen.nsf|Please cc the fol...| al
len-p| NULL| NULL|
NULL| NULL| NULL|
NULL| NULL| NULL| N
ULL| NULL| NULL| NULL| NULL|
NULL| NULL| NULL| NULL| NULL| NU
LL| NULL| NULL| NULL| NULL| NU
LL| NULL| NULL| NULL| NULL| NU
NULL| NULL| False|
|<17189699.1075863...|2000-07-14 13:59:00|frozensest({'phill...|frozensest({'joyce...|
Re: PRC review - ...|Phillip K Allen| Joyce Teixeira| NULL| NULL

|\\Phillip_Allen_De...| Allen-P| pallen.nsf|any morning betwe...
allen-p	NULL	NULL	NULL
NULL	NULL	NULL	False
<20641191.1075855...	2000-10-17 09:26:00	frozensest({'phill...	frozensest({'mark....
Re: High Speed In...	Phillip K Allen	Mark Scott	NULL
\\Phillip_Allen_De...	Allen-P	pallen.nsf	1. login: pallen ...
allen-p	NULL	NULL	NULL
NULL	NULL	NULL	False
<30795301.1075855...	2000-10-16 13:44:00	frozensest({'phill...	frozensest({'zimam...
FW: fixed forward...	Phillip K Allen	zimam@enron.com	NULL
\\Phillip_Allen_De...	Allen-P	pallen.nsf	-----...
Buck"" <buck.buc...	As discussed d...	In a Parallon 75...	I am developing ...
I need your > be...	3	5	7 and 10 years f...
truly you. > > W...	during peak > el...	give me a call.	this is real dea...
P.E.	MBA > Manager	Business Develop...	Inc. > 8725 Pan ...
113 > 505-7...	allen-p	NULL	NULL
NULL	NULL	NULL	NULL
ULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
<33076797.1075855...	2000-10-16 13:42:00	frozensest({'phill...	frozensest({'buck....
Re: FW: fixed for...	Phillip K Allen	""Buckner	Buck"" <buck.buc...
NULL	\\Phillip_Allen_De...	Allen-P	pallen.nsf
Mr. Buckner, For ...	allen-p	NULL	NULL
NULL	NULL	NULL	NULL
<25459584.1075855...	2000-10-13 13:45:00	frozensest({'phill...	frozensest({'stage...
NULL	Phillip K Allen	stagecoachmama@ho...	NULL
De...	Allen-P	pallen.nsf	Lucy, Here are th...
len-p	NULL	NULL	NULL
NULL	NULL	False	
<13116875.1075855...	2000-10-09 14:16:00	frozensest({'phill...	frozensest({'keith...
Consolidated posi...	Phillip K Allen	Keith Holst	NULL
\\Phillip_Allen_De...	Allen-P	pallen.nsf	-----...

