Steps to Implement Word Count in Python on Hadoop:

1. Create a Python Mapper Script (mapper.py):

The mapper reads the input line by line and emits a key-value pair for each word (word, 1).

```
#!/usr/bin/env python

import sys

# Input comes from standard input
for line in sys.stdin:
    # Remove leading and trailing whitespace
    line = line.strip()
    # Split the line into words
    words = line.split()
    # Emit each word, followed by a count of 1
    for word in words:
        print(f"{word}\t1")
```

2. Create a Python Reducer Script (reducer.py):

The reducer reads the key-value pairs from the mapper output, sums the values for each key (word), and emits the word with its total count.

```
#!/usr/bin/env python

import sys

current_word = None
current_count = 0
word = None

# Input comes from standard input
for line in sys.stdin:
```

```
    # Remove leading and trailing whitespace
    line = line.strip()

    # Parse the input we got from mapper.py
    word, count = line.split('\t', 1)

    # Convert count (currently a string) to int
    try:
        count = int(count)
    except ValueError:
        # If count is not a number, skip this line
        continue

    # This IF-switch only works because Hadoop sorts map output
    # by key (here: word) before it is passed to the reducer
    if current_word == word:
        current_count += count
    else:
        if current_word:
            # Write the result to standard output
            print(f"{current_word}\t{current_count}")
        current_count = count
        current_word = word

# Output the last word if needed
if current_word == word:
    print(f"{current_word}\t{current_count}")
```

3. **Make Python Scripts Executable:**

```
chmod +x mapper.py
chmod +x reducer.py
```

4. **Upload Input File to HDFS:**

If you already have your input file (input.txt) in HDFS, you can skip this step. Otherwise:

hdfs dfs -put input.txt /user/<your-username>/wordcount/input

5. Run the Hadoop Streaming Job:

Run the Hadoop Streaming command to execute your Python scripts as the mapper and reducer.

```
hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-*.jar \
  -input /user/<your-username>/wordcount/input/input.txt \
  -output /user/<your-username>/wordcount/output \
  -mapper mapper.py \
  -reducer reducer.py \
  -file mapper.py \
  -file reducer.py
```

-input: Path to the input file in HDFS.

-output: Path to the output directory in HDFS (this directory must not exist before running the job).

-mapper: Specifies the mapper script.

-reducer: Specifies the reducer script.

-file: Uploads the Python script to the cluster.

6. Check the Output:

After the job completes, check the output using the following command:

hdfs dfs -cat /user/<your-username>/wordcount/output/part-00000

This will display the word counts as calculated by your Python scripts.

7. Clean Up:

If you want to rerun the job, make sure to remove the previous output directory from HDFS:

hdfs dfs -rm -r /user/<your-username>/wordcount/output