Search

Searching for multiple words only shows matches that contain all words.

search

Search Results

Search finished, found 532 page(s) matching the search query.

- **Debugging PySpark**

  ...After that, you should install the corresponding version of the pydevd-pycharm package in all the machines which will connect to your PyCharm debugger. In the previous dialog, it shows the command to install. pip install pydevd-pycharm~=...

- **Installation**

  .... For Python users, PySpark also provides pip installation from PyPI. This is usually for local usage or as a client to connect to a cluster instead of setting up a cluster itself. This page includes instructions for installing PySpark by u...

- **Quickstart: Spark Connect**

  ...background: #f5f5f5; } div.rendered_html tbody tr:hover { background: rgba(66, 165, 245, 0.2); } Quickstart: Spark Connect Spark Connect introduced a decoupled client-server architecture for Spark that allows remote connectivity to Spa...

- **Contributing to PySpark**

  ...needs a proper JDK installed, etc. See Building Spark for more details. Note that if you intend to contribute to Spark Connect in Python, buf version 1.24.0 is required, see Buf Installation for more details. Conda If you are using Conda...

- **Functions**

  ...llections of builtin functions available for DataFrame operations. From Apache Spark 3.5.0, all functions support Spark Connect. Normal Functions col(col) Returns a Column based on the given column name. column(col) Returns a Column...

- **Getting Started**

  ...are live notebooks where you can try PySpark out without any other step: Live Notebook: DataFrame Live Notebook: Spark Connect Live Notebook: pandas API on Spark The list below is the contents of this quickstart page: Installation Pytho...

- **pyspark.ml.functions.array_to_vector**

  ...o a column of pyspark.ml.linalg.DenseVector instances New in version 3.1.0. Changed in version 3.5.0: Supports Spark Connect. Parameters colpyspark.sql.Column or strInput column Returns pyspark.sql.ColumnThe converted column of de...

- **pyspark.ml.functions.vector_to_array**

  ...b sparse/dense vectors into a column of dense arrays. New in version 3.0.0. Changed in version 3.5.0: Supports Spark Connect. Parameters colpyspark.sql.Column or strInput column dtypestr, optionalThe data type of the output array. Va...

- **pyspark.SparkContext**

  ...ass 'pyspark.profiler.MemoryProfiler'>)[source] Main entry point for Spark functionality. A SparkContext represents the connection to a Spark cluster, and can be used to create RDD and broadcast variables on that cluster. When you create a...

- **pyspark.sql.avro.functions.from_avro**

  ...ted Avro schema can be set via the option avroSchema. New in version 3.0.0. Changed in version 3.5.0: Supports Spark Connect. Parameters dataColumn or strthe binary column. jsonFormatSchemastrthe avro schema in JSON string format. o...

- **pyspark.sql.avro.functions.to_avro**

  ...source] Converts a column into binary of avro format. New in version 3.0.0. Changed in version 3.5.0: Supports Spark Connect. Parameters dataColumn or strthe data column. jsonFormatSchemastr, optionaluser-specified output avro schema...

- **pyspark.sql.Catalog**

  ...wrapper around its Scala implementation org.apache.spark.sql.catalog.Catalog. Changed in version 3.4.0: Supports Spark Connect. Methods cacheTable(tableName[, storageLevel]) Caches the specified table in-memory or with given storage...

- **pyspark.sql.Catalog.registerFunction**

  ...n 2.0.0. Deprecated since version 2.3.0: Use spark.udf.register() instead. Changed in version 3.4.0: Supports Spark Connect. previous...

- **pyspark.sql.Column**

  ..._gateway.JavaObject)[source] A column in a DataFrame. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Examples Column instances can be created by >>> df = spark.createDataFrame(... [(2, "Alice"), (5, "Bob")...

- **pyspark.sql.Column.__getattr__**

  ...out of a list, or gets an item by key out of a dict. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters itema literal value. Returns ColumnColumn representing the item got by key out of a dict....

- **pyspark.sql.Column.__getitem__**

  ...out of a list, or gets an item by key out of a dict. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters ka literal value, or a slice object without step. Returns ColumnColumn representing the item...

- **pyspark.sql.Column.alias**

  ...s that return more than one column, such as explode). New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters aliasstrdesired column names (collects all positional arguments passed) Returns ColumnColumn...

- **pyspark.sql.Column.asc**

  ....Column Returns a sort expression based on the ascending order of the column. Changed in version 3.4.0: Supports Spark Connect. Examples >>> from pyspark.sql import Row >>> df = spark.createDataFrame([('Tom', 80), ('Alice', None)], ["name...

- **pyspark.sql.Column.asc_nulls_first**

  ...olumn, and null values return before non-null values. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Examples >>> from pyspark.sql import Row >>> df = spark.createDataFrame([('Tom', 80), (None, 60), ('Alice', No...

- **pyspark.sql.Column.asc_nulls_last**

  ...column, and null values appear after non-null values. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Examples >>> from pyspark.sql import Row >>> df = spark.createDataFrame([('Tom', 80), (None, 60), ('Alice', No...

- **pyspark.sql.Column.between**

  ...s between the lower bound and upper bound, inclusive. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters lowerBoundColumn, int, float, string, bool, datetime, date or Decimala boolean expression that bo...

- **pyspark.sql.Column.bitwiseAND**

  ...al]) → Column Compute bitwise AND of this expression with another expression. Changed in version 3.4.0: Supports Spark Connect. Parameters othera value or Column to calculate bitwise and(&) with this Column. Examples >>> from pyspar...

- **pyspark.sql.Column.bitwiseOR**

  ...ral]) → Column Compute bitwise OR of this expression with another expression. Changed in version 3.4.0: Supports Spark Connect. Parameters othera value or Column to calculate bitwise or(|) with this Column. Examples >>> from pyspark...

- **pyspark.sql.Column.bitwiseXOR**

  ...al]) → Column Compute bitwise XOR of this expression with another expression. Changed in version 3.4.0: Supports Spark Connect. Parameters othera value or Column to calculate bitwise xor(^) with this Column. Examples >>> from pyspar...

- **pyspark.sql.Column.cast**

  ...n.Column[source] Casts the column into type dataType. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters dataTypeDataType or stra DataType or Python string literal with a DDL-formatted string to use whe...

- **pyspark.sql.Column.contains**

  ...Contains the other element. Returns a boolean Column based on a string match. Changed in version 3.4.0: Supports Spark Connect. Parameters otherstring in line. A value as a literal or a Column. Examples >>> df = spark.createDataFram...

- **pyspark.sql.Column.desc**

  ...pression based on the descending order of the column. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Examples >>> from pyspark.sql import Row >>> df = spark.createDataFrame([('Tom', 80), ('Alice', None)], ["name...

- **pyspark.sql.Column.desc_nulls_first**

  ...olumn, and null values appear before non-null values. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Examples >>> from pyspark.sql import Row >>> df = spark.createDataFrame([('Tom', 80), (None, 60), ('Alice', No...

- **pyspark.sql.Column.desc_nulls_last**

  ...column, and null values appear after non-null values. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Examples >>> from pyspark.sql import Row >>> df = spark.createDataFrame([('Tom', 80), (None, 60), ('Alice', No...

- **pyspark.sql.Column.dropFields**

  ...a no-op if the schema doesn’t contain field name(s). New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters fieldNamesstrDesired field names (collects all positional arguments passed) The result will drop...

- **pyspark.sql.Column.endswith**

  ...→ Column String ends with. Returns a boolean Column based on a string match. Changed in version 3.4.0: Supports Spark Connect. Parameters otherColumn or strstring at end of line (do not use a regex $) Examples >>> df = spark.create...

- **pyspark.sql.Column.eqNullSafe**

  ...→ Column Equality test that is safe for null values. New in version 2.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters othera value or Column Notes Unlike Pandas, PySpark doesn’t consider NaN values to be NULL. S...

- **pyspark.sql.Column.getField**

  ...expression that gets a field by name in a StructType. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters namea literal value, or a Column expression. The result will only be true at a location if the fi...

- **pyspark.sql.Column.getItem**

  ...out of a list, or gets an item by key out of a dict. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters keya literal value, or a Column expression. The result will only be true at a location if the ite...

- **pyspark.sql.Column.ilike**

  ...s a boolean Column based on a case insensitive match. New in version 3.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters otherstra SQL LIKE pattern Returns ColumnColumn of booleans showing whether each element in t...

- **pyspark.sql.Column.isin**

  ...s contained by the evaluated values of the arguments. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colsThe result will only be true at a location if any value matches in the Column. Returns Co...

- **pyspark.sql.Column.isNotNull**

  ...ull() → pyspark.sql.column.Column True if the current expression is NOT null. Changed in version 3.4.0: Supports Spark Connect. Examples >>> from pyspark.sql import Row >>> df = spark.createDataFrame([Row(name='Tom', height=80), Row(name=...

- **pyspark.sql.Column.isNull**

....isNull() → pyspark.sql.column.Column True if the current expression is null. Changed in version 3.4.0: Supports Spark Connect. Examples >>> from pyspark.sql import Row >>> df = spark.createDataFrame([Row(name='Tom', height=80), Row(name=...

- **pyspark.sql.Column.like**

  ...rce] SQL like expression. Returns a boolean Column based on a SQL LIKE match. Changed in version 3.4.0: Supports Spark Connect. Parameters otherstra SQL LIKE pattern Returns ColumnColumn of booleans showing whether each element in t...

- **pyspark.sql.Column.otherwise**

  ...t invoked, None is returned for unmatched conditions. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters valuea literal value, or a Column expression. Returns ColumnColumn representing whether each...

- **pyspark.sql.Column.over**

  ...ndowSpec) → Column[source] Define a windowing column. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters windowWindowSpec Returns Column Examples >>> from pyspark.sql import Window >>> window = (...

- **pyspark.sql.Column.rlike**

  ...xpression (LIKE with Regex). Returns a boolean Column based on a regex match. Changed in version 3.4.0: Supports Spark Connect. Parameters otherstran extended regex expression Returns ColumnColumn of booleans showing whether each el...

- **pyspark.sql.Column.startswith**

  ...Column String starts with. Returns a boolean Column based on a string match. Changed in version 3.4.0: Supports Spark Connect. Parameters otherColumn or strstring at start of line (do not use a regex ^) Examples >>> df = spark.crea...

- **pyspark.sql.Column.substr**

  ...] Return a Column which is a substring of the column. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters startPosColumn or intstart position lengthColumn or intlength of the substring Returns Colum...

- **pyspark.sql.Column.when**

  ...t invoked, None is returned for unmatched conditions. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters conditionColumna boolean Column expression. valuea literal value, or a Column expression. Ret...

- **pyspark.sql.Column.withField**

  ...ion that adds/replaces a field in StructType by name. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters fieldNamestra literal value. The result will only be true at a location if any field matches in t...

- **pyspark.sql.conf.RuntimeConfig**

  ...set here are automatically propagated to the Hadoop configuration during I/O. Changed in version 3.4.0: Supports Spark Connect. Methods get(key[, default]) Returns the value of Spark runtime configuration property for the given key,...

- **pyspark.sql.DataFrame**

  ...ibuted collection of data grouped into named columns. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Notes A DataFrame should only be created as described above. It should not be directly created via using the c...

- **pyspark.sql.DataFrame.__getattr__**

  ...mn.Column[source] Returns the Column denoted by name. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters namestrColumn name to return as Column. Returns ColumnRequested column. Examples >>> df =...

- **pyspark.sql.DataFrame.__getitem__**

  ...me.DataFrame][source] Returns the column as a Column. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters itemint, str, Column, list or tuplecolumn index, column name, column, or a list or tuple of colum...

- **pyspark.sql.DataFrame.agg**

  ...me without groups (shorthand for df.groupBy().agg()). New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters exprsColumn or dict of key and value stringsColumns or expressions to aggregate DataFrame by....

- **pyspark.sql.DataFrame.alias**

  ...me[source] Returns a new DataFrame with an alias set. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters aliasstran alias name to be set for the DataFrame. Returns DataFrameAliased DataFrame. Ex...

- **pyspark.sql.DataFrame.approxQuantile**

  ...tion of Quantile Summaries]] by Greenwald and Khanna. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col: str, tuple or listCan be a single column name, or a list of names for multiple columns. Cha...

- **pyspark.sql.DataFrame.cache**

  ...th the default storage level (MEMORY_AND_DISK_DESER). New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Returns DataFrameCached DataFrame. Notes The default storage level has changed to MEMORY_AND_DISK_DESER...

- **pyspark.sql.DataFrame.coalesce**

  ...parallel (per whatever the current partitioning is). New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters numPartitionsintspecify the target number of partitions Returns DataFrame Examples >>> df...

- **pyspark.sql.DataFrame.collect**

  ...ow][source] Returns all the records as a list of Row. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Returns listList of rows. Examples >>> df = spark.createDataFrame(... [(14, "Tom"), (23, "Alice"),...

- **pyspark.sql.DataFrame.colRegex**

  ...n name specified as a regex and returns it as Column. New in version 2.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colNamestrstring, column name specified as a regex. Returns Column Examples >>> df = spark...

- **pyspark.sql.DataFrame.columns**

  ...es in the list reflects their order in the DataFrame. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Returns listList of column names in the DataFrame. Examples Example 1: Retrieve column names of a DataFr...

