streamparse Documentation

Release 4.0.0

Parsely

Oct 07, 2020

Contents

streamparse lets you run Python code against real-time streams of data. Integrates with Apache Storm.

CHAPTER 1

Quickstart

1.1 Dependencies

1.1.1 Java and Clojure

To run local and remote computation clusters, streamparse relies upon a JVM technology called Apache Storm. The integration with this technology is lightweight, and for the most part, you don't need to think about it.

However, to get the library running, you'll need

1. JDK 7+, which you can install with apt-get, homebrew, or an installler; and
2. lein, which you can install from the Leiningen project page or github
3. Apache Storm development environment, which you can install from the Storm project page

   You will need to have at least Apache Storm version 0.10.0 to cooperate with Streamparse.

Confirm that you have `lein` installed by running:

```
> lein version
```

You should get output similar to this:

```
Leiningen 2.3.4 on Java 1.7.0_55 Java HotSpot(TM) 64-Bit Server VM
```

Confirm that you have `storm` installed by running:

```
> storm version
```

You should get output similar to this:

```
Running: java -client -Ddaemon.name= -Dstorm.options= -Dstorm.home=/opt/apache-storm-
↪1.0.1 -Dstorm.log.dir=/opt/apache-storm-1.0.1/logs -Djava.library.path=/usr/local/
↪lib:/opt/local/lib:/usr/lib -Dstorm.conf.file= -cp /opt/apache-storm-1.0.1/lib/
↪reflectasm-1.10.1.jar:/opt/apache-storm-1.0.1/lib/kryo-3.0.3.jar:/opt/apache-storm-
↪1.0.1/lib/log4j-over-slf4j-1.6.6.jar:/opt/apache-storm-1.0.1/lib/clojure-1.7.0.jar:/
↪opt/apache-storm-1.0.1/lib/log4j-slf4j-impl-2.1.jar:/opt/apache-storm-1.0.1/lib/
↪servlet-api-2.5.jar:/opt/apache-storm-1.0.1/lib/disruptor-3.3.2.jar:/opt/apache-
↪storm-1.0.1/lib/objenesis-2.1.jar:/opt/apache-storm-1.0.1/lib/storm-core-1.0.1.jar:/
↪opt/apache-storm-1.0.1/lib/slf4j-api-1.7.7.jar:/opt/apache-storm-1.0.1/lib/storm-
↪rename-hack-1.0.1.jar:/opt/apache-storm-1.0.1/lib/log4j-api-2.1.jar:/opt/apache-
↪storm-1.0.1/lib/log4j-core-2.1.jar:/opt/apache-storm-1.0.1/lib/minlog-1.3.0.jar:/
↪opt/apache-storm-1.0.1/lib/asm-5.0.3.jar:/opt/apache-storm-1.0.1/conf org.apache.
↪storm.utils.VersionInfo
Storm 1.0.1
URL https://git-wip-us.apache.org/repos/asf/storm.git -r␣
↪b5c16f919ad4099e6fb25f1095c9af8b64ac9f91
Branch (no branch)
Compiled by tgoetz on 2016-04-29T20:44Z
From source with checksum 1aea9df01b9181773125826339b9587e
```

If `lein` isn't installed, follow these directions to install it.

If `storm` isn't installed, follow these directions.

Once that's all set, you install streamparse using `pip`:

```
> pip install streamparse
```

1.2 Your First Project

When working with streamparse, your first step is to create a project using the command-line tool, `sparse`:

```
> sparse quickstart wordcount

Creating your wordcount streamparse project...
    create    wordcount
    create    wordcount/.gitignore
    create    wordcount/config.json
    create    wordcount/fabfile.py
    create    wordcount/project.clj
    create    wordcount/README.md
    create    wordcount/src
    create    wordcount/src/bolts/
    create    wordcount/src/bolts/__init__.py
    create    wordcount/src/bolts/wordcount.py
    create    wordcount/src/spouts/
    create    wordcount/src/spouts/__init__.py
    create    wordcount/src/spouts/words.py
    create    wordcount/topologies
    create    wordcount/topologies/wordcount.py
    create    wordcount/virtualenvs
    create    wordcount/virtualenvs/wordcount.txt
Done.
```

Try running your topology locally with:

```
> cd wordcount
  sparse run
```

The quickstart project provides a basic wordcount topology example which you can examine and modify. You can inspect the other commands that `sparse` provides by running:

```
> sparse -h
```

If you see an error like:

```
Local Storm version, 1.0.1, is not the same as the version in your project.clj, 0.10.
↪0. The versions must match.
```

You will have to edit your wordcount/project.clj file and change Apache Storm library version to match the one you have installed.

