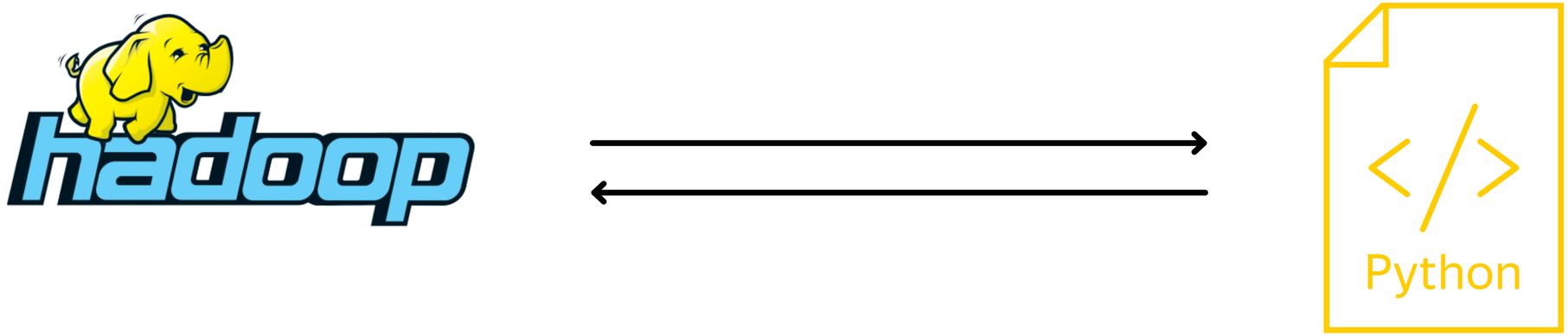
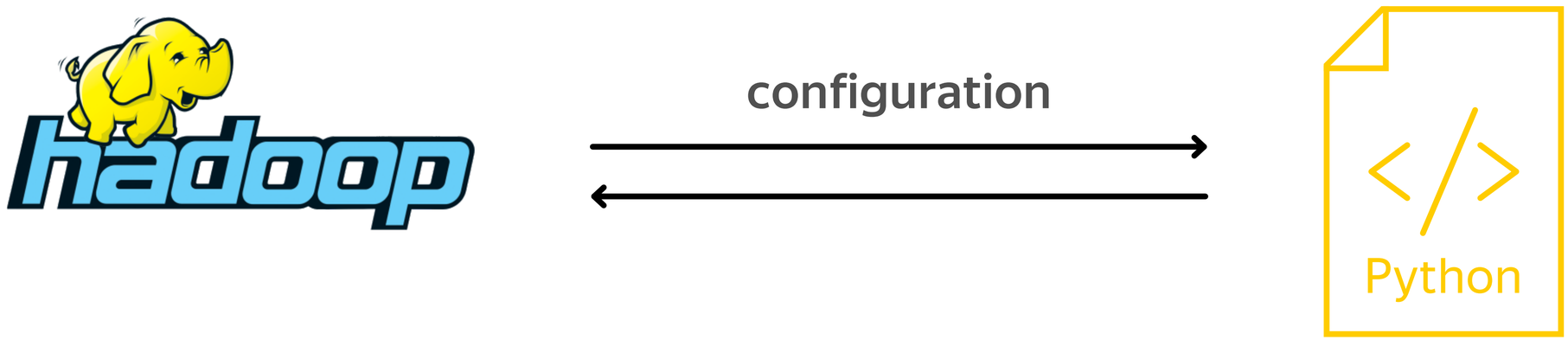
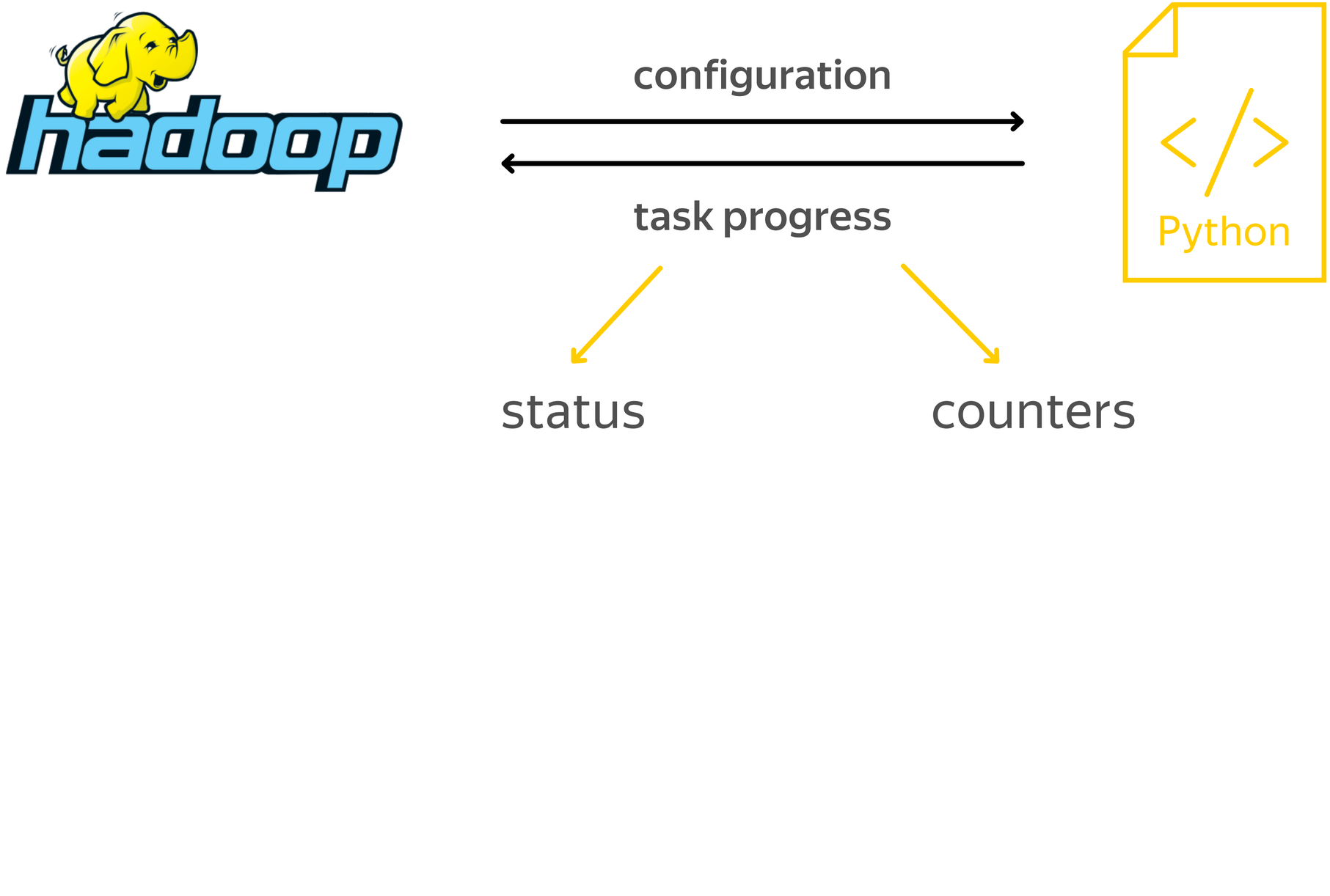
Yandex