	allen-p		NULL		NULL		NULL
	NULL		NULL		NULL		NULL
	NULL		NULL		NULL		NULL
	NULL	NULL	NULL		NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NU
LL	NULL	NULL	NULL	NULL	NULL	NULL	NUL
L	NULL	NULL	NULL	NULL	NULL	NULL	NU
NULL	NULL	NULL	False				
<2707340.10758556... 2000-10-09 14:00:00 frozense({'phill... frozense({'keith...							
Consolidated posi... Phillip K Allen				Keith Holst			
\Phillip_Allen_De...		Allen-P		pallen.nsf -----...			
	allen-p		NULL		NULL		NULL
	NULL		NULL		NULL		NULL
	NULL		NULL		NULL		NULL
	NULL	NULL	NULL		NULL	NULL	
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NU
LL	NULL	NULL	NULL	NULL	NULL	NULL	NUL
L	NULL	NULL	NULL	NULL	NULL	NULL	NU
NULL	NULL	NULL	False				
<2465689.10758556... 2000-10-05 13:26:00 frozense({'phill... frozense({'david...							
NULL Phillip K Allen		David W Delainey			NULL	NULL	\Phillip_Allen_
De...		Allen-P		pallen.nsf Dave, Here are th...			al
len-p		NULL		NULL		NULL	
NULL		NULL		NULL		NULL	
NULL		NULL		NULL		NULL	N
ULL		NULL		NULL		NULL	
NULL		NULL		NULL		NULL	NU
LL		NULL		NULL		NULL	NU
LL		NULL		NULL		NULL	NU
NULL		NULL	False				
<1115198.10758556... 2000-10-05 12:55:00 frozense({'phill... frozense({'paula...							
Re: 2001 Margin Plan Phillip K Allen				Paula Harris			
\Phillip_Allen_De...		Allen-P		pallen.nsf Paula, 35 million...			
	allen-p		NULL		NULL		NULL
	NULL		NULL		NULL		NULL
	NULL		NULL		NULL		NULL
	NULL	NULL	NULL		NULL	NULL	
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NU
LL	NULL	NULL	NULL	NULL	NULL	NULL	NUL
L	NULL	NULL	NULL	NULL	NULL	NULL	NU
NULL	NULL	NULL	False				
<19773657.1075855... 2000-10-04 16:23:00 frozense({'phill... frozense({'ina.r...							
Var, Reporting an... Phillip K Allen				Ina Rangel			
\Phillip_Allen_De...		Allen-P		pallen.nsf -----...			
	allen-p		NULL		NULL		NULL
	NULL		NULL		NULL		NULL
	NULL		NULL		NULL		NULL
	NULL	NULL	NULL		NULL	NULL	
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NU
LL	NULL	NULL	NULL	NULL	NULL	NULL	NUL
L	NULL	NULL	NULL	NULL	NULL	NULL	NU
NULL	NULL	NULL	False				
<7391389.10758553... 2001-05-04 18:26:00 frozense({'phill... frozense({'tim.h...							
NULL Phillip K Allen Tim Heizenrader <...					NULL	NULL	\Phillip_Allen_
Ja...		Allen-P pallen (Non-Privi... Tim, mike grigsby...					al
len-p		NULL		NULL		NULL	

NULL	NULL	NULL	NULL	NULL	NULL	N
NULL	NULL	NULL	NULL	NULL	NULL	NU
ULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NU
LL	NULL	NULL	NULL	NULL	NULL	NU
LL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	False				
<12759088.1075855... 2000-10-03 16:30:00 frozenset({'phill... frozenset({'palle...						
Westgate Phillip K Allen pallen70@hotmail.com					NULL	NULL \Phillip_A
len_De...	Allen-P	pallen.nsf -----				let me k
now. Then let me know your... which now has a ... roughly \$1.25 pe... but by ind						
ividua... just as the dupl... please let me kn... very short windo... but it wou						
ld be ... 000 to secure the... George W. Richar... Creekside Builders LLC - winm						
ail.dat" allen-p		NULL		NULL		NU
LL	NULL	NULL	NULL	NULL	NULL	NULL
	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL				
<29177675.1075855... 2000-10-03 16:15:00 frozenset({'phill... frozenset({'ina.r...						
Meeting re: Stora... Phillip K Allen		Ina Rangel			NULL	NULL
\Phillip_Allen_De...	Allen-P	pallen.nsf -----				
	allen-p	NULL		NULL		NULL
	NULL	NULL		NULL		NULL
	NULL	NULL		NULL		NULL
	NULL	NULL		NULL		NULL
NULL	NULL	NULL	NULL	NULL	NULL	NU
LL	NULL	NULL	NULL	NULL	NULL	NU
L	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	False			
+-----+-----+-----+-----+-----+-----+						
-+-----+-----+-----+-----+-----+-----+						
+-----+-----+-----+-----+-----+-----+						
+-----+-----+-----+-----+-----+-----+						
+-----+-----+-----+-----+-----+-----+						
+-----+-----+-----+-----+-----+-----+						
-+-----+-----+-----+-----+-----+-----+						
-+-----+-----+-----+-----+-----+-----+						
-+-----+-----+-----+-----+-----+-----+						
+-----+-----+-----+-----+-----+-----+						
only showing top 20 rows						