1.3 Project Structure

streamparse projects expect to have the following directory layout:

File/Folder	Contents
config.json	Configuration information for all of your topologies.
fabfile.py	Optional custom fabric tasks.
project.clj	leiningen project file (can be used to add external JVM dependencies).
src/	Python source files (bolts/spouts/etc.) for topologies.
tasks.py	Optional custom invoke tasks.
topologies/	Contains topology definitions written using the *Topology DSL*.
virtualenvs/	Contains pip requirements files used to install dependencies on remote Storm servers.

1.4 Defining Topologies

Storm's services are Thrift-based and although it is possible to define a topology in pure Python using Thrift. For details see *Topology DSL*.

Let's have a look at the definition file created by using the `sparse quickstart` command.

```
"""
Word count topology
"""

from streamparse import Grouping, Topology

from bolts.wordcount import WordCountBolt
from spouts.words import WordSpout


class WordCount(Topology):
    word_spout = WordSpout.spec()
    count_bolt = WordCountBolt.spec(inputs={word_spout: Grouping.fields("word")},␣
↪par=2)
```

In the `count_bolt` bolt, we've told Storm that we'd like the stream of input tuples to be grouped by the named field `word`. Storm offers comprehensive options for stream groupings, but you will most commonly use a **shuffle** or **fields** grouping:

- **Shuffle grouping**: Tuples are randomly distributed across the bolt's tasks in a way such that each bolt is guaranteed to get an equal number of tuples. This is the default grouping if no other is specified.
- **Fields grouping**: The stream is partitioned by the fields specified in the grouping. For example, if the stream is grouped by the "user-id" field, tuples with the same "user-id" will always go to the same task, but tuples with different "user-id"'s may go to different tasks.

There are more options to configure with spouts and bolts, we'd encourage you to refer to our *Topology DSL* docs or Storm's Concepts for more information.

1.5 Spouts and Bolts

The general flow for creating new spouts and bolts using streamparse is to add them to your `src` folder and update the corresponding topology definition.

Let's create a spout that emits sentences until the end of time:

```
import itertools

from streamparse.spout import Spout


class SentenceSpout(Spout):
    outputs = ['sentence']

    def initialize(self, stormconf, context):
        self.sentences = [
            "She advised him to take a long holiday, so he immediately quit work and␣
↪took a trip around the world",
            "I was very glad to get a present from her",
            "He will be here in half an hour",
            "She saw him eating a sandwich",
        ]
        self.sentences = itertools.cycle(self.sentences)

    def next_tuple(self):
        sentence = next(self.sentences)
        self.emit([sentence])

    def ack(self, tup_id):
        pass  # if a tuple is processed properly, do nothing

    def fail(self, tup_id):
        pass  # if a tuple fails to process, do nothing
```

The magic in the code above happens in the `initialize()` and `next_tuple()` functions. Once the spout enters the main run loop, streamparse will call your spout's `initialize()` method. After initialization is complete, streamparse will continually call the spout's `next_tuple()` method where you're expected to emit tuples that match whatever you've defined in your topology definition.

Now let's create a bolt that takes in sentences, and spits out words:

```
import re

from streamparse.bolt import Bolt

class SentenceSplitterBolt(Bolt):
    outputs = ['word']

    def process(self, tup):
        sentence = tup.values[0]  # extract the sentence
        sentence = re.sub(r"[,.;!\?]", "", sentence)  # get rid of punctuation
        words = [[word.strip()] for word in sentence.split(" ") if word.strip()]
        if not words:
            # no words to process in the sentence, fail the tuple
            self.fail(tup)
            return

        for word in words:
            self.emit([word])
        # tuple acknowledgement is handled automatically
```

The bolt implementation is even simpler. We simply override the default `process()` method which streamparse calls when a tuple has been emitted by an incoming spout or bolt. You are welcome to do whatever processing you would like in this method and can further emit tuples or not depending on the purpose of your bolt.