MapReduce

Testing


hadoop
Python


hadoop
configuration
Python


hadoop
configuration
task progress
Python
status
counters

Distributed Cache

configuration

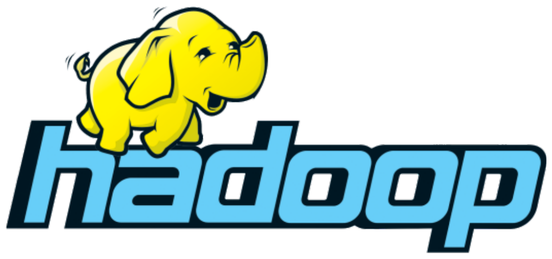
task progress

Python

status

counters

Distributed Cache



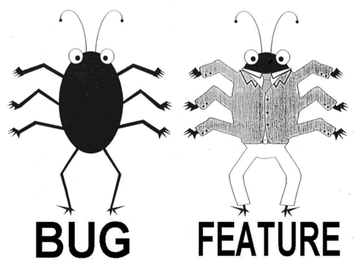
configuration

task progress

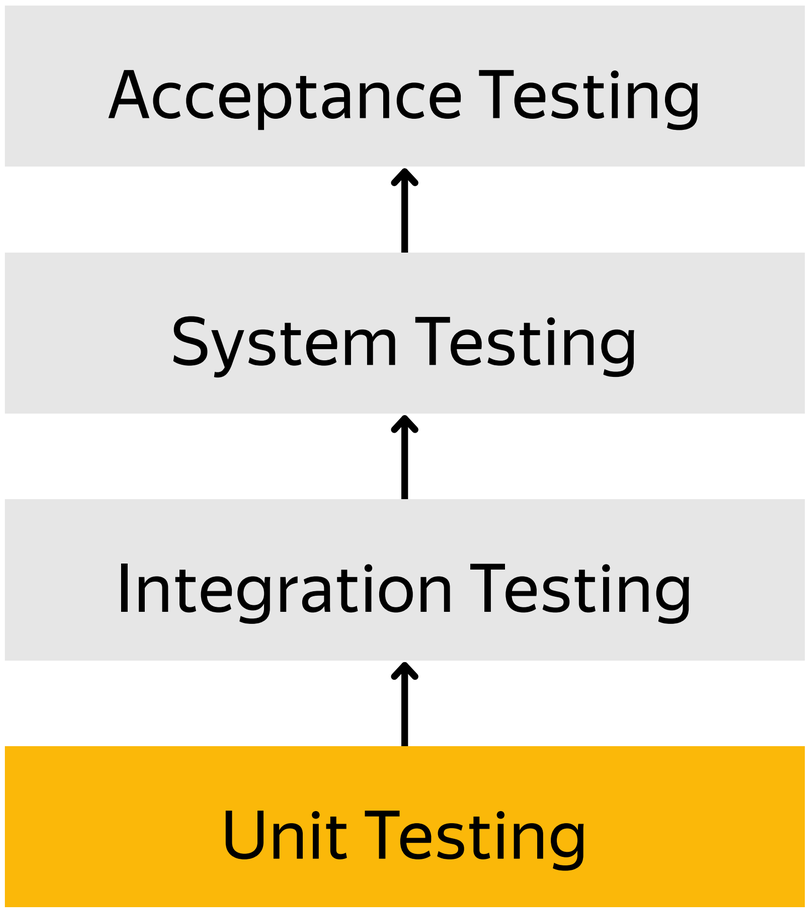
status

counters

Python




Testing
ctrl
alt


Acceptance Testing
System Testing
Integration Testing
Unit Testing

Acceptance Testing

↑

System Testing

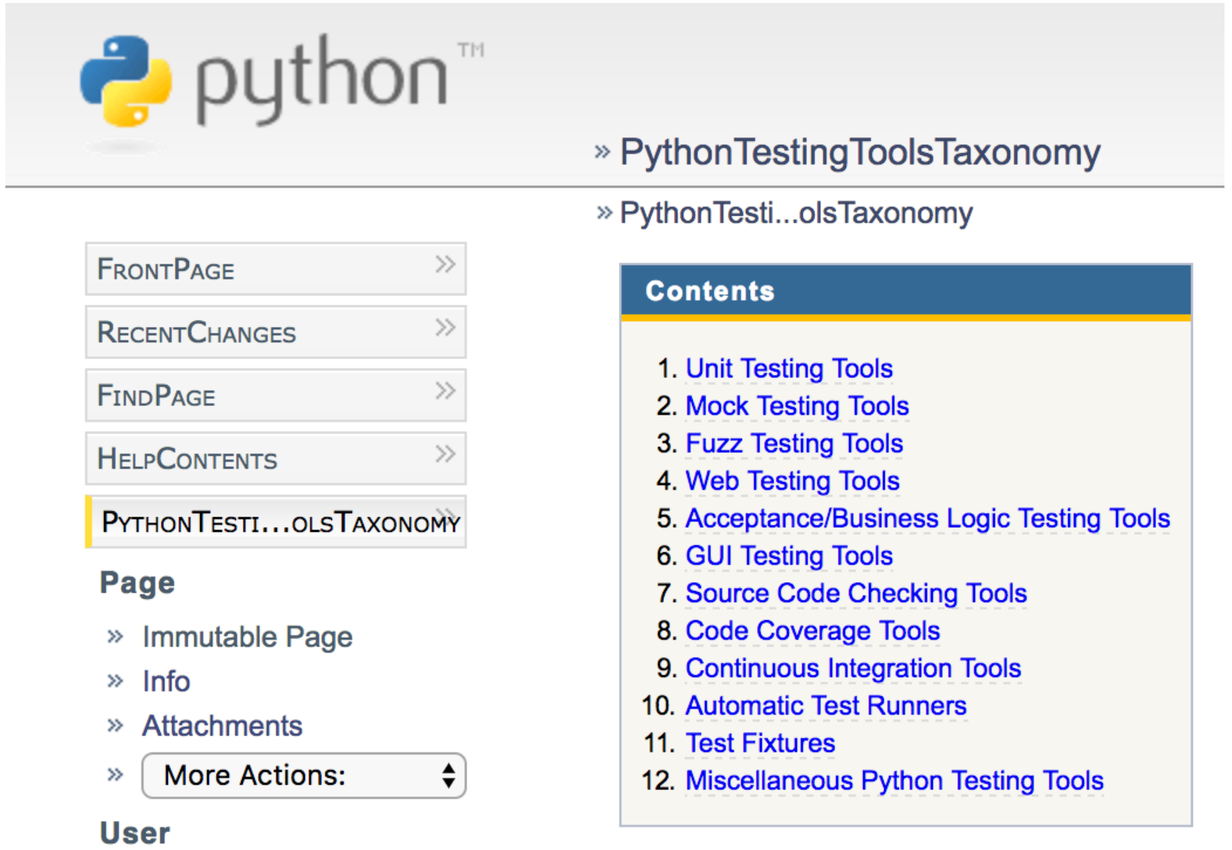
↑

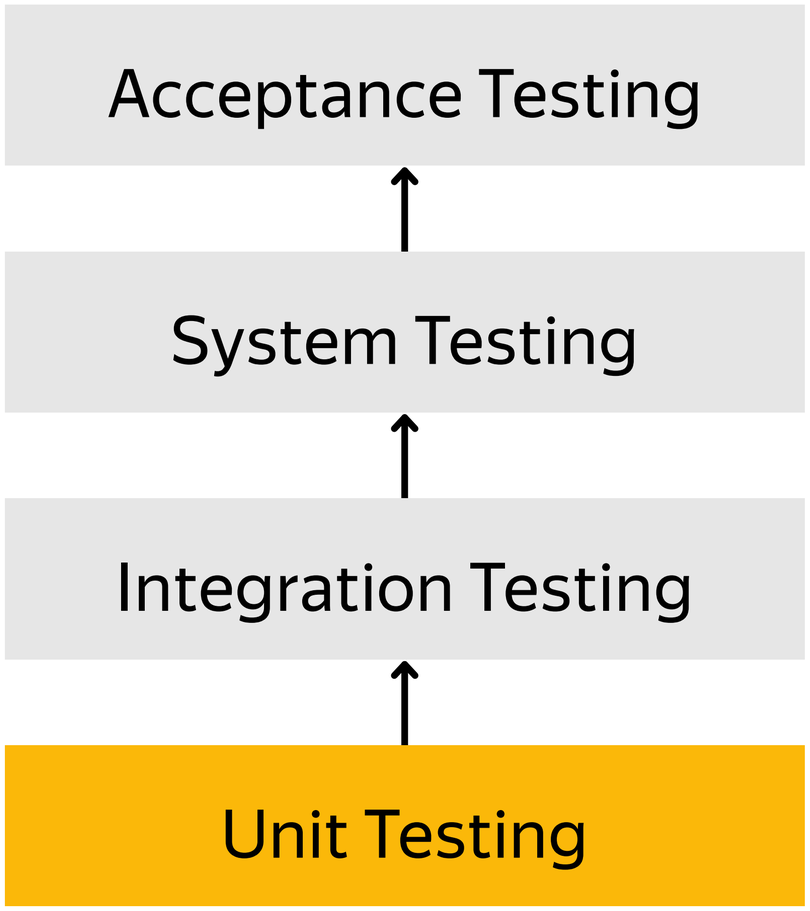
Integration Testing

↑

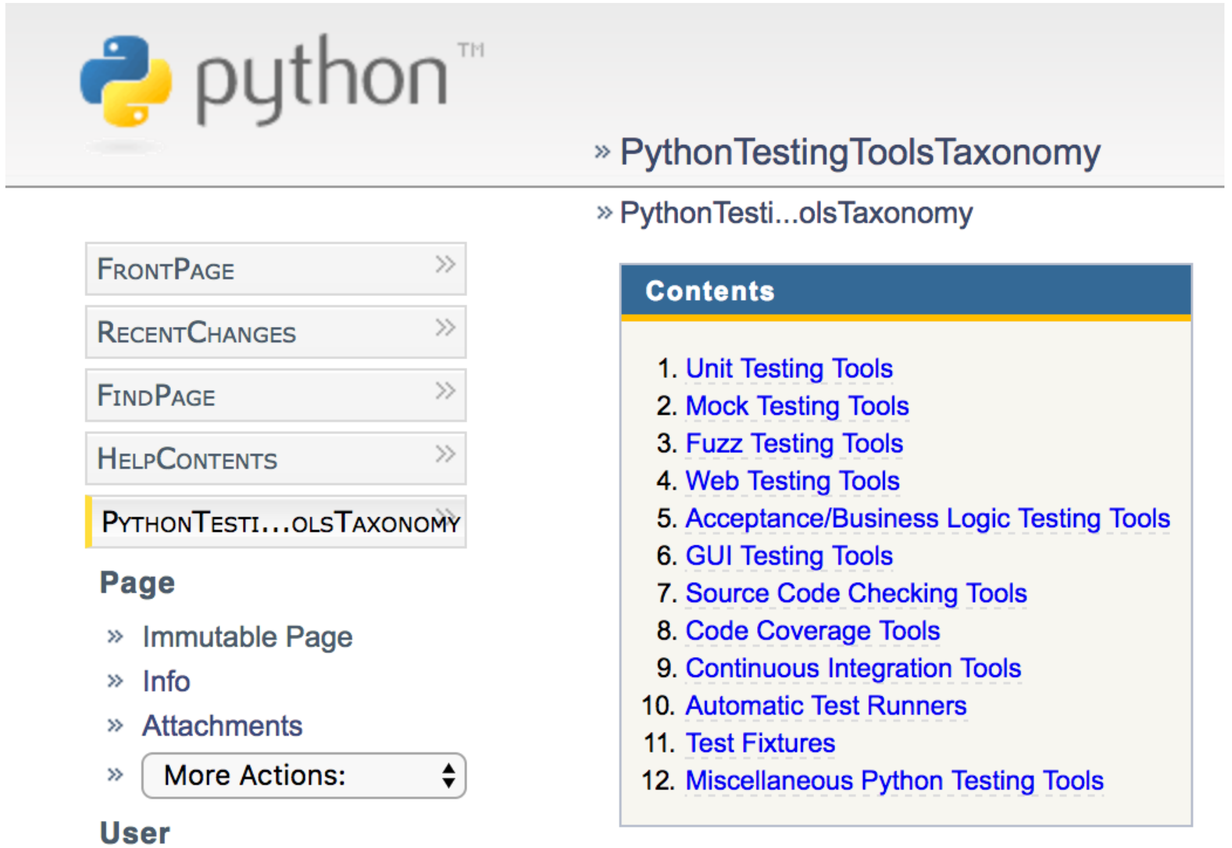
Unit Testing

https://wiki.python.org/moin/PythonTestingToolsTaxonomy

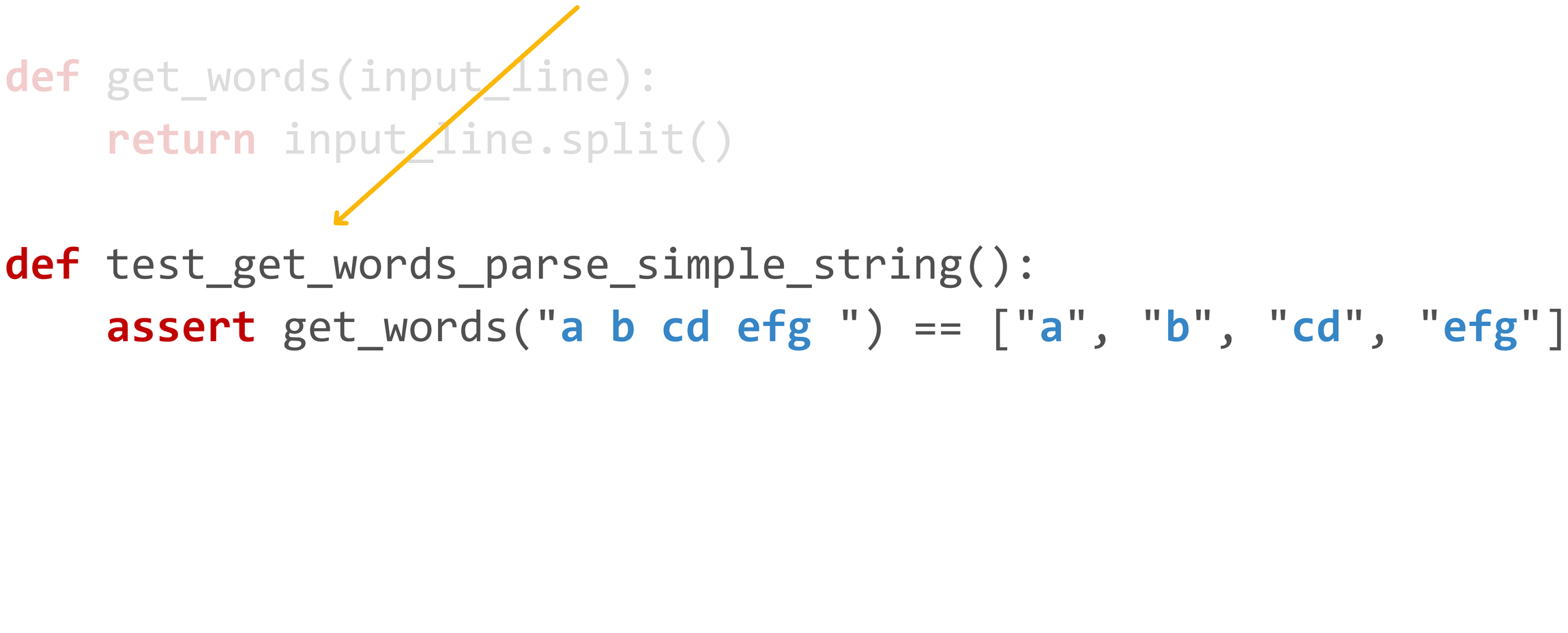




https://wiki.python.org/moin/PythonTestingToolsTaxonomy



```
def get_words(input_line):
    return input_line.split()
```



```
def get_words(input_line):
    return input_line.split()

def test_get_words_parse_simple_string():
