

PigUnit - Pig script testing simplified.

Table of contents

1 Overview.....	2
2 PigUnit Example.....	2
3 Running PigUnit.....	3
4 Building PigUnit.....	4
5 Troubleshooting Tips.....	4
6 Future Enhancements.....	5

1. Overview

PigUnit is a simple xUnit framework that enables you to easily test your Pig scripts. With PigUnit you can perform unit testing, regression testing, and rapid prototyping. No cluster set up is required if you run Pig in local mode.

2. PigUnit Example

We want to compute a top N of the most common queries. The Pig script is basic and very similar to the Query Phrase Popularity in the Pig tutorial. It expects in input a file of queries and a parameter n (n is 2 in our case in order to do a top 2).

Setting up a test for this script is simple as the argument and the input data are specified by just two arrays of text. It is the same for the expected output of the script that will be compared to the actual result of the execution of the Pig script.

Many examples are available in the [PigUnit tests](#).

2.1. Java test

```
@Test
public void testTop2Queries() {
    String[] args = {
        "n=2",
    };

    PigTest test = new PigTest("top_queries.pig", args);

    String[] input = {
        "yahoo",
        "yahoo",
        "yahoo",
        "twitter",
        "facebook",
        "facebook",
        "linkedin",
    };

    String[] output = {
        "(yahoo,3)",
        "(facebook,2)",
    };

    test.assertOutput("data", input, "queries_limit", output);
}
```

2.2. top_queries.pig

```
data =
  LOAD 'input'
  AS (query:CHARARRAY);

queries_group =
  GROUP data
  BY query;

queries_count =
  FOREACH queries_group
  GENERATE
    group AS query,
    COUNT(data) AS total;

queries_ordered =
  ORDER queries_count
  BY total DESC, query;

queries_limit =
  LIMIT queries_ordered $n;

STORE queries_limit INTO 'output';
```

2.3. Run

Then the test can be executed by JUnit (or any other Java testing framework). It requires:

1. pig.jar
2. pigunit.jar

It takes about 25s to run and should pass. In case of error (for example change the parameter `n` to `n=3`), the diff of output is displayed:

```
junit.framework.ComparisonFailure: null expected:<...ahoo,3)
(facebook,2)[ ]> but was:<...ahoo,3)
(facebook,2)[
(linkedin,1)]>
    at junit.framework.Assert.assertEquals(Assert.java:81)
    at junit.framework.Assert.assertEquals(Assert.java:87)
    at org.apache.pig.pigunit.PigTest.assertEquals(PigTest.java:272)
```

3. Running PigUnit

3.1. Local Mode

PigUnit runs in Pig's local mode by default. Local mode is fast and enables you to use your

local file system as the HDFS cluster. Local mode does not require a real cluster but a new local one is created each time.

3.2. Mapreduce Mode

PigUnit also runs in Pig's mapreduce mode. Mapreduce mode requires you to use a Hadoop cluster and HDFS installation. It is enabled when the Java system property `pigunit.exectype.cluster` is set to any value: e.g. `-Dpigunit.exectype.cluster=true` or `System.getProperties().setProperty("pigunit.exectype.cluster", "true")`. The cluster you select must be specified in the CLASSPATH (similar to the HADOOP_CONF_DIR variable).

4. Building PigUnit

To compile PigUnit (`pigunit.jar`), run this command from the Pig trunk:

```
$pig_trunk ant pigunit-jar
```

5. Troubleshooting Tips

Common problems you may encounter are discussed below.

5.1. Classpath in Mapreduce mode

When using PigUnit in mapreduce mode, be sure to include the `$HADOOP_CONF_DIR` of the cluster in your CLASSPATH.

The default value is `~/pigtest/conf`.

```
org.apache.pig.backend.executionengine.ExecException: ERROR 4010: Cannot
find hadoop configurations in classpath (neither hadoop-site.xml nor
core-site.xml was found in the classpath).If you plan to use local mode,
please put -x local option in command line
```

5.2. UDF jars Not Found

This error means that you are missing some jars in your test environment.

```
WARN util.JarManager: Couldn't find the jar for
org.apache.pig.piggybank.evaluation.string.LOWER, skip it
```

5.3. Storing data

Pig currently drops all STORE and DUMP commands. You can tell PigUnit to keep the commands and execute the script:

```
test = new PigTest(PIG_SCRIPT, args);  
test.unoverride("STORE");  
test.runScript();
```

5.4. Cache archive

For cache archive to work, your test environment needs to have the cache archive options specified by Java properties or in an additional XML configuration in its CLASSPATH.

If you use a local cluster, you need to set the required environment variables before starting it:

```
export LD_LIBRARY_PATH=/home/path/to/lib
```

6. Future Enhancements

Improvement and other components based on PigUnit that could be built later.

For example, we could build a PigTestCase and PigTestSuite on top of PigTest to:

1. Add the notion of workspaces for each test.
2. Remove the boiler plate code appearing when there is more than one test methods.
3. Add a standalone utility that reads test configurations and generates a test report.