- **pyspark.sql.DataFrame.corr**

  ...aFrameStatFunctions.corr() are aliases of each other. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col1strThe name of the first column col2strThe name of the second column methodstr, optionalThe...

- **pyspark.sql.DataFrame.count**

  ...source] Returns the number of rows in this DataFrame. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Returns intNumber of rows. Examples >>> df = spark.createDataFrame(... [(14, "Tom"), (23, "Alice"),...

- **pyspark.sql.DataFrame.cov**

  ...e.cov() and DataFrameStatFunctions.cov() are aliases. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col1strThe name of the first column col2strThe name of the second column Returns floatCovari...

- **pyspark.sql.DataFrame.createGlobalTempView**

  ...tion, if the view name already exists in the catalog. New in version 2.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters namestrName of the view. Examples Create a global temporary view. >>> df = spark.createDataFr...

- **pyspark.sql.DataFrame.createOrReplaceGlobalTempView**

  ...his temporary view is tied to this Spark application. New in version 2.2.0. Changed in version 3.4.0: Supports Spark Connect. Parameters namestrName of the view. Examples Create a global temporary view. >>> df = spark.createDataFr...

- **pyspark.sql.DataFrame.createOrReplaceTempView**

  ...SparkSession that was used to create this DataFrame. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters namestrName of the view. Examples Create a local temporary view named ‘people’. >>> df = spar...

- **pyspark.sql.DataFrame.createTempView**

...tion, if the view name already exists in the catalog. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters namestrName of the view. Examples Create a local temporary view. >>> df = spark.createDataFra...

- **pyspark.sql.DataFrame.crossJoin**

  ...Returns the cartesian product with another DataFrame. New in version 2.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters otherDataFrameRight side of the cartesian product. Returns DataFrameJoined DataFrame. Exa...

- **pyspark.sql.DataFrame.crosstab**

  ...() and DataFrameStatFunctions.crosstab() are aliases. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col1strThe name of the first column. Distinct items will make the first item of each row. col2st...

- **pyspark.sql.DataFrame.cube**

  ...pecified columns, so we can run aggregations on them. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colslist, str or Columncolumns to create cube by. Each element should be a column name (string) o...

- **pyspark.sql.DataFrame.describe**

  ...utes basic statistics for numeric and string columns. New in version 1.3.1. Changed in version 3.4.0: Supports Spark Connect. This includes count, mean, stddev, min, and max. If no columns are given, this function computes statistics fo...

- **pyspark.sql.DataFrame.distinct**

  ...Frame containing the distinct rows in this DataFrame. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Returns DataFrameDataFrame with distinct records. Examples >>> df = spark.createDataFrame(... [(14,...

- **pyspark.sql.DataFrame.drop**

  ...the schema doesn’t contain the given column name(s). New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters cols: str or :class:`Column`a name of the column, or the Column to drop Returns DataFrameData...

- **pyspark.sql.DataFrame.dropDuplicates**

  .... drop_duplicates() is an alias for dropDuplicates(). New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters subsetList of column names, optionalList of columns to use for duplicate comparison (default All...

- **pyspark.sql.DataFrame.dropDuplicatesWithinWatermark**

  ...duplicate comparison (default All columns). Returns DataFrameDataFrame without duplicates. Notes Supports Spark Connect. Examples >>> from pyspark.sql import Row >>> from pyspark.sql.functions import timestamp_seconds >>> df = spark...

- **pyspark.sql.DataFrame.dropna**

  ...ataFrameNaFunctions.drop() are aliases of each other. New in version 1.3.1. Changed in version 3.4.0: Supports Spark Connect. Parameters howstr, optional‘any’ or ‘all’. If ‘any’, drop a row if it contains any nulls. If ‘all’, drop a r...

- **pyspark.sql.DataFrame.dtypes**

  ...urns all column names and their data types as a list. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Returns listList of columns as tuple pairs. Examples >>> df = spark.createDataFrame(... [(14, "Tom"...

- **pyspark.sql.DataFrame.exceptAll**

  ...function resolves columns by position (not by name). New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters otherDataFrameThe other DataFrame to compare to. Returns DataFrame Examples >>> df1 = spar...

- **pyspark.sql.DataFrame.explain**

  ...hysical) plans to the console for debugging purposes. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters extendedbool, optionaldefault False. If False, prints only the physical plan. When this is a stri...

- **pyspark.sql.DataFrame.fillna**

  ...ataFrameNaFunctions.fill() are aliases of each other. New in version 1.3.1. Changed in version 3.4.0: Supports Spark Connect. Parameters valueint, float, string, bool or dictValue to replace null values with. If the value is a dict, t...

- **pyspark.sql.DataFrame.filter**

  ...he given condition. where() is an alias for filter(). New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters conditionColumn or stra Column of types.BooleanType or a string of SQL expressions. Returns D...

- **pyspark.sql.DataFrame.first**

  ...ql.types.Row][source] Returns the first row as a Row. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Returns RowFirst row if DataFrame is not empty, otherwise None. Examples >>> df = spark.createDataFrame(...

- **pyspark.sql.DataFrame.freqItems**

  ...) and DataFrameStatFunctions.freqItems() are aliases. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colslist or tupleNames of the columns to calculate frequent items for as a list or tuple of strin...

- **pyspark.sql.DataFrame.groupBy**

  ...egate functions. groupby() is an alias for groupBy(). New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colslist, str or Columncolumns to group by. Each element should be a column name (string) or an e...

- **pyspark.sql.DataFrame.head**

  ...ark.sql.types.Row]][source] Returns the first n rows. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters nint, optionaldefault 1. Number of rows to return. Returns If n is greater than 1, return a l...

- **pyspark.sql.DataFrame.hint**

  ...source] Specifies some hint on the current DataFrame. New in version 2.2.0. Changed in version 3.4.0: Supports Spark Connect. Parameters namestrA name of the hint. parametersstr, list, float or intOptional parameters. Returns Dat...

- **pyspark.sql.DataFrame.inputFiles**

  ...may not find all input files. Duplicates are removed. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Returns listList of file paths. Examples >>> import tempfile >>> with tempfile.TemporaryDirectory() as d...

- **pyspark.sql.DataFrame.intersect**

  ...e removed. To preserve duplicates use intersectAll(). New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters otherDataFrameAnother DataFrame that needs to be combined. Returns DataFrameCombined DataFram...

- **pyspark.sql.DataFrame.intersectAll**

  ...function resolves columns by position (not by name). New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters otherDataFrameAnother DataFrame that needs to be combined. Returns DataFrameCombined DataFram...

- **pyspark.sql.DataFrame.isEmpty**

  ...f the DataFrame is empty and returns a boolean value. New in version 3.3.0. Changed in version 3.4.0: Supports Spark Connect. Returns boolReturns True if the DataFrame is empty, False otherwise. See also DataFrame.countCounts th...

- **pyspark.sql.DataFrame.isLocal**

  ...ods can be run locally (without any Spark executors). New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Returns bool Examples >>> df = spark.sql("SHOW TABLES") >>> df.isLocal() True...

- **pyspark.sql.DataFrame.isStreaming**

  ...isException when there is a streaming source present. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Returns boolWhether it’s streaming DataFrame or not. Notes This API is evolving. Examples >>> df = spark...

- **pyspark.sql.DataFrame.join**

  ...h another DataFrame, using the given join expression. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters otherDataFrameRight side of the join onstr, list or Column, optionala string for the join column...

- **pyspark.sql.DataFrame.limit**

...rce] Limits the result count to the number specified. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters numintNumber of records to return. Will return this number of records or all records if the DataF...

- **pyspark.sql.DataFrame.mapInPandas**

  ...of the function's input and output can be different. New in version 3.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters funcfunctiona Python native function that takes an iterator of pandas.DataFrames, and outputs an...

- **pyspark.sql.DataFrame.melt**

  ...of the value column. Returns DataFrameUnpivoted DataFrame. See also DataFrame.unpivot Notes Supports Spark Connect. previous...

- **pyspark.sql.DataFrame.na**

  ...s a DataFrameNaFunctions for handling missing values. New in version 1.3.1. Changed in version 3.4.0: Supports Spark Connect. Returns DataFrameNaFunctions Examples >>> df = spark.sql("SELECT 1 AS c1, int(NULL) AS c2") >>> type(df.n...

- **pyspark.sql.DataFrame.observe**

  ...sql.util.QueryExecutionListener to the spark session. New in version 3.3.0. Changed in version 3.5.0: Supports Spark Connect. Parameters observationObservation or strstr to specify the name, or an Observation instance to obtain the me...

- **pyspark.sql.DataFrame.orderBy**

  ...ns a new DataFrame sorted by the specified column(s). New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colsstr, list, or Column, optionallist of Column or column names to sort by. Returns DataFram...

- **pyspark.sql.DataFrame.pandas_api**

  ...existing DataFrame into a pandas-on-Spark DataFrame. New in version 3.2.0. Changed in version 3.5.0: Supports Spark Connect. If a pandas-on-Spark DataFrame is converted to a Spark DataFrame and then back to pandas-on-Spark, it will los...

- **pyspark.sql.DataFrame.persist**

  ...evel is specified defaults to (MEMORY_AND_DISK_DESER) New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters storageLevelStorageLevelStorage level to set for persistence. Default is MEMORY_AND_DISK_DESER....

- **pyspark.sql.DataFrame.printSchema**

  ...specify how many levels to print if schema is nested. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters levelint, optional, default NoneHow many levels to print for nested schemas. Changed in version...

- **pyspark.sql.DataFrame.randomSplit**

  ...omly splits this DataFrame with the provided weights. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters weightslistlist of doubles as weights with which to split the DataFrame. Weights will be normaliz...

- **pyspark.sql.DataFrame.registerTempTable**

  ...SparkSession that was used to create this DataFrame. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Deprecated since version 2.0.0: Use DataFrame.createOrReplaceTempView() instead. Parameters namestrName of...

- **pyspark.sql.DataFrame.repartition**

  ...essions. The resulting DataFrame is hash partitioned. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters numPartitionsintcan be an int to specify the target number of partitions or a Column. If it is a...

- **pyspark.sql.DataFrame.repartitionByRange**

  ...ssions. The resulting DataFrame is range partitioned. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters numPartitionsintcan be an int to specify the target number of partitions or a Column. If it is a...

- **pyspark.sql.DataFrame.replace**

  ...-1, 42.0: 1}) and arbitrary replacement will be used. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters to_replacebool, int, float, string, list or dictValue to be replaced. If the value is a dict, the...

- **pyspark.sql.DataFrame.rollup**

  ...specified columns, so we can run aggregation on them. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colslist, str or ColumnColumns to roll-up by. Each element should be a column name (string) or an...

- **pyspark.sql.DataFrame.sameSemantics**

  ...ames are equal and therefore return the same results. New in version 3.1.0. Changed in version 3.5.0: Supports Spark Connect. Parameters otherDataFrameThe other DataFrame to compare against. Returns boolWhether these two DataFrame...

- **pyspark.sql.DataFrame.sample**

  ...e[source] Returns a sampled subset of this DataFrame. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters withReplacementbool, optionalSample with replacement or not (default False). fractionfloat, opti...

- **pyspark.sql.DataFrame.sampleBy**

  ...lacement based on the fraction given on each stratum. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strcolumn that defines strata Changed in version 3.0.0: Added sampling by a column...

- **pyspark.sql.DataFrame.schema**

  ...of this DataFrame as a pyspark.sql.types.StructType. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Returns StructType Examples >>> df = spark.createDataFrame(... [(14, "Tom"), (23, "Alice"), (16, "Bo...

- **pyspark.sql.DataFrame.select**

  ...cts a set of expressions and returns a new DataFrame. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colsstr, Column, or listcolumn names (string) or expressions (Column). If one of the column names...

- **pyspark.sql.DataFrame.selectExpr**

  ...s a variant of select() that accepts SQL expressions. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Returns DataFrameA DataFrame with new/old columns transformed by expressions. Examples >>> df = spark.cr...

- **pyspark.sql.DataFrame.semanticHash**

  ...ode of the logical query plan against this DataFrame. New in version 3.1.0. Changed in version 3.5.0: Supports Spark Connect. Returns intHash value. Notes Unlike the standard hash code, the hash is calculated against the query pla...

- **pyspark.sql.DataFrame.show**

  ...None[source] Prints the first n rows to the console. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters nint, optionalNumber of rows to show. truncatebool or int, optionalIf set to True, truncate stri...

- **pyspark.sql.DataFrame.sort**

  ...ns a new DataFrame sorted by the specified column(s). New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colsstr, list, or Column, optionallist of Column or column names to sort by. Returns DataFram...

- **pyspark.sql.DataFrame.sortWithinPartitions**

  ...ith each partition sorted by the specified column(s). New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colsstr, list or Column, optionallist of Column or column names to sort by. Returns DataFrame...

- **pyspark.sql.DataFrame.sparkSession**

  ...on Returns Spark session that created this DataFrame. New in version 3.3.0. Changed in version 3.4.0: Supports Spark Connect. Returns SparkSession Examples >>> df = spark.range(1) >>> type(df.sparkSession) <class '...session.SparkS...

- **pyspark.sql.DataFrame.stat**

  ...rns a DataFrameStatFunctions for statistic functions. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Returns DataFrameStatFunctions Examples >>> import pyspark.sql.functions as f >>> df = spark.range(3).wit...