    assert get_words("a b cd efg ") == ["a", "b", "cd", "efg"]
```

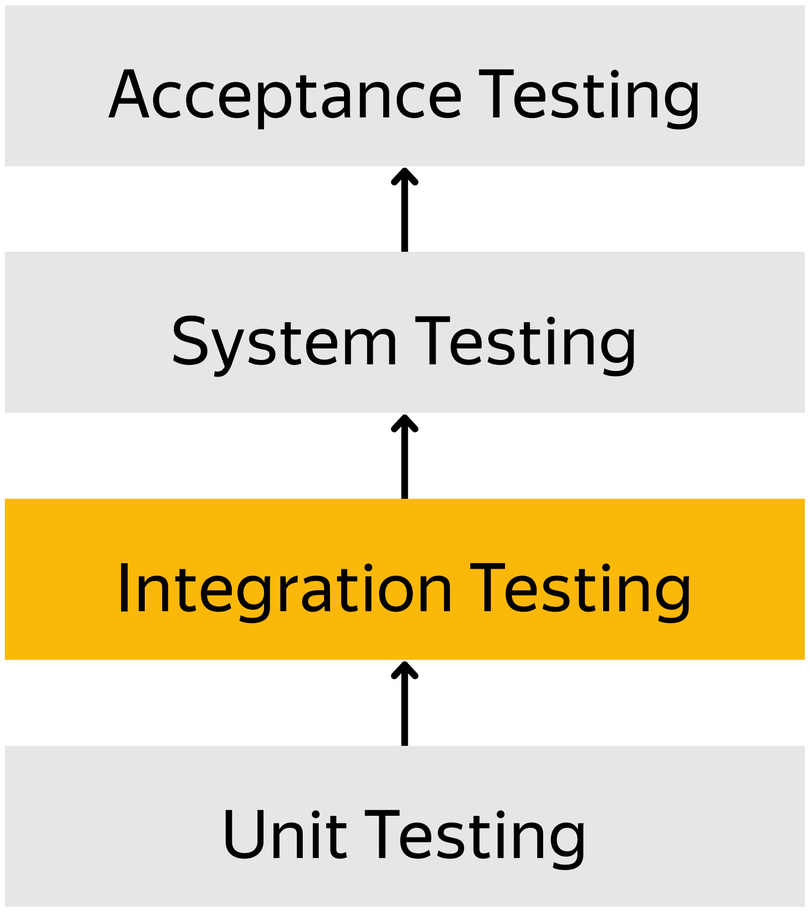
```
def get_words(input_line):
    return input_line.split()

def test_get_words_parse_simple_string():
    assert get_words("a b cd efg ") == ["a", "b", "cd", "efg"]

def test_get_words_parse_empty_string():
    assert get_words("") == []
```

```python
def get_words(input_line):
    return input_line.split()

def test_get_words_parse_simple_string():
    assert get_words("a b cd efg ") == ["a", "b", "cd", "efg"]

def test_get_words_parse_empty_string():
    assert get_words("") == []

def test_get_words_raise_exception_if_no_input():
    with pytest.raises(AttributeError):
        get_words(None)
```

```python
def get_words(input_line):
    return input_line.split()

def test_get_words_parse_simple_string():
    assert get_words("a b cd efg ") == ["a", "b", "cd", "efg"]

def test_get_words_parse_empty_string():
    assert get_words("") == []

def test_get_words_raise_exception_if_no_input():
    with pytest.raises(AttributeError):
        get_words(None)
```

