1. while working on batch migration project we developed our scripts for the small datasets i.e. sample data in our development cluster. you designed your algorithm by handling all the scenarios that can occur in sample data but the moment your logic ran over million of records there might be a chance of few scenarios which you dint handle and resulting into your job failure. This runtime error could be some garbage value or may be some special character anything could make your job vulnerable. **Advice Never rely on sample data for developing your logic or model.**before sending it to validation always run it with actual data.
2. Another issue again the sample data set versus the bigdata. **Heap space error**though it is something related to infra but anyhow it was production issue. heap space error generally occur when your JVM runs out of memory then you have to allocate more resources to your JVM. It used to happen few production jobs get failed due to this error
3. **Sqoop**: Job getting failed due to database server connection timeout and low disk space issues. Most of the time partial loading due to job failures

# Sqoop incremental load using "lastmodified" fails when going against MSSQL

You can obviously just use a manual filter ( WHERE condition ) in Sqoop. For example run sqoop hourly in oozie and get the past hour of data

1. If we get java heap space error and we have already given the maximum memory, what is the possible solution?

A) increase mappers by -m 100

1. How can we resolve a Communications Link Failure when connecting to MySQL?

Verify that we can connect to the database from the node where we are running Sqoop: $ mysql –host= –database=test –user= –password= Add the network port for the server to your my.cnf file. Set up a user account to connect via Sqoop. Grant permissions to the user to access the

database over the network:

Log into MySQL as root mysql -u root -p

Issue the following command: mysql> grant all privileges on \*.\* to ‘user’@’%’ identified by ‘testpassword’

mysql> grant all privileges on \*.\* to ‘user’@” identified by ‘testpassword’

1. hive (madhu\_db)> insert overwrite table part\_part\_t partition(category,txndate)

> select t.txnno,t.custno,t.amount,t.product,t.city,t.state,t.spendby,t.category,t.txndate

> from transactions t distribute by (category,txndate);

this query will give error of set max directory per node 200

Solution : We have to set hive.exec.max.dynamic.partitions.pernode=500 then it will work

set hive.exec.max.dynamic.partitions.pernode=500;

2.Oracle database and sqoop

1. oraclejar permission denied in sqoop lib

used chmod 777 to oraclejar

2. No commit in sql

error : all map reduce process run properly but 0 record retrieved

Details of Software /System life cycle fro hadoop projects

1. Sqoop – Incremental vs last modified relate to your project
2. What are the challenges you’ve faced in your project? Give 2 examples.
3. How do you check Data Integrity (log files)
4. Tell me about your project? work.
5. How do you use Partitioning/Bucketing in your project? (Examples from your project)
6. How to debug Production issue?      Give example. (logs, script counters, JVM)
7. What is the file size you’ve used?
8. What is the
9. file size for production environment?
10. How long does it take to run your script in Production cluster?
11. Are you planning for anything to improve the performance?
12. What size of file do you use for Development?
13. What scenario do you think you can use Java for?
14. Hive partitioning – your project example? Why?
15. Hive – What file format do you use in your work? (Avro, Parquet, Sequence file)
16. Hadoop – What is the challenge or difficulty you’ve faced?

**Project Devlopment Life Cycle**

CMMI is very helpful!

•If abstraction is needed, use modified Fibonacci for Agile software projects

**Quality : Quality is easier, but expertise is required to identify requirements**

•Data validation

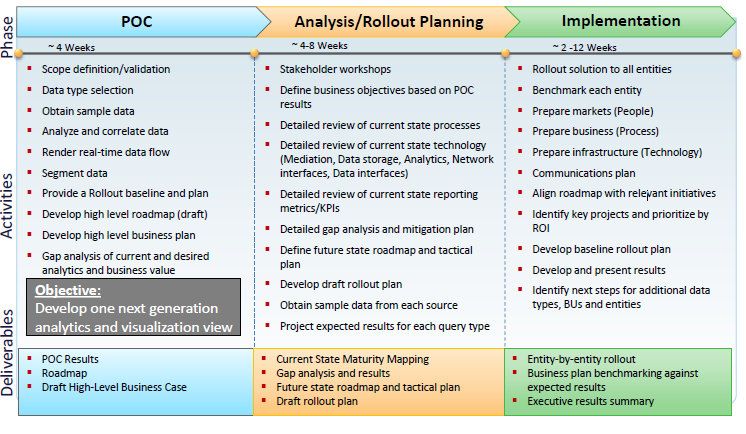
•Data reconciliation and deduplication

•All-or-nothing transactions

•Data privacy

•ACID and BASE

•Dedicated nodes



14. While importing tables from Oracle database, Sometimes I am getting java.lang.IllegalArgumentException: Attempted to generate class with no columns! or NullPointerException what might be the root cause and fix for this error scenario?

While dealing with Oracle database from Sqoop, Case sensitivity of table names and user names matters highly. Most probably by specifying these two values in UPPER case will solve the issue unless actual names are mixed with Lower/Upper cases. If these are mixed, then we need to provide them within double quotes.

In case, the source table is created under different user namespace, then we need to provide table name as USERNAME.TABLENAME as shown below.

sqoop import \

--connect jdbc:oracle:thin:@oracle.example.com/ORACLE \

--username SQOOP \

--password sqoop \

--table SIVA.EMPLOYEES

5. How to access sub directories recursively in Hive queries?  
  
To process directories recursively in Hive, we need to set below two commands in hive session. These two parameters work in conjunction.

hive> Set mapred.input.dir.recursive=true;  
hive> Set hive.mapred.supports.subdirectories=true