- **pyspark.sql.DataFrame.storageLevel**

...orageLevel Get the DataFrame’s current storage level. New in version 2.1.0. Changed in version 3.4.0: Supports Spark Connect. Returns StorageLevelCurrently defined storage level. Examples >>> df1 = spark.range(10) >>> df1.storageL...

- **pyspark.sql.DataFrame.subtract**

  ...rows in this DataFrame but not in another DataFrame. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters otherDataFrameAnother DataFrame that needs to be subtracted. Returns DataFrameSubtracted Data...

- **pyspark.sql.DataFrame.summary**

  ...uartiles (percentiles at 25%, 50%, and 75%), and max. New in version 2.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters statisticsstr, optionalColumn names to calculate statistics by (default All columns). Returns...

- **pyspark.sql.DataFrame.tail**

  ...m can crash the driver process with OutOfMemoryError. New in version 3.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters numintNumber of records to return. Will return this number of records or all records if the DataF...

- **pyspark.sql.DataFrame.take**

  ...[source] Returns the first num rows as a list of Row. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters numintNumber of records to return. Will return this number of records or all records if the DataF...

- **pyspark.sql.DataFrame.to**

  ...the column and/or inner fieldis nullable but the specified schema requires them to be not nullable. Supports Spark Connect. Examples >>> from pyspark.sql.types import StructField, StringType >>> df = spark.createDataFrame([("a", 1)], [...

- **pyspark.sql.DataFrame.toDF**

  ...a new DataFrame that with new specified column names New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters *colstuplea tuple of string new column name. The length of the list needs to be the same as the n...

- **pyspark.sql.DataFrame.toLocalIterator**

  ...consume up to the memory of the 2 largest partitions. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters prefetchPartitionsbool, optionalIf Spark should pre-fetch the next partition before it is needed....

- **pyspark.sql.DataFrame.toPandas**

  ...only available if Pandas is installed and available. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Notes This method should only be used if the resulting Pandas pandas.DataFrame is expected to be small, as all...

- **pyspark.sql.DataFrame.transform**

  .... Concise syntax for chaining custom transformations. New in version 3.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters funcfunctiona function that takes and returns a DataFrame. *argsPositional arguments to pass to...

- **pyspark.sql.DataFrame.union**

  ...ning the union of rows in this and another DataFrame. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters otherDataFrameAnother DataFrame that needs to be unioned. Returns DataFrameA new DataFrame co...

- **pyspark.sql.DataFrame.unionAll**

  ...ning the union of rows in this and another DataFrame. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters otherDataFrameAnother DataFrame that needs to be combined Returns DataFrameA new DataFrame co...

- **pyspark.sql.DataFrame.unionByName**

  ...ns is True, missing columns will be filled with null. New in version 2.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters otherDataFrameAnother DataFrame that needs to be combined. allowMissingColumnsbool, optional, de...

- **pyspark.sql.DataFrame.unpersist**

  ...t, and remove all blocks for it from memory and disk. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters blockingboolWhether to block until all blocks are deleted. Returns DataFrameUnpersisted DataF...

- **pyspark.sql.DataFrame.unpivot**

  ...ame of the value column. Returns DataFrameUnpivoted DataFrame. See also DataFrame.melt Notes Supports Spark Connect. Examples >>> df = spark.createDataFrame(... [(1, 11, 1.1), (2, 12, 1.2)], ... ["id", "int", "double"],...

- **pyspark.sql.DataFrame.withColumn**

  ...column from some other DataFrame will raise an error. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colNamestrstring, name of the new column. colColumna Column expression for the new column. Re...

- **pyspark.sql.DataFrame.withColumnRenamed**

  ...if the schema doesn’t contain the given column name. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters existingstrstring, name of the existing column to rename. newstrstring, new name of the column....

- **pyspark.sql.DataFrame.withColumns**

  ...er Dataset. New in version 3.3.0: Added support for multiple columns adding Changed in version 3.4.0: Supports Spark Connect. Parameters colsMapdicta dict of column name and Column. Currently, only a single map is supported. Return...

- **pyspark.sql.DataFrame.withColumnsRenamed**

  ...upported. Returns DataFrameDataFrame with renamed columns. See also withColumnRenamed() Notes Support Spark Connect Examples >>> df = spark.createDataFrame([(2, "Alice"), (5, "Bob")], schema=["age", "name"]) >>> df = df.withColu...

- **pyspark.sql.DataFrame.withMetadata**

  ...taFrame by updating an existing column with metadata. New in version 3.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters columnNamestrstring, name of the existing column to update the metadata. metadatadictdict, new m...

- **pyspark.sql.DataFrame.withWatermark**

  ...ss records that arrive more than delayThreshold late. New in version 2.1.0. Changed in version 3.5.0: Supports Spark Connect. Parameters eventTimestrthe name of the column that contains the event time of the row. delayThresholdstrthe...

- **pyspark.sql.DataFrame.write**

  ...he non-streaming DataFrame out into external storage. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Returns DataFrameWriter Examples >>> df = spark.createDataFrame([(2, "Alice"), (5, "Bob")], schema=["age"...

- **pyspark.sql.DataFrame.writeStream**

  ...of the streaming DataFrame out into external storage. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Returns DataStreamWriter Notes This API is evolving. Examples >>> import time >>> import tempfile >>> df...

- **pyspark.sql.DataFrame.writeTo**

  ...mple, to append or create or replace existing tables. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters tablestrTarget table name to write to. Returns DataFrameWriterV2DataFrameWriterV2 to use furt...

- **pyspark.sql.DataFrameNaFunctions**

  ...tionality for working with missing data in DataFrame. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Methods drop([how, thresh, subset]) Returns a new DataFrame omitting rows with null values. fill(value[...

- **pyspark.sql.DataFrameNaFunctions.drop**

  ...ataFrameNaFunctions.drop() are aliases of each other. New in version 1.3.1. Changed in version 3.4.0: Supports Spark Connect. Parameters howstr, optional‘any’ or ‘all’. If ‘any’, drop a row if it contains any nulls. If ‘all’, drop a r...

- **pyspark.sql.DataFrameNaFunctions.fill**

  ...ataFrameNaFunctions.fill() are aliases of each other. New in version 1.3.1. Changed in version 3.4.0: Supports Spark Connect. Parameters valueint, float, string, bool or dictValue to replace null values with. If the value is a dict, t...

- **pyspark.sql.DataFrameNaFunctions.replace**

...-1, 42.0: 1}) and arbitrary replacement will be used. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters to_replacebool, int, float, string, list or dictValue to be replaced. If the value is a dict, the...

- **pyspark.sql.DataFrameReader**

  ...e stores, etc). Use SparkSession.read to access this. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Methods csv(path[, schema, sep, encoding, quote, …]) Loads a CSV file and returns the result as a DataF...

- **pyspark.sql.DataFrameReader.csv**

  ...option or specify the schema explicitly using schema. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters pathstr or liststring, or list of strings, for input path(s), or RDD of Strings storing CSV rows....

- **pyspark.sql.DataFrameReader.format**

  ...eader[source] Specifies the input data source format. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters sourcestrstring, name of the data source, e.g. 'json', 'parquet'. Examples >>> spark.read.for...

- **pyspark.sql.DataFrameReader.jdbc**

  ...) → DataFrame[source] Construct a DataFrame representing the database table named table accessible via JDBC URL url and connection properties. Partitions of the table will be retrieved in parallel if either column or predicates is specified...

- **pyspark.sql.DataFrameReader.json**

  ...through the input once to determine the input schema. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters pathstr, list or RDDstring represents path to the JSON dataset, or a list of paths, or RDD of Str...

- **pyspark.sql.DataFrameReader.load**

  ...ata from a data source and returns it as a DataFrame. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters pathstr or list, optionaloptional string or a list of string for file-system backed data sources....

- **pyspark.sql.DataFrameReader.option**

  ...Adds an input option for the underlying data source. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters keystrThe key for the option to set. valueThe value for the option to set. Examples >>> spar...

- **pyspark.sql.DataFrameReader.options**

  ...e] Adds input options for the underlying data source. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters **optionsdictThe dictionary of string keys and prmitive-type values. Examples >>> spark.read....

- **pyspark.sql.DataFrameReader.orc**

  ...Loads ORC files, returning the result as a DataFrame. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters pathstr or list Other Parameters Extra optionsFor the extra options, refer to Data Source Opti...

- **pyspark.sql.DataFrameReader.parquet**

  ...s Parquet files, returning the result as a DataFrame. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters pathsstr Other Parameters **optionsFor the extra options, refer to Data Source Option for the...

- **pyspark.sql.DataFrameReader.schema**

  ...chema inference step, and thus speed up data loading. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters schemapyspark.sql.types.StructType or stra pyspark.sql.types.StructType object or a DDL-formatted...

- **pyspark.sql.DataFrameReader.table**

  ...e[source] Returns the specified table as a DataFrame. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters tableNamestrstring, name of the table. Examples >>> df = spark.range(10) >>> df.createOrRepla...

- **pyspark.sql.DataFrameReader.text**

  ...he text file is a new row in the resulting DataFrame. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters pathsstr or liststring, or list of strings, for input path(s). Other Parameters Extra options...

- **pyspark.sql.DataFrameStatFunctions**

  ...Functionality for statistic functions with DataFrame. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Methods approxQuantile(col, probabilities, relativeError) Calculates the approximate quantiles of numeri...

- **pyspark.sql.DataFrameStatFunctions.approxQuantile**

  ...tion of Quantile Summaries]] by Greenwald and Khanna. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col: str, tuple or listCan be a single column name, or a list of names for multiple columns. Cha...

- **pyspark.sql.DataFrameStatFunctions.corr**

  ...aFrameStatFunctions.corr() are aliases of each other. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col1strThe name of the first column col2strThe name of the second column methodstr, optionalThe...

- **pyspark.sql.DataFrameStatFunctions.cov**

  ...e.cov() and DataFrameStatFunctions.cov() are aliases. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col1strThe name of the first column col2strThe name of the second column Returns floatCovari...

- **pyspark.sql.DataFrameStatFunctions.crosstab**

  ...() and DataFrameStatFunctions.crosstab() are aliases. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col1strThe name of the first column. Distinct items will make the first item of each row. col2st...

- **pyspark.sql.DataFrameStatFunctions.freqItems**

  ...) and DataFrameStatFunctions.freqItems() are aliases. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colslist or tupleNames of the columns to calculate frequent items for as a list or tuple of strin...

- **pyspark.sql.DataFrameStatFunctions.sampleBy**

  ...lacement based on the fraction given on each stratum. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strcolumn that defines strata Changed in version 3.0.0: Added sampling by a column...

- **pyspark.sql.DataFrameWriter**

  ...lue stores, etc). Use DataFrame.write to access this. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Methods bucketBy(numBuckets, col, *cols) Buckets the output by the given columns. csv(path[, mode, comp...

- **pyspark.sql.DataFrameWriter.bucketBy**

  ...function and is not compatible with Hive’s bucketing. New in version 2.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters numBucketsintthe number of buckets to save colstr, list or tuplea name of a column, or a list of...

- **pyspark.sql.DataFrameWriter.csv**

  ...of the DataFrame in CSV format at the specified path. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters pathstrthe path in any Hadoop supported file system modestr, optionalspecifies the behavior of t...

- **pyspark.sql.DataFrameWriter.format**

  ...[source] Specifies the underlying output data source. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters sourcestrstring, name of the data source, e.g. ‘json’, ‘parquet’. Examples >>> spark.range(1)...

- **pyspark.sql.DataFrameWriter.insertInto**

  ...the DataFrame is the same as the schema of the table. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters overwritebool, optionalIf true, overwrites existing data. Disabled by default Notes Unlike Da...

- **pyspark.sql.DataFrameWriter.jdbc**

  ...the DataFrame to an external database table via JDBC. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters tablestrName of the table in the external database. modestr, optionalspecifies the behavior of t...

- **pyspark.sql.DataFrameWriter.json**

...mat or newline-delimited JSON) at the specified path. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters pathstrthe path in any Hadoop supported file system modestr, optionalspecifies the behavior of t...

- **pyspark.sql.DataFrameWriter.mode**

  ...lently ignore this operation if data already exists. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Examples Raise an error when writing to an existing path. >>> import tempfile >>> with tempfile.TemporaryDirec...

- **pyspark.sql.DataFrameWriter.option**

  ...Adds an output option for the underlying data source. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters keystrThe key for the option to set. valueThe value for the option to set. Examples >>> spar...

- **pyspark.sql.DataFrameWriter.options**

  ...] Adds output options for the underlying data source. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters **optionsdictThe dictionary of string keys and primitive-type values. Examples >>> spark.rang...

- **pyspark.sql.DataFrameWriter.orc**

  ...of the DataFrame in ORC format at the specified path. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters pathstrthe path in any Hadoop supported file system modestr, optionalspecifies the behavior of t...

- **pyspark.sql.DataFrameWriter.parquet**

  ...he DataFrame in Parquet format at the specified path. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters pathstrthe path in any Hadoop supported file system modestr, optionalspecifies the behavior of t...

- **pyspark.sql.DataFrameWriter.partitionBy**

  ...he file system similar to Hive’s partitioning scheme. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colsstr or listname of columns Examples Write a DataFrame into a Parquet file in a partitione...