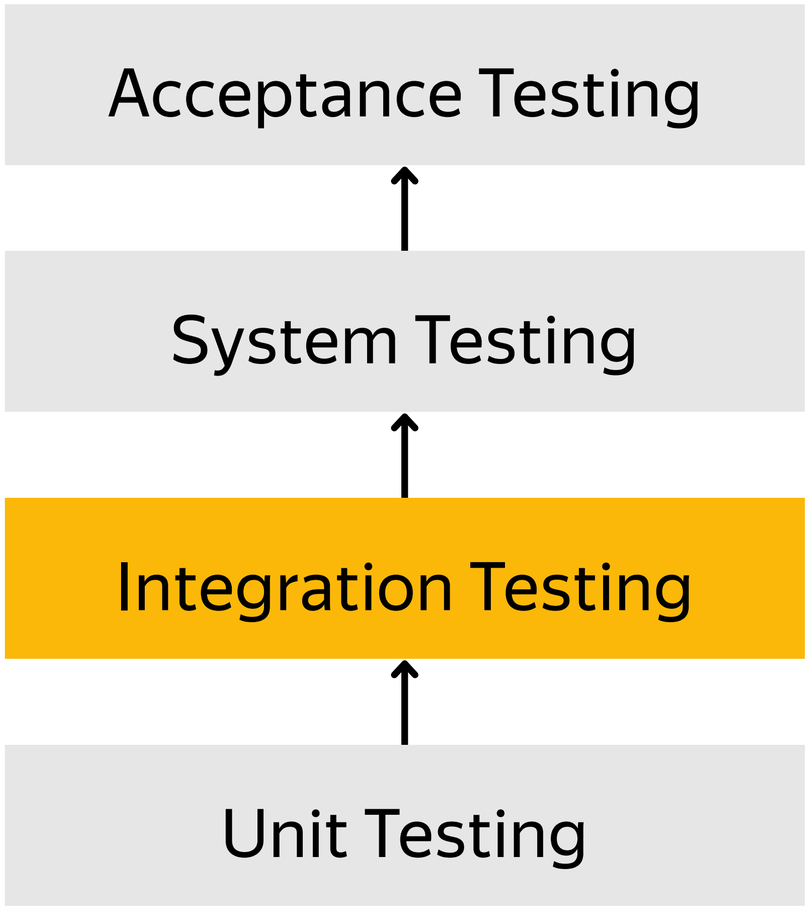
```
(venv) ➜ coursera pytest -v test_function.py
============================= test session starts ==============================
platform darwin -- Python 2.7.12, pytest-3.0.7, py-1.4.33, pluggy-0.4.0 -- /Users/aadral/workspace/personal/mipt/coursera/venv/bin/python2.7
cachedir: .cache
rootdir: /Users/aadral/workspace/personal/mipt/coursera, inifile:
collected 3 items

test_function.py::test_get_words_parse_simple_string PASSED
test_function.py::test_get_words_parse_empty_string PASSED
test_function.py::test_get_words_raise_exception_if_no_input PASSED

=========================== 3 passed in 0.01 seconds ===========================
```

```
$ cat -n wikipedia.dump | tr ' ' '\n' | sort | uniq -c
```



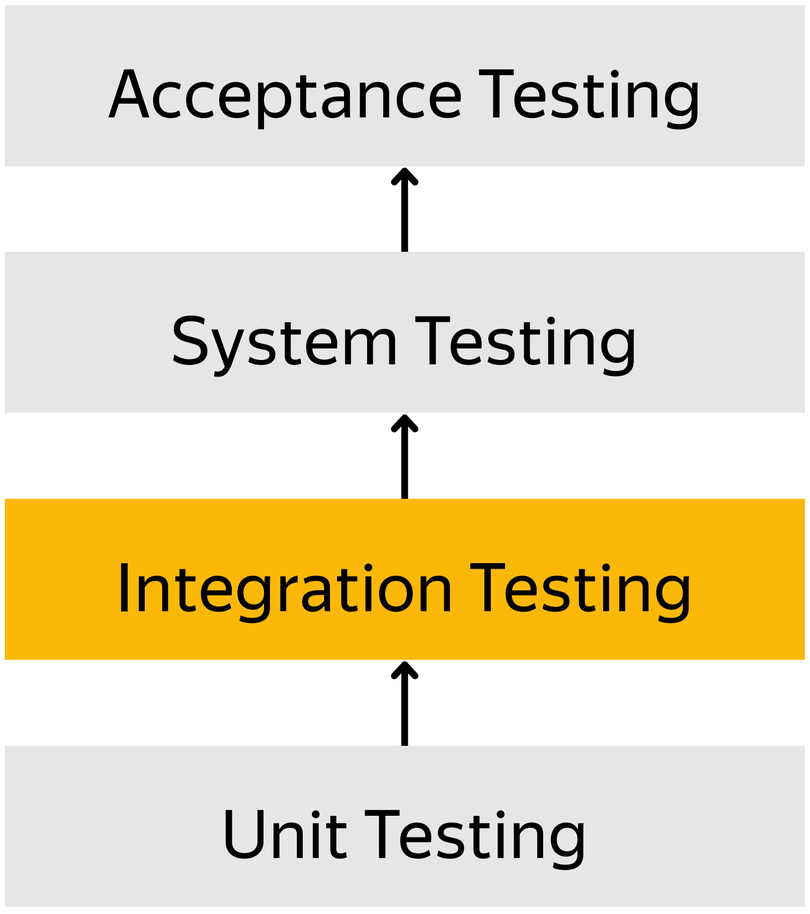
```
$ cat -n wikipedia.dump | tr ' ' '\n' | sort | uniq -c
```

```
$ cat | mapper | sort | reducer
```



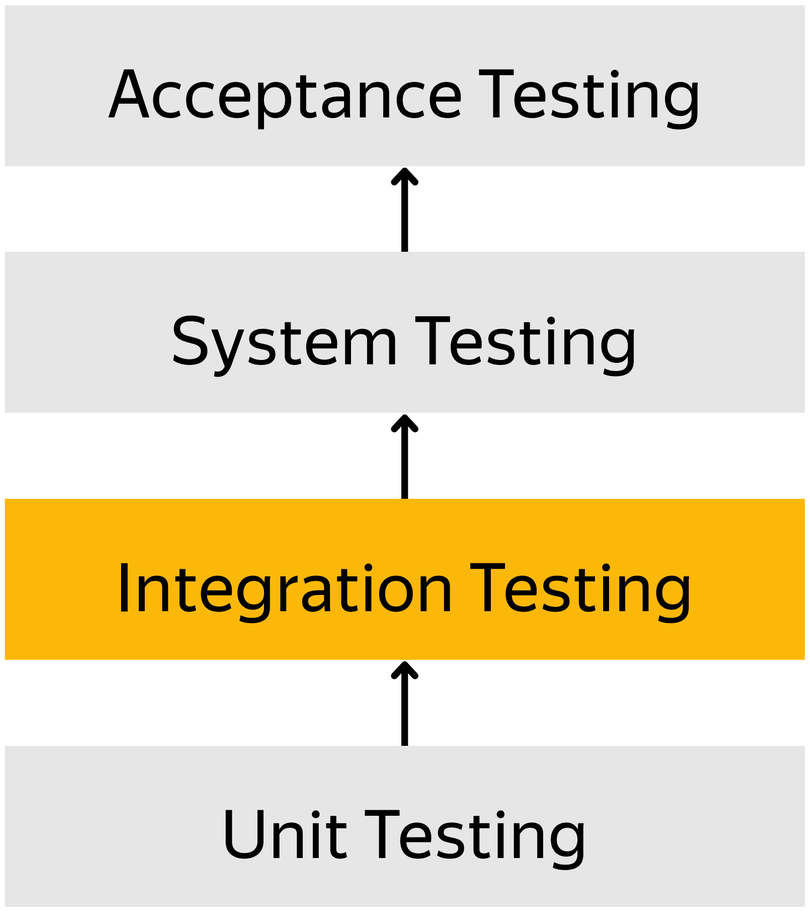
```
$ cat -n wikipedia.dump | tr ' ' '\n' | sort | uniq -c
```

```
$ cat | mapper | sort | reducer
```



