

Join GitHub today

Dismiss

GitHub is home to over 50 million developers working together to host and review code, manage projects, and build software together.

Sign up

[SPARK-32180][PYTHON][DOCS] Installation page in Getting Started in PySpark documentation #29410

New issue

Open rohitmishr1484 wants to merge 170 commits into apache:master from rohitmishr1484:SPARK-32180-Getting-Started-Installation

Conversation 58 Commits 170 Checks 12 Files changed 1,056 +142,051 -11,833



rohitmishr1484 commented 23 days ago • edited

What changes were proposed in this pull request?

This PR proposes to add getting started- installation to new PySpark docs.

Why are the changes needed?

Better documentation.

Does this PR introduce any user-facing change?

No. Documentation only.

How was this patch tested?

Generating documents locally.

Reviewers

HyukjinKwon

Assignees

No one assigned

Labels

CORE DOCS INFRA PYTHON R SQL

Projects

None yet

Milestone

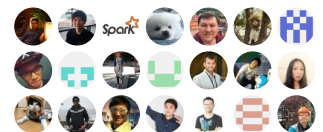
No milestone

Linked issues

Successfully merging this pull request may close these issues.

None yet

61 participants



and others

Installation

Package Overview

Quickstart

Using Conda

PySpark installation using [Conda](#) can be performed using the below command:

```
conda install -c conda-forge pyspark
```

Using PyPI

PySpark installation using [PyPI](#):

```
pip install pyspark
```

Official release channel

Different flavor of PySpark is available in the [official release channel](#). Any suitable version can be downloaded and extracted as below:

```
tar xzvf spark-3.0.0-bin-hadoop2.7.tgz
```

An important step is to ensure `SPARK_HOME` environment variable points to the directory where the code has been extracted. The next step is to properly define `PYTHONPATH` such that it can find the PySpark and Py4J under `$$SPARK_HOME/python/lib`:

```
cd spark-3.0.0-bin-hadoop2.7
export SPARK_HOME="pwd"
export PYTHONPATH=$(ZIPS=$(("$SPARK_HOME"/python/lib/*.zip);IFS=:; echo "${ZIPS[*]}");$PYTHONPATH
```

Installing from source

To install PySpark from source, refer [Building Spark](#).

- Steps for defining `PYTHONPATH` is same as described in [Official release channel](#) section above.

On this page

Using Conda

Using PyPI

Official release channel

Installing from source

Installation

Package Overview

Quickstart

Official release channel

Different flavor of PySpark is available in the [official release channel](#). Any suitable version can be downloaded and extracted as below:

```
tar xzvf spark-3.0.0-bin-hadoop2.7.tgz
```

An important step is to ensure `SPARK_HOME` environment variable points to the directory where the code has been extracted. The next step is to properly define `PYTHONPATH` such that it can find the PySpark and Py4J under `$$SPARK_HOME/python/lib`:

```
cd spark-3.0.0-bin-hadoop2.7
export SPARK_HOME="pwd"
export PYTHONPATH=$(ZIPS=$(("$SPARK_HOME"/python/lib/*.zip);IFS=:; echo "${ZIPS[*]}");$PYTHONPATH
```

Installing from source

To install PySpark from source, refer [Building Spark](#).

- Steps for defining `PYTHONPATH` is same as described in [Official release channel](#) section above.

Dependencies

- Using PySpark requires the Spark JARs.
- At its core PySpark depends on Py4J, but some additional sub-packages have their own extra requirements for some features (including NumPy, pandas, and PyArrow).

[<< Getting Started](#)
[Package Overview >>](#)

On this page

Using Conda

Using PyPI

Official release channel

Installing from source

© Copyright .
Created using [Sphinx 3.1.2](#).

Initial Skeleton Setup

0f8db3a

probot-autolabeler bot added **DOCS** **PYTHON** labels 23 days ago

rohitmishr1484 changed the title ~~[WIP][SPARK-32180][PYTHON][DOCS] Getting started-Installation guide for pyspark doc~~ [WIP][SPARK-32180][PYSPARK][DOCS] Getting started-Installation guide for pyspark doc 23 days ago

rohitmishr1484 changed the title ~~[WIP][SPARK-32180][PYSPARK][DOCS] Getting started-Installation guide for pyspark doc~~ [SPARK-32180][PYSPARK][DOCS] Getting started-Installation guide for pyspark doc 23 days ago


Corrected typo in index.rst and added content in installation.rst







8b55608

Hi @HyukjinKwon,

I was not sure how to add you as a Reviewer for this Pull request, thus adding this comment. I would like to mention a few points:

1. Baseline description: I have used this pull request as a reference since this is my first pull request- [#29385](#)
2. Most of the information used for the "Installation Page" has come from Koalas documentation- https://koalas.readthedocs.io/en/latest/getting_started/install.html. Am I supposed to mention this as reference/credit anywhere in this documentation?
3. Personally I have used only PyPI as an installation mechanism thus haven't tried the other three. A prerequisite for PyPI installation was the availability of JAVA 8 path in JAVA_HOME environment variable. Please let me know if that's something I need to add explicitly in the "Dependencies" section.
4. Where ever necessary I have updated the links.

