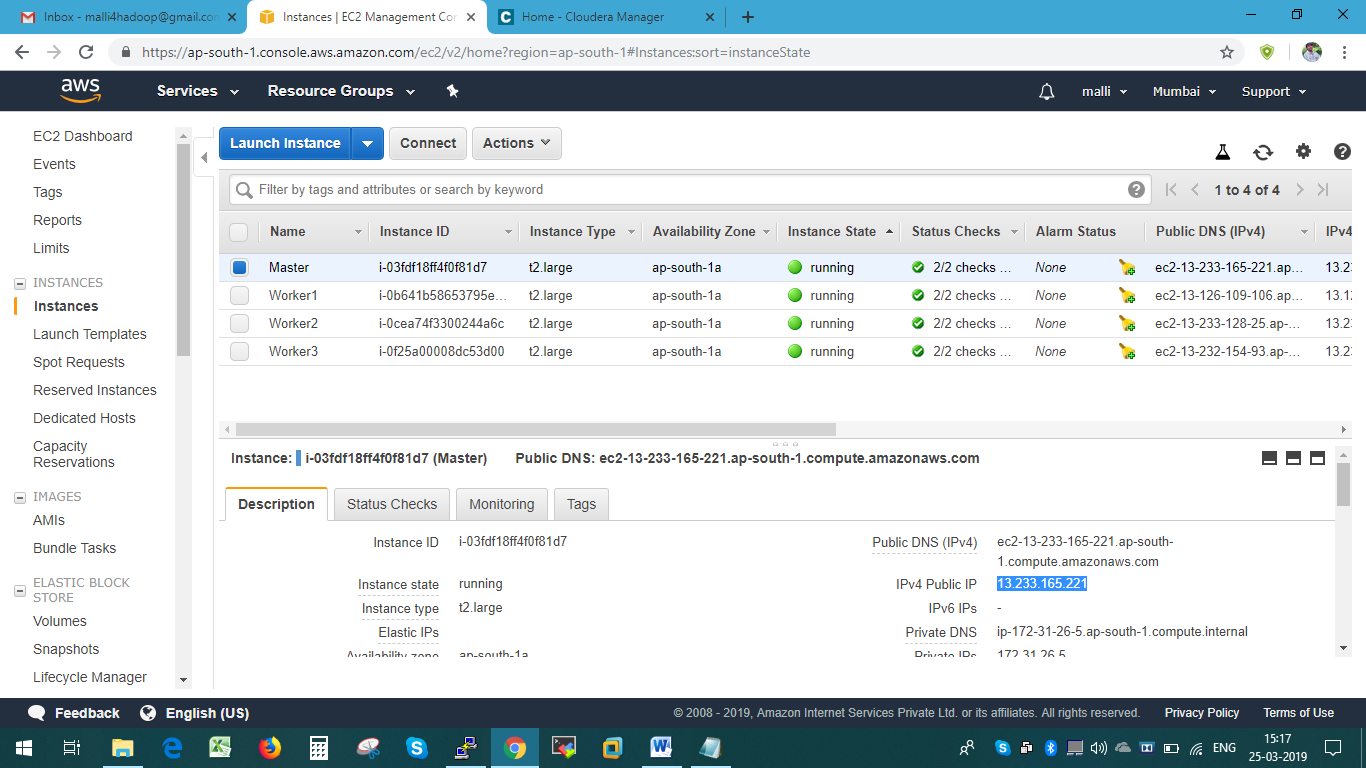
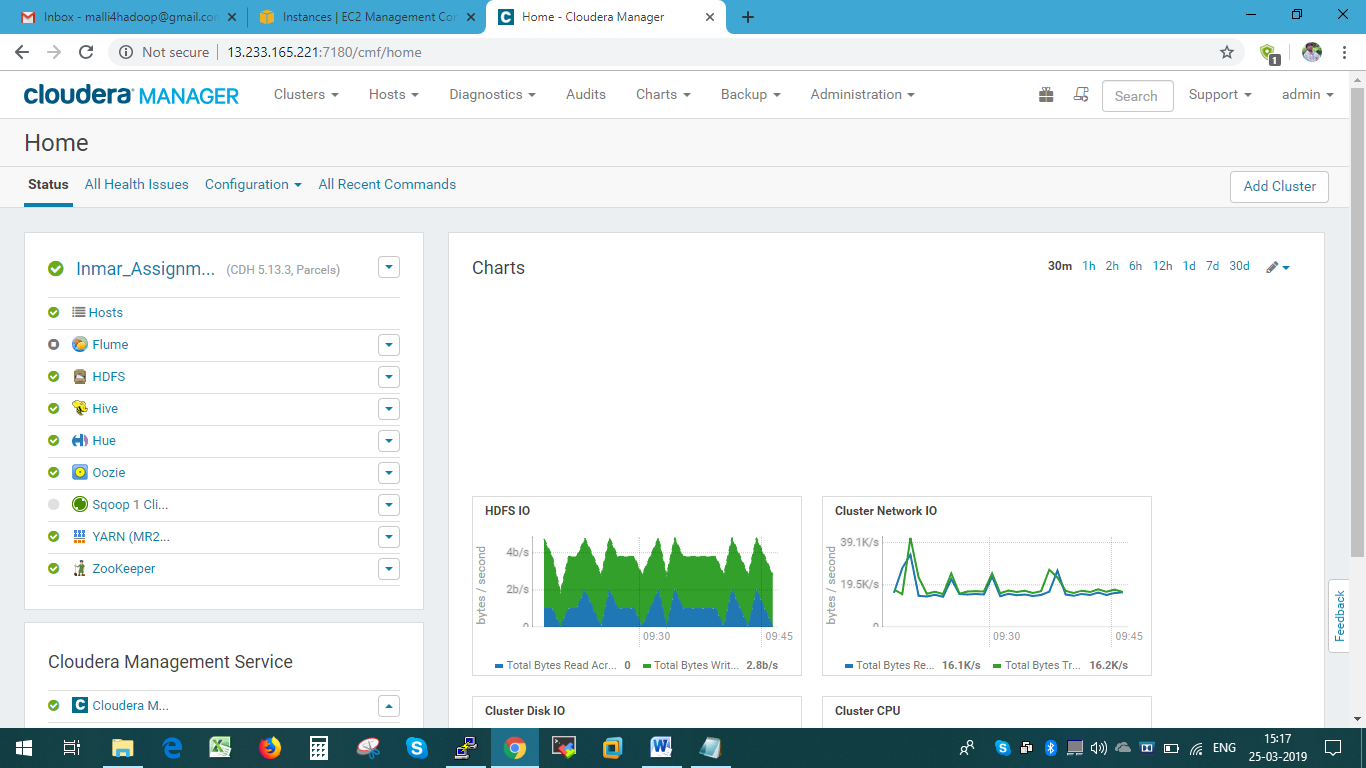
*Hadoop/DBA Technical Exercise*

