**Production Issues: In beginning almost every time!**

It was almost every time production deployment and running a job on production getting failed or killed or some random error. But indeed it was a good learning.

I am jotting down few issues and trying to recollect what all issues my team faced while running a job on production.

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. There was this situation happened with us for eg: you have newly developed model and it is scheduled to run weekly and it is productionalised now you have new requirement to add into that existing model lets say you have to create a UDF and add it to hive script and you add new dependencies to your xml file. when we went live with new addition our already existing job got failed due to some conflicts in the version of dependencies. You can avoid this by confirming that your newly added dependencies doesnt conflict with new one.

Now we are pretty much stable by using best practices in developing and deploying our models.

P.S: i have few deliverable pending and i am writing answer on quora. I will add more to it once i will recollect more issues till then  **Happy learning!**

One of Hive's biggest problems used to be skewed keys on joins. The way Hive is implemented, a join distributes different values of the key being joined across reducers such that a unique value is joined on a single reducer.  
  
Imagine two tables in twitter that look like the following

1. user\_id name
2. 1 markzuckerberg
3. 2 johndoe

1. user\_id follower\_id
2. 2 34
3. 2 55
4. 1 2
5. 1 3
6. 1 6
7. 1 6
8. ... (millions of people following markzuckerberg)

Then if you join the two into a single table of the schema  
user\_id name follower\_id

On the user\_id column, then everything with user\_id = 1 passes through a single reducer. That means that even if the John Does of the world join fairly quickly, you'll still blocked on the skewed keys for the final table to land.

Hadoop is bit challenging work on, as job may work for sample data however as data size increases job execution is prone to failures due various reasons.

Few issues which are challenging and are mostly related to distributed nature of system like network , memory and disk space issues are bound to happen as data size increases.

1. **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
2. **Pig and Hive :** The job unable to proceed further or getting killed, no further progress most of the memory issues and network issues.
3. Network issue like connection cannot be made to specific nodes.
4. Nodes frequently getting down.

Hope this helps.

Scenario: 1  
I have a table which has 10 records which are first imported into HDFS from MYSQL. Later i have two rows updated and 3 rows newly inserted into MYSQL.  
use lastmodified --------check-column date or timestamp last-value-mention it in qoute

Scenario: 2

I have a table with a primary key but not increasing/incrementing values. How do we handle on such cases as last value cannot help in this case.

either it must have column of date on which we can perform import or else primary key has to be set to auto increment

Scenario: 3

I have a table with just data and there is no date/timestamp available. How do we handle on such cases as lastmodified cannot help in this case.

Sqooping issues:

1) Data type conversion issue: will have to be very careful when we import the data from RDBMS to hadoop system , you will notice default conversion happening and are not suitable to the business need. we can use map-column-java/map-column-hive function to handle this issue

2)Sqoop does not support few hadoop file types like ORC,RC,Parquet

3)Delimiters : make sure the delimiter you are using is not part of the data you are importing/loading

4)mention schema and table names in Capital letters else you might face some issues.

5)import-all command will not work if your tables are not having primary key for sqoop to split the data

6)if you need multiple mappers(parallel import) you will have to provide split-by column

7)Table and column names can not have special characters . When importing data from legacy system you might face this issue, so we will have to write scripts to take care of this.

1. How will you get data from RDBMS into HDFS?

using sqoop with --target-dir

1. Can we store mysql table data as sequence file in hdfs via sqoop?

sqoop import --connect jdbc:mysql://quickstart:3306/tempdb1 --username root --password cloudera --table seqsqooptext --m 1 --as-sequencefile --target-dir=/user/cloudera/sqoop/seqimport

*If we create a simple external table, as we would for textfile format, the table would be created but it would not be able to read any data*

hive> ADD JAR /home/cloudera/Downloads/hive-sqoop-serde.jar ;

hive> CREATE EXTERNAL TABLE seqsqooptext (  
id int,  
name String  
)   
ROW FORMAT SERDE 'com.cloudera.sqoop.contrib.FieldMappableSerDe'  
WITH SERDEPROPERTIES (  
  "fieldmappable.classname" = "seq.seqsqooptext"  
)  
STORED AS SEQUENCEFILE  
LOCATION 'hdfs://quickstart.cloudera:8020/user/cloudera/sqoop/seqimport';

1. Does sqoop support compression techniques to store data in HDFS? [adsense]

yes we can and data will be stored in ORC file format

1. Can we load all the tables in a database into hdfs in a single shot? A) import-all-tables
2. Can we copy a subset of data from a table in RDBMS into HDFS?(based on some criteria) A) Using –where “country=’us'” condition in import command
3. How many reduce tasks will be run by default for a sqoop import command?How many mappers?
4. 0 , 4 [adsense]
5. 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. What is the default port for connecting to MySQL server?