- **pyspark.sql.DataFrameWriter.save**

  ...configured by spark.sql.sources.default will be used. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters pathstr, optionalthe path in a Hadoop supported file system formatstr, optionalthe format used t...

- **pyspark.sql.DataFrameWriter.saveAsTable**

  ...lently ignore this operation if data already exists. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters namestrthe table name formatstr, optionalthe format used to save modestr, optionalone of append...

- **pyspark.sql.DataFrameWriter.sortBy**

  ...each bucket by the given columns on the file system. New in version 2.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colstr, tuple or lista name of a column, or a list of names. colsstradditional names (optional)....

- **pyspark.sql.DataFrameWriter.text**

  ...cified path. The text files will be encoded as UTF-8. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters pathstrthe path in any Hadoop supported file system Other Parameters Extra optionsFor the ext...

- **pyspark.sql.DataFrameWriterV2**

  ...frame.DataFrame to external storage using the v2 API. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Methods append() Append the contents of the data frame to the output table. create() Create a new table...

- **pyspark.sql.functions.abs**

  ...ql.column.Column[source] Computes the absolute value. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columncolumn for computed results. E...

- **pyspark.sql.functions.acos**

  ...[source] Computes inverse cosine of the input column. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columninverse cosine of col, as if compu...

- **pyspark.sql.functions.acosh**

  ...mputes inverse hyperbolic cosine of the input column. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnthe column for computed results....

- **pyspark.sql.functions.add_months**

  ...ese amount of months will be deducted from the start. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters startColumn or strdate column to work on. monthsColumn or str or inthow many months after the gi...

- **pyspark.sql.functions.aggregate**

  ...UserDefinedFunctions are not supported (SPARK-27052). New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression initialValueColumn or strinitial value. Name of col...

- **pyspark.sql.functions.approx_count_distinct**

  ...Column for approximate distinct count of column col. New in version 2.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or str rsdfloat, optionalmaximum relative standard deviation allowed (default = 0.05)....

- **pyspark.sql.functions.approxCountDistinct**

  ...al[float] = None) → pyspark.sql.column.Column[source] New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Deprecated since version 2.1.0: Use approx_count_distinct() instead....

- **pyspark.sql.functions.array**

  ...sql.column.Column[source] Creates a new array column. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colsColumn or strcolumn names or Columns that have the same data type. Returns Columna column...

- **pyspark.sql.functions.array_append**

  ...lumn expression. Returns Columnan array of values from first array along with the element. Notes Supports Spark Connect. Examples >>> from pyspark.sql import Row >>> df = spark.createDataFrame([Row(c1=["b", "a", "c"], c2="c")]) >>>...

- **pyspark.sql.functions.array_compact**

  ...mn or strname of column or expression Returns Columnan array by excluding the null values. Notes Supports Spark Connect. Examples >>> df = spark.createDataFrame([([1, None, 2, 3],), ([4, 5, None, 4],)], ['data']) >>> df.select(array...

- **pyspark.sql.functions.array_contains**

  ...array contains the given value, and false otherwise. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column containing array value :value or column to check for in array...

- **pyspark.sql.functions.array_distinct**

  ...on function: removes duplicate values from the array. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression Returns Columnan array of unique values. Exa...

- **pyspark.sql.functions.array_except**

  ...elements in col1 but not in col2, without duplicates. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col1Column or strname of column containing array col2Column or strname of column containing arra...

- **pyspark.sql.functions.array_insert**

  ...r a Column expression. Returns Columnan array of values, including the new specified value Notes Supports Spark Connect. Examples >>> df = spark.createDataFrame(... [(['a', 'b', 'c'], 2, 'd'), (['c', 'b', 'a'], -2, 'd')], ...

- **pyspark.sql.functions.array_intersect**

  ...he intersection of col1 and col2, without duplicates. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col1Column or strname of column containing array col2Column or strname of column containing arra...

- **pyspark.sql.functions.array_join**

...null_replacement if set, otherwise they are ignored. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. delimiterstrdelimiter used to concatenate elements nu...

- **pyspark.sql.functions.array_max**

  ...ion function: returns the maximum value of the array. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression Returns Columnmaximum value of an array. Exa...

- **pyspark.sql.functions.array_min**

  ...ion function: returns the minimum value of the array. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression Returns Columnminimum value of array. Exampl...

- **pyspark.sql.functions.array_position**

  ...ay. Returns null if either of the arguments are null. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. valueAnyvalue to look for. Returns Columnposition...

- **pyspark.sql.functions.array_remove**

  ...elements that equal to element from the given array. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column containing array element :element to be removed from the array...

- **pyspark.sql.functions.array_repeat**

  ...es an array containing a column repeated count times. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strcolumn name or column that contains the element to be repeated countColumn or st...

- **pyspark.sql.functions.array_sort**

  ...n version 2.4.0. Changed in version 3.4.0: Can take a comparator function. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression comparatorcallable, optionalA binary (Column, Colum...

- **pyspark.sql.functions.array_union**

  ...ts in the union of col1 and col2, without duplicates. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col1Column or strname of column containing array col2Column or strname of column containing arra...

- **pyspark.sql.functions.arrays_overlap**

  ...hem contains a null element; returns false otherwise. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Returns Columna column of Boolean type. Examples >>> df = spark.createDataFrame([(["a", "b"], ["b", "c"]...

- **pyspark.sql.functions.arrays_zip**

  ...truct type value will be a null for missing elements. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colsColumn or strcolumns of arrays to be merged. Returns Columnmerged array of entries. E...

- **pyspark.sql.functions.asc**

  ...ased on the ascending order of the given column name. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to sort by in the ascending order. Returns Columnthe column spe...

- **pyspark.sql.functions.asc_nulls_first**

  ...name, and null values return before non-null values. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to sort by in the ascending order. Returns Columnthe column spe...

- **pyspark.sql.functions.asc_nulls_last**

  ...n name, and null values appear after non-null values. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to sort by in the ascending order. Returns Columnthe column spe...

- **pyspark.sql.functions.ascii**

  ...ic value of the first character of the string column. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. Returns

Columnnumeric value. Examples >>> df =...

- **pyspark.sql.functions.asin**

  ...mn[source] Computes inverse sine of the input column. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columninverse sine of col, as if compute...

- **pyspark.sql.functions.asinh**

  ...Computes inverse hyperbolic sine of the input column. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnthe column for computed results....

- **pyspark.sql.functions.assert_true**

  ...exception with the provided error message otherwise. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strcolumn name or column that represents the input column to test errMsgColumn or s...

- **pyspark.sql.functions.atan**

  ...[source] Compute inverse tangent of the input column. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columninverse tangent of col, as if comp...

- **pyspark.sql.functions.atan2**

  ...mnOrName, float]) → pyspark.sql.column.Column[source] New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col1str, Column or floatcoordinate on y-axis col2str, Column or floatcoordinate on x-axis Ret...

- **pyspark.sql.functions.atanh**

  ...putes inverse hyperbolic tangent of the input column. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnthe column for computed results....

- **pyspark.sql.functions.avg**

  ...nction: returns the average of the values in a group. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnthe column for computed results....

- **pyspark.sql.functions.base64**

  ...of a binary column and returns it as a string column. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. Returns ColumnBASE64 encoding of string value....

- **pyspark.sql.functions.bin**

  ...presentation of the binary value of the given column. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. Returns Columnbinary representation of given value...

- **pyspark.sql.functions.bit_length**

  ...lates the bit length for the specified string column. New in version 3.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strSource column or strings Returns ColumnBit length of the col Examples >>>...

- **pyspark.sql.functions.bitwise_not**

  ...spark.sql.column.Column[source] Computes bitwise not. New in version 3.2.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnthe column for computed results....

- **pyspark.sql.functions.bitwiseNOT**

  ...spark.sql.column.Column[source] Computes bitwise not. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Deprecated since version 3.2.0: Use bitwise_not() instead....

- **pyspark.sql.functions.broadcast**

  ...DataFrame as small enough for use in broadcast joins. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Returns DataFrameDataFrame marked as ready for broadcast join. Examples >>> from pyspark.sql import type...

- **pyspark.sql.functions.bround**

...ode if scale >= 0 or at integral part when scale < 0. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strinput column to round. scaleint optional default 0scale value. Returns Colum...

- **pyspark.sql.functions.bucket**

  ...y type that partitions by a hash of the input column. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget date or timestamp column to work on. Returns Columndata partitioned by...

- **pyspark.sql.functions.cbrt**

  ...mn[source] Computes the cube-root of the given value. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnthe column for computed results....

- **pyspark.sql.functions.ceil**

  ...lumn[source] Computes the ceiling of the given value. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnthe column for computed results....

- **pyspark.sql.functions.ceiling**

  ...lumn[source] Computes the ceiling of the given value. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnthe column for computed results....

- **pyspark.sql.functions.coalesce**

  ...mn[source] Returns the first column that is not null. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colsColumn or strlist of columns to work on. Returns Columnvalue of the first column that is...

- **pyspark.sql.functions.col**

  ...rce] Returns a Column based on the given column name. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colstrthe name for the column Returns Columnthe corresponding column instance. Examples >...

- **pyspark.sql.functions.collect_list**

  ...function: returns a list of objects with duplicates. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnlist of objects with duplicates....

- **pyspark.sql.functions.collect_set**

  ...a set of objects with duplicate elements eliminated. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnlist of objects with no duplicates...

- **pyspark.sql.functions.column**

  ...lumn Returns a Column based on the given column name. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colstrthe name for the column Returns Columnthe corresponding column instance. Examples >...

- **pyspark.sql.functions.concat**

  ...trings, numeric, binary and compatible array columns. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colsColumn or strtarget column or columns to work on. Returns Columnconcatenated values. Type...

- **pyspark.sql.functions.concat_ws**

  ...to a single string column, using the given separator. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters sepstrwords separator. colsColumn or strlist of columns to work on. Returns Columnstring of...

- **pyspark.sql.functions.conv**

  ...a number in a string column from one base to another. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or stra column to convert base for. fromBase: intfrom base number. toBase: intto base...

- **pyspark.sql.functions.corr**

  ...he Pearson Correlation Coefficient for col1 and col2. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col1Column or strfirst column to calculate correlation.

col1Column or strsecond column to calcul...

- **pyspark.sql.functions.cos**

  ...n.Column[source] Computes cosine of the input column. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strangle in radians Returns Columncosine of the angle, as if computed by java.la...

- **pyspark.sql.functions.cosh**

  ...urce] Computes hyperbolic cosine of the input column. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strhyperbolic angle Returns Columnhyperbolic cosine of the angle, as if computed...

- **pyspark.sql.functions.cot**

  ...olumn[source] Computes cotangent of the input column. New in version 3.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strangle in radians. Returns Columncotangent of the angle. Examples >>> impo...

- **pyspark.sql.functions.count**

  ...ate function: returns the number of items in a group. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columncolumn for computed results. E...

- **pyspark.sql.functions.count_distinct**

  ...turns a new Column for distinct count of col or cols. New in version 3.2.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strfirst column to compute on. colsColumn or strother columns to compute on. Ret...

- **pyspark.sql.functions.countDistinct**

  ...nd it is encouraged to use count_distinct() directly. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. previous...

- **pyspark.sql.functions.covar_pop**

  ...olumn for the population covariance of col1 and col2. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col1Column or strfirst column to calculate covariance. col1Column or strsecond column to calcula...

- **pyspark.sql.functions.covar_samp**

  ...ew Column for the sample covariance of col1 and col2. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col1Column or strfirst column to calculate covariance. col1Column or strsecond column to calcula...

- **pyspark.sql.functions.crc32**

  ...cy check value (CRC32) of a binary column and returns the value as a bigint. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnthe column for computed results....

- **pyspark.sql.functions.create_map**

  ...k.sql.column.Column[source] Creates a new map column. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colsColumn or strcolumn names or Columns that are grouped as key-value pairs, e.g. (key1, value1,...

- **pyspark.sql.functions.csc**

  ...Column[source] Computes cosecant of the input column. New in version 3.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strangle in radians. Returns Columncosecant of the angle. Examples >>> impor...

- **pyspark.sql.functions.cume_dist**

  ...the fraction of rows that are below the current row. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Returns Columnthe column for calculating cumulative distribution. Examples >>> from pyspark.sql import W...

- **pyspark.sql.functions.current_date**

  ...ent_date within the same query return the same value. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Returns Columncurrent date. Examples >>> df = spark.range(1) >>> df.select(current_date()).show() +----...

- **pyspark.sql.functions.current_timestamp**

...imestamp within the same query return the same value. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Returns Columncurrent date and time. Examples >>> df = spark.range(1) >>> df.select(current_timestamp())...

- **pyspark.sql.functions.date_add**

  ...hen these amount of days will be deducted from start. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters startColumn or strdate column to work on. daysColumn or str or inthow many days after the given...

- **pyspark.sql.functions.date_format**

  ...All pattern letters of datetime pattern. can be used. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters dateColumn or strinput column of values to format. format: strformat to use to represent datetim...

- **pyspark.sql.functions.date_sub**

  ...lue then these amount of days will be added to start. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters startColumn or strdate column to work on. daysColumn or str or inthow many days before the given...