If your `process()` method completes without raising an Exception, streamparse will automatically ensure any emits you have are anchored to the current tuple being processed and acknowledged after `process()` completes.

If an Exception is raised while `process()` is called, streamparse automatically fails the current tuple prior to killing the Python process.

1.5.1 Failed Tuples

In the example above, we added the ability to fail a sentence tuple if it did not provide any words. What happens when we fail a tuple? Storm will send a "fail" message back to the spout where the tuple originated from (in this case `SentenceSpout`) and streamparse calls the spout's `fail()` method. It's then up to your spout implementation to decide what to do. A spout could retry a failed tuple, send an error message, or kill the topology. See *Dealing With Errors* for more discussion.

1.5.2 Bolt Configuration Options

You can disable the automatic acknowleding, anchoring or failing of tuples by adding class variables set to false for: `auto_ack`, `auto_anchor` or `auto_fail`. All three options are documented in `streamparse.bolt.Bolt`.

Example:

```
from streamparse.bolt import Bolt

class MyBolt(Bolt):

    auto_ack = False
    auto_fail = False

    def process(self, tup):
        # do stuff...
        if error:
```

(continues on next page)

(continued from previous page)

```
            self.fail(tup)  # perform failure manually
        self.ack(tup)  # perform acknowledgement manually
```

1.5.3 Handling Tick Tuples

Ticks tuples are built into Storm to provide some simple forms of cron-like behaviour without actually having to use cron. You can receive and react to tick tuples as timer events with your python bolts using streamparse too.

The first step is to override `process_tick()` in your custom Bolt class. Once this is overridden, you can set the storm option `topology.tick.tuple.freq.secs=<frequency>` to cause a tick tuple to be emitted every `<frequency>` seconds.

You can see the full docs for `process_tick()` in `streamparse.bolt.Bolt`.

Example:

```
from streamparse.bolt import Bolt

class MyBolt(Bolt):

    def process_tick(self, freq):
        # An action we want to perform at some regular interval...
        self.flush_old_state()
```

Then, for example, to cause `process_tick()` to be called every 2 seconds on all of your bolts that override it, you can launch your topology under `sparse run` by setting the appropriate -o option and value as in the following example:

```
$ sparse run -o "topology.tick.tuple.freq.secs=2" ...
```

1.6 Remote Deployment

1.6.1 Setting up a Storm Cluster

See Storm's Setting up a Storm Cluster.

1.6.2 Submit

When you are satisfied that your topology works well via testing with:

```
> sparse run -d
```

You can submit your topology to a remote Storm cluster using the command:

```
sparse submit [--environment <env>] [--name <topology>] [-dv]
```

Before submitting, you have to have at least one environment configured in your project's `config.json` file. Let's create a sample environment called "prod" in our `config.json` file:

```
{
    "serializer": "json",
    "topology_specs": "topologies/",
    "virtualenv_specs": "virtualenvs/",
    "envs": {
        "prod": {
            "user": "storm",
            "nimbus": "storm1.my-cluster.com",
            "workers": [
                "storm1.my-cluster.com",
                "storm2.my-cluster.com",
                "storm3.my-cluster.com"
            ],
            "log": {
                "path": "/var/log/storm/streamparse",
                "file": "pystorm_{topology_name}_{component_name}_{task_id}_{pid}.log
↪",
                "max_bytes": 100000,
                "backup_count": 10,
                "level": "info"
            },
            "use_ssh_for_nimbus": true,
            "virtualenv_root": "/data/virtualenvs/"
        }
    }
}
```

We've now defined a `prod` environment that will use the user `storm` when deploying topologies. Before submitting the topology though, streamparse will automatically take care of instaling all the dependencies your topology requires. It does this by sshing into everyone of the nodes in the `workers` config variable and building a virtualenv using the the project's local `virtualenvs/<topology_name>.txt` requirements file.

This implies a few requirements about the user you specify per environment:

1. Must have ssh access to all servers in your Storm cluster
2. Must have write access to the `virtualenv_root` on all servers in your Storm cluster

If you would like to use your system user for creating the SSH connection to the Storm cluster, you can omit the `user` setting from your `config.json`.