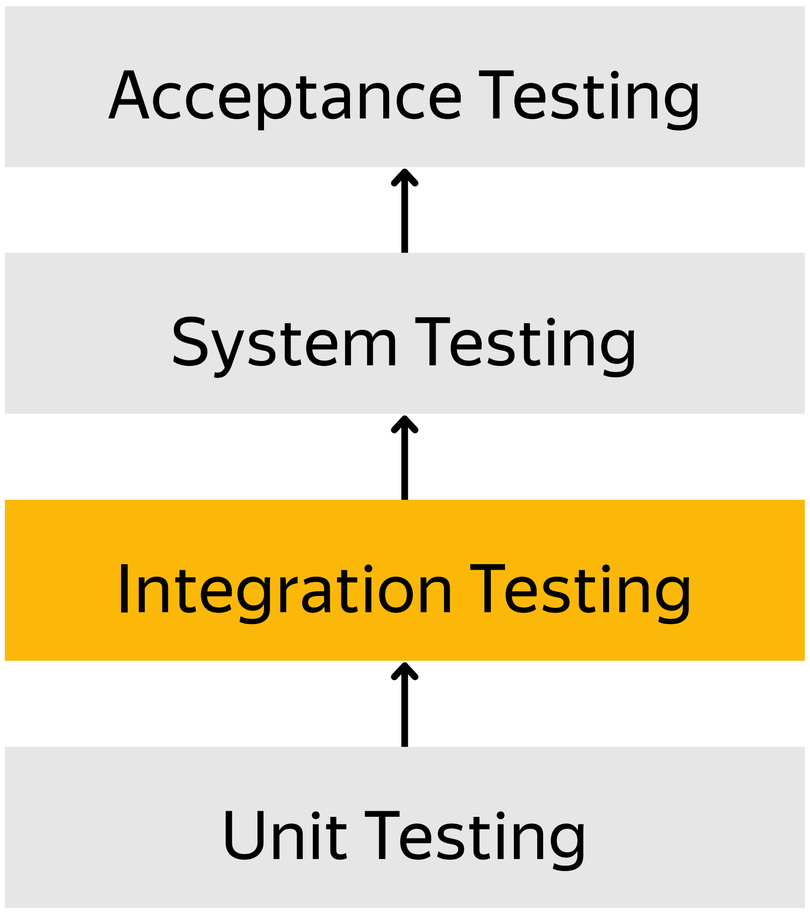
```
$ cat -n wikipedia.dump | tr ' ' '\n' | sort | uniq -c
```

```
$ cat | mapper | sort | reducer
```



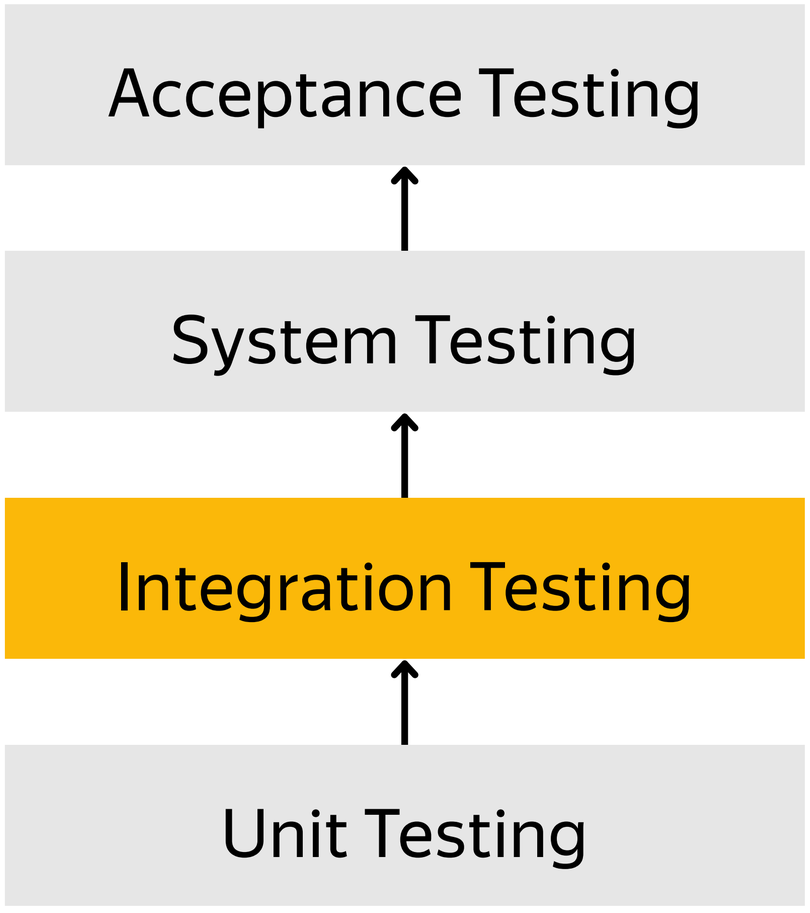
```
$ cat -n wikipedia.dump | tr ' ' '\n' | sort | uniq -c
```

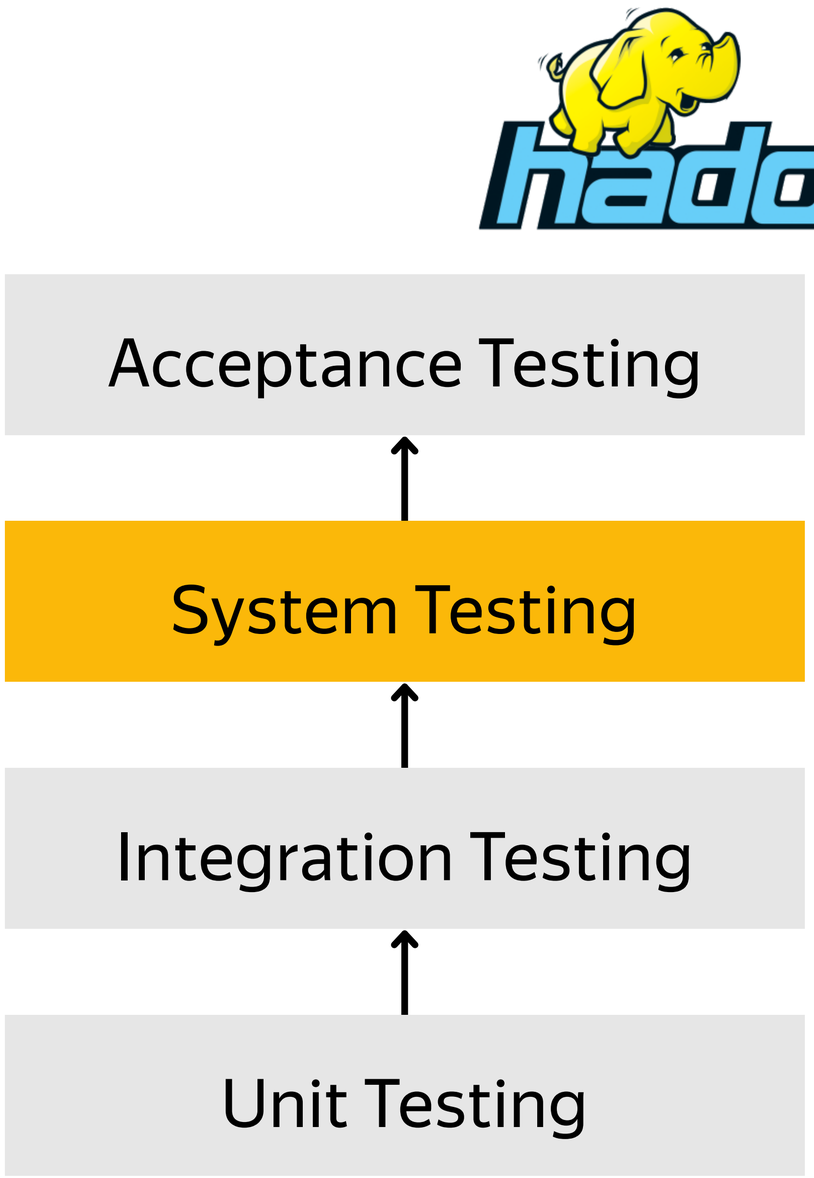
```
$ cat                                          reducer
```



