Hive UDF

Task 1: Code your function

1. Simple UDF

   Purpose:

   One primitive input and one output.

   Example:

   Convert an age to age group information

   Coding Approach:

   Create a class which extends org.apache.hadoop.hive.ql.exec.UDF

   Implement evaluate method

Task 2: Deploy it temporarily

1. Add JAR to classpath

   add JAR <Local path of your JAR file>
2. Create temporary function

   create temporary function agegroup as ‘udf.AgeGroup”
3. Use it

   select agegroup(14)

```
hive (mydb)> add jar /home/hduser/myudf.jar;
Added [/home/hduser/myudf.jar] to class path
Added resources: [/home/hduser/myudf.jar]
hive (mydb)> create temporary function agegroup as 'udf.AgeGroup';
OK
Time taken: 0.004 seconds
hive (mydb)> select agegroup(11) ;
OK
_c0
Children
Time taken: 0.032 seconds, Fetched: 1 row(s)
hive (mydb)>
```

Or just append it into .hiverc file

Following is the content of a sample .hiverc file. It should be within conf directory of hive. If not present then you can create one.

set hive.cli.print.header=true;
set hive.cli.print.current.db=true;
add JAR /home/s_kante/IdeaProjects/HiveUDF/out/artifacts/HiveUDF_jar/HiveUDF.jar;
create temporary function isaccepted as 'udf.IsAcceptedNew';

Task 3: Deploy it permanently

1. Copy JAR file to hdfs file system
   hadoop fs -copyFromLocal myudf.jar /udf/
2. Register function
   create function agegroup as ‘udf.AgeGroup’ using jar ‘hdfs://localhost:54310/udf/myudf.jar’;

```
hive (default)> create function agegroup as 'udf.AgeGroup' using jar 'hdfs://localhost:54310/udf/myudf.jar';
Added [/tmp/9d5d7e7b-7b0b-4dd0-a6ef-63e1aad8a09e_resources/myudf.jar] to class path
Added resources: [hdfs://localhost:54310/udf/myudf.jar]
OK
Time taken: 0.552 seconds
hive (default)> select agegroup(40);
OK
_c0
Adult
Time taken: 0.449 seconds, Fetched: 1 row(s)
hive (default)>
```

Possible errors:

- Case sensitive package and class name

```
hive (default)> create function agegroup as 'udf.agegroup' using jar 'hdfs://localhost:54310/udf/myudf.jar';
Added [/tmp/9d5d7e7b-7b0b-4dd0-a6ef-63e1aad8a09e_resources/myudf.jar] to class path
Added resources: [hdfs://localhost:54310/udf/myudf.jar]
Failed to register default.agegroup using class udf.agegroup
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.FunctionTask
```

- Using function in another database

```
hive (default)> use mydb;
OK
Time taken: 0.033 seconds
hive (mydb)> select agegroup(40);
FAILED: SemanticException [Error 10011]: Invalid function agegroup
hive (mydb)> select default.agegroup(40);
OK
_c0
Adult
Time taken: 0.043 seconds, Fetched: 1 row(s)
hive (mydb)>
```

Task 4: Deregister the function

```
hive (default)> drop function agegroup;
OK
Time taken: 0.407 seconds
hive (default)> select agegroup(40);
FAILED: SemanticException [Error 10011]: Invalid function agegroup
hive (default)>
```

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

Let's deploy a UDF IsAccepted that returns True if the student has scored above certain bands in each category and average is also greater than some minimum value.

Debugging Hive CLI

hive -hiveconf hive.log.file=debug_hive_20180403.log -hiveconf hive.log.dir=/tmp/hivedebug/ -hiveconf hive.root.logger=DEBUG,DRFA

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

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