The following is an exercise to show the ability to create a Hadoop cluster, maneuver and administer it and then a simple exercise to show some data engineering. Please have your cluster available while presenting to us so you can show it in real-time. Be prepared to talk about how you accomplished each task and demonstrate it.

Task 1 – Installation

Set up a 3-node (at least) cluster of the latest Hortonworks Data Platform via Ambari installation. You'll need to evaluate which components are actually required based on the rest of the exercise. Utilize whatever technologies you have at your disposal to do so. It is highly recommended that this is done using Amazon AWS services, but local laptop (Mac, PC, etc) are fine also if you feel comfortable using technologies such as VirtualBox.





Note: I am attaching Installation procedure to evaluate.



Task 2 – Hadoop Confidence Test

The following will demonstrate that you are able to maneuver around different parts of the cluster and operate both within Ambari and without it.

HDFS

Create a folder called consumercomplaints.

Use the following [link](https://s3.inmar.com/dbaexercise/Consumer_Complaints.csv) to download a CSV file from S3 and place it into your new Hadoop cluster in the folder you created above.

hdfs dfs -mkdir -p /consumercomplaints

Change permissions on the file so that it is readable by the world, but only can be updated by the owner.

hdfs dfs -copyFromLocal consumer\_complaints.csv /consumercomplaints

hdfs dfs -chown -R hdfs:hdfs /consumercomplaints/consumer\_complaints.csv

Note: S3 Link is not working, however for demo purpose I downloaded from internet

HIVE

Utilizing the field names provided in the CSV file above,

Create an external table over the file in the consumercomplaints folder called consumercomplaints.

Create another table called consumercomplaints\_partitioned and partition the table by state.

Query your new tables to find the following:

1. Total number of complaints for each company
2. Total number of Disputed vs. Undisputed consumer complaints
3. All consumer complaints centered around Mortgages that were being foreclosed upon from Bank of America

Create a new transactional table called consumercomplaints\_new and insert the data from the consumercomplaints table.