By default the `root` user is used for creating virtualenvs when you do not specify a `user` in your `config.json`. To override this, set the `sudo_user` option in your `config.json`. `sudo_user` will default to `user` if one is specified.

streamparse also assumes that virtualenv is installed on all Storm servers.

Once an environment is configured, we could deploy our wordcount topology like so:

```
> sparse submit
```

Seeing as we have only one topology and environment, we don't need to specify these explicitly. streamparse will now:

1. Package up a JAR containing all your Python source files
2. Build a virtualenv on all your Storm workers (in parallel)
3. Submit the topology to the `nimbus` server

1.6.3 Disabling & Configuring Virtualenv Creation

If you do not have ssh access to all of the servers in your Storm cluster, but you know they have all of the requirements for your Python code installed, you can set `"use_virtualenv"` to `false` in `config.json`.

If you have virtualenvs on your machines that you would like streamparse to use, but not update or manage, you can set `"install_virtualenv"` to `false` in `config.json`.

If you would like to pass command-line flags to virtualenv, you can set `"virtualenv_flags"` in `config.json`, for example:

```
"virtualenv_flags": "-p /path/to/python"
```

Note that this only applies when the virtualenv is created, not when an existing virtualenv is used.

If you would like to share a single virtualenv across topologies, you can set `"virtualenv_name"` in `config.json` which overrides the default behaviour of using the topology name for virtualenv. Updates to a shared virtualenv should be done after shutting down topologies, as code changes in running topologies may cause errors.

1.6.4 Using unofficial versions of Storm

If you wish to use streamparse with unofficial versions of storm (such as the HDP Storm) you should set `:repositories` in your `project.clj` to point to the Maven repository containing the JAR you want to use, and set the version in `:dependencies` to match the desired version of Storm.

For example, to use the version supplied by HDP, you would set `:repositories` to:

```
:repositories {"HDP Releases" "http://repo.hortonworks.com/content/
repositories/releases"}
```

1.6.5 Local Clusters

Streamparse assumes that your Storm cluster is not on your local machine. If it is, such as the case with VMs or Docker images, change `"use_ssh_for_nimbus"` in `config.json` to `false`.

1.6.6 Setting Submit Options in config.json

If you frequently use the same options to `sparse submit` in your project, you can set them in `config.json` using the `options` key in your environment settings. For example:

```
{
    "topology_specs": "topologies/",
    "virtualenv_specs": "virtualenvs/",
    "envs": {
        "vagrant": {
            "user": "vagrant",
            "nimbus": "streamparse-box",
            "workers": [
                "streamparse-box"
            ],
            "virtualenv_root": "/data/virtualenvs",
            "options": {
                "topology.environment": {
                    "LD_LIBRARY_PATH": "/usr/local/lib/"
                }
```

(continues on next page)

(continued from previous page)

```
            }
        }
    }
}
```

You can also set the `--worker` and `--acker` parameters in `config.json` via the `worker_count` and `acker_count` keys in your environment settings.

```
{
    "topology_specs": "topologies/",
    "virtualenv_specs": "virtualenvs/",
    "envs": {
        "vagrant": {
            "user": "vagrant",
            "nimbus": "streamparse-box",
            "workers": [
                "streamparse-box"
            ],
            "virtualenv_root": "/data/virtualenvs",
            "acker_count": 1,
            "worker_count": 1
        }
    }
}
```

1.6.7 Logging

The Storm supervisor needs to have access to the `log.path` directory for logging to work (in the example above, `/var/log/storm/streamparse`). If you have properly configured the `log.path` option in your config, streamparse will use the value for the `log.file` option to set up log files for each Storm worker in this path. The filename can be customized further by using certain named placeholders. The default filename is set to:

```
pystorm_{topology_name}_{component_name}_{task_id}_{pid}.log
```

Where:

- `topology_name`: is the `topology.name` variable set in Storm
- `component_name`: is the name of the currently executing component as defined in your topology definition file (.clj file)
- `task_id`: is the task ID running this component in the topology
- `pid`: is the process ID of the Python process

streamparse uses Python's `logging.handlers.RotatingFileHandler` and by default will only save 10 1 MB log files (10 MB in total), but this can be tuned with the `log.max_bytes` and `log.backup_count` variables.

