**Interview Preparation**

**WORK BASED QUESTIONS**

1. tell me about yourself
2. How did I enter into Hadoop? Undergrad opening in sansco services
3. tell me about your project?
4. what is ur team size?
5. Explain me about the infrastructure? All resources software+HR
6. What is ur role in your team / Explain your involvement in the projects
7. how long have you been working for this company?
8. Who are your clients?
9. What is your current CTC?
10. why do you looking for job change?
11. what do you know about this company?
12. do you have any feedback or question for me?
13. What are your roles and responsibilities?
14. Tell me about ur production cluster?
15. what are your hadoop components / what are the hadoop stack in your cluster
16. how many cluster do you manage?
17. how many nodes of cluster u have worked on?
18. how many projects you have handled
19. Data workflow?
20. capacity planning
21. Tell me about the cluster planning and capacity planning from scratch what are things an admin needs to take into consideration?
22. How will you calculate the size of your hadoop cluster?
23. How is data being stored on HDFS?
24. How is data processed from HDFS?
25. Where is data being dumped into HDFS?
26. Can you give some client requirements in your Project?
27. How you contact your client everyday ?
28. size of ur cluster
29. On what servers Hadoop is installed?
30. What are the minimum file size and maximum file size you have handled?
31. What is the directory of DN & HDFS?
32. Cluster is put into maintenance
33. User Management
34. Monitoring tools
35. Alerting mechanism

DATA INGESTION / MIGRATION [admin is also involved in it]

1. In Hadoop, how we get a data from the client from which channel, the data are on the server, if yes the name of server how we can access the data?
2. how does distcp work?
3. What is the amount of data you receive on daily basis?
4. Is WINSCP is capable of transferring 4GB data at a time?
5. What are the key challenges you faced while importing data in Sqoop?
6. how will i copy a file without changing its user n superuser and with same properties to another destination n without using distcp.
7. From how many input sources you are getting data?
8. Tell me file naming scheme?
9. In your project ur using Sqoop, Flume

HIVE

1. if hive comes in where will it rest
2. what are you using for hive-shell
3. which connector for hive
4. what is the hive port number
5. which DB are you using?how many option do we have for RDBMS DB & which ones have you worked on?Do you have mySQL database's HA?
6. my HMS went down then what will happen to my data?
7. Can we setup HA for hive server? explain us the process
8. What about HMS tell us about that too?
9. I have 10 HS2 and 2 HMS and i have configured 1st 5 HS2 pointing HMS1 and rest 5 HS2 pointing to HMS2, so can my HS2 servers cross communicate with other HMS?
10. if data comes from hive where will data rest
11. Advantages of Columnar data stores over Row-oriented data stores.
12. How can you implement Updates in Hive?
13. How to increase hive performance?
14. What are managed tables and external tables in hive?
15. Can we change the default location of Managed tables?
16. What are the factors that we consider while creating a hive table?
17. Tell me 15 table names?
18. Can you suggest some sample tables used?
19. Tell me the number of tables you have handled?
20. Tell me the number of columns you have handled?
21. How did you choose the columns for partitioning?
22. Partitioned data is divided into how many buckets.
23. On which column you have applied Partitioning and Bucketing?
24. How did you conclude on the optimal bucket size?

PERFORMANCE TUNING

1. Were you involved in performance tuning? How did you do that?

COMPRESSION

1. Can you explain compression techniques in Hadoop/Which technique did you use?
2. How does ORC implement compression, what encoding techniques does it use?
3. What are the compression techniques and how do you decide which one to use?

YARN & JOB SCHEDULING [tuning running running ur own jobs and

benchmarking]

1. YARN architecture
2. Explain Yarn Child process in Yarn Architecture n how it works
3. What is a container? (JVM process)
4. Who creates containers? RM/ AM/ NM?
5. Why Container will fail?
6. If i have 8 DataNodes , how many Application Master & Application Manager will be there?
7. Where will the subsequent containers be allocated once AM's container is spawned?
8. How many type of scheduler
9. Which scheduler is being used in your cluster?
10. Among the 3 schedulers which one is the best?
11. What is FIFO scheduler?
12. Tell me about DRF(Dynamic resource fairness) scheduler? will DRF dominantly take allocated cpu or memory or both from other que?
13. Explain about Capacity scheduler?
14. How do we change the number of mapper and reducers from CLI and GUI in runtime?
15. What will be reason if Yarn Job Stuck
16. Have you performed yarn tuning
17. Explain me the 4 step procedure which needs to be performed at the time of long running job (not sure about the exact question here i couldn't understand it fully but manager did ask about the 4 commands in that context over and over)
18. what should be done before killing a Job?
19. can i kill the specific process?
20. if the job is running and job with a high priority comes then what should we do?
21. MR1 and MR2 difference and working
22. what is the command to locate the directory of intermediate data
23. What is mapper counter
24. What is cmd to see mapper counter
25. i have job 00of 99 mapers and 1 reducer how do you optimize the performance?
26. Where we can see in Web UI mapper counter
27. Data spilling map reduce
28. what is preemption? have you set the preemptions
29. what are the parameters that needs to be tweaked to fine-grain the performance?
30. How can i list the zombie process from CLI?
31. can you schedule dynamic pool allocation? how explain?
32. What is bloom filter, how is it different from distributed cache, how do you implement it in MR?