A) 3306

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. Can we provide SQL queries in SQOOP Import command? [adsense]

s

**3) I have around 300 tables in a database. I want to import all the tables from the database except the tables named Table298, Table 123, and Table299. How can I do this without having to import the tables one by one?**

This can be accomplished using the import-all-tables import command in Sqoop and by specifying the exclude-tables option with it as follows-

sqoop import-all-tables

--connect –username –password --exclude-tables Table298, Table 123, Table 299

**6) How can you execute a free form SQL query in Sqoop to import the rows in a sequential manner?**

This can be accomplished using the –m 1 option in the Sqoop import command. It will create only one MapReduce task which will then import rows serially.

**9) What is Sqoop metastore?**

Sqoop metastore is a shared metadata repository for remote users to define and execute saved jobs created using sqoop job defined in the metastore. The sqoop –site.xml should be configured to connect to the metastore.

**4. Is Sqoop similar to distcp in hadoop?**

Partially yes, hadoop’s **distcp** command is similar to Sqoop Import command. Both submits parallel map-only jobs but **distcp** is used to copy any type of files from Local FS/HDFS to HDFS and Sqoop is for transferring the data records only between RDMBS and Hadoop eco system services, HDFS, Hive and HBase.

**6. When Importing tables from MySQL to what are the precautions that needs to be taken care w.r.t to access?**

In MySQL, we need to make sure that we have granted all privileges on the databases, that needs to be accessed, should be given to all users at destination hostname. If Sqoop is being run under localhost and MySQL is also present on the same then we can grant the permissions with below two commands from MySQL shell logged in with ROOT user.

MySQL

$ mysql -u root -p

mysql> GRANT ALL PRIVILEGES ON \*.\* TO '%'@'MachineB hostname or Ip address';

mysql> GRANT ALL PRIVILEGES ON \*.\* TO ''@'MachineB hostname or Ip address';

**11. While loading tables from MySQL into HDFS, if we need to copy tables with maximum possible speed, what can you do ?**

We need to use **–direct** argument in import command to use direct import fast path and this –direct can be used only with MySQL and PostGreSQL as of now.

**7) How will you list all the columns of a table using Apache Sqoop?**

Unlike sqoop-list-tables and sqoop-list-databases, there is no direct command like sqoop-list-columns to list all the columns. The indirect way of achieving this is to retrieve the columns of the desired tables and redirect them to a file which can be viewed manually containing the column names of a particular table.

Sqoop import --m 1 --connect 'jdbc: sqlserver: //nameofmyserver; database=nameofmydatabase; username=DeZyre; password=mypassword' --query "SELECT column\_name, DATA\_TYPE FROM INFORMATION\_SCHEMA.Columns WHERE table\_name='mytableofinterest' AND \$CONDITIONS" --target-dir 'mytableofinterest\_column\_name'

**13) You successfully imported a table using Apache Sqoop to HBase but when you query the table it is found that the number of rows is less than expected. What could be the likely reason?**

If the imported records have rows that contain null values for all the columns, then probably those records might have been dropped off during import because HBase does not allow null values in all the columns of a record

**14) The incoming value from HDFS for a particular column is NULL. How will you load that row into RDBMS in which the columns are defined as NOT NULL?**

Using the –input-null-string parameter, a default value can be specified so that the row gets inserted with the default value for the column that it has a NULL value in HDFS.

**3. How can we import data from particular row or column? What is the destination types allowed in Sqoop import command?**

Sqoop allows to Export and Import the data from the data table based on the where clause. The syntax is

--columns

<col1,col2……> --where

--query

Example:

sqoop import –connect jdbc:mysql://db.one.com/corp --table INTELLIPAAT\_EMP --where “start\_date> ’2016-07-20’ ”

sqoopeval --connect jdbc:mysql://db.test.com/corp --query “SELECT \* FROM intellipaat\_emp LIMIT 20”

sqoop import –connect jdbc:mysql://localhost/database --username root --password aaaaa –columns “name,emp\_id,jobtitle”

**19. When to use --target-dir and when to use --warehouse-dir while importing data?**   
  
 To specify a particular directory in HDFS use --target-dir but to specify the parent directory of all the sqoop jobs use --warehouse-dir. In this case under the parent directory sqoop will cerate a directory with the same name as th e table.

**69. How to import new data in a particular table every day?**  
It’a one of the main problems for Hadoop developers. Let example, you had downloaded 1TB data yesterday, today you got another 1gb data, if you import the data, again sqoop import 1TB+1GB data. So to get only use this command. Let example, you have already downloaded 1TB data which stored in the hive $Lastimport file. Now you can run it.

sqoop import --incremental lastmodified --check-column lastmodified --last-value "$LASTIMPORT --connect jdbc:mysql://localhost:3306/database\_name --table table\_name --username user\_name