master 6 branches 15 tags Go to file Add file Code

scala-steward Update scala-library to 2.13.8 (#593) d95713d 4 hours ago 888 commits

.github	Update scala-library to 2.13.8 (#593)	4 hours ago
cats/src	Establish Spark 3.x & Scala 2.13 cross publish (#573)	3 months ago
core/src/main/scala/frameless	negate	3 years ago
dataset/src	Resolve a possible explodeMap column names collision (#582)	2 months ago
docs	Flatten docs and fix scripts	7 months ago
ml/src	Establish Spark 3.x & Scala 2.13 cross publish (#573)	3 months ago
project	Update sbt-scoverage to 1.9.3 (#592)	8 days ago
refined/src	Establish Spark 3.x & Scala 2.13 cross publish (#573)	3 months ago
scripts	Flatten docs and fix scripts	7 months ago
.gitignore	Add the CI release (#551)	5 months ago
LICENSE	Initial commit	6 years ago
README.md	Update README.md	2 months ago
build.sbt	Update scala-library to 2.13.8 (#593)	4 hours ago

README.md

Frameless

Frameless is a Scala library for working with Spark using more expressive types. It consists of the following modules:

- `frameless-dataset` for a more strongly typed `Dataset` / `DataFrame` API
- `frameless-ml` for a more strongly typed Spark ML API based on `frameless-dataset`
- `frameless-cats` for using Spark's `RDD` API with cats

Note that while Frameless is still getting off the ground, it is very possible that breaking changes will be made for at least the next few versions.

The Frameless project and contributors support the Typelevel Code of Conduct and want all its associated channels (e.g. GitHub, Discord) to be a safe and friendly environment for contributing and learning.

Versions and dependencies

The compatible versions of Spark and cats are as follows:

Frameless	Spark	Cats	Cats-Effect	Scala
0.4.0	2.2.0	1.0.0-IF	0.4	2.11
0.4.1	2.2.0	1.x	0.8	2.11
0.5.2	2.2.1	1.x	0.8	2.11
0.6.1	2.3.0	1.x	0.8	2.11
0.7.0	2.3.1	1.x	1.x	2.11
0.8.0	2.4.0	1.x	1.x	2.11 / 2.12
0.9.0	3.0.0	1.x	1.x	2.12
0.10.1	3.1.0	2.x	2.x	2.12
0.11.0*	3.2.0 / 3.1.2 / 3.0.1	2.x	2.x	2.12 / 2.13
0.11.1	3.2.0 / 3.1.2 / 3.0.1	2.x	2.x	2.12 / 2.13

** 0.11.0 has broken Spark 3.1.2 and 3.0.1 artifacts published.*

Starting 0.11 we introduced Spark cross published artifacts:

- By default, frameless artifacts depend on the most recent Spark version
- Suffix `-spark{major}{minor}` is added to artifacts that are released for the previous Spark version(s)

Artifact names examples:

- `frameless-dataset` (the latest Spark dependency)
- `frameless-dataset-spark31` (Spark 3.1.x dependency)
- `frameless-dataset-spark30` (Spark 3.0.x dependency)

Versions 0.5.x and 0.6.x have identical features. The first is compatible with Spark 2.2.1 and the second with 2.3.0.

The **only** dependency of the `frameless-dataset` module is on shapeless 2.3.2. Therefore, depending on `frameless-dataset`, has a minimal overhead on your Spark's application jar. Only the `frameless-cats` module depends on cats and cats-effect, so if you prefer to work just with `Datasets` and not with `RDD`s, you may choose not to depend on `frameless-cats`.

Frameless intentionally **does not** have a compile dependency on Spark. This essentially allows you to use any version of Frameless with any version of Spark. The aforementioned table simply provides the versions of Spark we officially compile and test Frameless with, but other versions may probably work as well.

Breaking changes in 0.9

- Spark 3 introduces a new ExpressionEncoder approach, the schema for single value DataFrame's is now "value" not "_1".

Why?

Frameless introduces a new Spark API, called `TypedDataset`. The benefits of using `TypedDataset` compared to the standard Spark `Dataset` API are as follows:

- Typesafe columns referencing (e.g., no more runtime errors when accessing non-existing columns)
- Customizable, typesafe encoders (e.g., if a type does not have an encoder, it should not compile)
- Enhanced type signature for built-in functions (e.g., if you apply an arithmetic operation on a non-numeric column, you get a compilation error)
- Typesafe casting and projections

Click here for a detailed comparison of `TypedDataset` with Spark's `Dataset` API.

Documentation

- TypedDataset: Feature Overview
- Typed Spark ML
- Comparing TypedDatasets with Spark's Datasets
- Typed Encoders in Frameless
- Injection: Creating Custom Encoders
- Job[A]
- Using Cats with RDDs
- Proof of Concept: TypedDataFrame

Quick Start

Since the 0.9.x release, Frameless is compiled only against Scala 2.12.x.

To use Frameless in your project add the following in your `build.sbt` file as needed:

```
val framelessVersion = "0.9.0" // for Spark 3.0.0

libraryDependencies ++= List(
  "org.typelevel" %% "frameless-dataset" % framelessVersion,
  "org.typelevel" %% "frameless-ml"      % framelessVersion,
  "org.typelevel" %% "frameless-cats"    % framelessVersion
)
```

An easy way to bootstrap a Frameless sbt project:

- if you have Giter8 installed then simply:

```
g8 imarios/frameless.g8
```

- with sbt >= 0.13.13:

```
sbt new imarios/frameless.g8
```

Typing `sbt console` inside your project will bring up a shell with Frameless and all its dependencies loaded (including Spark).

Need help?

Feel free to messages us on our discord channel for any issues/questions.

Development

We require at least *one* sign-off (thumbs-up, +1, or similar) to merge pull requests. The current maintainers (people who can merge pull requests) are:

- adelbertc
- imarios
- kanterov
- non
- OlivierBlanvillain

Testing

Frameless contains several property tests. To avoid `OutOfMemoryError`s, we tune the default generator sizes. The following environment variables may be set to adjust the size of generated collections in the `TypedDataSet` suite:

Property	Default
FRAMELESS_GEN_MIN_SIZE	0
FRAMELESS_GEN_SIZE_RANGE	20

License

Code is provided under the Apache 2.0 license available at http://opensource.org/licenses/Apache-2.0, as well as in the LICENSE file. This is the same license used as Spark.

About

Expressive types for Spark.

scala spark functional-programming typelevel fp

Readme
Apache-2.0 License
Code of conduct
786 stars
34 watching
129 forks

Releases 13

v0.11.1 Latest
on 11 Nov 2021

+ 12 releases

Packages

No packages published

Contributors 50

+ 39 contributors

Environments 1

github-pages Active

Languages

Scala 99.8% Shell 0.2%