Bug-Fix Dataset Description

This dataset brings information regarding the static and dynamic perspectives of 10 years (since 2009-01-01 until 2019-01-02) of bug fixing activity of Apache Hadoop[[1]](#footnote-0). The static perspective describes, for each bug opened and fixed in this period of time, a snapshot of the bug fixing report and the changes performed in the source code. The second one brings information regarding the bug fixing activity over time, tracking the events (changes in the bug report, added comments, and performed commits) occurred since the bug was first reported until it was considered fixed.

1. Project List

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **List of Target Projects** | | | | |
| **Name** | **Id** | **Category** | **1st Release** | **Site** |
| Hadoop Core | HADOOP | big-data | 2006 | http://hadoop.apache.org/ |
| Hadoop Yarn | YARN | big-data | 2012 | http://hadoop.apache.org/ |
| Hadoop HDFS | HDFS | big-data | 2009 | http://hadoop.apache.org/ |
| Hadoop MapReduce | MAPREDUCE | big-data | 2009 | http://hadoop.apache.org/ |

1. Static - Bug-Fix Dataset

Dataset location:

./snapshot/

Dataset file name format:

<jira-project-id>-full-bug-fix-dataset.csv (jira+git infos)

|  |  |  |  |
| --- | --- | --- | --- |
| **Fields Description of the Static Bug-Fix Dataset** | | | |
| **Field** | **From** | **Type** | **Description** |
| Project | Jira | General | The project to which the bug belongs |
| Owner | Jira | General | The owner/maintainer of the project |
| Manager | Jira | General | The committee responsible for the project |
| Category | Jira | General | The project domain category |
| Key | Jira | General | A unique identifier for the bug |
| Priority | Jira | General | The importance of the issue in relation to other issues. Blocker — Highest priority. Indicates that this issue takes precedence over all others. | Critical — Indicates that this issue is causing a problem and requires urgent attention. | Major — Indicates that this issue has a significant impact. | Minor — Indicates that this issue has a relatively minor impact. | Trivial — Lowest priority. |
| Status | Jira | General | The stage the bug is currently at in its lifecycle |
| Reporter | Jira | General | The person who entered the bug into the system |
| Assignee | Jira | General | The person who was assigned to fix the bug |
| Components | Jira | General | The project component(s) to which the bug relates |
| SummaryTopWords | Jira | Text | 1000 top most frequent words of a brief one-line summary of the bug |
| DescriptionTopWords | Jira | Text | 1000 top most frequent words of a detailed description of the bug |
| CommentsTopWords | Jira | Text | 1000 top most frequent words of the comments of the bug |
| CreationDate | Jira | Time | The date and time on which the bug was entered into Jira |
| ResolutionDate | Jira | Time | The date and time on which the bug was resolved |
| LastUpdateDate | Jira | Time | The date and time on which the bug report was updated for the last time |
| AffectsVersions | Jira | Versioning | The project version(s) for which the bug is (or was) manifesting |
| FixVersions | Jira | Versioning | The project version(s) in which the bug was (or will be) fixed |
| NoComments | Jira | Summation | The number of all comments added in the bug |
| FirstCommentDate | Jira | Time | The date and time on which the first comment was added in the bug report |
| LastCommentDate | Jira | Time | The date and time on which the last comment was added in the bug report |
| NoWatchers | Jira | Summation | The number of how many people are watching (interested in) the bug |
| NoAttachments | Jira | Summation | The number of attachments (e.g., documents, images, and screenshots) added in the bug report |
| FirstAttachmentDate | Jira | Time | The date and time on which the first attachment was added in the bug report |
| LastAttachmentDate | Jira | Time | The date and time on which the last attachment was added in the bug report |
| NoAttachedPatches | Jira | Summation | The number of patches added in the bug report |
| FirstAttachedPatchDate | Jira | Time | The date and time on which the first patch was attached in the bug report |
| LastAttachedPatchDate | Jira | Time | The date and time on which the last patch was attached in the bug report |
| InwardIssueLinks | Jira | Link | The issue on which the bug depends on. (type:inwardIssue). Ex. "the bug is blocked by issue X" |
| OutwardIssueLinks | Jira | Link | The issues that depend on the bug. (type:outwardIssue). Ex. "the bug blocks issue X" |
| HasMergeCommit | Git | Versioning | This field informs whether a bug fix had or not a merge commit |
| CommitsMessagesTopWords | Git | Text | 1000 top most frequent words of the commit messages related to the bug |
| NoCommits | Git | Summation | The number of commits related to the bug |
| NoAuthors | Git | Summation | The number of different authors who performed commits related to the bug (The author is the person who originally wrote the code) |
| NoCommitters | Git | Summation | The number of different committers who performed commits related to the bug (The committer is assumed to be the person who committed the code on behalf of the original author) |
| AuthorsFirstCommitDate | Git | Time | The date and time on which the author performed the first commit to fix the bug |
| AuthorsLastCommitDate | Git | Time | The date and time on which the author performed the last commit to fix the bug |
| CommittersFirstCommitDate | Git | Time | The date and time on which the committer performed the first commit to fix the bug |
| CommittersLastCommitDate | Git | Time | The date and time on which the committer performed the last commit to fix the bug |
| NonSrcAddFiles | Git | SRC | The number of non-source code files added by commits related to the bug |
| NonSrcDelFiles | Git | SRC | The number of non-source code files deleted by commits related to the bug |
| NonSrcModFiles | Git | SRC | The number of non-source code files modified by commits related to the bug |
| NonSrcAddLines | Git | SRC | The number of non-source code lines added by commits related to the bug |
| NonSrcDelLines | Git | SRC | The number of non-source code lines deleted by commits related to the bug |
| SrcAddFiles | Git | SRC | The number of source code files added by commits related to the bug |
| SrcDelFiles | Git | SRC | The number of source code files deleted by commits related to the bug |
| SrcModFiles | Git | SRC | The number of source code files modified by commits related to the bug |
| SrcAddLines | Git | SRC | The number of source code lines added by commits related to the bug |
| SrcDelLines | Git | SRC | The number of deleted source code lines by commits related to the bug |
| TestAddFiles | Git | SRC | The number of source code test files added by commits related to the bug |
| TestDelFiles | Git | SRC | The number of source code test files deleted by commits related to the bug |
| TestModFiles | Git | SRC | The number of source code test files modified by commits related to the bug |
| TestAddLines | Git | SRC | The number of source code test lines added by commits related to the bug |
| TestDelLines | Git | SRC | The number of source code test lines deleted by commits related to the bug |