only showing top 20 rows

4. Using Spark to load Data into Hive Table

```
In [ ]: # Extract schema from DataFrame
schema = df1.schema
print(schema)
```

```
StructType([StructField('Message_ID', StringType(), True), StructField('Date', StringType(), True), StructField('From', StringType(), True), StructField('To', StringType(), True), StructField('Subject', StringType(), True), StructField('X_From', StringType(), True), StructField('X_To', StringType(), True), StructField('X_cc', StringType(), True), StructField('X_bcc', StringType(), True), StructField('X_Folder', StringType(), True), StructField('X-Origin', StringType(), True), StructField('X_FileName', StringType(), True), StructField('content', StringType(), True), StructField('user', StringType(), True), StructField('Cat_1_level_1', StringType(), True), StructField('Cat_1_level_2', StringType(), True), StructField('Cat_2_level_1', StringType(), True), StructField('Cat_2_level_2', StringType(), True), StructField('Cat_3_level_1', StringType(), True), StructField('Cat_3_level_2', StringType(), True), StructField('Cat_4_level_1', StringType(), True), StructField('Cat_4_level_2', StringType(), True), StructField('Cat_5_level_1', StringType(), True), StructField('Cat_5_level_2', StringType(), True), StructField('Cat_6_level_1', StringType(), True), StructField('Cat_6_level_2', StringType(), True), StructField('Cat_7_level_1', StringType(), True), StructField('Cat_7_level_2', StringType(), True), StructField('Cat_8_level_1', StringType(), True), StructField('Cat_8_level_2', StringType(), True), StructField('Cat_8_weight', StringType(), True), StructField('Cat_9_level_1', StringType(), True), StructField('Cat_9_weight', StringType(), True), StructField('Cat_10_level_1', StringType(), True), StructField('Cat_10_level_2', StringType(), True), StructField('Cat_10_weight', StringType(), True), StructField('Cat_11_level_1', StringType(), True), StructField('Cat_11_level_2', StringType(), True), StructField('Cat_11_weight', StringType(), True), StructField('Cat_12_level_1', StringType(), True), StructField('Cat_12_level_2', StringType(), True), StructField('label', StringType(), True)])
```

```
In [ ]: # Create Hive table with extracted schema
table_name = "enron"
schema_ddl = ', '.join([f'{field.name} {field.dataType.simpleString()}' for field in
create_table_query = f"CREATE TABLE IF NOT EXISTS {table_name} ({schema_ddl}) STORE
spark.sql(create_table_query)
```

24/03/24 11:46:32 WARN SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.security.authorization.manager is set to instance of HiveAuthorizerFactory.

Out[]: DataFrame[]

```
In [ ]: # Load data into Hive table
df1.write.mode("overwrite").saveAsTable(table_name)
```

```
In [ ]: spark.stop()
```

5. Performing Hive Queries

```
In [ ]: # Create SparkSession
spark = SparkSession.builder \
    .appName("Load Hive Table") \
    .config("hive.metastore.uris", "thrift://localhost:9083") \
    .enableHiveSupport() \
    .getOrCreate()

# Load Hive table into DataFrame
df = spark.sql("SELECT * FROM enron")

# Show DataFrame
df.show()
```

```
24/03/24 11:55:41 WARN SparkSession: Using an existing Spark session; only runtime S
QL configurations will take effect.
[Stage 3:> (0 + 1) / 1]
```

	Message_ID	Date	From	To
Subject	X_From	X_To	X_cc	
X_bcc	X_Folder	X_Origin	X_FileName	co
ntent	user	Cat_1_level_1	Cat_1_level_2	Cat_1_w
eight	Cat_2_level_1	Cat_2_level_2	Cat_2_weight	Cat_3_le
vel_1	Cat_3_level_2	Cat_3_weight	Cat_4_level_1	Cat_4_le
vel_2	Cat_4_weight	Cat_5_level_1	Cat_5_level_2	Cat_5_w
eight	Cat_6_level_1	Cat_6_level_2	Cat_6_weight	Cat_7_le
vel_1	Cat_7_level_2	Cat_7_weight	Cat_8_level_1	Cat_8_le
vel_2	Cat_8_weight	Cat_9_level_1	Cat_9_level_2	Cat_9_w
eight	Cat_10_level_1	Cat_10_level_2	Cat_10_weight	Cat_11_le
vel_1	Cat_11_level_2	Cat_11_weight	Cat_12_level_1	Cat_12_le
vel_2	Cat_12_weight	labeled		
<25542984.1075854... 2000-04-20 10:11:00 frozenset({'eric.... frozenset({'lwbth...				
Fwd: Qualify for ...	Eric Bass lwbthemarine@bigp...			NULL
NULL \Eric_Bass_Dec200...	Bass-E	ebass.nsf "Thought you mig		
h... 19 Apr 2000 20:5... queueup Copyright... Inc. 2000)with s... 19 Apr 2000 >2				
0:... 000+ FREE FREE > \$25	000 - \$99 999 FREE \$14.95 >... 000 \$14.95 \$14.			
95... 7 >days a week. ...	November 18 1999. Barron's i... money market			
funds mutual funds	stocks and >bonds >held...			000 s
hares >each additional... Inc. Member NASD... Inc. and is >not... just hit the r				
ep... review or >chang... Free Email at ht...	bass-e			
NULL	NULL	NULL	NULL	
NULL	NULL	NULL	NULL	
NULL	NULL	NULL	NULL	
NULL	NULL	NULL	NULL	
<5541885.10758546... 2000-04-20 16:27:00 frozenset({'eric.... frozenset({'timot...				
NULL	Eric Bass Timothy Blanchard			NULL
NULL \Eric_Bass_Dec200...	Bass-E	ebass.nsf -----		
-...	bass-e	NULL	NULL	

NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	False	
<28915064.1075854... 2000-04-24 08:31:00 frozense({'eric.... frozense({""o'n... david.baumbach@... jeffrey.gossett... amir.ahanchian@... kenneth.shulkla... timothy.blancha... nick.hiemstra@e... matthew.lenhart... bryan.hull@enro... luis.mena@enron... brian.hoskins@e... Game Tonight Eric Bass B rian Hoskins, Lu... NULL NULL \Eric_Bass_Dec200...			
Bass-E	ebass.nsf We have a game to...	bass-e	
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
<6649833.10758546... 2000-04-26 08:55:00 frozense({'eric.... frozense({'georg... WH DATA Eric Bass George Weissman NULL			
NULL \Eric_Bass_Dec200...	Bass-E	ebass.nsf Here is the wh d	
a... bass-e	NULL	NULL	
NULL NULL	NULL	NULL	
NULL NULL	NULL	NULL	
NULL NULL	NULL	NULL	
NULL NULL	NULL	NULL	
NULL NULL	NULL	NULL	
NULL NULL	NULL	NULL	
NULL NULL	NULL	NULL	
NULL NULL	NULL	NULL	
NULL NULL	False		
<27501004.1075854... 2000-04-26 11:12:00 frozense({'eric.... frozense({'timot... I'm Having a Party! Eric Bass Timothy Blanchard... NULL			
NULL \Eric_Bass_Dec200...	Bass-E	ebass.nsf ----- -... ""Michael Galun... ""Robert Garcia"... ""Kara Lynn Holm... ""Ho ward Patrick"" <Howar... ""Robinson Jesse"" <Jesse.R... Kevin Swantkows k... ""Curtis Willefo... bwoodwar@central... ""Lee Zieben"" <... jhcl@dyneg y.com Chad Starnes/Cor... Billy Braddock/E... dwighthelms@aol.com Michael Stockt on... Janel Guerrero/C... April 29th- All ... Duck Soup will b... the more the m er... please call. If ... let me know by F... bass-e	
NULL NULL	NULL	NULL	
NULL NULL	NULL	NULL	
NULL NULL	NULL	NULL	
NULL NULL	NULL	NULL	
<30641730.1075854... 2000-04-26 16:46:00 frozense({'eric.... frozense({'lwbth... Fw: Corruption Test Eric Bass "lwbthemarine@big... mballases@hotmail... Jason.Bass2@COMP... NULL NULL \Eric_Bass_Dec200...			
Bass-E	ebass.nsf -----	bass-e	
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL

NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
<2149245.10758546... 2000-04-26 16:47:00 frozense({'eric.... frozense({'shuss...			
Fw: Corruption Test	Eric Bass	Shusser@enron.com	NULL
NULL \Eric_Bass_Dec200...	Bass-E	ebass.nsf i bet you score	
o... bass-e	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	False		
<29739912.1075854... 1999-12-14 09:40:00 frozense({'eric.... frozense({'danie...			
Fw: FROGAPULT, EL...	Eric Bass	danielles@jonesgr...	NULL
NULL \Eric_Bass_Dec200...	Bass-E	ebass.nsf -----	
-... M.S." <rparry@b...	M.S." <rparry@b...	ELFBOWL	Y2KGAME Virus H
o... ELFBOWL	Y2KGAME Virus Ho...	December 13	1999 11:37 AM S
u... ELFBOWL	Y2KGAME Virus Ho...	Liesha -----Orig...	Decembe
r 13 1999 10:50 AM To...	ELFBOWL	Y2KGAME Virus Ho...	ELF
BOWL Y2KGAME Virus Ho...	bass-e	NULL	
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
<17794525.1075854... 2000-04-26 16:47:00 frozense({'eric.... "frozense({""o'n...			
'david.baumbach@... 'bryan.hull@enro... 'michael.walters... Fw: Corruption Test			
Eric Bass Bryan Hull, O'Nea...		NULL	NULL \Eric_Bass_
Dec200... Bass-E		ebass.nsf -----	
bass-e NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
<1944629.10758546... 2000-04-27 07:38:00 frozense({'eric.... frozense({'timot...			
Fw: Corruption Test	Eric Bass	Timothy Blanchard	NULL
NULL \Eric_Bass_Dec200...	Bass-E	ebass.nsf -----	
-... bass-e	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL
NULL NULL	NULL	NULL	NULL