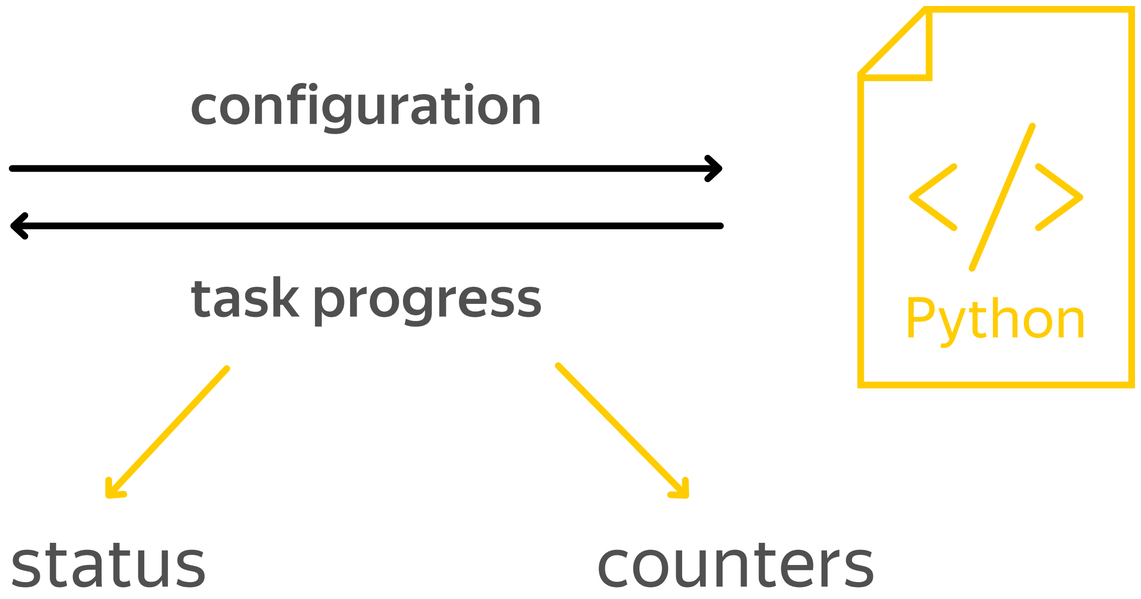
```
$ cat -n wikipedia.dump | tr ' ' '\n' | sort | uniq -c
```