BACKUP AND RECOVERY [must know metadata,nn,dn,disk backup]

1. What do u take the backup of your data?
2. how do you take backup and how often
3. What is BDR and how to take backup using BDR? Steps? can you schedule BDR for later time? how?
4. i have prod kerberized cluster and non-kerberized DR cluster, can i schedule backup?
5. where is your prod cluster and DR cluster located?
6. my Prod cluster has 50 nodes and we added another 20 nodes similarly DR cluster has 50 nodes i have added another 20 nodes will i be able perfom BDR?

**HADOOP**

1. what is port ntp listen on?
2. are your clusters behind firewalls?
3. what is purpose of firewalls why its useful?
4. what do you mean by checksum?
5. What is rack awareness?
6. have you worked on vanila hadoop?
7. what are the configuration files?
8. Explain about HDFS architecture?
9. what is the core feature of hadoop over other systems?
10. Can we implement raid in HDFS? Datanode as well as Namenode
11. What will happen if my standby namenode goes down? will active namenode fail or it'll continue to work?
12. what are your static pool allocation values and ques
13. how would i ascertain which block is been picked for processing
14. What is httpfs?
15. How Checkpointing works without Secondary Name node in HA.
16. In HA how Checkpointing happens Standby Namenode

**ADMIN TASK**

1. decommissioning
2. What happens to data when we decommission a node where it goes
3. Commissioning [testing and benchmarking]
4. copy data from one node to another? Can we use scp and how?
5. copy from one cluster to another how ll u do it? Can we use scp here?
6. what is cache? How u will check cache.
7. Trash Config

**HIGH AVAILABILITY**

1. HA Environment
2. Zookeeper on separate node
3. How Zookeeper works?
4. Role of Journal Quorum Manager
5. can i manage the cluster without zookeeper services
6. what is zkfc
7. How Automatic Failover
8. why there's a need for zookeeper in hadoop cluster what difficulties will i have to face in that situation
9. Epoc Fencing how it works
10. Types of znode
11. Which type of znode is used for zookeeper election
12. How zookeeper leader and follower election works
13. Explain Resource Manager HA n how it works
14. If active NN and standby NN both goes down, what could be the reason,how to resolve it?

**UPGRADE**

1. Types of upgrade?
2. Pre-conditions for cluster upgrade
3. CM upgrade
4. CDH upgrade
5. Version previous to current (CM/ CDH)
6. All services/ roles should be running or stopped? Which services /roles should be stopped?
7. Do we need to upgrade JDK or is it optional?
8. What mandates the upgradation of JDK?
9. What kind of issues will arise if JDK not upgraded?
10. After JDK upgrade is important to restart the Cloudera Manager Server & Agent?
11. How will you install Cloudera Manager Server & Agents?
12. Which files & properties needs to updated?
13. How to verify if the upgrade has succeeded or not?
14. What is host inspector?
15. How to restart the services once the upgrade is completed?
16. Is it mandatory to upgrade CDH post CM upgrade?
17. If we've TLS/SSL configured in our cluster then which properties need to be edited?
18. On which node Cloudera Manager Server runs? (on CM host)
19. What if upgrade fails?
20. After upgrade if some of keytabs are missing then how to resolve?

**AWS**

1. object storage and block storage
2. Type of Instance.
3. Do u know AWS Cloud S3,VPC,Glaciers
4. emr and cloudera difference
5. what is IAM in cloud AWS
6. Assume i am your client and convince me to use AWS Cloud

**LINUX**

1. how would you rate yourself in linux out of 10?
2. How do you list ONLY directories in Linux ?
3. How do you list files excluding your name from a directory ?
4. Hadoop command to see disk usage ?
5. Commands to copy local file to HDFS ?
6. What is the command to check up utilization?
7. Command to check OS version?
8. what is fstab
9. How will check running services from command line?
10. How will you check ram, cpu, storage?
11. How will you check filesystem?
12. How will you check a process ID of a service?
13. What if a lot of memory is used in swappiness?
14. How will you check swappiness memory?
15. How will drop swappiness memory?
16. Roles and responsibilities as Linux Admin
17. How can you directly adopt Hadoop on a Linux environment?