 wangyum and others added 3 commits 22 days ago



-   [SPARK-32586][SQL] Fix NumberFormatException error message when ansi ... ✓5d130f0
-   Fixed Title underline too short error in packahe overview rst file ✓e93b4e1
-   [SPARK-32594][SQL] Fix serialization of dates inserted to Hive tables ... ✓0477d23



HyukjinKwon commented 22 days ago

Member

ok to test

  HyukjinKwon changed the title ~~[SPARK-32180][PYSPARK][DOCS] Getting started-Installation guide for pyspark doc~~ [SPARK-32180][PYTHON][DOCS] Installation page in Getting Started in PySpark documentation 22 days ago



 HyukjinKwon reviewed 22 days ago

[View changes](#)

HyukjinKwon left a comment


Member

Thanks @rohitmishr1484 for working on this. In general, let's put some more flesh. We can also describe how to set up the development environment by using Conda like pandas from the beginning (https://pandas.pydata.org/pandas-docs/stable/getting_started/install.html#installing-with-miniconda).

We can use `pip` installation under Conda environment as the official guide, and just mention that PySpark is available in Conda although it's not the part of an official release.

python/docs/source/getting_started/index.rst Outdated  Show resolved

python/docs/source/getting_started/index.rst Outdated  Show resolved

python/docs/source/getting_started/installation.rst Outdated  Show resolved

python/docs/source/getting_started/installation.rst Outdated  Show resolved

python/docs/source/getting_started/installation.rst Outdated  Show resolved



This comment was marked as off-topic.

[Sign in to view](#)

Venkata krishnan Sowrirajan and others added 16 commits 22 days ago

- [SPARK-32596][CORE] Clear Ivy resolution files as part of finally block ... ✓ 2d6eb00
- [SPARK-32400][SQL] Improve test coverage of HiveScriptTransformationExec ... ✓ 4cf8c1d
- [SPARK-31703][SQL] Parquet RLE float/double are read incorrectly on b... ✓ a418548
- [SPARK-32599][SQL][TESTS] Check the TEXTFILE file format in `HiveSerD... ✓ f664aaa
- [SPARK-32250][SPARK-27510][CORE][TEST] Fix flaky MasterSuite.test(... ✗ c6ea983
- [SPARK-32352][SQL] Partially push down support data filter if it mixe... ✓ 643cd87
- [SPARK-31694][SQL] Add SupportsPartitions APIs on DataSourceV2 ... ✗ 60fa8e3
- [SPARK-31198][CORE] Use graceful decommissioning as part of dynamic s... ✓ 548ac7c
- [MINOR] Update URL of the parquet project in code comment ... ✓ 08d86eb
- [SPARK-32511][SQL] Add dropFields method to Column class ... ✓ 0c850c7
- [SPARK-32526][SQL] Fix some test cases of `sql/catalyst` module in sc... ✗ 6ae2cb2
- [SPARK-32357][INFRA] Publish failed and succeeded test reports in Git... ✓ 5debde9
- [SPARK-20680][SQL][FOLLOW-UP] Add HiveVoidType in HiveClientImpl ... ✓ 339eec5
- [SPARK-32590][SQL] Remove fullOutput from RowDataSourceScanExec ... ✓ 14003d4
- [MINOR][SQL] Fixed approx_count_distinct rsd param description ... ✓ 10edeaf
- [SPARK-32616][SQL] Window operators should be added determinedly ... ✓ c6be207

162 hidden items

[Load more...](#)

sarutak and others added 2 commits 17 hours ago

- [SPARK-32772][SQL] Reduce log messages for spark-sql CLI ... ✓ ad6b887
- [SPARK-32781][SQL] Non-ASCII characters are mistakenly omitted in the... ✓ 1fba286



HyukjinKwon commented 7 hours ago

Member

@rohitmishr1484, thanks. Looks getting there. Can you resolve the conflicts? See also "The Review Process" at <https://spark.apache.org/contributing.html> to make this PR synced to the latest master.



HyukjinKwon reviewed 7 hours ago

[View changes](#)

python/docs/source/getting_started/installation.rst

97 +



HyukjinKwon 7 hours ago Member

Let's remove empty newlines here.



HyukjinKwon reviewed 7 hours ago

[View changes](#)

python/docs/source/getting_started/installation.rst

```

117 + `Py4J`          0.10.9          Required
118 + =====
119 +
120 + Note: A prerequisite for PySpark installation is the availability of
      ``JAVA 8`` and ``JAVA 8`` path in ``JAVA_HOME`` environment variable.

```



HyukjinKwon 7 hours ago Member

- ``JAVA 8`` ... -> Java 8 or 11 and JAVA_HOME properly set.



