# A package for useful pyspark functions

## Directory

* [A. Setup] (#markdown-header-a-setup)
* [B. Resources for data conversion] (#markdown-header-b-resources-for-data-conversion)
* [C. Documentation] (#markdown-header-c-documentation)
* [D. Areas for improvement]( #markdown-header-d-areas-for-improvement)

## A. Setup

#### 1. Clone the repo

* go to jupyter terminal ![jupyter terminal] (assets/jup\_terminal.png)
* enter in terminal

```

git clone #repo-name

``’

#### 2. Configure your Environment

- Add environment keys: value pairs

- key names

- username

- password

- hadoop1

- hadoop2

- for cdsw, you can add these in the project's advanced settings as shown in the images below

- ![env keys] (assets/env\_keys.png)

- install requirements in terminal or jupyter

‘’’

%cd pyspark\_worker

‘’’

‘’’

! pip3 install -r requirements.txt

‘’’

#### 3. Import into your notebooks python

‘’’

import pyspark\_worker.core.general as G

from pyspark\_worker.core.main\_import import \*

spark = Sparky()

‘’’

‘’’

query = f"""( # enter your sql)"""

df = spark.read\_table\_presto (query)

’’’

## B. Resources for data conversion

### check/ find data sources

Nil

## C. Documentation

Includes examples

* [ Documentation](examples)

## D. Areas for improvement

* Pip packaging