NULL	NULL	False
<1373091.10758546... 2000-04-28 07:38:00 frozense({'eric.... frozense({'phill...		
Fw: Corruption Test	Eric Bass	Phillip M Love
NULL \Eric_Bass_Dec200...	Bass-E	ebass.nsf -----
-...	bass-e	NULL
NULL	NULL	False
<25662208.1075854... 2000-05-01 07:53:00 frozense({'eric.... "frozense({""o'n...		
'david.baumbach@... 'jeffrey.gossett... 'amir.ahanchian@... 'kenneth.shulkla...		
'timothy.blancha... 'nick.hiemstra@e... 'matthew.lenhart... 'bryan.hull@enro...		
'luis.mena@enron... 'brian.hoskins@e... GAME TONIGHT @ 7:45 Eric Bass B		
rian Hoskins, Lu...	NULL	NULL \Eric_Bass_Dec200...
Bass-E	ebass.nsf If the weather ho...	bass-e
NULL	NULL	NULL
<29617661.1075854... 2000-05-01 09:02:00 frozense({'eric.... frozense({'dale....		
New Product	Eric Bass	Dale Neuner Melba Lozano, Tod...
NULL \Eric_Bass_Dec200...	Bass-E	ebass.nsf Dale, On the new
...	bass-e	NULL
ULL	NULL	False
<4041756.10758546... 2000-05-01 13:40:00 frozense({'eric.... "frozense({""o'n...		
'bryan.hull@enro... Re: Wellhead Accr...	Eric Bass Bryan Hull, O'Nea...	
NULL	NULL \Eric_Bass_Dec200...	Bass-E
ns.nsf "Here are the num... 000 of accrual va... accrual term deals actual volumes		ebas
g... half cent bid to... incorrect markings and value withhe... Greg Sharp will		
... actual volumes g... and deals incorr... 504 Greg's Value ... 750 Bid to Mid =		
... 000 Accrual Value... 000 Accrual Costs... 000 Actual Volume... 000		
\$118 000 Incorrect Mar... 000) Total Value ... 750 February- Tot... 305 Greg's Value		
... 000 Bid to Mid = ... 000 Accrual Value... 000 Accrual Costs... 000 Actual Volum		
e... 000 \$113 000 Incorrect Mar... 000) Total Value ... 000 March - Tota		
1... 145 Greg's Value ... 000 Bid to Mid = ... 000 Accrual Value... 000 Accrual Cost		
s... 000 Actual Volume... 000 \$119 000 Incorrect Mar... 000) Total Value		
... 000 Let me know i... Here are the acc...		
<32439370.1075854... 2000-10-24 09:57:00 frozense({'eric.... frozense({'shann...		
Re: It could happ...	Eric Bass	Shanna Husser
NULL \Eric_Bass_Jun200...	Bass-E	ebass.nsf "Thought you mig