HyukjinKwon reviewed 7 hours ago

[View changes](#)

python/docs/source/getting_started/installation.rst

```

109 + Dependencies
110 + -----
111 + =====
112 + Package      Minimum supported version Required or Optional

```



HyukjinKwon 7 hours ago Member

Let's convert Required or Optional to Note.

NumPy is an optional dependency for ML module in PySpark.



HyukjinKwon reviewed 7 hours ago

[View changes](#)

python/docs/source/getting_started/installation.rst

```

99 +
100 + export PYTHONPATH=$(ZIPS=("$SPARK_HOME"/python/lib/*.zip); IFS=:; echo
      "${ZIPS[*]}"):PYTHONPATH
101 +
102 + Installing From Source

```



HyukjinKwon 7 hours ago Member

From -> from



HyukjinKwon reviewed 7 hours ago

[View changes](#)

python/docs/source/getting_started/installation.rst

```

26 + Python Version Supported
27 + ~~~~~
28 +
29 + Python 3.6, 3.7 and 3.8

```



HyukjinKwon 7 hours ago Member

I think you can just say Python 3.6 and above.



HyukjinKwon reviewed 7 hours ago

[view changes](#)

python/docs/source/getting_started/installation.rst

```
19 + Installation
20 + =====
21 +
22 + The official release channel is to download it from
    https://spark.apache.org/downloads.html but we can install it via ``pip`` as
    well from PyPI. PyPI installation is usually to use standalone locally or as
    a client to connect to a cluster.
```



HyukjinKwon 7 hours ago Member

Let's make it properly linked and slightly reword, for example:

The official release channel is to download it from `the Apache Spark website <
https://spark.apache.org/downloads.html>`_.
Alternatively, you can also install it via pip from PyPI. PyPI installation is
usually to use
standalone locally or as a client to connect to a cluster instead of setting a
cluster up.



HyukjinKwon reviewed 7 hours ago

[View changes](#)

python/docs/source/getting_started/installation.rst

```
21 +
22 + The official release channel is to download it from
    https://spark.apache.org/downloads.html but we can install it via ``pip`` as
    well from PyPI. PyPI installation is usually to use standalone locally or as
    a client to connect to a cluster.
23 +
24 + Instruction for downloading PySpark using PyPI, Conda, Official Release
    Channel and Source are available in this document.
```



HyukjinKwon 7 hours ago Member

I would write it like such as:

This page includes the instructions for installing PySpark by using pip, Conda,
downloading manually, and building it from the source.

Feel free to rephrase.



HyukjinKwon reviewed 7 hours ago

[View changes](#)

python/docs/source/getting_started/installation.rst

```
80 + Official Release Channel
81 + ~~~~~
82 +
83 + Different flavor of PySpark is available in `the official release channel
    <https://spark.apache.org/downloads.html>`_.
```



HyukjinKwon 7 hours ago Member

I think we can have one underscore for the links. __ -> _. You'd have to change other
links in this page as well.

python/docs/source/getting_started/installation.rst

```
6 + "License"); you may not use this file except in compliance
7 + with the License. You may obtain a copy of the License at
8 +
9 + .. http://www.apache.org/licenses/LICENSE-2.0
```



HyukjinKwon 7 hours ago Member

Let's add two leading spaces here just to match with other files.



HyukjinKwon reviewed 7 hours ago

[View changes](#)

python/docs/source/getting_started/installation.rst

```
100 + export PYTHONPATH=$(ZIPS=("$SPARK_HOME"/python/lib/*.zip); IFS=:; echo
    + "${ZIPS[*]}"):PYTHONPATH
101 +
102 + Installing From Source
103 + ~~~~~
```



HyukjinKwon 7 hours ago Member

I think we should use - for all subsections here consistently.

xuanyuanking and others added 11 commits 6 hours ago

- [SPARK-32782][SS] Refactor StreamingRelationV2 and move it to catalyst ... ✓95f1e95
- Initial Skeleton Setup 792592c
- Corrected typo in index.rst and added content in installation.rst 1b08fc3
- Fixed Title underline too short error in packahe overview rst file f08f63a
- Removed unwanted files and undated index.rst and installation.rst 828f598
- added the suggested changes 8b5f745
- conflict removal a6dfaf7
- Removed quickstart from index as its not part of the PR 274e419
- Merge conflict removal 7ca3425
- merge conflict error 1 4f8627c
- Merge branch 'SPARK-32180-Getting-Started-Installation' of <https://gi...> ... ●001814f

probot-autolabeler bot added CORE INFRA R SQL labels 1 hour ago



HyukjinKwon commented 31 minutes ago

Member

Oops, looks something went wrong during syncing. Feel free to close this and create new PR :-)



SparkQA commented 27 minutes ago

[Test build #128258 has finished](#) for PR 29410 at commit [001814f](#) .

- This patch passes all tests.