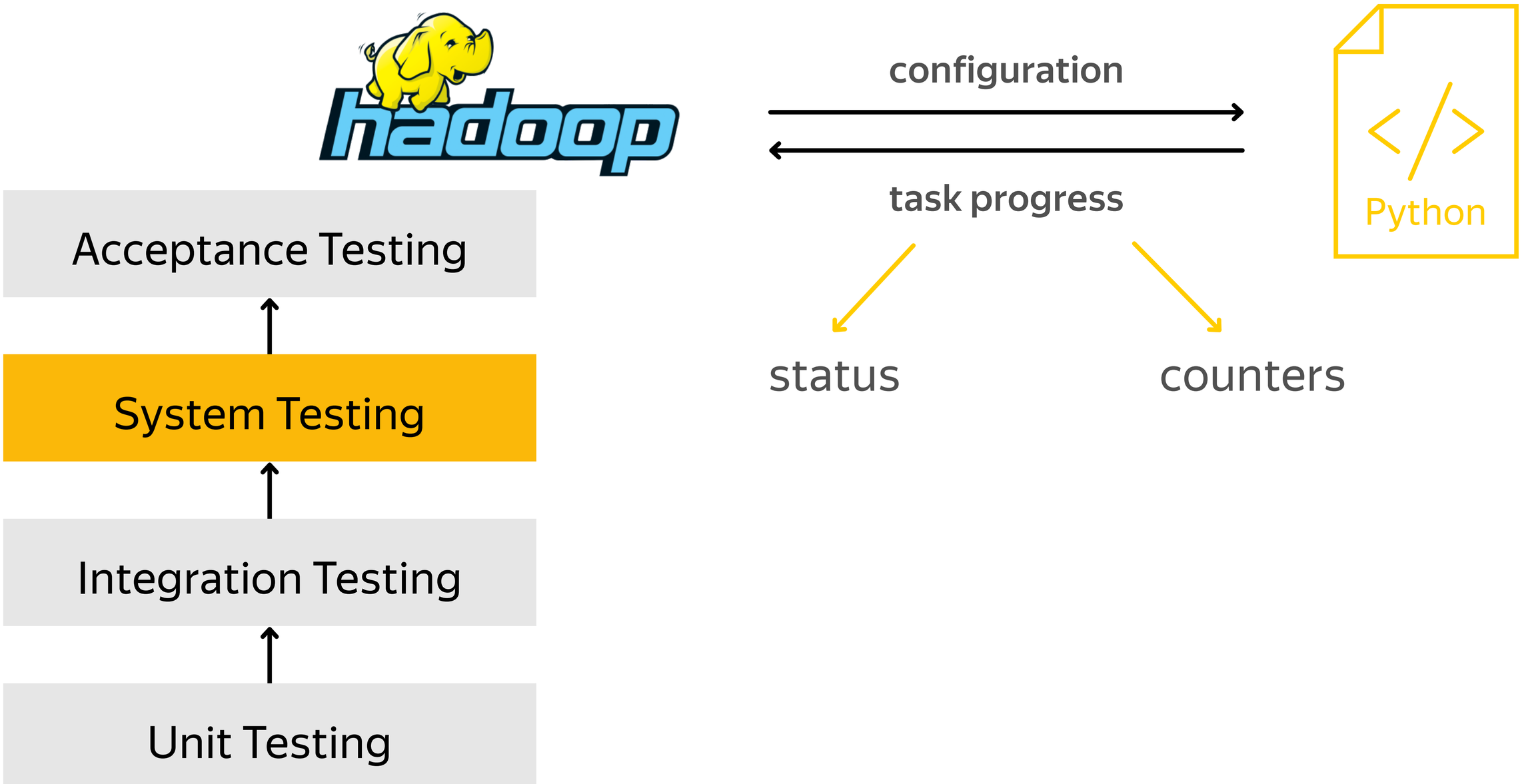
```
$ cat | mapper | sort | reducer
```



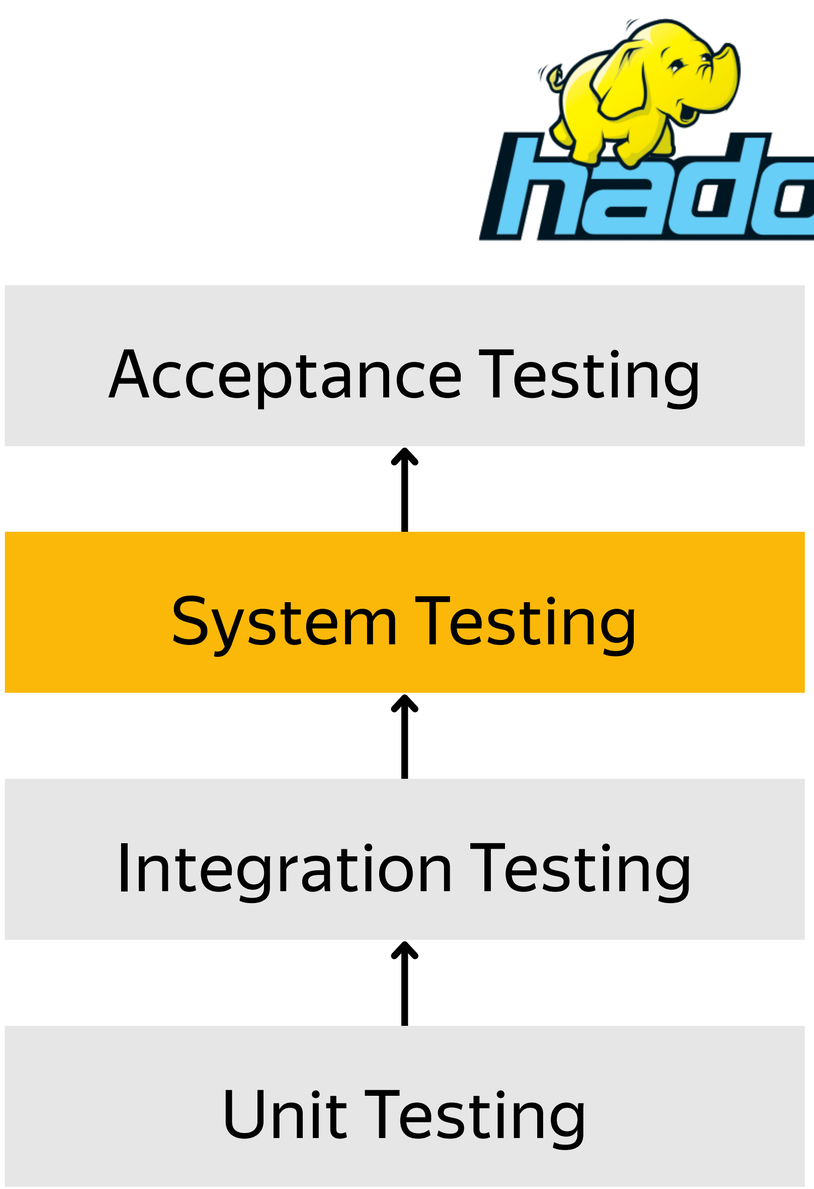

Acceptance Testing
System Testing
Integration Testing
Unit Testing

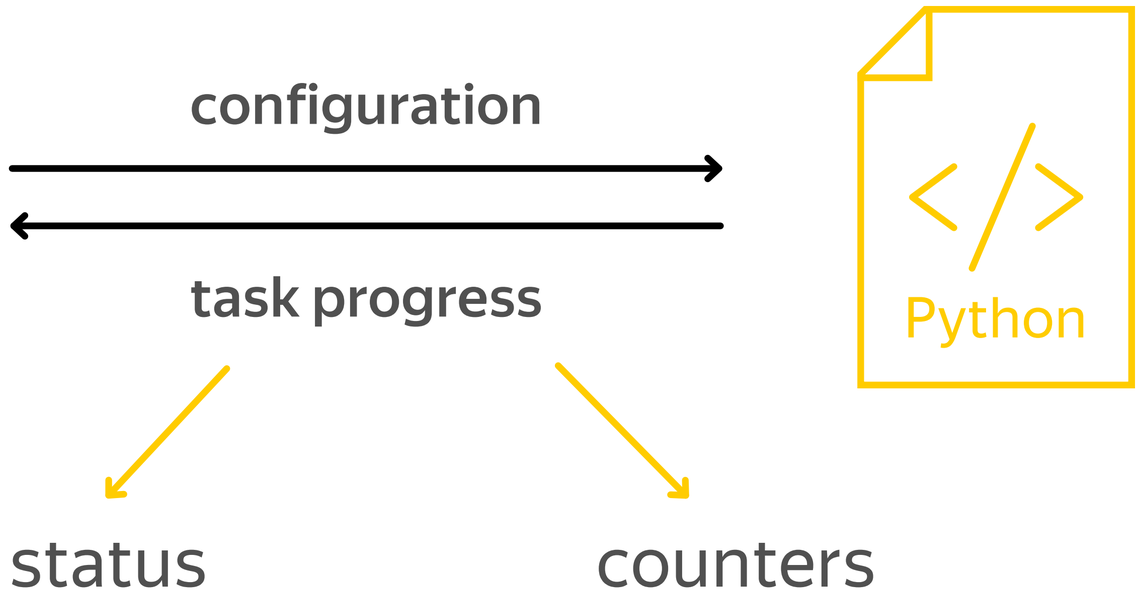
hadoop


configuration
task progress
Python
status
counters



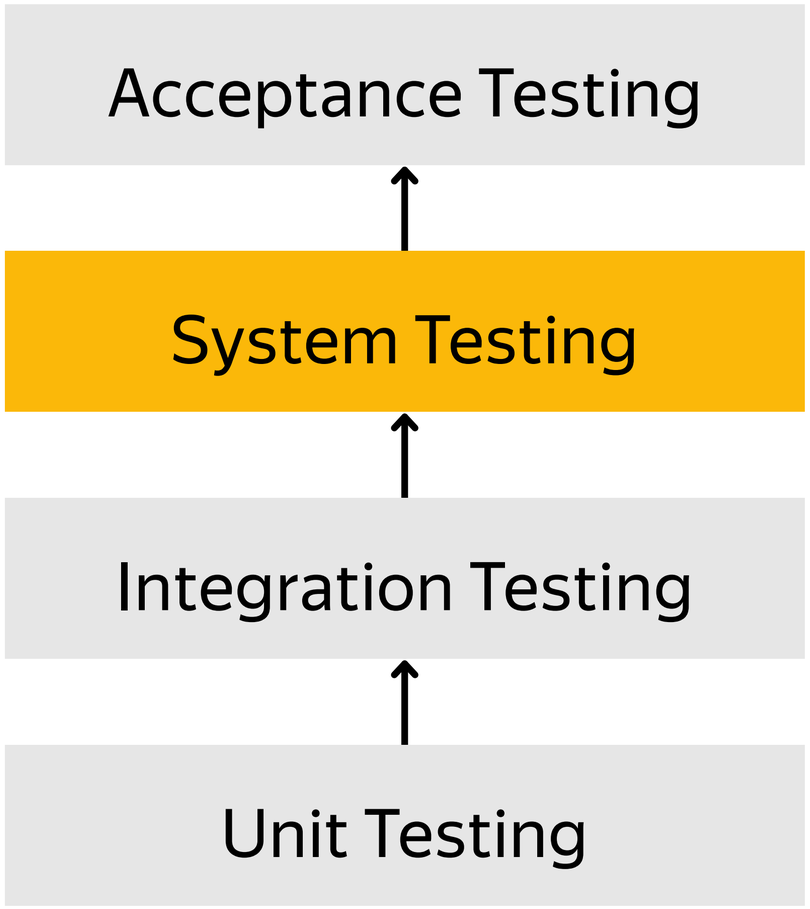

```
hdfs --config $HADOOP_EMPTY_CONFIG dfs -rm -r word_count
yarn --config $HADOOP_EMPTY_CONFIG jar $HADOOP_STREAMING_JAR
```

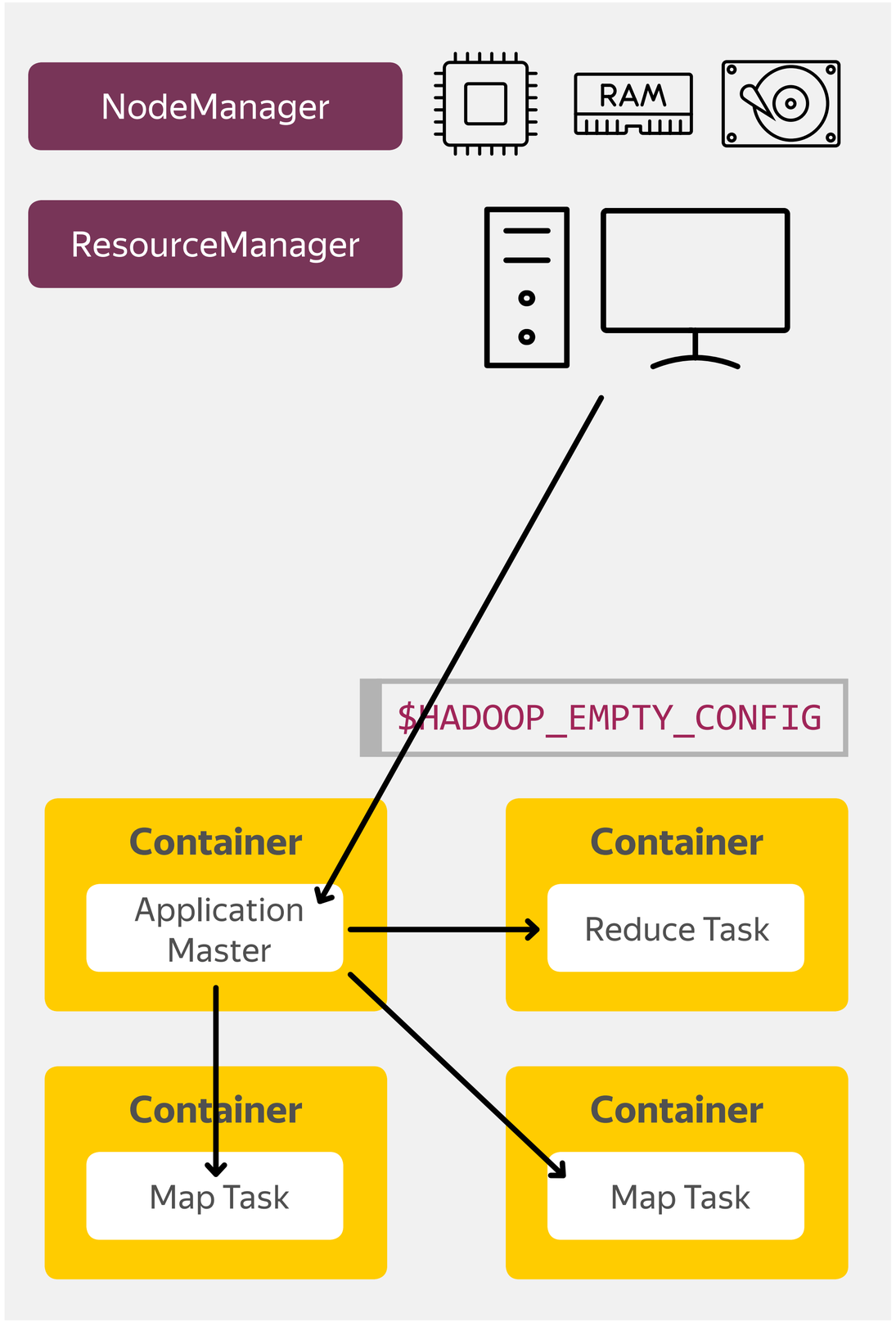





```
hdfs --config $HADOOP_EMPTY_CONFIG dfs -rm -r word_count
yarn --config $HADOOP_EMPTY_CONFIG jar $HADOOP_STREAMING_JAR
```

```
$ locate "hadoop/conf.empty"
```


