Project Selection

We want to select 10-30 sizable projects that are fairly young (for mature CI usage) and that have been using CI from the beginning to ensure every commits are covered by CI.

Requirements

- NLOC >= 5,000
- Use of Travis CI
- Have a baseline of less than 6 years of CI use
  - Have been using CI from the beginning of the project
- The project has at least one of each of the 16 bugType template.

Methodology

Information on data format – https://travistorrent.testroots.org/page_dataformat/

Phase 1

Select projects that are in both datasets (Sstubs and TravisTorrent) 1. Query the TravisTorrent dataset to obtain all the project names and output to JSON (SELECT * is too big and right now we only want to select the project names) from GoogleCloud Platform (BigQuery) 2. Dump the JSON data into my existing SQLITe database (sstubs.db). 3. Query the tables to select projects that are in both tables.

```
SELECT DISTINCT projectName FROM commits WHERE projectName IN(SELECT * from names)
```

Output – 34 results

Projects both in Sstubs and TravisTorrent
Graylog2.graylog2-server
apache.flink
apache.storm
aws.aws-sdk-java
brettwooldridge.HikariCP
brianfrankcooper.YCSB
checkstyle.checkstyle
code4craft.webmagic
deeplearning4j.deeplearning4j
dropwizard.dropwizard
dropwizard.metrics
druid-io.druid
facebook.presto
google.auto
google.closure-compiler
google.guava
google.guice
iluwatar.java-design-patterns
javaee-samples.javaee7-samples
jhy.jsoup
joelittlejohn.jsonschema2pojo
junit-team.junit
knightliao.disconf
mybatis.mybatis-3
naver.pinpoint
perwendel.spark

Projects both in Sstubs and TravisTorrent
roboguice.roboguice
springside.springside4
square.dagger
square.javapoet
square.okhttp
square.retrofit
thinkaurelius.titan
xetorthio.jedis

Phase 2

Filter the list according to the requirements: manual verification of the GitHub repo - `.travis.yml` history - Initial commit - Current CI pipeline - Number of distinct bugType

Projects both in Sstubs and TravisTorrent

Project Age

Travis History

Still using Travis?

Nb of (distinct) bugType

Graylog2.graylog2-server

10 years

8 years

no

12

apache.flink

8 years

7 years

no

15

apache.storm

9 years

5 years

yes

12

aws.aws-sdk-java

11 years

7 years

yes

6

brettwooldridge.HikariCP

7 years

7 years

yes

9

brianfrankcooper.YCSB

11 years

6 years

yes

7

checkstyle.checkstyle

20 years

7 years

yes

12

code4craft.webmagic

8 years

7 years

yes

5

deeplearning4j.deeplearning4j

-

-

-

dropwizard.dropwizard

9 years

7 years

yes

7

dropwizard.metrics

9 years

7 years

yes

8

druid-io.druid

8 years

6 years

yes

14

facebook.presto

8 years

7 years

yes

15

google.auto

8 years

7 years

yes

5

google.closure-compiler

11 years

6 years

no

15

google.guava

11 years

6 years

yes

8

google.guice

15 years

6 years

yes

7

iluwatar.java-design-patterns

6 years

6 years

no

2

javaee-samples.javaee7-samples

7 year

6 years

yes

4

jhy.jsoup

10 years

6 years

no

6

joelittlejohn.jsonschema2pojo

10 years

7 years

yes

4

junit-team.junit

19 years

7 years

yes

7

knightliao.disconf

6 years

6 years

yes

1

mybatis.mybatis-3

11 years

7 years

no

8

naver.pinpoint

8 years

5 years

yes

9

perwendel.spark

10 years

7 years

yes

6

roboguice.roboguice

12 years

8 years

yes

2

springside.springside4

8 years

7 years

yes

5

square.dagger

9 years

8 years

yes

1

square.javapoet

8 years

7 years

yes

4

square.okhttp

9 years

8 years

no

4

square.retrofit

9 years

8 years

no

2

thinkaurelius.titan

9 years

8 years

no

7

xetorthio.jedis

10 years

6 years

no

10

deeplearning4j did not have any build history on Travis, so it will be excluded from further selection phases.

It is difficult to know if all commits are covered by CI. Once we narrow down the project list, we can: 1. Create a smaller TravisTorrent dataset containing the projects we selected. 2. Create a smaller Sstubs dataset containing commits related to the projects we selected 3. Verify if the sstubs commits are in the travis dataset. 4. Create new table with sstubs commits that are also in travis torrent.

