FAQ

Q: What shall I do if I saw "Failed to create DataStorage"?

This usually happens when you are connecting hadoop cluster other than standard Apache hadoop 20.2 release. Pig bundles standard hadoop 20.2 jars in release. If you want to connect to other version of hadoop cluster, you need to replace bundled hadoop 20.2 jars with compatible jars. You can try:

1. do "ant"
2. copy hadoop jars from your hadoop installation to overwrite ivy/lib/Pig/hadoop-core-0.20.2.jar and ivy/lib/Pig/hadoop-test-0.20.2.jar
3. do "ant" again
4. cp pig.jar to overwrite pig-*-core.jar

Some other tricks is also possible. You can use "bin/pig -secretDebugCmd" to inspect the command line of Pig. Make sure you are using the right version of hadoop.
This issue will be solved in Pig 0.9.1 and beyond.

Q: How can I pass a specific hadoop configuration parameter to Pig?

There are multiple places you can pass hadoop configuration parameter to Pig. Here is a list from high priority to low priority (configuration in high priority will override the configuration in low priority):
1. set command
2. -P properties_file
3. pig.properties
4. java system property/environmental variable
5. Hadoop configuration file: hadoop-site.xml/core-site.xml/hdfs-site.xml/mapred-site.xml, or Pig specific hadoop configuration file: pig-cluster-hadoop-site.xml)

Both 3 and 5 require the configuration file in classpath.

Q: I already register my LoadFunc/StoreFunc jars in "register" statement, but why I still get "Class Not Found" exception?

Try to put your jars in PIG_CLASSPATH as well. "register" guarantees your jar will be shipped to backend. But in the frontend, you still need to put the jars in CLASSPATH by setting "PIG_CLASSPATH" environment variable.

Q: How can I load data using Unicode control characters as delimiters?

The first parameter to PigStorage is the dataset name, the second is a regular expression to describe the delimiter. We used `String.split(regex, -1)` to extract fields from lines. See java.util.regex.Pattern for more information on the way to use special characters in regex.

If you are loading a file which contains Ctrl+A as separators, you can specify this to PigStorage using the Unicode notation.

```
LOAD 'input.dat' USING PigStorage('\u0001')as (x,y,z);
```

Q: How do I control the number of mappers?

It is determined by your InputFormat. If you are using PigStorage, FileInputFormat will allocate at least 1 mapper for each file. If the file is large, FileInputFormat will split the file into smaller trunks. You can control this process by two hadoop setting: "mapred.min.split.size", "mapred.max.split.size". In addition, after InputFormat tells Pig all the splits information, Pig will try to combine small input splits into one mapper. This process can be controlled by "pig.noSplitCombination" and "pig.maxCombinedSplitSize".

Q: How do I make my Pig jobs run on a specified number of reducers?

You can achieve this with the PARALLEL clause. For example:

```
C = JOIN A by url, B by url PARALLEL 50.
```

Besides PARALLEL clause, you can also use "set default_parallel" statement in Pig script, or set "mapred.reduce.tasks" system property to specify default parallel to use. If none of these values are set, Pig will only use 1 reducers. (In Pig 0.8, we change the default reducer from 1 to a number calculated by a simple heuristic for foolproof purpose)

More details can be found at http://pig.apache.org/docs/r0.9.0/perf.html#parallel.

Q: Can I do a numerical comparison while filtering?

Yes, you can choose between numerical and string comparison. For numerical comparison use the operators =, <>, < etc. and for string comparisons use eq, neq etc.

Q: Does Pig support regular expressions?

Pig does support regular expression matching via the `matches` keyword. It uses java.util.regex matches which means your pattern has to match the entire string (e.g. if your string is `"hi fred"` and you want to find `"fred"` you have to give a pattern of `".*fred"` not `"fred"`).

Q: How do I prevent failure if some records don't have the needed number of columns?

You can filter away those records by including the following in your Pig program:

```
A = LOAD 'foo' USING PigStorage('\t');
B = FILTER A BY ARITY(*) < 5;
.....
```

This code would drop all records that have fewer than five (5) columns.

Q: Is there any difference between `==` and `eq` for numeric comparisons?

There is no difference when using integers. However, `11.0` and `11` will be equal with `==` but not with `eq`.

Q: Is there an easy way for me to figure out how many rows exist in a dataset from it's alias?

You can run the following set of commands, which are equivalent to `SELECT COUNT⭐` in SQL:

```
a = LOAD 'mytestfile.txt';
b = GROUP a ALL;
c = FOREACH b GENERATE COUNT(a.$0);
```

Q: Does Pig allow grouping on expressions?

Pig allows grouping of expressions. For example:

```
grunt> a = LOAD 'mytestfile.txt' AS (x,y,z);
grunt> DUMP a;
(1,2,3)
(4,2,1)
(4,3,4)
(4,3,4)
(7,2,5)
(8,4,3)

b = GROUP a BY (x+y);
(3.0,{(1,2,3)})
(6.0,{(4,2,1)})
(7.0,{(4,3,4),(4,3,4)})
(9.0,{(7,2,5)})
(12.0,{(8,4,3)})
```

If the grouping is based on constants, the result is the same as GROUP ALL except the group-id is replaced by the constant.

```
grunt> b = GROUP a BY 4;
(4,{(1,2,3),(4,2,1),(4,3,4),(4,3,4),(7,2,5),(8,4,3)})
```

Q: Is there a way to check if a map is empty?

In Pig 2.0 you can test the existence of values in a map using the null construct:
m#'key' is not null

Q: I load data from a directory which contains different file. How do I find out where the data comes from?

You can write a LoadFunc which append filename into the tuple you load.

Eg,

```
A = load '*.txt' using PigStorageWithInputPath();
```

Here is the LoadFunc:

```
public class PigStorageWithInputPath extends PigStorage {
    Path path = null;

    @Override
    public void prepareToRead(RecordReader reader, PigSplit split) {
        super.prepareToRead(reader, split);
        path = ((FileSplit)split.getWrappedSplit()).getPath();
    }

    @Override
    public Tuple getNext() throws IOException {
        Tuple myTuple = super.getNext();
        if (myTuple != null)
            myTuple.append(path.toString());
        return myTuple;
    }
}
```

In Pig 0.8/0.9.0/0.9.1, you need to set "pig.splitCombination" to false for PigStorageWithInputPath work correctly. 0.9.2 fix the issue.

Q: How can I calculate a percentage (partial aggregate / total aggregate)?

The challenge here is to get the total aggregate into the same statement as the partial aggregate. The key is to cast the relation for the total aggregate to a scalar:

```
A = LOAD 'sample.txt' AS (x:int, y:int);
-- calculate the denominator
B = foreach (group A all) generate COUNT(A) as total;
-- cacluate the percentage
C = foreach (group A by x) generate group as x, (double)COUNT(A) / (double) B.total as percentage;
```

Q: How can I pass a parameter with space to a pig script?

```
# Following should work
-p "NAME='Firstname Lastname'"
-p "NAME=Firstname\ Lastname"
# Following are incorrect
-p "NAME=Firstname Lastname"
-p NAME="Firstname Lastname"
```