- **pyspark.sql.functions.date_trunc**

  ...estamp truncated to the unit specified by the format. New in version 2.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters formatstr‘year’, ‘yyyy’, ‘yy’ to truncate by year, ‘month’, ‘mon’, ‘mm’ to truncate by month, ‘da...

- **pyspark.sql.functions.datediff**

  ...source] Returns the number of days from start to end. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters endColumn or strto date column to work on. startColumn or strfrom date column to work on. Ret...

- **pyspark.sql.functions.dayofmonth**

  ...ay of the month of a given date/timestamp as integer. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget date/timestamp column to work on. Returns Columnday of the month for g...

- **pyspark.sql.functions.dayofweek**

  ...anges from 1 for a Sunday through to 7 for a Saturday New in version 2.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget date/timestamp column to work on. Returns Columnday of the week for gi...

- **pyspark.sql.functions.dayofyear**

  ...day of the year of a given date/timestamp as integer. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget date/timestamp column to work on. Returns Columnday of the year for gi...

- **pyspark.sql.functions.days**

  ...for timestamps and dates to partition data into days. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget date or timestamp column to work on. Returns Columndata partitioned by...

- **pyspark.sql.functions.decode**

  ...-8859-1’, ‘UTF-8’, ‘UTF-16BE’, ‘UTF-16LE’, ‘UTF-16’). New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. charsetstrcharset to use to decode to. Returns Col...

- **pyspark.sql.functions.degrees**

  ...n approximately equivalent angle measured in degrees. New in version 2.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strangle in radians Returns Columnangle in degrees, as if computed by java.lang....

- **pyspark.sql.functions.dense_rank**

  ...This is equivalent to the DENSE_RANK function in SQL. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Returns Columnthe column for calculating ranks. Examples >>> from pyspark.sql import Window, types >>> d...

- **pyspark.sql.functions.desc**

  ...sed on the descending order of the given column name. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to sort by in the descending order. Returns Columnthe column sp...

- **pyspark.sql.functions.desc_nulls_first**

  ...name, and null values appear before non-null values. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to sort by in the descending order. Returns Columnthe column sp...

- **pyspark.sql.functions.desc_nulls_last**

  ...n name, and null values appear after non-null values. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to sort by in the descending order. Returns Columnthe column sp...

- **pyspark.sql.functions.element_at**

  ...side the array boundaries then None will be returned. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column containing array or map extraction :index to check for in array o...

- **pyspark.sql.functions.encode**

  ...-8859-1', 'UTF-8', 'UTF-16BE', 'UTF-16LE', 'UTF-16'). New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. charsetstrcharset to use to encode. Returns Column...

- **pyspark.sql.functions.exists**

  ...redicate holds for one or more elements in the array. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression ffunction(x: Column) -> Column: ... returning the...

- **pyspark.sql.functions.exp**

  ...[source] Computes the exponential of the given value. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strcolumn to calculate exponential for. Returns Columnexponential of the given v...

- **pyspark.sql.functions.explode**

  ...e for elements in the map unless specified otherwise. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. Returns Columnone row per array item or map key val...

- **pyspark.sql.functions.explode_outer**

  ...e for elements in the map unless specified otherwise. New in version 2.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. Returns Columnone row per array item or map key val...

- **pyspark.sql.functions.expm1**

  ...omputes the exponential of the given value minus one. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strcolumn to calculate exponential for. Returns Columnexponential less one....

- **pyspark.sql.functions.expr**

  ...expression string into the column that it represents New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters strstrexpression defined in string. Returns Columncolumn representing the expression. Exa...

- **pyspark.sql.functions.factorial**

  ...mn[source] Computes the factorial of the given value. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or stra column to calculate factorial for. Returns Columnfactorial of given value....

- **pyspark.sql.functions.filter**

  ...lements for which a predicate holds in a given array. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression ffunctionA function that returns the Boolean expres...

- **pyspark.sql.functions.first**

  ...true. If all values are null, then null is returned. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strcolumn to fetch first value for. ignorenullsColumn or strif first value is null...

- **pyspark.sql.functions.flatten**

...han two levels, only one level of nesting is removed. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression Returns Columnflattened array. Examples >>>...

- **pyspark.sql.functions.floor**

  ...Column[source] Computes the floor of the given value. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strcolumn to find floor for. Returns Columnnearest integer that is less than or...

- **pyspark.sql.functions.forall**

  ...her a predicate holds for every element in the array. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression ffunction(x: Column) -> Column: ... returning the...

- **pyspark.sql.functions.format_number**

  ..._EVEN round mode, and returns the result as a string. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strthe column name of the numeric value to be formatted dintthe N decimal places...

- **pyspark.sql.functions.format_string**

  ...intf-style and returns the result as a string column. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters formatstrstring that can contain embedded format tags and used as result column’s value colsColu...

- **pyspark.sql.functions.from_csv**

  .... Returns null, in the case of an unparseable string. New in version 3.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or stra column or column name in CSV format schema :class:`~pyspark.sql.Column` or str...

- **pyspark.sql.functions.from_json**

  .... Returns null, in the case of an unparseable string. New in version 2.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or stra column or column name in JSON format schemaDataType or stra StructType, ArrayT...

- **pyspark.sql.functions.from_unixtime**

  ...in the current system time zone in the given format. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters timestampColumn or strcolumn of unix time values. formatstr, optionalformat to use to convert to...

- **pyspark.sql.functions.from_utc_timestamp**

  ...mp to string according to the session local timezone. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters timestampColumn or strthe column that contains timestamps tzColumn or strA string detailing the...

- **pyspark.sql.functions.get**

  ...e at given position. See also element_at() Notes The position is not 1 based, but 0 based index. Supports Spark Connect. Examples >>> df = spark.createDataFrame([(["a", "b", "c"], 1)], ['data', 'index']) >>> df.select(get(df.data, 1...

- **pyspark.sql.functions.get_json_object**

  ...will return null if the input json string is invalid. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strstring column in json format pathstrpath to the json object to extract Return...

- **pyspark.sql.functions.greatest**

  ...ters. It will return null if all parameters are null. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strcolumns to check for gratest value. Returns Columngratest value. Examples...

- **pyspark.sql.functions.grouping**

  ...aggregated or 0 for not aggregated in the result set. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strcolumn to check if it’s aggregated. Returns Columnreturns 1 for aggregated or...

- **pyspark.sql.functions.grouping_id**

  ...(n-1)) + (grouping(c2) << (n-2)) + … + grouping(cn) New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colsColumn or strcolumns to check for. Returns

Columnreturns level of the grouping it relates...

- **pyspark.sql.functions.hash**

  ...ven columns, and returns the result as an int column. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colsColumn or strone or more columns to compute on. Returns Columnhash value as int column....

- **pyspark.sql.functions.hex**

  ....sql.types.IntegerType or pyspark.sql.types.LongType. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. Returns Columnhexadecimal representation of given v...

- **pyspark.sql.functions.hour**

  ...e] Extract the hours of a given timestamp as integer. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget date/timestamp column to work on. Returns Columnhour part of the times...

- **pyspark.sql.functions.hours**

  ...ransform for timestamps to partition data into hours. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget date or timestamp column to work on. Returns Columndata partitioned by...

- **pyspark.sql.functions.hypot**

  ...^2 + b^2) without intermediate overflow or underflow. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col1str, Column or floata leg. col2str, Column or floatb leg. Returns Columnlength of the hy...

- **pyspark.sql.functions.initcap**

  ...st letter of each word to upper case in the sentence. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. Returns Columnstring with all first letters are upp...

- **pyspark.sql.functions.inline**

  ...Returns Columngenerator expression with the inline exploded result. See also explode() Notes Supports Spark Connect. Examples >>> from pyspark.sql import Row >>> df = spark.createDataFrame([Row(structlist=[Row(a=1, b=2), Row(a=3...

- **pyspark.sql.functions.inline_outer**

  ...umngenerator expression with the inline exploded result. See also explode_outer() inline() Notes Supports Spark Connect. Examples >>> from pyspark.sql import Row >>> df = spark.createDataFrame([... Row(id=1, structlist=[Row(a=1...

- **pyspark.sql.functions.input_file_name**

  ...g column for the file name of the current Spark task. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Returns Columnfile names. Examples >>> import os >>> path = os.path.abspath(__file__) >>> df = spark.rea...

- **pyspark.sql.functions.instr**

  ...ng. Returns null if either of the arguments are null. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters strColumn or strtarget column to work on. substrstrsubstring to look for. Returns Columnloca...

- **pyspark.sql.functions.isnan**

  ...An expression that returns true if the column is NaN. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns ColumnTrue if value is NaN and False oth...

- **pyspark.sql.functions.isnull**

  ...n expression that returns true if the column is null. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns ColumnTrue if value is null and False ot...

- **pyspark.sql.functions.json_tuple**

  ...for a json column according to the given field names. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strstring column in json format fieldsstra field or fields to extract Returns C...

- **pyspark.sql.functions.kurtosis**

...ction: returns the kurtosis of the values in a group. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnkurtosis of given column. Exam...

- **pyspark.sql.functions.lag**

  ...ition. This is equivalent to the LAG function in SQL. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression offsetint, optional default 1number of row to exten...

- **pyspark.sql.functions.last**

  ...true. If all values are null, then null is returned. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strcolumn to fetch last value for. ignorenullsColumn or strif last value is null th...

- **pyspark.sql.functions.last_day**

  ...ast day of the month which the given date belongs to. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters dateColumn or strtarget column to compute on. Returns Columnlast day of the month. Exampl...

- **pyspark.sql.functions.lead**

  ...tion. This is equivalent to the LEAD function in SQL. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression offsetint, optional default 1number of row to exten...

- **pyspark.sql.functions.least**

  ...ters. It will return null if all parameters are null. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colsColumn or strcolumn names or columns to be compared Returns Columnleast value. Exampl...

- **pyspark.sql.functions.length**

  ...ces. The length of binary data includes binary zeros. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. Returns Columnlength of the value. Examples >>>...

- **pyspark.sql.functions.levenshtein**

  ...es the Levenshtein distance of the two given strings. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters leftColumn or strfirst column value. rightColumn or strsecond column value. thresholdint, optio...

- **pyspark.sql.functions.lit**

  ...umn.Column[source] Creates a Column of literal value. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn, str, int, float, bool or list, NumPy literals or ndarray.the value to make it as a PyS...

- **pyspark.sql.functions.localtimestamp**

  ...imestamp within the same query return the same value. New in version 3.4.0. Changed in version 3.4.0: Supports Spark Connect. Returns Columncurrent local date and time. Examples >>> df = spark.range(1) >>> df.select(localtimestamp...

- **pyspark.sql.functions.locate**

  ...nce of substr in a string column, after position pos. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters substrstra string strColumn or stra Column of pyspark.sql.types.StringType posint, optionalstar...

- **pyspark.sql.functions.log**

  ...hen this takes the natural logarithm of the argument. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters arg1Column, str or floatbase number or actual number (in this case base is e) arg2Column, str or...

- **pyspark.sql.functions.log10**

  ...Computes the logarithm of the given value in Base 10. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strcolumn to calculate logarithm for. Returns Columnlogarithm of the given value...

- **pyspark.sql.functions.log1p**

  ...the natural logarithm of the “given value plus one”. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strcolumn to calculate natural logarithm for. Returns Columnnatural logarithm of...

- **pyspark.sql.functions.log2**

  ...source] Returns the base-2 logarithm of the argument. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or stra column to calculate logariphm for. Returns Columnlogariphm of given value....

- **pyspark.sql.functions.lower**

  ...n[source] Converts a string expression to lower case. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. Returns Columnlower case values. Examples >>> d...

- **pyspark.sql.functions.lpad**

  ...ce] Left-pad the string column to width len with pad. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. lenintlength of the final string. padstrchars to prep...

- **pyspark.sql.functions.ltrim**

  ...spaces from left end for the specified string value. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. Returns Columnleft trimmed values. Examples >>>...

- **pyspark.sql.functions.make_date**

  ...th a date built from the year, month and day columns. New in version 3.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters yearColumn or strThe year to build the date monthColumn or strThe month to build the date dayCo...

- **pyspark.sql.functions.map_concat**

  ...lumn[source] Returns the union of all the given maps. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colsColumn or strcolumn names or Columns Returns Columna map of merged entries from other map...

- **pyspark.sql.functions.map_contains_key**

  ...umn[source] Returns true if the map contains the key. New in version 3.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression value :a literal value Returns ColumnTrue if ke...

- **pyspark.sql.functions.map_entries**

  ...s an unordered array of all entries in the given map. New in version 3.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression Returns Columnan array of key value pairs as a s...

- **pyspark.sql.functions.map_filter**

  ...urns a map whose key-value pairs satisfy a predicate. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression ffunctiona binary function (k: Column, v: Column) -...

- **pyspark.sql.functions.map_from_arrays**

  ...umn.Column[source] Creates a new map from two arrays. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col1Column or strname of column containing a set of keys. All elements should not be null col2Co...

- **pyspark.sql.functions.map_from_entries**

  ...entries (key value struct types) to a map of values. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression Returns Columna map created from the given array...

- **pyspark.sql.functions.map_keys**

  ...ns an unordered array containing the keys of the map. New in version 2.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression Returns Columnkeys of the map as an array. E...