The default logging level is set to `INFO`, but if you can tune this with the `log.level` setting which can be one of critical, error, warning, info or debug. **Note** that if you perform `sparse run` or `sparse submit` with the `--debug` set, this will override your `log.level` setting and set the log level to debug.

When running your topology locally via `sparse run`, your log path will be automatically set to `/path/to/your/streamparse/project/logs`.

New in version 3.0.0.

CHAPTER 2

Topologies

Storm topologies are described as a Directed Acyclic Graph (DAG) of Storm components, namely *bolts* and *spouts*.

2.1 Topology DSL

To simplify the process of creating Storm topologies, streamparse features a Python Topology DSL. It lets you specify topologies as complex as those you can in Java or Clojure, but in concise, readable Python.

Topology files are located in `topologies` in your streamparse project folder. There can be any number of topology files for your project in this directory.

- `topologies/my_topology.py`
- `topologies/my_other_topology.py`
- `topologies/my_third_topology.py`
- ...

A valid `Topology` may only have `Bolt` and `Spout` attributes.

2.1.1 Simple Python Example

The first step to putting together a topology, is creating the bolts and spouts, so let's assume we have the following bolt and spout:

```
from collections import Counter

from redis import StrictRedis

from streamparse import Bolt


class WordCountBolt(Bolt):
```

(continues on next page)

(continued from previous page)

```
    outputs = ["word", "count"]

    def initialize(self, conf, ctx):
        self.counter = Counter()
        self.total = 0

    def _increment(self, word, inc_by):
        self.counter[word] += inc_by
        self.total += inc_by

    def process(self, tup):
        word = tup.values[0]
        self._increment(word, 10 if word == "dog" else 1)
        if self.total % 1000 == 0:
            self.logger.info("counted %i words", self.total)
        self.emit([word, self.counter[word]])


class RedisWordCountBolt(Bolt):
    outputs = ["word", "count"]
```

```
from itertools import cycle

from streamparse import Spout


class WordSpout(Spout):
    outputs = ["word"]

    def initialize(self, stormconf, context):
        self.words = cycle(["dog", "cat", "zebra", "elephant"])

    def next_tuple(self):
        word = next(self.words)
        self.emit([word])
```

One important thing to note is that we have added an `outputs` attribute to these classes, which specify the names of the output fields that will be produced on their `default` streams. If we wanted to specify multiple streams, we could do that by specifying a list of `Stream` objects.

Now let's hook up the bolt to read from the spout:

```
"""
Word count topology (in memory)
"""

from bolts import WordCountBolt
from spouts import WordSpout
from streamparse import Grouping, Topology


class WordCount(Topology):
    word_spout = WordSpout.spec()
    count_bolt = WordCountBolt.spec(inputs={word_spout: Grouping.fields("word")},␣
↪par=2)
```

Note: Your project's `src` directory gets added to `sys.path` before your topology is imported, so you should use absolute imports based on that.

As you can see, `streamparse.Bolt.spec()` and `streamparse.Spout.spec()` methods allow us to specify information about the components in your topology and how they connect to each other. Their respective docstrings outline all of the possible ways they can be used.

2.1.2 Java Components

The topology DSL fully supports JVM-based bolts and spouts via the `JavaBolt` and `JavaSpout` classes.

Here's an example of how we would use the Storm Kafka Spout:

```
"""
Pixel count topology
"""

from streamparse import Grouping, JavaSpout, Topology

from bolts.pixel_count import PixelCounterBolt
from bolts.pixel_deserializer import PixelDeserializerBolt


class PixelCount(Topology):
    pixel_spout = JavaSpout.spec(
        name="pixel-spout",
        full_class_name="pixelcount.spouts.PixelSpout",
        args_list=[],
        outputs=["pixel"],
    )
    pixel_deserializer = PixelDeserializerBolt.spec(
        name="pixel-deserializer-bolt", inputs=[pixel_spout]
    )
    pixel_counter = PixelCounterBolt.spec(
        name="pixel-count-bolt",
        inputs={pixel_deserializer: Grouping.fields("url")},
        config={"topology.tick.tuple.freq.secs": 1},
    )
```

One limitation of the Thrift interface we use to send the topology to Storm is that the constructors for Java components can only be passed basic Python data types: *bool*, *bytes*, *float*, *int*, and *str*.

