Q)How to answer for the question what is the size of your cluster?

ANS

Size of the cluster,

80 PROD

48 STAGE

24 DEV

80 nodes

48 nodes

24 nodes

Cluster is composed of nodes not TBs

Q)What about the size of ur data ?

Size of the data -

1 GB transactional file

200/300 Master data file

daily around 15 files to process

Sqooped data,

SFTP server data files.

It's just your communication that matters ans how you present yourself

**Oh..and we dont use flume?**

Nope, You use Kafka, NiFi

or

The files are shared by users in a shared folder, you download and load them into HDFS

**So..i have one que the cluster is setup at user side na..**

**We login to user environment by vpn**

You use VPN to connect to Your client's network where your Servers are located.

You use Putty to connect to the hadoop cluster edge node.

Alternately, HUE access is given to you.

Admin creates your user, You are given access to particular databases in HIVE, particular folders in HDFS,

You are expected to work as told by your Manager a.k.a. BigData Architecture.

**So...we dont deploy hivr scripts or jar files**

**These work is done by admins**

You just don't go on spree to create any Random tables, databases or scripts.

Your PIG statements run in Script, Hive script.

You or other team member creates either Oozie job or writes a shell script that calls all the jar registration, script execution.

**What about the daily task**

1. Get files,

2. Verify the name of the file matches the format is requires to be in.

3. Load the files in HDFS at specified location.

4. Launch Oozie coordinator job.

Alternative, run shell script.

5. Verify the output.

6. Note the time it took.

7. Get the count of records.

8. Verify the count.

9. Run additional scripts via Oozie or shell script to apply transformations and load data into ODS database.

10. Perform testing.

11. Testing passes, load into DATAMART database.

12. Let the Tableau team know.

13. meeting with team, let them know what you did, the time it took, any tips for performance improvement.

14. Listen to them.

15. Confirm with the team if everything went smooth

13, 14, 15 are for every thing in IT sector, Sorry a

**Great..bro..**

**For data ingestion we dont do autmation like scripts or oozie jobs which get data daily..?**

Your data ingestion happens automatically with scripts.

People choose data ingestion as a single task. It's mostly not included in Oozie jobs.

Ingestion is done and then for all your business logic to be applied, You have scripts, queries written which are then ran using Oozie.

**Data ingestion with which scripts ? Unix shell**

SFTP to Unix , shell command

Unix to HDFS, shell command.

some use Kafka for first step

[13:32, 8/25/2017] +91 99107 74870: These for unstructured files..

[13:32, 8/25/2017] +91 99107 74870: For structured rdbms

[13:32, 8/25/2017] +91 99107 74870: We use sqoop

[13:32, 8/25/2017] +91 99107 74870: Right

[13:33, 8/25/2017] +91 72076 59780: Yes ! You are smart 😊

[13:33, 8/25/2017] +91 72076 59780: If your data is coming from RDBMS, Sqoop

**Which one is used**

**Pig and hive?**

People use HIVE and PIG.

Imagine this,

Your data is in database, there are business rules applied on it.

Now you need to use Hadoop, You take your data using Sqoop as is. The business rules are required to be applied on the data in hadoop.

You choose HIVE to be used. So for business rules, you write UDFs in HIVE. UDFs are written in Java

**What business use cases are written in hive UDf?**

Which RDBMS do.you use?

If you are using Sql Server, you have SPs that you may need to write in Java.

If you are using Oracle then you have Functions, Procedures as business rules that will be required to write in Java

may be calculation, may be lookup conditioning

Usually this part is covered wrt calculations. If you have binary column to handle that you may write your udf.

**And what is the archi of ur project how u get data..what automation u have done and all**

As a developer, you should focus on explaining Development related explanation.

Have a fair understanding of everything.

While explaining Admin parts, don't say you did it.

If the interviewer is experienced Hadoop guy, He might sense you and shoe you out.

For Developers, concentrate on Development explanation.

Don't explain him how use are using Zookeeper, sentry and all.

**So..these things which u told me also include archi que na?**

Yes.

Architecture explanation simply means

1. From where the data is coming.

2. How are you loading it into HDFS

3. What tools are you using after data loading

4. Are you pre-processing the data

5. What tool are you using for pre-processing

6. what is next step.

7. Are you using HIVE

8. If yes, How many databases in total

8. Where the pre-processing happening nad which database is getting populated with data after pre-processing

9. Where are you applying your transformation.

10. which database gets populated after transformation are done.

11. Which is PROD database.

12. Size of the datanodes

13. Occupied space till date.

**[14:01, 8/25/2017] +91 99107 74870: One que more**

**[14:01, 8/25/2017] +91 99107 74870: In projects do we use agile?**

**[14:01, 8/25/2017] +91 99107 74870: Like in hadoop**

**[14:01, 8/25/2017] +91 99107 74870: Bcz in my project**

**[14:01, 8/25/2017] +91 99107 74870: We use agile**

**[14:02, 8/25/2017] +91 99107 74870: We have a sprint for 1 mth in which all dev testing is done**

**[14:02, 8/25/2017] +91 99107 74870: And delivered to the client**

**[14:03, 8/25/2017] +91 99107 74870: So for hadoop if someone needs insights of 2 days data**

**[14:03, 8/25/2017] +91 99107 74870: Is that possible**

**[14:03, 8/25/2017] +91 99107 74870: For do all the testing dev in 2 days**

Your Hadoop project is no different than any project in this sector. It goes through "Development-Developer testing-PreUAT-Performance Improvement-UAT-PROD

And testing time depends in the things/requirements you will be testing against.

It may take 1 week of testing, 2 weeks or 3 days.

It's the data, the requirements that Matter.

You need to actually understand client's requirements.

One of the projects we had was.

1. Get data from Oracle database to HDFS.

2. Write UDFs to meet business logics.

3. Load data into HIVE, process it there.

4. Connect Hive/Impala using JDBC to generate reports out of it.

5. Load the data back to Oracle.

6. Connect Oracle to Front end JSP page.

7. User queries Something there and data is fetched from Oracle.

8. It was faster until data started increasing.

9. We used Apache Solr as search engine behind JSP page.

10. Connect Solr to Oracle and fetch data into it.

11. Display the data in JSP page.

12. Solr helped us speed up the process.

13. User could download and view the data at much faster speed than that when we were using Oracle as direct connection to JSP.

**Hive udf**

Create a file where you have records with one column as binary data. You can generate such file using a simple java program. Load in hive table. Lets say binary data is your credit card number. Assuming first 4 digits say if it master or visa write a query to fetch number of users of visa and master card. Here you may want to wtite a udf to fetch first 4 digits and then do the do.