Acceptance Testing
System Testing
Integration Testing
Unit Testing


NodeManager
RAM
ResourceManager
$HADOOP_EMPTY_CONFIG
Container
Application Master
Container
Reduce Task
Container
Map Task
Container
Map Task

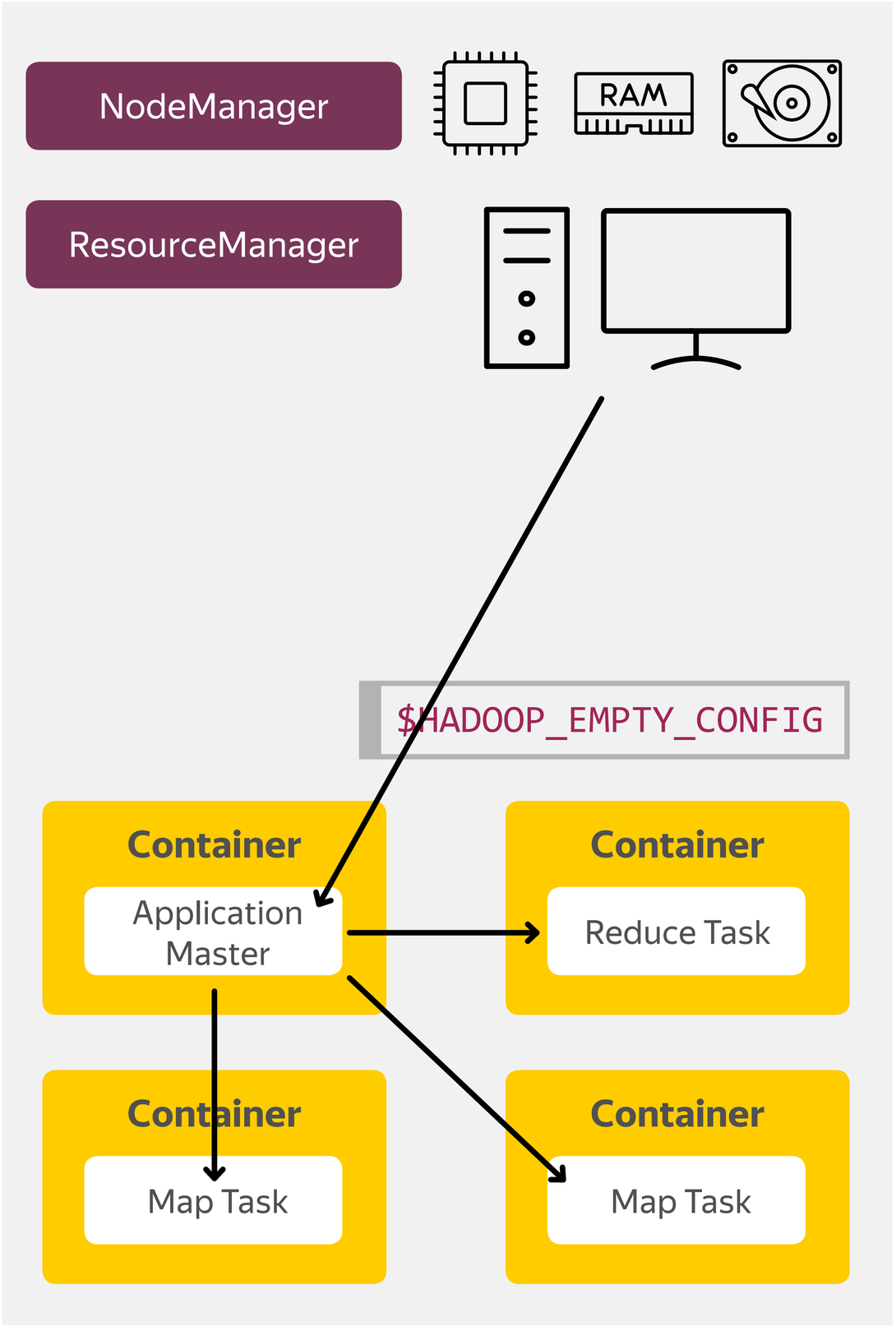
Acceptance Testing

System Testing

Integration Testing

Unit Testing

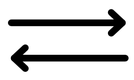
```
        Personal Counters
                word found=182
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=197193
        File Output Format Counters
                Bytes Written=769
17/03/26 18:10:13 INFO streaming.StreamJob: Output directory: w
```



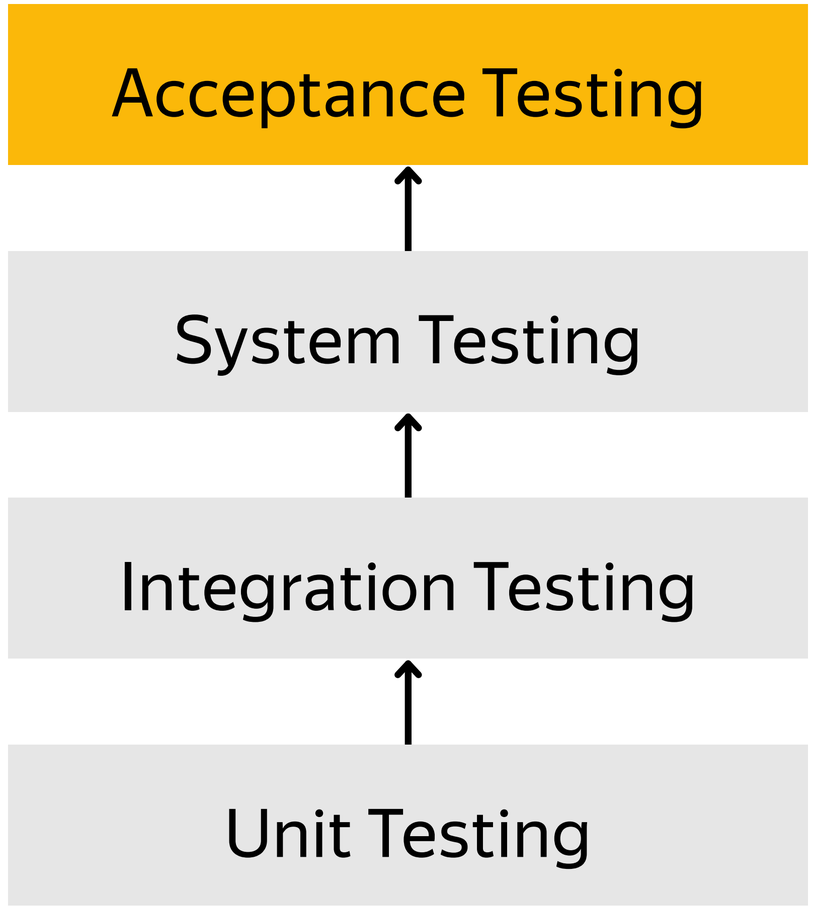

WIKIPEDIA
The Free Encyclopedia


WIKIPEDIA
The Free Encyclopedia


hadoop




Python


Acceptance Testing
System Testing
Integration Testing
Unit Testing

WIKIPEDIA
The Free Encyclopedia

Don't become this guy.

Acceptance Testing

System Testing

Integration Testing

Unit Testing

WIKIPEDIA
The Free Encyclopedia

hadoop

Python

Summary

Summary

- you know **how to use** Python **unit testing**

Summary

- you know **how to use** Python **unit testing**
- you know **how to emulate** MapReduce **locally** with (cat | map | sort | reduce)

Summary

- you know **how to use** Python **unit testing**
- you know **how to emulate** MapReduce **locally** with (cat | map | sort | reduce)
- you know **how to run** MapReduce in a **standalone** mode (hadoop/conf.empty)

Summary

- you know **how to use** Python **unit testing**
- you know **how to emulate** MapReduce **locally** with (cat | map | sort | reduce)
- you know **how to run** MapReduce in a **standalone** mode (hadoop/conf.empty)
- **you know why** you need to execute MapReduce against **sample datasets**

BigDATAteam