2.1.3 Components in Other Languages

If you have components that are written in languages other than Java or Python, you can have those as part of your topology as well—assuming you're using the corresponding multi-lang library for that language.

To do that you just need to use the `streamparse.ShellBolt.spec()` and `streamparse.ShellSpout.spec()` methods. They take `command` and `script` arguments to specify a binary to run and its string-separated arguments.

2.1.4 Multiple Streams

To specify that a component has multiple output streams, instead of using a list of strings for `outputs`, you must specify a list of `Stream` objects, as shown below.

```
class FancySpout(Spout):
    outputs = [Stream(fields=['good_data'], name='default'),
               Stream(fields=['bad_data'], name='errors')]
```

To select one of those streams as the input for a downstream `Bolt`, you simply use `[]` to specify the stream you want. Without any stream specified, the `default` stream will be used.

```
class ExampleTopology(Topology):
    fancy_spout = FancySpout.spec()
    error_bolt = ErrorBolt.spec(inputs=[fancy_spout['errors']])
    process_bolt = ProcessBolt.spec(inputs=[fancy_spout])
```

2.1.5 Groupings

By default, Storm uses a `SHUFFLE` grouping to route tuples to particular executors for a given component, but you can also specify other groupings by using the appropriate `Grouping` attribute. The most common grouping is probably the `fields()` grouping, which will send all the tuples with the same value for the specified fields to the same executor. This can be seen in the prototypical word count topology:

```
"""
Word count topology (in memory)
"""

from bolts import WordCountBolt
from spouts import WordSpout
from streamparse import Grouping, Topology


class WordCount(Topology):
    word_spout = WordSpout.spec()
    count_bolt = WordCountBolt.spec(inputs={word_spout: Grouping.fields("word")},␣
↪par=2)
```

2.1.6 Topology Cycles

On rare occasions, you may want to create a cyclical topology. This may not seem easily done with the current topology DSL, but there is a workaround you can use: manually declaring a temporary lower-level *:class:~streamparse.thrift.GlobalStreamId* that you can refer to in multiple places.

The following code creates a `Topology` with a cycle between its two Bolts.

```
from streamparse.thrift import GlobalStreamId

# Create a reference to B's output stream before we even declare Topology
b_stream = GlobalStreamId(componentId='b_bolt', streamId='default')

class CyclicalTopology(Topology):
    some_spout = SomeSpout.spec()
    # Include our saved stream in your list of inputs for A
```

(continues on next page)

(continued from previous page)

```
    a_bolt = A.spec(name="A", inputs=[some_spout, b_stream])
    # Have B get input from A like normal
    b_bolt = B.spec(name="B", inputs=[a_bolt])
```

2.1.7 Topology-Level Configuration

If you want to set a config option for all components in your topology, like `topology.environment`, you can do that by adding a `config` class attribute to your `Topology` that is a *dict* mapping from option names to their values. For example:

```
class WordCount(Topology):
    config = {'topology.environment': {'LD_LIBRARY_PATH': '/usr/local/lib/'}}
    ...
```

2.2 Running Topologies

2.2.1 What Streamparse Does

When you run a topology either locally or by submitting to a cluster, streamparse will

1. Bundle all of your code into a JAR
2. Build a Thrift Topology struct out of your Python topology definition.
3. Pass the Thrift Topology struct to Nimbus on your Storm cluster.

If you invoked streamparse with `sparse run`, your code is executed directly from the `src/` directory.

If you submitted to a cluster with `sparse submit`, streamparse uses `lein` to compile the `src` directory into a jar file, which is run on the cluster. Lein uses the `project.clj` file located in the root of your project. This file is a standard lein project file and can be customized according to your needs.

2.2.2 Dealing With Errors

When detecting an error, bolt code can call its `fail()` method in order to have Storm call the respective spout's `fail()` method. Known error/failure cases result in explicit callbacks to the spout using this approach.

Exceptions which propagate without being caught will cause the component to crash. On `sparse run`, the entire topology will stop execution. On a running cluster (i.e. `sparse submit`), Storm will auto-restart the crashed component and the spout will receive a `fail()` call.

