

---

# pyddf Documentation

*Release 1.2.0.dev0*

**Adatao**

July 03, 2015



<b>1</b>	<b>ddf package</b>	<b>3</b>
1.1	Submodules . . . . .	3
1.2	ddf.conf module . . . . .	3
1.3	ddf.dataframe module . . . . .	3
1.4	ddf.ddf_manager module . . . . .	4
1.5	ddf.gateway module . . . . .	4
<b>2</b>	<b>Indices and tables</b>	<b>7</b>
	<b>Python Module Index</b>	<b>9</b>
	<b>Index</b>	<b>11</b>



Contents:



## 1.1 Submodules

## 1.2 ddf.conf module

`ddf.conf.find_ddf()`

## 1.3 ddf.dataframe module

Created on Jun 22, 2014

@author: nhanitvn

**class** `ddf.dataframe.DistributedDataFrame` (*jddf*)

Bases: `object`

A Distributed Data Frame, the basic abstraction in DistributedDataFrame library.

**aggregate** (*aggr\_columns, by\_columns*)

Split the DistributedDataFrame into sub-sets by some columns and perform aggregation on some columns within each sub-set

**colnames**

List the column names of this DDF

**Returns** a list of strings

**cols**

Get number of columns of this DDF

**Returns** an int

**correlation** (*col1, col2*)

Correlation coefficient of a DistributedDataFrame's two numeric columns

**Parameters**

- **col1** – a numeric column
- **col2** – a numeric column

**Returns** a float

**drop\_na** ()

**five\_nums** ()

Calculate Turkey five number for numeric columns

**head** (*n=10*)

Return this DistributedDataFrame's some first rows

**project** (*column\_names*)

Project on some columns and return a new DistributedDataFrame

**rows**

Get number of rows of this DDF

**Returns** an int

**sample** (*size, replacement=False, seed=123*)

Get a sample of this DistributedDataFrame and return a list of strings

**Parameters**

- **size** – number of samples
- **replacement** – sample with or without replacement
- **seed** – random seed

**Returns** WRITE ME

**sample2ddf** (*fraction, replacement=False, seed=123*)

Get a sample of this DistributedDataFrame and return a new DistributedDataFrame

**Parameters**

- **fraction** – fraction to take sample, has to be in the (0, 1] range
- **replacement** – sample with or without replacement
- **seed** – random seed

**Returns** a new DistributedDataFrame

**summary** ()

Calculate this DistributedDataFrame's columns'summary numbers

## 1.4 ddf.ddf\_manager module

Created on Jun 22, 2014

@author: nhanitvn

**class** ddf.ddf\_manager.DDFManager (*engine\_name*)

Bases: object

Main entry point for DDF functionality. A SparkDDFManager can be used to create DDFs that are implemented for Spark framework.

**shutdown** ()

Shut down the DDF Manager

**sql** (*command*)

Execute a sql command and return a list of strings :param command: the sql command to run

**sql2ddf** (*command*)

Create a DistributedDataFrame from an sql command. :param command: the sql command to run

## 1.5 ddf.gateway module

ddf.gateway.compute\_classpath (*root\_path*)

ddf.gateway.list\_jar\_files (*path*)



```
ddf.gateway.pre_exec_func()  
ddf.gateway.start_gateway_server()
```



---

## Indices and tables

---

- `genindex`
- `modindex`
- `search`



## d

`ddf`, 3  
`ddf.conf`, 3  
`ddf.dataframe`, 3  
`ddf.ddf_manager`, 4  
`ddf.gateway`, 4



## A

aggregate() (ddf.dataframe.DistributedDataFrame method), 3

## C

colnames (ddf.dataframe.DistributedDataFrame attribute), 3

cols (ddf.dataframe.DistributedDataFrame attribute), 3

compute\_classpath() (in module ddf.gateway), 4

correlation() (ddf.dataframe.DistributedDataFrame method), 3

## D

ddf (module), 3

ddf.conf (module), 3

ddf.dataframe (module), 3

ddf.ddf\_manager (module), 4

ddf.gateway (module), 4

DDFManager (class in ddf.ddf\_manager), 4

DistributedDataFrame (class in ddf.dataframe), 3

drop\_na() (ddf.dataframe.DistributedDataFrame method), 3

## F

find\_ddf() (in module ddf.conf), 3

five\_nums() (ddf.dataframe.DistributedDataFrame method), 3

## H

head() (ddf.dataframe.DistributedDataFrame method), 3

## L

list\_jar\_files() (in module ddf.gateway), 4

## P

pre\_exec\_func() (in module ddf.gateway), 4

project() (ddf.dataframe.DistributedDataFrame method), 4

## R

rows (ddf.dataframe.DistributedDataFrame attribute), 4

## S

sample() (ddf.dataframe.DistributedDataFrame method), 4

sample2ddf() (ddf.dataframe.DistributedDataFrame method), 4

shutdown() (ddf.ddf\_manager.DDFManager method), 4

sql() (ddf.ddf\_manager.DDFManager method), 4

sql2ddf() (ddf.ddf\_manager.DDFManager method), 4

start\_gateway\_server() (in module ddf.gateway), 5

summary() (ddf.dataframe.DistributedDataFrame method), 4