- **pyspark.sql.functions.map_values**

  ...an unordered array containing the values of the map. New in version 2.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression Returns Columnvalues of the map as an array....

- **pyspark.sql.functions.map_zip_with**

...en maps, key-wise into a single map using a function. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col1Column or strname of the first column or expression col2Column or strname of the second colu...

- **pyspark.sql.functions.max**

  ...turns the maximum value of the expression in a group. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columncolumn for computed results. E...

- **pyspark.sql.functions.max_by**

  ...s the value associated with the maximum value of ord. New in version 3.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. ordColumn or strcolumn to be maximized Returns...

- **pyspark.sql.functions.md5**

  ...t and returns the value as a 32 character hex string. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnthe column for computed results....

- **pyspark.sql.functions.mean**

  ...average of the values in a group. An alias of avg(). New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnthe column for computed results....

- **pyspark.sql.functions.median**

  ...lumn or strtarget column to compute on. Returns Columnthe median of the values in a group. Notes Supports Spark Connect. Examples >>> df = spark.createDataFrame([... ("Java", 2012, 20000), ("dotNET", 2012, 5000), ... ("Java...

- **pyspark.sql.functions.min**

  ...turns the minimum value of the expression in a group. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columncolumn for computed results. E...

- **pyspark.sql.functions.min_by**

  ...s the value associated with the minimum value of ord. New in version 3.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. ordColumn or strcolumn to be minimized Returns...

- **pyspark.sql.functions.minute**

  ...Extract the minutes of a given timestamp as integer. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget date/timestamp column to work on. Returns Columnminutes part of the ti...

- **pyspark.sql.functions.mode**

  ...olumn or strtarget column to compute on. Returns Columnthe most frequent value in a group. Notes Supports Spark Connect. Examples >>> df = spark.createDataFrame([... ("Java", 2012, 20000), ("dotNET", 2012, 5000), ... ("Java...

- **pyspark.sql.functions.monotonically_increasing_id**

  ..., and each partition has less than 8 billion records. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Returns Columnlast value of the group. Notes The function is non-deterministic because its result depend...

- **pyspark.sql.functions.month**

  ...tract the month of a given date/timestamp as integer. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget date/timestamp column to work on. Returns Columnmonth part of the date...

- **pyspark.sql.functions.months**

  ...r timestamps and dates to partition data into months. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget date or timestamp column to work on. Returns Columndata partitioned by...

- **pyspark.sql.functions.months_between**

  ...nded off to 8 digits unless roundOff is set to False. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters date1Column or strfirst date column. date2Column or strsecond date column. roundOffbool, option...

- **pyspark.sql.functions.nanvl**

  ...be floating point columns (DoubleType or FloatType). New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col1Column or strfirst column to check. col2Column or strsecond column to return if first is NaN...

- **pyspark.sql.functions.next_day**

  ...of the date column based on second week day argument. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters dateColumn or strtarget column to compute on. dayOfWeekstr day of the week, case-insensitive, ac...

- **pyspark.sql.functions.nth_value**

  ...This is equivalent to the nth_value function in SQL. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression offsetintnumber of row to use as the value ignoreN...

- **pyspark.sql.functions.ntile**

  ...t 4. This is equivalent to the NTILE function in SQL. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters nintan integer Returns Columnportioned group id. Examples >>> from pyspark.sql import Win...

- **pyspark.sql.functions.octet_length**

  ...ates the byte length for the specified string column. New in version 3.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strSource column or strings Returns ColumnByte length of the col Examples >>...

- **pyspark.sql.functions.overlay**

  ...yte position pos of src and proceeding for len bytes. New in version 3.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters srcColumn or strcolumn name or column containing the string that will be replaced replaceColumn...

- **pyspark.sql.functions.pandas_udf**

  ...behaves as a regular PySpark function API in general. New in version 2.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters ffunction, optionaluser-defined function. A python function if used as a standalone function ret...

- **pyspark.sql.functions.percent_rank**

  ...(i.e. percentile) of rows within a window partition. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Returns Columnthe column for calculating relative rank. Examples >>> from pyspark.sql import Window, typ...

- **pyspark.sql.functions.percentile_approx**

  ...values is less than the value or equal to that value. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strinput column. percentageColumn, float, list of floats or tuple of floatspercenta...

- **pyspark.sql.functions.pmod**

  ...isor, or the specified divisor value Returns Columnpositive value of dividend mod divisor. Notes Supports Spark Connect. Examples >>> from pyspark.sql.functions import pmod >>> df = spark.createDataFrame([... (1.0, float('nan')...

- **pyspark.sql.functions.posexplode**

  ...e for elements in the map unless specified otherwise. New in version 2.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. Returns Columnone row per array item or map key val...

- **pyspark.sql.functions.posexplode_outer**

  ...e for elements in the map unless specified otherwise. New in version 2.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. Returns Columnone row per array item or map key val...

- **pyspark.sql.functions.pow**

  ...argument raised to the power of the second argument. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col1str, Column or floatthe base number. col2str, Column or floatthe exponent number. Returns...

- **pyspark.sql.functions.power**

...argument raised to the power of the second argument. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters col1str, Column or floatthe base number. col2str, Column or floatthe exponent number. Returns...

- **pyspark.sql.functions.product**

  ...nction: returns the product of the values in a group. New in version 3.2.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colstr, Columncolumn containing values to be multiplied together Returns Columnthe column for...

- **pyspark.sql.functions.quarter**

  ...act the quarter of a given date/timestamp as integer. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget date/timestamp column to work on. Returns Columnquarter of the date/ti...

- **pyspark.sql.functions.radians**

  ...n approximately equivalent angle measured in radians. New in version 2.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strangle in degrees Returns Columnangle in radians, as if computed by java.lang....

- **pyspark.sql.functions.raise_error**

  ...Throws an exception with the provided error message. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters errMsgColumn or strA Python string literal or column containing the error message Returns Col...

- **pyspark.sql.functions.rand**

  ...(i.i.d.) samples uniformly distributed in [0.0, 1.0). New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters seedint (default: None)seed value for random generator. Returns Columnrandom values. Note...

- **pyspark.sql.functions.randn**

  ....i.d.) samples from the standard normal distribution. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters seedint (default: None)seed value for random generator. Returns Columnrandom values. Note...

- **pyspark.sql.functions.rank**

  ...ifth. This is equivalent to the RANK function in SQL. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Returns Columnthe column for calculating ranks. Examples >>> from pyspark.sql import Window, types >>> d...

- **pyspark.sql.functions.regexp_extract**

  ...ied group did not match, an empty string is returned. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters strColumn or strtarget column to work on. patternstrregex pattern to apply. idxintmatched group...

- **pyspark.sql.functions.regexp_replace**

  ...fied string value that match regexp with replacement. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters stringColumn or strcolumn name or column containing the string value patternColumn or strcolumn...

- **pyspark.sql.functions.repeat**

  ...olumn n times, and returns it as a new string column. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. nintnumber of times to repeat value. Returns Colum...

- **pyspark.sql.functions.reverse**

  ...ed string or an array with reverse order of elements. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression Returns Columnarray of elements in reverse order...

- **pyspark.sql.functions.rint**

  ...the argument and is equal to a mathematical integer. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnthe column for computed results....

- **pyspark.sql.functions.round**

  ...ode if scale >= 0 or at integral part when scale < 0. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strinput column to round. scaleint optional default 0scale value. Returns Colum...

- **pyspark.sql.functions.row_number**

  ...ntial number starting at 1 within a window partition. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Returns Columnthe column for calculating row numbers. Examples >>> from pyspark.sql import Window >>> df...

- **pyspark.sql.functions.rpad**

  ...e] Right-pad the string column to width len with pad. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. lenintlength of the final string. padstrchars to appe...

- **pyspark.sql.functions.rtrim**

  ...spaces from right end for the specified string value. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. Returns Columnright trimmed values. Examples >>...

- **pyspark.sql.functions.schema_of_csv**

  ...ses a CSV string and infers its schema in DDL format. New in version 3.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters csvColumn or stra CSV string or a foldable string column containing a CSV string. optionsdict, o...

- **pyspark.sql.functions.schema_of_json**

  ...es a JSON string and infers its schema in DDL format. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters jsonColumn or stra JSON string or a foldable string column containing a JSON string. optionsdict...

- **pyspark.sql.functions.sec**

  ...n.Column[source] Computes secant of the input column. New in version 3.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strAngle in radians Returns ColumnSecant of the angle. Examples >>> df = spa...

- **pyspark.sql.functions.second**

  ...urce] Extract the seconds of a given date as integer. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget date/timestamp column to work on. Returns Columnseconds part of the ti...

- **pyspark.sql.functions.sentences**

  ...optional, and if omitted, the default locale is used. New in version 3.2.0. Changed in version 3.4.0: Supports Spark Connect. Parameters stringColumn or stra string to be split languageColumn or str, optionala language of the locale...

- **pyspark.sql.functions.sequence**

  ...if start is less than or equal to stop, otherwise -1. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters startColumn or strstarting value (inclusive) stopColumn or strlast values (inclusive) stepColum...

- **pyspark.sql.functions.session_window**

  ...and ‘end’ will be of pyspark.sql.types.TimestampType. New in version 3.2.0. Changed in version 3.4.0: Supports Spark Connect. Parameters timeColumnColumn or strThe column name or column to use as the timestamp for windowing by time. T...

- **pyspark.sql.functions.sha1**

  ...olumn[source] Returns the hex string result of SHA-1. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnthe column for computed results....

- **pyspark.sql.functions.sha2**

  ...24, 256, 384, 512, or 0 (which is equivalent to 256). New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. numBitsintthe desired bit length of the result, whi...

- **pyspark.sql.functions.shiftleft**

  ...mn.Column[source] Shift the given value numBits left. New in version 3.2.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strinput column of values to shift. numBitsintnumber of bits to shift. Returns...

- **pyspark.sql.functions.shiftright**

...source] (Signed) shift the given value numBits right. New in version 3.2.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strinput column of values to shift. numBitsintnumber of bits to shift. Returns...

- **pyspark.sql.functions.shiftrightunsigned**

  ...source] Unsigned shift the given value numBits right. New in version 3.2.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strinput column of values to shift. numBitsintnumber of bits to shift. Returns...

- **pyspark.sql.functions.shuffle**

  ...n: Generates a random permutation of the given array. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression Returns Columnan array of elements in random ord...

- **pyspark.sql.functions.sign**

  ...olumn[source] Computes the signum of the given value. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnthe column for computed results....

- **pyspark.sql.functions.signum**

  ...olumn[source] Computes the signum of the given value. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnthe column for computed results....

- **pyspark.sql.functions.sin**

  ...umn.Column[source] Computes sine of the input column. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnsine of the angle, as if computed...

- **pyspark.sql.functions.sinh**

  ...source] Computes hyperbolic sine of the input column. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strhyperbolic angle. Returns Columnhyperbolic sine of the given value, as if com...

- **pyspark.sql.functions.size**

  ...the length of the array or map stored in the column. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression Returns Columnlength of the array/map. Examp...

- **pyspark.sql.functions.skewness**

  ...ction: returns the skewness of the values in a group. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnskewness of given column. Exam...

- **pyspark.sql.functions.slice**

  ...end if start is negative) with the specified length. New in version 2.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters xColumn or strcolumn name or column containing the array to be sliced startColumn or str or intc...

- **pyspark.sql.functions.sort_array**

  ...at the end of the returned array in descending order. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression ascbool, optionalwhether to sort in ascending or de...

- **pyspark.sql.functions.soundex**

  ...umn[source] Returns the SoundEx encoding for a string New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. Returns ColumnSoundEx encoded string. Examples...

- **pyspark.sql.functions.spark_partition_id**

  ....sql.column.Column[source] A column for partition ID. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Returns Columnpartition id the record belongs to. Notes This is non deterministic because it depends on...

- **pyspark.sql.functions.split**

  ...urce] Splits str around matches of the given pattern. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters strColumn or stra string expression to split patternstra string representing a regular expressio...

- **pyspark.sql.functions.sqrt**

  ...omputes the square root of the specified float value. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columncolumn for computed results. E...

- **pyspark.sql.functions.stddev**

  ...mn[source] Aggregate function: alias for stddev_samp. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnstandard deviation of given column...

- **pyspark.sql.functions.stddev_pop**

  ...tion standard deviation of the expression in a group. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnstandard deviation of given column...

- **pyspark.sql.functions.stddev_samp**

  ...mple standard deviation of the expression in a group. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnstandard deviation of given column...

- **pyspark.sql.functions.struct**

  ...ql.column.Column[source] Creates a new struct column. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colslist, set, str or Columncolumn names or Columns to contain in the output struct. Returns...

- **pyspark.sql.functions.substring**

  ...in byte and is of length len when str is Binary type. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters strColumn or strtarget column to work on. posintstarting position in str. lenintlength of chars...

- **pyspark.sql.functions.substring_index**

  ...orms a case-sensitive match when searching for delim. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters strColumn or strtarget column to work on. delimstrdelimiter of values. countintnumber of occurr...

- **pyspark.sql.functions.sum**

  ...ion: returns the sum of all values in the expression. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnthe column for computed results....

- **pyspark.sql.functions.sum_distinct**

  ...returns the sum of distinct values in the expression. New in version 3.2.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnthe column for computed results....