1. Dynamic - Bug-Fix Changelog Dataset

Dataset location:

./changelog/

Dataset file name format:

<jira-project-id>-bug-fix-changelog-dataset.csv

|  |  |  |  |
| --- | --- | --- | --- |
| **Fields Description of the Bug-Fix Change Log Dataset** | | | |
| **Field** | **From** | **Type** | **Description** |
| Project | Jira | General | The parent project to which the bug belongs |
| Manager | Jira | General | The owner/maintainer of the project |
| Category | Jira | General | The project domain category (big-data, database, cloud, network-server, web-framework, security, build-management, library, and machine-learning) |
| Key | Jira | General | A unique identifier for the bug |
| Author | Jira | General | The people who made the comment |
| ChangeDate | Jira | Time | The date and time when the change was made in the bug report |
| Field | Jira | Location | The bug report field where the change happened |
| From | Jira | Text | The content stored in the field before the change |
| To | Jira | Text | The content stored in the field after the change |

1. Dynamic - Bug-Fix Comments Log Dataset

Dataset location:

./comment-log/

Dataset file name format:

<jira-project-id>-bug-fix-comment-log-dataset.csv

|  |  |  |  |
| --- | --- | --- | --- |
| **Fields Description of the Bug-Fix Comments Log Dataset** | | | |
| **Field** | **From** | **Type** | **Description** |
| Project | Jira | General | The parent project to which the bug belongs |
| Manager | Jira | General | The owner/maintainer of the project |
| Category | Jira | General | The project domain category (big-data, database, cloud, network-server, web-framework, security, build-management, library, and machine-learning) |
| Key | Jira | General | A unique identifier for the bug |
| Author | Jira | General | The people who made the comment |
| CreationDate | Jira | Time | The date and time when the comment was added in the bug report. |
| Content | Jira | Text | 1000 top most frequent words of the added comment |

1. Dynamic - Bug-Fix Commits Log Dataset

Dataset location:

./commit-log/

Dataset file name format:

<jira-project-id>-bug-fix-commit-log-dataset.csv

|  |  |  |  |
| --- | --- | --- | --- |
| **Fields Description of the Bug-Fix Commits Log Dataset** | | | |
| **Field** | **From** | **Type** | **Description** |
| Project | Jira | General | The parent project to which the bug belongs |
| Manager | Jira | General | The owner/maintainer of the project |
| Category | Jira | General | The project domain category (big-data, database, cloud, network-server, web-framework, security, build-management, library, and machine-learning) |
| Key | Jira | General | A unique identifier for the bug |
| CommitHash | Git | Versioning | The hash code of the commit related to the bug |
| IsMergeCommit | Git | Versioning | This field informs whether a the commit is or not a merge commit |
| Author | Git | General | The name of the person who originally wrote the code to fix the bug |
| AuthorDate | Git | Time | The date and time when the author performed the commit |
| Committer | Git | General | The name of the person who committed the bug fix code on behalf of the original author |
| CommitterDate | Git | Time | The date and time on which the committer performed the commit |
| CommitMessageTopWords | Git | Text | 1000 top most frequent words of the commit message |
| FileName | Git | SRC | The name of the file affected by the commit |
| FilePath | Git | SRC | The path of the file affected by the commit |
| ChangeType | Git | SRC | The type of the change made in the file (Added, Deleted, Modified, or Renamed) |
| IsSrcFile | Git | SRC | This field informs whether a the modified file is or not a source code file |
| IsTestFile | Git | SRC | This field informs whether a the modified file is or not a source code test file |
| AddLines | Git | SRC | The number of lines added from the modified file |
| DelLines | Git | SRC | The number of lines deleted from the modified file |
| NoMethods | Git | SRC | The number of methods in the source code file |
| LoC | Git | SRC | The number of lines of code of the modified source code file |
| CyC | Git | SRC | The cyclomatic complexity of the modified source code file |
| NoTokens | Git | SRC | The number of tokens of the modified file |

1. It is a subset of “10 Years Bug-Fix Dataset (PROMISE'19)” available at https://doi.org/10.6084/m9.figshare.8852084.v4 [↑](#footnote-ref-0)