Complex data types in Hive

1. **Array**

   create table result(student_id int, bands array<double>) row format delimited fields terminated by '|' collection items terminated by ','

   insert into result select 10 , array(cast(4.5 as double),cast(6.7 as double));

   hive (mydb)> *select * from result;*
   OK
   result.student_id result.bands
   10 [4.5,6.7]
   Time taken: 0.662 seconds, Fetched: 1 row(s)

   hive (mydb)> *select student_id, bands[0] from result;*
   OK
   student_id _c1
   10 4.5
   Time taken: 0.821 seconds, Fetched: 1 row(s)

   Hdfs view
   hduser@shyam:/usr/local/hadoop/etc/hadoop$ *hdfs dfs -cat /user/hive/warehouse/mydb.db/result/000000_0*

   10|4.5,6.7

2. **Map**

   create table resultMap(student_id int, bands map<string,double>) row format delimited fields terminated by '|' collection items terminated by ',' map keys terminated by ':';

   Input file

   hduser@shyam:~$ *cat tempfile*

   10|reading:4.5,speaking:6.7,listening:7.5,writing:7.0
   20|reading:5.5,speaking:6.5,listening:6.5,writing:8.0

   hduser@shyam:~$ *hdfs dfs -copyFromLocal tempfile /user/hive/warehouse/mydb.db/resultmap/*

   hive (mydb)> *select * from resultMap;*

OK
resultmap.student_id resultmap.bands
10 {"reading":4.5,"speaking":6.7,"listening":7.5,"writing":7.0}
20 {"reading":5.5,"speaking":6.5,"listening":6.5,"writing":8.0}
Time taken: 0.126 seconds, Fetched: 2 row(s)

hive (mydb)> *select student_id, bands["listening"] from resultMap;*
OK
student_id _c1
10 7.5
20 6.5
Time taken: 0.107 seconds, Fetched: 2 row(s)

3. **Struct**

create table rider(name string, age int, vehicle_conf struct<reg_no:string, top_speed:int, cc:int, brand:string>) row format delimited fields terminated by '|' collection items terminated by ',';

hduser@shyam:~$ *cat tempfile*

Sukhajinder|27|QC123,250,600,YAMAHA
Cesar|35|ON123,300,700,HONDA

hduser@shyam:~$ *hdfs dfs -copyFromLocal tempfile /user/hive/warehouse/mydb.db/rider/*

hive (mydb)> *select * from rider;*
OK
rider.name rider.age rider.vehicle_conf
Sukhajinder 27 {"reg_no":"QC123","top_speed":250,"cc":600,"brand":"YAMAHA"}
Cesar 35 {"reg_no":"ON123","top_speed":300,"cc":700,"brand":"HONDA"}
Time taken: 0.119 seconds, Fetched: 2 row(s)

hive (mydb)> *select vehicle_conf.brand from rider;*
OK
brand
YAMAHA
HONDA
Time taken: 0.111 seconds, Fetched: 2 row(s)

Read XML file in Hive Table

Create TABLE xmltable(xmldata string) STORED AS TEXTFILE;

hduser@shyam:~$ *cat tempfile*

<dependency><groupId>org.apache.hive</groupId><artifactId>hive-exec</artifactId><version>0.8.0</version></dependency>
<dependency><groupId>org.apache.hadoop</groupId><artifactId>hadoop-core</artifactId><version>1.2.1</version></dependency>
<dependency><groupId>junit</groupId><artifactId>junit</artifactId><version>4.5</version><scope>test</scope></dependency>

hduser@shyam:~$ *hdfs dfs -copyFromLocal tempfile /user/hive/warehouse/mydb.db/xmltable/*

hive (mydb)> *select xpath(xmldata, 'dependency/groupId/text()') from xmltable;*
OK
_c0
["org.apache.hive"]
["org.apache.hadoop"]
["junit"]
Time taken: 0.109 seconds, Fetched: 3 row(s)

GitHub Branching

- Get in to dev branch
  git checkout dev
- Do all your changes
- Stage those changes to dev branch
  git add .
- Commit those changes in dev branch
  git commit -m “pushing changes in dev”
- Push changes in dev
  git push origin dev
- Let the testing team does all test cases
- If everything looks good then testing team will merge these changes to master
  git checkout master
  git merge dev
- Delete dev branch from local system
  git branch -d dev
- Delete dev branch from remote (web)
  git push origin --delete dev

Reference:

https://blog.matthewrathbone.com/2013/08/10/guide-to-writing-hive-udfs.html
https://cwiki.apache.org/confluence/display/Hive/GenericUDAFCaseStudy
https://cwiki.apache.org/confluence/display/Hive/GenericUDAFCaseStudy#GenericUDAFCaseStudy-WritingGenericUDAFs:ATutorial
https://community.hortonworks.com/content/supportkb/150214/how-to-enable-debug-hive-cli-logging.html
https://www.edureka.co/blog/apache-hive-installation-on-ubuntu
http://www.bogotobogo.com/Hadoop/BigData_hadoop_Hive_Install_On_Ubuntu_16_04.php

Hive Serde:
https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL#LanguageManualDDL-RowFormat,StorageFormat,andSerDe
https://cwiki.apache.org/confluence/display/Hive/DeveloperGuide#DeveloperGuide-CodeOrganizationandaBriefArchitecture
https://stackoverflow.com/questions/24607685/loading-xml-data-into-hive-table-org-apache-hadoop-hive-ql-metadata-hiveexcepti

Fun to Learn:
https://stackoverflow.com/questions/20208696/hadoop-restart-datanode-and-tasktracker
https://github.com/apache/hive/blob/master/contrib/src/java/org/apache/hadoop/hive/contrib/udaf/example/UDAFExampleAvg.java
https://courses.cs.ut.ee/MTAT.08.036/2013_fall/uploads/Main/slides5.pdf
https://cwiki.apache.org/confluence/display/Hive/Tutorial#Tutorial-QueryingandInsertingData

Java API:
https://stackoverflow.com/questions/3724415/maven-artifact-and-groupid-naming

Github:
https://git-scm.com/book/en/v2/Git-Branching-Basic-Branching-and-Merging
https://makandracards.com/makandra/621-git-delete-a-branch-local-or-remote