- **pyspark.sql.functions.sumDistinct**

  ...returns the sum of distinct values in the expression. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Deprecated since version 3.2.0: Use sum_distinct() instead....

- **pyspark.sql.functions.tan**

  ....Column[source] Computes tangent of the input column. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strangle in radians Returns Columntangent of the given value, as if computed by...

- **pyspark.sql.functions.tanh**

  ...rce] Computes hyperbolic tangent of the input column. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strhyperbolic angle Returns Columnhyperbolic tangent of the given value as if co...

- **pyspark.sql.functions.timestamp_seconds**

  ...the Unix epoch (1970-01-01T00:00:00Z) to a timestamp. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strunix time values. Returns Columnconverted timestamp value. Examples >>> f...

- **pyspark.sql.functions.to_csv**

...ows an exception, in the case of an unsupported type. New in version 3.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column containing a struct. options: dict, optionaloptions to control co...

- **pyspark.sql.functions.to_date**

  ...he format is omitted. Equivalent to col.cast("date"). New in version 2.2.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strinput column of values to convert. format: str, optionalformat to use to convert...

- **pyspark.sql.functions.to_json**

  ...ows an exception, in the case of an unsupported type. New in version 2.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column containing a struct, an array or a map. optionsdict, optionalopti...

- **pyspark.sql.functions.to_timestamp**

  ...rmat is omitted. Equivalent to col.cast("timestamp"). New in version 2.2.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strcolumn values to convert. format: str, optionalformat to use to convert timestam...

- **pyspark.sql.functions.to_utc_timestamp**

  ...mp to string according to the session local timezone. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters timestampColumn or strthe column that contains timestamps tzColumn or strA string detailing the...

- **pyspark.sql.functions.toDegrees**

  ...ol: ColumnOrName) → pyspark.sql.column.Column[source] New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Deprecated since version 2.1.0: Use degrees() instead....

- **pyspark.sql.functions.toRadians**

  ...ol: ColumnOrName) → pyspark.sql.column.Column[source] New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Deprecated since version 2.1.0: Use radians() instead....

- **pyspark.sql.functions.transform**

  ...a transformation to each element in the input array. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression ffunctiona function that is applied to each element...

- **pyspark.sql.functions.transform_keys**

  ...of those applications as the new keys for the pairs. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression ffunctiona binary function (k: Column, v: Column) -...

- **pyspark.sql.functions.transform_values**

  ...f those applications as the new values for the pairs. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strname of column or expression ffunctiona binary function (k: Column, v: Column) -...

- **pyspark.sql.functions.translate**

  ...tring is matching with the character in the matching. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters srcColColumn or strSource column or strings matchingstrmatching characters. replacestrcharacter...

- **pyspark.sql.functions.trim**

  ...paces from both ends for the specified string column. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. Returns Columntrimmed values from both sides. E...

- **pyspark.sql.functions.trunc**

  ...s date truncated to the unit specified by the format. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters dateColumn or strinput column of values to truncate. formatstr‘year’, ‘yyyy’, ‘yy’ to truncate b...

- **pyspark.sql.functions.udf**

  ...Like]][source] Creates a user defined function (UDF). New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters ffunctionpython function if used as a standalone function returnTypepyspark.sql.types.DataType o...

- **pyspark.sql.functions.unbase64**

...oded string column and returns it as a binary column. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. Returns Columnencoded string value. Examples >>...

- **pyspark.sql.functions.unhex**

  ...er and converts to the byte representation of number. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. Returns Columnstring representation of given hexade...

- **pyspark.sql.functions.unix_timestamp**

  ...timestamp is None, then it returns current timestamp. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters timestampColumn or str, optionaltimestamps of string values. formatstr, optionalalternative form...

- **pyspark.sql.functions.unwrap_udt**

  ...olumn.Column[source] Unwrap UDT data type column into its underlying type. New in version 3.4.0. Notes Supports Spark Connect. previous...

- **pyspark.sql.functions.upper**

  ...n[source] Converts a string expression to upper case. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to work on. Returns Columnupper case values. Examples >>> d...

- **pyspark.sql.functions.var_pop**

  ...rns the population variance of the values in a group. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnvariance of given column. Exam...

- **pyspark.sql.functions.var_samp**

  ...he unbiased sample variance of the values in a group. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnvariance of given column. Exam...

- **pyspark.sql.functions.variance**

  ...Column[source] Aggregate function: alias for var_samp New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget column to compute on. Returns Columnvariance of given column. Exam...

- **pyspark.sql.functions.weekofyear**

  ...st week with more than 3 days, as defined by ISO 8601 New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget timestamp column to work on. Returns Columnweek of the year for given...

- **pyspark.sql.functions.when**

  ...t invoked, None is returned for unmatched conditions. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Parameters conditionColumna boolean Column expression. value :a literal value, or a Column expression. R...

- **pyspark.sql.functions.window**

  ...and ‘end’ will be of pyspark.sql.types.TimestampType. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters timeColumnColumnThe column or the expression to use as the timestamp for windowing by time. The t...

- **pyspark.sql.functions.window_time**

  ...indow column of a window aggregate records. Returns Columnthe column for computed results. Notes Supports Spark Connect. Examples >>> import datetime >>> df = spark.createDataFrame(... [(datetime.datetime(2016, 3, 11, 9, 0, 7),...

- **pyspark.sql.functions.xxhash64**

  ...umn. The hash computation uses an initial seed of 42. New in version 3.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colsColumn or strone or more columns to compute on. Returns Columnhash value as long column....

- **pyspark.sql.functions.year**

  ...xtract the year of a given date/timestamp as integer. New in version 1.5.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget date/timestamp column to work on. Returns Columnyear part of the date/...

- pyspark.sql.functions.years

  ...or timestamps and dates to partition data into years. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colColumn or strtarget date or timestamp column to work on. Returns Columndata partitioned by...

- pyspark.sql.functions.zip_with

  ...th of the longer array, before applying the function. New in version 3.1.0. Changed in version 3.4.0: Supports Spark Connect. Parameters leftColumn or strname of the first column or expression rightColumn or strname of the second col...

- pyspark.sql.GroupedData

  ...tions on a DataFrame, created by DataFrame.groupBy(). New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Methods agg(*exprs) Compute aggregates and returns the result as a DataFrame. apply(udf) It is an alias...

- pyspark.sql.GroupedData.agg

  ...s can also be a list of aggregate Column expressions. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters exprsdicta dict mapping from column name (string) to aggregate functions (string), or a list of C...

- pyspark.sql.GroupedData.apply

  ...edData.applyInPandas() takes a Python native function. New in version 2.3.0. Changed in version 3.4.0: Support Spark Connect. Parameters udfpyspark.sql.functions.pandas_udf()a grouped map user-defined function returned by pyspark.sql....

- pyspark.sql.GroupedData.applyInPandas

  ...gth of the returned pandas.DataFrame can be arbitrary. New in version 3.0.0. Changed in version 3.4.0: Support Spark Connect. Parameters funcfunctiona Python native function that takes a pandas.DataFrame and outputs a pandas.DataFrame...

- pyspark.sql.GroupedData.applyInPandasWithState

  ...e in both the input and output can also be arbitrary. New in version 3.4.0. Changed in version 3.5.0: Supports Spark Connect. Parameters funcfunctiona Python native function to be called on every group. It should take parameters (key,...

- pyspark.sql.GroupedData.avg

  ...columns for each group. mean() is an alias for avg(). New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colsstrcolumn names. Non-numeric columns are ignored. Examples >>> df = spark.createDataFrame...

- pyspark.sql.GroupedData.cogroup

  ...another group so that we can run cogrouped operations. New in version 3.0.0. Changed in version 3.4.0: Support Spark Connect. See PandasCogroupedOps for the operations that can be run....

- pyspark.sql.GroupedData.count

  ...[source] Counts the number of records for each group. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Examples >>> df = spark.createDataFrame(... [(2, "Alice"), (3, "Alice"), (5, "Bob"), (10, "Bob")], ["age...

- pyspark.sql.GroupedData.max

  ...he max value for each numeric columns for each group. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Examples >>> df = spark.createDataFrame([... (2, "Alice", 80), (3, "Alice", 100), ... (5, "Bob", 120)...

- pyspark.sql.GroupedData.mean

  ...columns for each group. mean() is an alias for avg(). New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colsstrcolumn names. Non-numeric columns are ignored....

- pyspark.sql.GroupedData.min

  ...the min value for each numeric column for each group. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colsstrcolumn names. Non-numeric columns are ignored. Examples >>> df = spark.createDataFrame...

- pyspark.sql.GroupedData.pivot

...first compute the list of distinct values internally. New in version 1.6.0. Changed in version 3.4.0: Supports Spark Connect. Parameters pivot_colstrName of the column to pivot. valueslist, optionalList of values that will be transla...

- **pyspark.sql.GroupedData.sum**

  ...utes the sum for each numeric columns for each group. New in version 1.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters colsstrcolumn names. Non-numeric columns are ignored. Examples >>> df = spark.createDataFrame...

- **pyspark.sql.PandasCogroupedOps**

  ...of two GroupedData, created by GroupedData.cogroup(). New in version 3.0.0. Changed in version 3.4.0: Support Spark Connect. Notes This API is experimental. Methods applyInPandas(func, schema) Applies a function to each cogroup u...

- **pyspark.sql.PandasCogroupedOps.applyInPandas**

  ...gth of the returned pandas.DataFrame can be arbitrary. New in version 3.0.0. Changed in version 3.4.0: Support Spark Connect. Parameters funcfunctiona Python native function that takes two pandas.DataFrames, and outputs a pandas.DataF...

- **pyspark.sql.protobuf.functions.from_protobuf**

  ...3.4.0. Changed in version 3.5.0: Supports binaryDescriptorSet arg to pass binary descriptor directly. Supports Spark Connect. Parameters dataColumn or strthe binary column. messageName: str, optionalthe protobuf message name to look...

- **pyspark.sql.protobuf.functions.to_protobuf**

  ...3.4.0. Changed in version 3.5.0: Supports binaryDescriptorSet arg to pass binary descriptor directly. Supports Spark Connect. Parameters dataColumn or strthe data column. messageName: str, optionalthe protobuf message name to look fo...

- **pyspark.sql.SparkSession**

  ...d parquet files. To create a SparkSession, use the following builder pattern: Changed in version 3.4.0: Supports Spark Connect. builder[source] Examples Create a Spark session. >>> spark = (... SparkSession.builder ... .m...

- **pyspark.sql.SparkSession.addArtifact**

  ...Spark job on every node. The path passed can only be a local file for now. Notes This is an API dedicated to Spark Connect client only. With regular Spark Session, it throws an exception....

- **pyspark.sql.SparkSession.addArtifacts**

  ...Spark job on every node. The path passed can only be a local file for now. Notes This is an API dedicated to Spark Connect client only. With regular Spark Session, it throws an exception....

- **pyspark.sql.SparkSession.builder.appName**

  ...name is set, a randomly generated name will be used. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters namestran application name Returns SparkSession.Builder Examples >>> SparkSession.builder....

- **pyspark.sql.SparkSession.builder.config**

  ...both SparkConf and SparkSession's own configuration. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters keystr, optionala key name string for configuration property valuestr, optionala value for confi...

- **pyspark.sql.SparkSession.builder.getOrCreate**

  ...s a new one based on the options set in this builder. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Returns SparkSession Examples This method first checks whether there is a valid global default SparkSessi...

- **pyspark.sql.SparkSession.catalog**

  ...r query underlying databases, tables, functions, etc. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Returns Catalog Examples >>> spark.catalog <...Catalog object ...> Create a temp view, show the list, a...

- **pyspark.sql.SparkSession.conf**

  ...the value set in the underlying SparkContext, if any. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Returns pyspark.sql.conf.RuntimeConfig Examples >>> spark.conf <pyspark...RuntimeConf...> Set a runtime...

- **pyspark.sql.SparkSession.copyFromLocalToFs**

...o. The path must be an an absolute path. Notes This API is a developer API. Also, this is an API dedicated to Spark Connect client only. With regular Spark Session, it throws an exception....

- **pyspark.sql.SparkSession.createDataFrame**

  ...n RDD, a list, a pandas.DataFrame or a numpy.ndarray. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters dataRDD or iterablean RDD of any kind of SQL data representation (Row, tuple, int, boolean, etc.)...

- **pyspark.sql.SparkSession.getActiveSession**

  ...ssion for the current thread, returned by the builder New in version 3.0.0. Changed in version 3.5.0: Supports Spark Connect. Returns SparkSessionSpark session if an active session exists for the current thread Examples >>> s = Sp...

- **pyspark.sql.SparkSession.range**

  ...e from start to end (exclusive) with step value step. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters startintthe start value endint, optionalthe end value (exclusive) stepint, optionalthe incremen...

- **pyspark.sql.SparkSession.read**

  ...ader that can be used to read data in as a DataFrame. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Returns DataFrameReader Examples >>> spark.read <...DataFrameReader object ...> Write a DataFrame into...

- **pyspark.sql.SparkSession.readStream**

  ...e used to read data streams as a streaming DataFrame. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Returns DataStreamReader Notes This API is evolving. Examples >>> spark.readStream <pyspark...DataStreamR...

- **pyspark.sql.SparkSession.sql**

  ...amed and positional parameters in the same SQL query. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect and parameterized SQL. Changed in version 3.5.0: Added positional parameters. Parameters sqlQuerystrSQL que...

