Apache Hive with Apache Hivemall

Hivemall: Machine Learning on Hive, Pig and Spark SQL

Install HiveMall https://github.com/myui/hivemall/wiki/Installation

Pick latest release https://github.com/myui/hivemall/releases

Setup Your Environment $HOME/.hiverc add jar /home/myui/tmp/hivemall-core-xxx-with-dependencies.jar; source /home/myui/tmp/define-all.hive;

Create a directory in HDFS for the JAR

adoop fs -mkdir -p /apps/hivemall hdfs dfs -chmod -R 777 /apps/hivemall cp hivemall-core-0.4.2-rc.2-with-dependencies.jar hivemall-with-dependencies.jar hdfs dfs -put hivemall-with-dependencies.jar /apps/hivemall/ hdfs dfs -put hivemall-with-dependencies.jar /apps/hive/warehouse/ hdfs dfs -put hivemall-core-0.4.2-rc.2-with-dependencies.jar /apps/hivemall

```
show functions "hivemall.*";
+-----------------------------------------+--+
|                tab_name                 |
+-----------------------------------------+--+
| hivemall.add_bias                       |
| hivemall.add_feature_index              |
| hivemall.amplify                        |
| hivemall.angular_distance               |
| hivemall.angular_similarity             |
| hivemall.argmin_kld                     |
| hivemall.array_avg                      |
...
| hivemall.x_rank                         |
| hivemall.zscore                         |
+-----------------------------------------+--+
149 rows selected (0.054 seconds)
```

Once installed the hivemall database will be filled with great functions to use for general processing as well as machine learning via SQL.

An example function is for Base91 encoding:

```
select hivemall.base91(hivemall.deflate('aaaaaaaaaaaaaaaabbbbccc'));
+----------------------+--+
|         _c0          |
+----------------------+--+
| AA+=kaIM|WTt!+wbGAA  |
+----------------------+--+
```

A more useful example is I ran tokenize on messages in a Hive table that I store tweets in.

select hivemall.tokenize(tweets.msg) from tweets limit 10;

|
["water","pipe","break","#TEST","#TEST","#WATERMAINBREAK","FakeMockTown","NJ","https","//t","c
| |

["RT","@CNNNewsource","Main","water","pipe","break","causes","flooding","sinkhole","swallows","car","ir
| |
["RT","@PaaSDev","#TEST","water","pipe","break","#TEST","Water","Main","Break","in","Fakeville","NJ",
| |
["Water","break","on","a","mountain","run","tonight","#saopaulo","#correr","#run","sdfdf,"https","//t","co/dvN
| |
["RT","@PaaSDev","water","pipe","break","#TEST","#TEST","#WATERMAINBREAK","FakeMockTown",
| |
["Route","33","In","Wilton","Closed","Due","To","Water","Main","Break","https","//t","co/UQMksljRUm","h
| |
["water","pipe","break","nj","#TEST","#TEST","#WATERMAINBREAK","https","//t","co/kvYNTG7wHf"]
| | ["water","pipe","break","nj","#TEST","test","https","//t","co/zjgjSaNvUz"] | |
["#TEST","#watermainbreak","water","main","break","pipe","test","nj","https","//t","co/qZEdnhlgYG"] | |
["Customers","of","Langley","Water","and","Sewer","District","under","boil","water","advisory","-","Aiken","
| 10 rows selected (4.848 seconds)

For more examples of usage: https://github.com/myui/hivemall/wiki/webspam-dataset I will be using HiveMall in future projects, I am expecting to include into an NiFi workflow for process NLP and other machine learning operations. The project has just joined Apache.