DVS Real Time Interview Questions

1. What is your Cluster Configuration?

Production Cluster Configuration	
How many nodes cluster it is?	**32** **For Hadoop 1X:** • Name Nodes : 1 • Secondary Name Nodes : 1 • Data Nodes : 30 **For Hadoop 2X:** • Name Nodes : 2 (NN +Standby NN) • Secondary Name Nodes : 1 • Data Nodes : 29
How many JT in cluster?	1 (Running on NN itself or separate machine)
What is NN configuration in cluster?	Hex Core System (16 Cores in one machine) 1 X 2 TB HDD, 16 X 16 GB RAM individual **Total Capacity: 2 TB HDD, 256 GB RAM**
What is DN configuration in cluster?	Hex Core System (16 Cores in one machine) 16 X 2 TB HDD, 16 X 4 GB RAM individual **Total Capacity: 32 TB HDD, 64 GB RAM**

What is SNN configuration in cluster?	Quad Core System (4 Cores in one machine) 1 X 2 TB HDD, 4 X 16 GB RAM individual **Total Capacity: 2 TB HDD, 64 GB RAM**
What is your total production data size?	2PB
UAT Cluster Configuration	
How many nodes cluster it is?	8 **For Hadoop 1X:** • Name Nodes : 1 • Secondary Name Nodes : 1 • Data Nodes : 6 **For Hadoop 2X:** • Name Nodes : 2 (NN +Standby NN) • Secondary Name Nodes : 1 • Data Nodes : 5
What is NN configuration in cluster?	Quad Core System (4 Cores in one machine) 1 X 1 TB HDD, 4 X 16 GB RAM individual **Total Capacity: 1 TB HDD, 64 GB RAM**
What is DN configuration in cluster?	Quad Core System (4 Cores in one machine) 4 X 2 TB HDD, 4 X 4 GB RAM individual

	Total Capacity: 8 TB HDD, 16 GB RAM
What is your total UAT data size?	15 TB (this will be 45TB after replication)
What is your UAT cluster size?	Around 50 TB (This includes UAT data + some free space)

2. Other Real-time Questions?

How you connect to your cluster?	Putty+ LDAP
How do you take back-up of your cluster?	• Distcp • Copy "Data Store" folder to backup cluster
How much data you receive on daily basis?	1.5 – 2 GB
How you get your input data?	Flat Files & different RDBMS systems
What is your input data format?	|Pipe, CSV
What will be your input data file size?	200 - 500 MB
How do you test your code?	Unit Testing & Testers even do Load & performance testing
How do you handle data updates or deletes?	We can handle in Hbase, However DML is supported after hive-0.14 onwards
How do you receive	Business Analyst will provide

requirements from client?	requirements. (I only handle this task in my project)
What LOBs you support?	Workers' Compensation, Auto Liability, General Liability, Disability, Managed Care, Professional Liability, Warranty and Credit Card Claims, fraud & investigation, Medicare compliance solutions

3. What is your Hadoop distribution, HIVE & PIG version?

   CDH 5.3 (Stable)

   Hadoop 2.5.0
   Pig 0.13.0
   HBase 0.98.6
   Sqoop 1.4.5
   Hive 0.13.0

   Latest is 5.5
   Hadoop-2.7
   Hive-1.2.1
   Hbase 1.0
   Pig-0.15.0
   Sqoop 1.4.6

4. Do we have VIEW concept in HIVE?

   Yes hive 7 is not having but 9 on words,we have the views

5. Do we have Primary Key, Foreign Key concepts in HIVE?

   No constraints in hive

6. How to improve HIVE query performance?
   Using Partitions, Buckets, map join, no: reducers, skewed by, instead of order by we can go for distribute by & sort by etc.,
7. How frequently NN transfers metadata to SNN?
   Check point for every 60 min. This can be set to any value like 30 min, 15 min or 1, 2, 4 hrs.
8. Can we set replication factor by file?
   Yes by file level replication
9. Can we set block size by file?
   Yes .. By file level replication
10. Can we set heartbeat interval time?
    Yes configuration file hdfs-site.xml and the property name is dfs.heartbeat.interval
11. Drawback with 1.X & how to overcome ?
    NN is SPOF & needs Active & passive NN concept
12. Difference between Edit-Log and FS-Image?
    Edit Log –current ….FSI Image previous
13. Do we have replication concept for SNN?
    NO
14. How do you do Process Automation?
    Oozie, crontab, scheduling tools like autosys, control M etc.,
15. What is Thrift server?

Thrift server provides JDBC/ODBC interface between JAVA application and HIVE.

16. How to retrieve the metadata if we change the database from MySQL to any other?

Mysql backup

17. Major properties in hive-site.xml?

- Hive.metastore.dir=/user/hive/warehouse
- set hive.exec.dynamic.partition.mode=nonstrict;
- set hive.exec.max.dynamic.partitions.pernode=300;
- set hive.enforce.bucketing = true;

20. Maximum velocity can be achieved by Hadoop?

4.8 TB/sec

Calculation:

1 HD = 100 MB/sec

1 Server = 12 HD = 12 * 100 = 1200 MB/Sec

1 Rack = 20 Servers = 20 * 1.2 GB = 24 GB/Sec

Average Cluster = 6 Racks = 6 * 24 = 144 GB/Sec

Large Cluster – 200 Racks = 200 * 24 = 4800 GB/Sec = 4.8 TB/Sec

18. What happens if a empty file is copied to the cluster. Where does the file gets stored ?

Metadata gets stored in Name Node & no entry in Data Nodes

19. What if the warehouse path gets changed in hive ?
What wil happen to previous tables & databases?
New Databases & Tables data will be moved to new warehouse path & old tables data still point to the same old warehouse path
20. When do we go for HBASE Managed Hive tables ?
21. Can we sqoop data to hdfs or hive ?
Both
22. Can do incremental updates using sqoop A: yes using - -append command
23. What is parameter file in Sqoop ?A : All static parameters we keep in this & call when required.
24. How many reducers will work when a sqoop job runs ?
No reducer when a sqoop job runs. Only map task because importing data to hadoop as it is.
25. If we don't specify target directory during sqooping, where the data gets stored by default ?
/user/<username> i.e., /user/training
26. Where does the table gets stored by default ?
In default database
27. Types of TCL commands in HIVE ? A : No TCL in HIVE
28. Advantages of Using Pig ?

i) Pig can be treated as a higher level language a) Increases Programming Productivity b) Decreases duplication of Effort c) Opens the M/R Programming system to more users

29. What is tokenize & what is the default delimiter ?
30. What command in pig is used to process each tuple in PIG ?

FOREACH

31. Find the highest occurred word in the given file ?Explain the sequence of steps involved
32. How do we pass the parameters in Hive &pig ?
33. How can we do incremental load in hadoop from RDBMS ?
34. What is Difference Between Mapreduce and Pig ?

•In MR Need to write entire logic for operations like join,group,filter,sumetc ..

•In Pig Bulit in functions are available

•In MR Number of lines of code required is too much even for a simple functionality

•In Pig 10 lines of pig latin equal to 200 lines of java

•In MR Time of effort in coding is high

•In Pig What took 4hrs to write in java took 15 mins in pig latin (approx)

•In MRLess productivity

•In PIG High Productivity

35. What is the difference in run and exec commands in PIG ?

used for executing a pig script from grunt shell. Both r same.

36. What is the difference between partition and bucketing ?

Partitions are for segregating the data based on particular columns and used to avoid full table scan. Buckets on the other used for join optimization & table sampling. We decide & control no:of buckets but not partitions

37. Whatis data localisation ?

Executing the logic wherever data is located.

38. What are virtual columns in hive ?

INPUT__FILE__NAME & BLOCK__OFFSET__INSIDE__FILE

One is INPUT__FILE__NAME, which is the input file's name for a mapper task.

the other is BLOCK__OFFSET__INSIDE__FILE, which is the current block's first byte's file offset.

39. What is speculative execution in Hadoop?

If user committed a job , If task tracker running slow(or down at that point of time) task tracker will create same mapper task on different data node with is having same data block(backup copy). I willwait till the output from any one of the mapper once its get the mapper output it will the job on second data node

40. What is a combiner ?

It's an intermediate reducer that can aggregate the mapper output. Ex: if you have generate the 1 cr (key,value) pair out of 60lks are duplicate then the transfer(mapper to reducer) will take much time, instead of that we create reducer and will aggregate the results at mapper level then we can transfer only 40lk of (key,value) pair.

41. How to kill hadoopjob ?

Hadoop job –kill <job_id>

42. How to see list of all jobs in hadoop through CLI ?

hadoop job –list

43. How to see the free space in HDFS ?

hadoop fs –df –h or hadoopdfsadmin –report

hadoopdistcp -f hdfs://nn1:8020/dvs_hdfs/emp.csv hdfs://nn2:8020/dvs_hdfs1

44. what is the flatten in pig?

Un-nests the fields in a tuple

45. How to register UDF in hive and PIG ?

Add jar &register(explain the commands)

46. How to change the replication factor for a existing file in hdfs ?

usingsetrep command.

47. Different between sort by and order by ?

order by is global sort by sorts data in each reducer

48. what is dfs.namenode.name.dir and dfs.datanode.data.dir ?

Belongs to admin but it is the place where NN stores Metadata and DN stores the blocks respectively.

49. Does hive supports updates ?

Yes from hive 0.14.0 onwards

50. What is msckcommad in hive ?

msck repair table <table_name> checks all the table meta data in meta store DB & verifies in hdfs table directories & partition directories. If mismatch, it adds the meta data for the newly added directories.

51. What is Intermediate data in Map Reduce & where it gets stored ?

Mapper o/p & stored in the DNs

52. What is a Outer Bag & Inner Bag ?

Outer Bag is main bag & Inner Bag is a subset of that.

53. How to Add/rename/exchange/drop/repair a partition ?

```
alter table std1 add partition(yop=2011);
alter table std1 exchange partition(yop=2011) with table std;
alter table std1 partition(yop=2011) rename to partition(yop=2013);
alter table drop partition (yop=2013);
msck repair tblestd;
```

54. What is skewed table & difference between Partition table & skewed table.

When there are more partitions for a table & if the data in the partitions is unevenly distributed like few partitions contain huge data and many partitions contain only few records, partitions will kill the performance wherein very few records are pushed to a partition. The more mappers will be invoked & more intermediate files will be created thereby increasing the no: of partitions. This is the concept of skewed table will come. In this concept, we can specify particular column values which are having huge data & make sure, the partitions are created only for these values & all the other data is pushed in to a single partition. In this case, no: partitions are reduced, no: of mappers are reduced, no: of intermediate files are reduced.

55. What are Table & DB Properties in hive?

Hope we have discussed in the class in detail.

56. Optimization techniques in hive & pig

Wil discuss in class

57. How to pass parameters in hive & pig

Discussed in class

58. How to skip headers in the file & load

59. Default delimiter in pig

60. Default delimiter in hive