If the spout's fail handling logic is to hold back the tuple and not re-emit it, then things will keep going. If it re-emits it, then it may crash that component again. Whether the topology is tolerant of the failure depends on how you implement failure handling in your spout.

Common approaches are to:

- Append errant tuples to some sort of error log or queue for manual inspection later, while letting processing continue otherwise.
- Attempt 1 or 2 retries before considering the tuple a failure, if the error was likely an transient problem.
- Ignore the failed tuple, if appropriate to the application.

2.3 Parallelism and Workers

In general, use the ‘‘par‘‘ “parallelism hint” parameter per spout and bolt in your configuration to control the number of Python processes per component.

Reference: Understanding the Parallelism of a Storm Topology

Storm parallelism entities:

- A *worker process* is a JVM, i.e. a Java process.
- An *executor* is a thread that is spawned by a worker process.
- A *task* performs the actual data processing. (To simplify, you can think of it as a Python callable.)

Spout and bolt specs take a `par` keyword to provide a parallelism hint to Storm for the number of executors (threads) to use for the given spout/bolt; for example, `par=2` is a hint to use two executors. Because streamparse implements spouts and bolts as independent Python processes, setting `par=N` results in N Python processes for the given spout/bolt.

Many streamparse applications will need only to set this parallelism hint to control the number of resulting Python processes when tuning streamparse configuration. For the underlying topology workers, streamparse sets a default of 2 workers, which are independent JVM processes for Storm. This allows a topology to continue running when one worker process dies; the other is around until the dead process restarts.

Both `sparse run` and `sparse submit` accept a `-p N` command-line flag to set the number of topology workers to N. For convenience, this flag also sets the number of Storm's underlying messaging reliability *acker bolts* to the same N value. In the event that you need it (and you understand Storm ackers), use the `-a` and `-w` command-line flags instead of `-p` to control the number of acker bolts and the number of workers, respectively. The `sparse` command does not support Storm's rebalancing features; use `sparse submit -f -p N` to kill the running topology and redeploy it with N workers.

Note that the underlying Storm thread implementation, LMAX Disruptor, is designed with high-performance inter-thread messaging as a goal. Rule out Python-level issues when tuning your topology:

- bottlenecks where the number of spout and bolt processes are out of balance
- serialization/deserialization overhead of more data emitted than you need
- slow routines/callables in your code

CHAPTER 3

API

3.1 Tuples

You should never have to instantiate an instance of a `streamparse.Tuple` yourself as streamparse handles this for you prior to, for example, a `streamparse.Bolt`'s `process()` method being called.

None of the emit methods for bolts or spouts require that you pass a `streamparse.Tuple` instance.

3.2 Components

Both `streamparse.Bolt` and `streamparse.Spout` inherit from a common base-class, `streamparse.storm.component.Component`. It extends pystorm's code for handling Multi-Lang IPC between Storm and Python and adds suport for our Python *Topology DSL*.

3.2.1 Spouts

Spouts are data sources for topologies, they can read from any data source and emit tuples into streams.

3.2.2 Bolts

3.3 Logging

3.4 Topology DSL

CHAPTER 4

Developing Streamparse

4.1 Lein

Install Leiningen according to the instructions in the quickstart.

4.2 Local pip installation

In your virtualenv for this project, go into `~/repos/streamparse` (where you cloned streamparse) and simply run:

```
python setup.py develop
```

This will install a streamparse Python version into the virtualenv which is essentially symlinked to your local version.

NOTE: streamparse currently pip installs streamparse's **released** version on remote clusters automatically. Therefore, though this will work for local development, you'll need to push streamparse somewhere pip installable (or change requirements.txt) to make it pick up that version on a remote cluster.

4.3 Installing Storm pre-releases

You can clone Storm from Github here:

```
git clone git@github.com:apache/storm.git
```

There are tags available for releases, e.g.:

```
git checkout v1.0.1
```

To build a local Storm release, use:

```
mvn install
cd storm-dist/binary
mvn package
```

These steps will take awhile as they also run Storm's internal unit and integration tests.

The first line will actually install Storm locally in your maven (.m2) repository. You can confirm this with:

```
ls ~/.m2/repository/org/apache/storm/storm-core/1.0.1
```

You should now be able to change your `project.clj` to include a reference to this new release.