h...	bass-e	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL
NULL	NULL	False	
<15626432.1075854... 2000-05-01 17:13:00 frozense({'eric.... frozense({'timot...	Fwd: Fw: Somethin...	Eric Bass Brian Hoskins, Ja...	NULL
NULL \Eric_Bass_Dec200...	Bass-E	ebass.nsf -----	
-... John9375@aol.com CHMARGAU@aol.com srhea@acm.org Eric Bass/HOU/E			
C... ""MONTY WEEKS"" ... ""John Steiert""... ""Spillar >Steve"" <mrmrmax			
4... ""Short Michael J"" <bec... >""Dwayne Shellh... ""John Schofiel			
d... ""Matthew Rub"" ... ""RD"" <sabbath@... ""Cecil Poe"" ><... ""Pal			
umbo Donaldq"" <hardhe... ""Barry >Mullin".... ""Gary Monday"" ... ""Shane >Mille			
r"... ""Mike McBride"".... >""Doug Marshall... ""Brett Lawler""....			
""LaHue Robert"" <RLaHue... ""JettyCat"" ><f... ""Vernon Hooks""... ""Aaron >Hom			
mel".... ""Mike Henderson... ""John >Handley".... ""Mike Greene"" ... >""Roy V. Cl			
ark".... ""Allan Carruthe... >""Saul Carrillo... ""Jeff Campbell"..." ""BREW"" <BR			
EWMA... ""Johnny Breed"".... ""BeachBum"" <Be...			
<16080707.1075854... 2000-10-24 10:22:00 frozense({'eric.... frozense({'shann...			
Re: It could happ...	Eric Bass	Shanna Husser	NULL
NULL \Eric_Bass_Jun200...	Bass-E	ebass.nsf -----	
-... Christopher Coff... William Kelly/HO... Kyle Etter/HOU/E... Kam Keiser/HOU/			
E... Jay Reitmeyer/HO... Jeff Coates/HOU/... William Keeney/H... Jeffrey C Gosse			
t... John King/HOU/EC... Luis Mena/NA/Enr... @ ENRON Lisa Gillette/H			
O... Susan M Scott/HO... Dawn C Kenne/HOU... Nick Hiemstra/HO... Benjamin Thomas			
o... David Marks/HOU/... Timothy Blanchar... you still have t... none of those a			
r... LSU to the Cotto... the Sharpie Inde... a sweep would ne...			
RBs WRs and Defense. Bes... we have turned t... the SEC is havi			
n... but which school... Matthew Lenhart/... Christopher Coff... William Kelly/H			
O... Kyle Etter/HOU/E... Kam Keiser/HOU/E... Jay Reitmeyer/HO... Jeff Coates/HO			
U/... William Keeney/H... Jeffrey C Gosset...			
<2906108.10758546... 2000-05-04 10:29:00 frozense({'eric.... frozense({'dale....			
5X24 Eric Bass Dale Neuner NULL			
NULL \Eric_Bass_Dec200...	Bass-E	ebass.nsf Any news? we wou	
1... bass-e NULL NULL			
NULL NULL NULL NULL			
NULL NULL NULL NULL			
NULL NULL NULL NULL			
NULL NULL NULL NULL			
NULL NULL NULL NULL			
NULL NULL NULL NULL			
NULL NULL NULL NULL			
NULL NULL NULL NULL			
NULL NULL False			
<21265183.1075854... 2000-05-08 07:08:00 frozense({'eric.... "frozense({""o'n...			
'david.baumbach@... 'jeffrey.gossett... 'amir.ahanchian@... 'kenneth.shulkla...			
'timothy.blancha... 'nick.hiemstra@e... 'matthew.lenhart... 'bryan.hull@enro...			
'luis.mena@enron... 'brian.hoskins@e... GAME TONIGHT @ 8:45 Eric Bass B			
rian Hoskins, Lu... NULL NULL \Eric_Bass_Dec200...			
Bass-E ebass.nsf Please let me kno... bass-e			
NULL NULL NULL NULL			

6. Linking Neo4j to the Underlying Datastore

```
In [ ]: # connect to neo4j
uri = "bolt://localhost:7687"
neo4j_user = "neo4j"
neo4j_password = "password"
driver = GraphDatabase.driver(uri, auth=("neo4j", "password"))

# write data to neo4j
df.write \
    .format("org.neo4j.spark.DataSource") \
    .option("url", uri) \
    .option("authentication.basic.username", neo4j_user) \
    .option("authentication.basic.password", neo4j_password) \
    .option("labels", ":Enron") \
    .option("node.keys", "id") \
    .mode("Overwrite") \
    .save()
```

CONCLUSION	In this experiment we learned how to use Hive to store and query data in a distributed environment. We also learned how to use Spark to load data into Hive tables and perform queries on the data. Finally, we learned how to use Neo4j to run graph queries on our data. This experiment helped us understand the power of big data analysis and how it can be used to derive insights from large datasets.
------------	---