Perform the following:

1. Copy all rows that have state = ‘NC’ or state = ‘TX’ out of the consumercomplaints\_new table into a separate table
2. Delete all rows out of the consumercomplaints\_new table that have state = ‘NC’ or state = ‘TX’
3. Insert only the rows where state = ‘NC’ into the consumercomplaints\_new table from the temporary table above

General Hadoop Administration

1. One of your data nodes has crashed and is not recoverable. Demonstrate how you would decommission the node so that no jobs are run against it.

Goto Host tab > Select All hosts

Select the host, we want to decommission

Click on Hosts Decommission > confirm > Wait for Decommissioning process complete

Check the status in Namenode UI

Select the host & Click on Remove from cluster

Select the host & Click on Remove from cloudera manager

Uninstall "cloudera-manager-agent" and "cloudera-manager- daemons"

service cloudera-scm-agent next\_stop\_hard

service cloudera-scm-agent stop

cleanup all the remnants

yum clean all

Run balancer carefully

Goto HDFS > Instances > click on checkbox of Balancer > click on start

1. Once fully decommissioned, show how you would bring that node back into the cluster and rebalance the cluster.

Provision the host

Execute the pre-requisites

Goto Host tab

Click on "Add new hosts to cluster" button

Use wizard to install the software & then configure the host with the services desired

Configuration changes b/n the role of new host(Master or Slave)

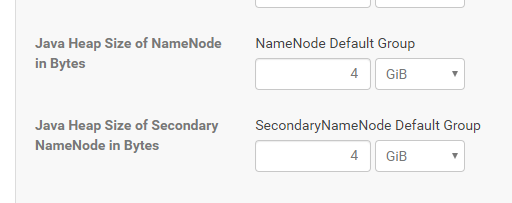
Validate the cluster

Run balancer carefully

Goto HDFS > Instances > click on checkbox of Balancer > click on start

1. Demonstrate changing the namenode java heap size in both Ambari as well as outside of Ambari via the command line.

Cloudera Dash board: Changing namenode java heap size



In CLI: we have to fire following command

+FIREHOSE\_JAVA\_OPTS=-Xms1073741824 -Xmx1073741824 -XX:+HeapDumpOnOutOfMemoryError -XX:HeapDumpPath=/tmp/mgmt\_mgmt-SERVICEMONITOR-958e98291878a4d3675a8438c2d6ad5b\_pid{{PID}}.hprof -XX:OnOutOfMemoryError={{AGENT\_COMMON\_DIR}}/killparent.sh

Install the Falcon service on your management node via Ambari and show that it will start successfully. (Adding Roles Cloudera Navigator & Cloudera Audit in CDM)

Sqoop

Utilizing the following mysql instance, sqoop the tables you deem are relevant in order to complete the below task. Sqoop them into HCatalog tables in a new database in Hive called AdventureWorks.

Database Server: hadooptest.cyrjpmsirks6.us-east-1.rds.amazonaws.com

Login: HadoopTest

Password: CanUSq00p1T?!

Schema: AdventureWorks

Write a query to determine the total sales for each product within ProductModel 'Racing Socks'.

Note : Database Server not working, however I installed sqoop client and its working fine.

I ran sample sample queries as well !!

sqoop list-databases --connect jdbc:mysql://ip-172-31-26-5.ap-south-1.compute.internal --username root --password admin@123

Oozie

Create a basic workflow to run one of the sqoop table jobs above and schedule it to run nightly at 1AM.

Initially I setup Oozie service and its working fine.

By checking that we use following command

oozie admin -oozie http://ip-172-31-26-5.ap-south-1.compute.internal:11000/oozie -status

Showing as Normal



Task 3 - Streaming Architecture w/ Hive

Hive

1. Store the twitter data called sample\_twitter\_data in HDFS. You can use [this link](https://s3.inmar.com/dbaexercise/sample_twitter_data.txt) to download it. The data is in json format and should not be altered.
2. Once the data is in HDFS, create an hcat/hive schema to be able to answer the following question: What are all the tweets by the twitter user "Aimee\_Cottle"? You will need to provide the query that answers this question.

Note: Above link is not working

*Hint: there are multiple ways to do this, the preferred method involves org.apache.hcatalog.data.JsonSerDe - if that doesn't work search for Json serde's in the www - there are some you can compile from source to get it to work*

Streaming (Flume Twitter)

1. Implement a storm topology (<https://hortonworks.com/apache/storm/>) that streams in tweets (<https://dev.twitter.com/streaming/overview>), does some interesting analytics in real-time on the tweets, and then persists into HDFS.

Initially I installed Flume service and its working fine

Update agents like below



Presentation

Prepare to speak to how you tackled this assignment and what you learned as you demo the solution. Assume that the panel includes potential customers evaluating Hadoop. You will be given one hour to present your solution and answer questions from the panel.