Once you change that, you can run:

```
lein deps :tree | grep storm
```

To confirm it is using the upgraded Clojure 1.5.1 (changed in 0.9.2), run:

```
lein repl
```

CHAPTER 5

Frequently Asked Questions (FAQ)

5.1 General Questions

5.1.1 Why use streamparse?

To lay your Python code out in topologies which can be automatically parallelized in a Storm cluster of machines. This lets you scale your computation horizontally and avoid issues related to Python's GIL. See *Parallelism and Workers*.

5.1.2 Is streamparse compatible with Python 3?

Yes, streamparse is fully compatible with Python 3 starting with version 3.3 which we use in our unit tests.

5.1.3 How can I contribute to streamparse?

Please see the CONTRIBUTING document in Github

5.1.4 How do I trigger some code before or after I submit my topology?

After you create a streamparse project using `sparse quickstart`, you'll have a `fabfile.py` in that directory. In that file, you can specify two functions (`pre_submit` and `post_submit`) which are expected to accept four arguments:

- `topology_name`: the name of the topology being submitted
- `env_name`: the name of the environment where the topology is being submitted (e.g. `"prod"`)
- `env_config`: the relevant config portion from the `config.json` file for the environment you are submitting the topology to
- `options`: the fully resolved Storm options

Here is a sample `fabfile.py` file that sends a message to IRC after a topology is successfully submitted to prod.

```
# my_project/fabfile.py
from __future__ import absolute_import, print_function, unicode_literals

from my_project import write_to_irc


def post_submit(topo_name, env_name, env_config):
    if env_name == "prod":
        write_to_irc("Deployed {} to {}".format(topo_name, env_name))
```

5.1.5 How should I install streamparse on the cluster nodes?

streamparse assumes your Storm servers have Python, pip, and virtualenv installed. After that, the installation of all required dependencies (including streamparse itself) is taken care of via the *config.json* file for the streamparse project and the `sparse submit` command.

5.1.6 Should I install Clojure?

No, the Java requirements for streamparse are identical to that of Storm itself. Storm requires Java and bundles Clojure as a requirement, so you do not need to do any separate installation of Clojure. You just need Java on all Storm servers.

How do I deploy into a VPC?

Update your `~/.ssh/config` to use a bastion host inside your VPC for your commands:

```
Host *.internal.example.com
    ProxyCommand ssh bastion.example.com exec nc %h %p
```

If you don't have a common subdomain you'll have to list all of the hosts individually:

```
Host host1.example.com
    ProxyCommand ssh bastion.example.com exec nc %h %p
...
```

Set up your streamparse config to use all of the hosts normally (without bastion host).

How do I override SSH settings?

It is highly recommended that you just modify your `~/.ssh/config` file if you need to tweak settings for setting up the SSH tunnel to your Nimbus server, but you can also set your SSH password or port in your `config.json` by setting the `ssh_password` or `ssh_port` environment settings.

```
{
    "topology_specs": "topologies/",
    "virtualenv_specs": "virtualenvs/",
    "envs": {
        "prod": {
            "user": "somebody",
            "ssh_password": "THIS IS A REALLY BAD IDEA",
            "ssh_port": 52,
            "nimbus": "streamparse-box",
            "workers": [
                "streamparse-box"
            ],
            "virtualenv_root": "/data/virtualenvs"
        }
    }
}
```

How do I dynamically generate the worker list?

In a small cluster it's sufficient to specify the list of workers in `config.json`. However, if you have a large or complex environment where workers are numerous or short-lived, `streamparse` supports querying the nimbus server for a list of hosts.

An undefined list (empty or None) of `workers` will trigger the lookup. Explicitly defined hosts are preferred over a lookup.

Lookups are configured on a per-environment basis, so the `prod` environment below uses the dynamic lookup, while `beta` will not.

```
{
    "topology_specs": "topologies/",
    "virtualenv_specs": "virtualenvs/",
    "envs": {
        "prod": {
            "nimbus": "streamparse-prod",
            "virtualenv_root": "/data/virtualenvs"
        },
        "beta": {
            "nimbus": "streamparse-beta",
            "workers": [
                "streamparse-beta"
            ],
            "virtualenv_root": "/data/virtualenvs"
        }
    }
}
```

CHAPTER 6

Indices and tables

- genindex
- modindex
- search