pyddf Documentation

Release 1.2.0.dev0

Adatao

July 03, 2015

Contents

Contents:

CHAPTER 1

ddf package

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**()

CHAPTER 2

Indices and tables

- genindex
- modindex
- search