Phase 3

Threshold for the number of distinct bugType: at least 10

Projects both in Sstubs and TravisTorrent

Project Age

Travis History

Still using Travis?

Nb of (distinct) bugType

Graylog2.graylog2-server

10 years

8 years

no

12

apache.flink

8 years

7 years

no

15

apache.storm

9 years

5 years

yes

12

checkstyle.checkstyle

20 years

7 years

yes

12

druid-io.druid

8 years

6 years

yes

14

facebook.presto

8 years

7 years

yes

15

google.closure-compiler

11 years

6 years

no

15

xetorthio.jedis

10 years

6 years

no

10

Phase 4

Step 1 – Query TravisTorrent dataset to collect all commits from projects we selected

Query for each project to avoid API limitation. Save each result to JSON file and dump through Python pipeline to go into our SQLite database.

The reason we query for each project is because querying for all 8 project at once results in exceeding the API limit to save the output.

```
SELECT * from travistorrent-bq.data.2017_01_11 WHERE gh_project_name = 'Graylog2/graylog2-server'

SELECT * from travistorrent-bq.data.2017_01_11 WHERE gh_project_name = 'apache/flink'

SELECT * from travistorrent-bq.data.2017_01_11 WHERE gh_project_name = 'apache/storm'

SELECT * from travistorrent-bq.data.2017_01_11 WHERE gh_project_name = 'checkstyle/checkstyle'

SELECT * from travistorrent-bq.data.2017_01_11 WHERE gh_project_name = 'druid-io/druid'

SELECT * from travistorrent-bq.data.2017_01_11 WHERE gh_project_name = 'facebook/presto'

SELECT * from travistorrent-bq.data.2017_01_11 WHERE gh_project_name = 'google/closure-compiler'

SELECT * from travistorrent-bq.data.2017_01_11 WHERE gh_project_name = 'xetorthio/jedis'
```

From the query results, I created a new table called `travis_builds`.

Step 2 – Query Sstubs dataset to collect all commits from projects we selected

```
SELECT * FROM commits WHERE projectName IN ('Graylog2.graylog2-server', 'apache.flink',
'apache.storm', 'checkstyle.checkstyle', 'druid-io.druid', 'facebook.presto', 'google.closure-compiler'
'xetorthio.jedis')
```

With the query results, I created a new table called `sstubs_commits`.

After step 1 and step 2, I created a new database to hold the two newly created tables.

Step 3 – Find out if Sstubs commits are in TravisTorrent builds

For this, I will do a `LEFT JOIN` on the two tables so I get all the columns from `sstubs_commits` and all the columns from `travis_builds`.

The query verifies that the Sstubs commit (fixCommitSha1 column) is equal to the git_trigger_commit column of a travis build. Then, it verifies that the row is not null for the travis_builds columns by checking that the git_project_name column is not null.

```
SELECT * FROM sstubs_commits LEFT JOIN travis_builds ON sstubs_commits.fixCommitSha1 =
travis_builds.git_trigger_commit  WHERE travis_builds.gh_project_name IS NOT NULL
```

This yields 1,199 rows containing multiple duplicates. This is because there is no unique way to group the results with one column. The only unique combination we can get is by group by (bugType, fixCommitSha1). Since a commit can have multiple bugType, this condition allows us to cover the whole set without having duplicates.

```
SELECT * FROM sstubs_commits LEFT JOIN travis_builds ON sstubs_commits.fixCommitSha1 =
travis_builds.git_trigger_commit  WHERE travis_builds.gh_project_name IS NOT NULL GROUP
BY sstubs_commits.bugType, sstubs_commits.fixCommitSHA1
```

This yields 131 bug fixes. New table created with the query results.

Commit Guru

All projects EXCEPT apache.flink were either found or added for analysis to Commit Guru. Hence, that leaves us with a total of 7 projects to work with.

It was impossible to add apache.flink for some odd reasons, the repo could not be found on commit guru and when trying to add it, it did not work.

Projects

Commit Guru Link

Graylog2.graylog2-server

http://commit.guru/repo/graylog2-server(master)

apache.storm

checkstyle.checkstyle

druid-io.druid

http://commit.guru/repo/druid(master)-1

facebook.presto

google.closure-compiler

xetorthio.jedis