- **pyspark.sql.SparkSession.stop**

  ...op() → None[source] Stop the underlying SparkContext. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Examples >>> spark.stop()...

- **pyspark.sql.SparkSession.streams**

  ...the StreamingQuery instances active on this context. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Returns StreamingQueryManager Notes This API is evolving. Examples >>> spark.streams <pyspark...Streaming...

- **pyspark.sql.SparkSession.table**

  ...e[source] Returns the specified table as a DataFrame. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Parameters tableNamestrthe table name to retrieve. Returns DataFrame Examples >>> spark.range(5).crea...

- **pyspark.sql.SparkSession.udf**

  ...n.udf Returns a UDFRegistration for UDF registration. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Returns UDFRegistration Examples Register a Python UDF, and use it in SQL. >>> strlen = spark.udf.registe...

- **pyspark.sql.SparkSession.udtf**

  ...ns a UDTFRegistration for UDTF registration. New in version 3.5.0. Returns UDTFRegistration Notes Supports Spark Connect. previous...

- **pyspark.sql.SparkSession.version**

  ...ersion of Spark on which this application is running. New in version 2.0.0. Changed in version 3.4.0: Supports Spark Connect. Returns strthe version of Spark in string. Examples >>> _ = spark.version...

- **pyspark.sql.streaming.DataStreamReader**

  ...es, etc). Use SparkSession.readStream to access this. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Notes This API is evolving. Examples >>> spark.readStream <...streaming.readwriter.DataStreamReader object ......

- **pyspark.sql.streaming.DataStreamReader.csv**

...-formatted string (For example col0 INT, col1 DOUBLE). .. versionadded:: 2.0.0 .. versionchanged:: 3.5.0Supports Spark Connect. Other Parameters Extra optionsFor the extra options, refer to Data Source Option in the version you use....

- **pyspark.sql.streaming.DataStreamReader.format**

  ...eader[source] Specifies the input data source format. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Parameters sourcestrname of the data source, e.g. ‘json’, ‘parquet’. Notes This API is evolving. Example...

- **pyspark.sql.streaming.DataStreamReader.json**

  ...through the input once to determine the input schema. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Parameters pathstrstring represents path to the JSON dataset, or RDD of Strings storing JSON objects. schem...

- **pyspark.sql.streaming.DataStreamReader.load**

  ...eam from a data source and returns it as a DataFrame. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Parameters pathstr, optionaloptional string for file-system backed data sources. formatstr, optionaloptiona...

- **pyspark.sql.streaming.DataStreamReader.option**

  ...Adds an input option for the underlying data source. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Notes This API is evolving. Examples >>> spark.readStream.option("x", 1) <...streaming.readwriter.DataStreamRe...

- **pyspark.sql.streaming.DataStreamReader.options**

  ...e] Adds input options for the underlying data source. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Notes This API is evolving. Examples >>> spark.readStream.options(x="1", y=2) <...streaming.readwriter.DataStr...

- **pyspark.sql.streaming.DataStreamReader.orc**

  ...ORC file stream, returning the result as a DataFrame. New in version 2.3.0. Changed in version 3.5.0: Supports Spark Connect. Other Parameters Extra optionsFor the extra options, refer to Data Source Option in the version you use....

- **pyspark.sql.streaming.DataStreamReader.parquet**

  ...uet file stream, returning the result as a DataFrame. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Parameters pathstrthe path in any Hadoop supported file system Other Parameters Extra optionsFor the ext...

- **pyspark.sql.streaming.DataStreamReader.schema**

  ...chema inference step, and thus speed up data loading. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Parameters schemapyspark.sql.types.StructType or stra pyspark.sql.types.StructType object or a DDL-formatted...

- **pyspark.sql.streaming.DataStreamReader.table**

  ...esponding to the table should support streaming mode. New in version 3.1.0. Changed in version 3.5.0: Supports Spark Connect. Parameters tableNamestrstring, for the name of the table. Returns DataFrame Notes This API is evolvin...

- **pyspark.sql.streaming.DataStreamReader.text**

  ...he text file is a new row in the resulting DataFrame. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Parameters pathstr or liststring, or list of strings, for input path(s). Other Parameters Extra optionsF...

- **pyspark.sql.streaming.DataStreamWriter**

  ...ores, etc). Use DataFrame.writeStream to access this. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Notes This API is evolving. Examples The example below uses Rate source that generates rows continuously. Afte...

- **pyspark.sql.streaming.DataStreamWriter.foreach**

  ...following methods. open(partition_id, epoch_id): Optional method that initializes the processing(for example, open a connection, start a transaction, etc). Additionally, you can use the partition_id and epoch_id to deduplicate regenerate...

- **pyspark.sql.streaming.DataStreamWriter.foreachBatch**

  ...uming all operations are deterministic in the query). New in version 2.4.0. Changed in version 3.5.0: Supports Spark Connect. Notes This API is evolving. This function behaves differently in

Spark Connect mode. See examples. In Connect,...

- **pyspark.sql.streaming.DataStreamWriter.format**

  ...[source] Specifies the underlying output data source. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Parameters sourcestrstring, name of the data source, which for now can be ‘parquet’. Notes This API is e...

- **pyspark.sql.streaming.DataStreamWriter.option**

  ...Adds an output option for the underlying data source. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Notes This API is evolving. Examples >>> df = spark.readStream.format("rate").load() >>> df.writeStream.option...

- **pyspark.sql.streaming.DataStreamWriter.options**

  ...] Adds output options for the underlying data source. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Notes This API is evolving. Examples >>> df = spark.readStream.format("rate").load() >>> df.writeStream.option...

- **pyspark.sql.streaming.DataStreamWriter.outputMode**

  ...ing DataFrame/Dataset is written to a streaming sink. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Options include: append: Only the new rows in the streaming DataFrame/Dataset will be written tothe sink...

- **pyspark.sql.streaming.DataStreamWriter.partitionBy**

  ...he file system similar to Hive’s partitioning scheme. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Parameters colsstr or listname of columns Notes This API is evolving. Examples >>> df = spark.readStream...

- **pyspark.sql.streaming.DataStreamWriter.queryName**

  ...rently active queries in the associated SparkSession. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Parameters queryNamestrunique name for the query Notes This API is evolving. Examples >>> import time >>...

- **pyspark.sql.streaming.DataStreamWriter.start**

  ...configured by spark.sql.sources.default will be used. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Parameters pathstr, optionalthe path in a Hadoop supported file system formatstr, optionalthe format used t...

- **pyspark.sql.streaming.DataStreamWriter.toTable**

  ...Query object can be used to interact with the stream. New in version 3.1.0. Changed in version 3.5.0: Supports Spark Connect. Parameters tableNamestrstring, for the name of the table. formatstr, optionalthe format used to save. outp...

- **pyspark.sql.streaming.DataStreamWriter.trigger**

  ...to setting the trigger to processingTime='0 seconds'. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Parameters processingTimestr, optionala processing time interval as a string, e.g. ‘5 seconds’, ‘1 minute’....

- **pyspark.sql.streaming.StreamingQuery**

  ...new data arrives. All these methods are thread-safe. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Notes This API is evolving. Methods awaitTermination([timeout]) Waits for the termination of this query,...

- **pyspark.sql.streaming.StreamingQuery.awaitTermination**

  ...ption, if this query has terminated with an exception New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Parameters timeoutint, optionaldefault None. The waiting time for specified streaming query to terminate....

- **pyspark.sql.streaming.StreamingQuery.exception**

  ...rors.exceptions.base.StreamingQueryException][source] New in version 2.1.0. Changed in version 3.5.0: Supports Spark Connect. Returns StreamingQueryExceptionthe StreamingQueryException if the query was terminated by an exception, or N...

- **pyspark.sql.streaming.StreamingQuery.explain**

  ...physical) plans to the console for debugging purpose. New in version 2.1.0. Changed in version 3.5.0: Supports Spark Connect. Parameters extendedbool, optionaldefault False. If False, prints only the physical plan. Examples >>> sd...

- **pyspark.sql.streaming.StreamingQuery.id**

...e same id active in a Spark cluster. Also see, runId. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Returns strThe unique id of query that persists across restarts from checkpoint data. Examples >>> sdf =...

- **pyspark.sql.streaming.StreamingQuery.isActive**

  ...ther this streaming query is currently active or not. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Returns boolThe result whether specified streaming query is currently active or not. Examples >>> sdf =...

- **pyspark.sql.streaming.StreamingQuery.lastProgress**

  ...aming query or None if there were no progress updates New in version 2.1.0. Changed in version 3.5.0: Supports Spark Connect. Returns dict, optionalThe most recent StreamingQueryProgress update of this streaming query or None if there...

- **pyspark.sql.streaming.StreamingQuery.name**

  ...me, if set, must be unique across all active queries. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Returns strThe user-specified name of the query, or null if not specified. Examples >>> sdf = spark.read...

- **pyspark.sql.streaming.StreamingQuery.processAllAvailable**

  ...ted to the sink. This method is intended for testing. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Notes In the case of continually arriving data, this method may block forever. Additionally, this method is on...

- **pyspark.sql.streaming.StreamingQuery.recentProgress**

  ...uration spark.sql.streaming.numRecentProgressUpdates. New in version 2.1.0. Changed in version 3.5.0: Supports Spark Connect. Returns listList of dict which is the most recent StreamingQueryProgress updates for this query. Example...

- **pyspark.sql.streaming.StreamingQuery.runId**

  ...started from checkpoint) will have a different runId. New in version 2.1.0. Changed in version 3.5.0: Supports Spark Connect. Returns strThe unique id of query that does not persist across restarts. Examples >>> sdf = spark.readSt...

- **pyspark.sql.streaming.StreamingQuery.status**

  ...Query.status Returns the current status of the query. New in version 2.1.0. Changed in version 3.5.0: Supports Spark Connect. Returns dictThe current status of the specified query. Examples >>> sdf = spark.readStream.format("rate"...

- **pyspark.sql.streaming.StreamingQuery.stop**

  ...uery.stop() → None[source] Stop this streaming query. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Examples >>> sdf = spark.readStream.format("rate").load() >>> sq = sdf.writeStream.format('memory').queryName(...

- **pyspark.sql.streaming.StreamingQueryManager**

  ...anage all the StreamingQuery StreamingQueries active. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Notes This API is evolving. Methods addListener(listener) Register a StreamingQueryListener to receive u...

- **pyspark.sql.streaming.StreamingQueryManager.active**

  ...t of active queries associated with this SparkSession New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Returns listThe active queries associated with this SparkSession. Examples >>> sdf = spark.readStream.fo...

- **pyspark.sql.streaming.StreamingQueryManager.addListener**

  ...ive up-calls for life cycle events of StreamingQuery. New in version 3.4.0. Changed in version 3.5.0: Supports Spark Connect. Parameters listenerStreamingQueryListenerA StreamingQueryListener to receive up-calls for life cycle events...

- **pyspark.sql.streaming.StreamingQueryManager.awaitAnyTermination**

  ...ption, if this query has terminated with an exception New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Parameters timeoutint, optionaldefault None. The waiting time for any streaming query to terminate. Retur...

- **pyspark.sql.streaming.StreamingQueryManager.get**

  ...urce] Returns an active query from this SparkSession. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Parameters idstrThe unique id of specified query. Returns

StreamingQueryAn active query with id from thi...

- **pyspark.sql.streaming.StreamingQueryManager.resetTerminated**

  ...ion() can be used again to wait for new terminations. New in version 2.0.0. Changed in version 3.5.0: Supports Spark Connect. Examples >>> spark.streams.resetTerminated()...

- **pyspark.sql.UDFRegistration.register**

  ...nction) or a user-defined function as a SQL function. New in version 1.3.1. Changed in version 3.4.0: Supports Spark Connect. Parameters namestr,name of the user-defined function in SQL statements. ffunction, pyspark.sql.functions.ud...

- **pyspark.sql.UDFRegistration.registerJavaFunction**

  ...pe is not specified we would infer it via reflection. New in version 2.3.0. Changed in version 3.4.0: Supports Spark Connect. Parameters namestrname of the user-defined function javaClassNamestrfully qualified name of java class ret...

- **pyspark.sql.UDFRegistration.registerJavaUDAF**

  ...va user-defined aggregate function as a SQL function. New in version 2.3.0. Changed in version 3.4.0: Supports Spark Connect. namestrname of the user-defined aggregate function javaClassNamestrfully qualified name of java class Exam...

- **pyspark.sql.Window**

  ...Utility functions for defining window in DataFrames. New in version 1.4.0. Changed in version 3.4.0: Supports Spark Connect. Notes When ordering is not defined, an unbounded window frame (rowFrame, unboundedPreceding, unboundedFollowin...

- **pyspark.testing.assertDataFrameEqual**

  ...d expected (DataFrames or lists of Rows), with optional parameters checkRowOrder, rtol, and atol. Supports Spark, Spark Connect, pandas, and pandas-on-Spark DataFrames. For more information about pandas-on-Spark DataFrame equality, see the...

- **TorchDistributor**

  ...ining on PyTorch and PyTorch Lightning using PySpark. New in version 3.4.0. Changed in version 3.5.0: Supports Spark Connect. Parameters num_processesint, optionalAn integer that determines how many different concurrent tasks are allo...

Created using Sphinx 3.0.4.