**SECURITY**

1. what are the security measure you have taken to secure your cluster

AUTHENTICATION

1. Working of Kerberos?
2. What is kerberization
3. how many KDC servers do you have?
4. Command ktutil used for?
5. Kadmin used for?
6. what is command to install kdc server?
7. is your KDC server in AD, do you have control over AD?
8. do you HA of AD?
9. why hadoop uses kerberos and not other mechanism for security
10. how to create principal and keytab?
11. can we create keytab with kadmin.local command?
12. how will see list of d principals
13. what is keytab. What is importance of keytab.
14. what is the command to create keytab
15. how can i find out SPN and keytabs given the situation there's no CM access to me neither i have running server access i.e. /var/run/process
16. how u will check user having valid kerberos ticket or not? Command?
17. if there is long running job (15-30 days) is this necessary to renew kerberos ticket,.... yes then how? (ans hint - delegation token)
18. How can i renew my ticket lifetime?
19. what is KTS?
20. what is kts used for?
21. how to connect AD with cluster

AUTHORIZATION

1. if i want to give access of one file to another user how will i do it
2. what is sentry why its is used?
3. Sentry is used at user level or groups level?
4. How will you check which roles is assigned to which group?
5. what is the command to create role?Write the full command to create sentry roles.
6. please explain the complete process of creating and granting permission?
7. do you assign permission to group or user?
8. what is the command to check roles
9. what is the command to check grants
10. i have user who is complaining for not having access to table, how will you troubleshoot that?
11. Explain Sentry Policies, and How Masking on Tables Permission works

ENCRYPTION

1. What are the encryption?
2. tell me about KMS how it works
3. what key trustee server and KMS? explain the difference?
4. how data at rest encryption is taken cared in cloudera?
5. how does user write to encrypted zone?
6. who create encryption zone user or admin?
7. where the encryption keys are stored?
8. how would user read his/her encrypted data which is stored in encryption zone?
9. what is the command to create encrypted zone?
10. who provides the encryption key to user?
11. how do we move data from EZ to another location
12. where does EZK and EDEK are stored?
13. why do we need to skip checksum while moving data from EZ?

TLS/SSL

1. do you have SSL implemented on your cluster
2. what is SSL is used for

**ISSUES**

1. Issues & how did you resolve them?
2. What are technical challenges you have faced?
3. What are the key challenges you faced and how did you resolve
4. On which condition should we have to stop the service
5. What will happen if “*hdfs dsf –ls /”* this cmd stuck what will be the reason
6. if data comes from hive where will data rest
7. user unable to connect to hive server, how would be trouble shoot this issue.
8. what is the most recent issue you have resolved on your cluster?
9. how often do you seek cloudera help and raise ticket to them?

**ACHIEVEMENTS**

1. What is your biggest accomplishment?
2. What are your key achievements?
3. Which technologies your strength and rate yourself?

**CLOUDERA DIRECT QUESTION**

1. How to stop Cloudera Manager Server, Database & Agent? (sudo service cloudera-scm-agent start)
2. Why should we need to stop these? Because we've to upgrade Cloudera Manager
3. cm ports
4. Why Oozie and Hue on edge node? What if my edge server is down then what about the services running on that server
5. are you using any scripted alert mechanism?
6. How will you check which stack is there if don't have webUI of CM?
7. Name the service which has web UI in cloudera?
8. Difference between Parcel and Packages

**WAY TO AVOID ANSWERS**

1. Do u know ranger,kafka
2. They will ask about hortonworks
3. have you worked with ansible?
4. tell something about kafka
5. tell me something about hbase
6. have you worked with spark how does it works
7. shell scripting ? have you written any script to automate your work?
8. explain about the script that you have written and how it benefited you?
9. do you know about spark have you worked on it? do you think spark can replace mapreduce? Why

**SITUATION BASED QUESTIONS**

1. Suppose CM is down,it will take 24 hour for troubleshooting how will you monitor.
2. Suppose a disk is down out of 10 what will you do.
3. What is my namenode is down and standby namenode is also not coming up, what can b the issue?
4. Suppose one day u start getting tremendous alerts and oozie jobs are failing in counts of hundred, what will u do acc. to ur responsibility?
5. Due to over commit of RAM system is going down again and again?

TROUBLESHOOTING

1. <https://github.com/wbsg/ldif/wiki/Hadoop-troubleshooting>
2. <http://www.cs.brandeis.edu//~cs147a/lab/hadoop-troubleshooting/>