

Date functions:-

- ① CURRENT-DATE:- Returns the current date in the system's time zone.
- ② CURRENT-TIME STAMP:- Returns the current date and time in the system time zone.
- ③ Extract:- Extracts a specific part (Year, month, day, hour, etc) from a date or time stamp.
- ④ Date-ADD:- Adds a specified interval to a date or timestamp.
- ⑤ Date-SUB:- Subtract a specific interval from a date or timestamp.
- ⑥ Date-Format:- Format a date or timestamp into a specific string format (YYYY-MM-DD).
- ⑦ Date-Diff:- Calculate the difference between two dates in terms of days, month, Year.

8) date_TRUNC:-

Truncates a date with time to a specified unit (year, month, day, hour, etc.).

9) date_PART:-

Similar to extract, it extracts a specific part from a date (or) time stamp.

10) now:- Returns the current date & time.

String functions

1) CONCAT:- Concatenates two (or) more strings together

2) Length (LEN):- Returns the length (number of characters) of a string

- 3) UPPER:- Converts a String to uppercase
- 4) LOWER:- Converts a String to lowercase
- 5) Substring (or) SUBSTR:-
Extract a portion of a string
based on the specified start
position and length.
- 6) REPLACE:- Replace all occurrences of a
substring with in a string
with another substring
- 7) TRIM:- Removes leading and
trailing spaces from a
string.
- 8) char_length :- Returns the length
(no of char) of a
string

9) left, :- no. of char from the left side of a string

Right 'r'

10) ASCII :- Code value of the first char in the string.

integer function:-

1) ABS :- Return the absolute value of an integer.

ABS(-5)

2) ROUND :- a numeric value to a specified no. of decimal places.

3) CEILING (or) CEIL :- Round a numeric value up to the nearest integer. (4.2)

4) FLOOR:- Rounds a numeric value down to the nearest integer
(-1)

5) MOD:- Return the remainder of the division of the two integers.

6) POW or POW:-
Raises a numeric value to a specified power

7) SQRT:- Return square root of a numeric value

8) GREATEST:- Return the largest value among the provided arguments.

9) LEAST:- Return the smallest value among the provided arguments.

10) RAND or RAND:-
generates a random float value between 0 and 1.

numpy functions:

① `np.array`:- create a numpy array from a python list or tuple

② `np.zeros`:- create an array filled with zeros of a specified shape.

③ `np.ones`:- create an array filled with ones of a specified shape

`arr = np.ones((2,2))`

④ `np.arange`:- creates array with regularly spaced values within a given interval.

`np.arange(0,10,2).`

⑤ `np.linspace`:- creates an array with specified number of equally spaced values within a given interval.

6) `np.random.randn` = generates random numbers from a standard normal distribution (mean, sd 1).

7) `np.max` or `np.amax`:- Return max value of an array or along a specified axis.

8) `np.min` or `np.amin`:- Return the min value of an array or along a specified axis.

9) `np.mean`:- Cal the arithmetic mean of an array or along a specified axis.

10) `np.reshape`:- reshape an array to specified shape without changing its data.

np.sum, np.size, np.mean, np.unique
np.count, np.log, np.std, np.prod.

Pandas Functions:

- ① pd.read_csv → Read CSV File
- ② df.head → first n rows of a dataframe
- ③ df.tail → Return the last n rows.
- ④ df.shape → return the dimensions
(row, column)
- ⑤ df.info → provide summary
- ⑥ df.describe → generates descriptive
statistics of numeric columns
in df
- ⑦ df.groupby → groups the Dataframe by
one or more columns.
- ⑧ df['column-name'] or df.column-name
Access the specific column

a) `df.drop na()` → remove rows (or) column

i.i) `df.to_csv` writes the dataframe to a CSV file.

RDD (Resilient Distributed Dataset) !
is a fundamental data structure
used for distributed computing.

→ parallel across a cluster.

→ RDD API operations.

① `map` :- ~~transform~~

`add. map (lambda x: x+1)`

Apply function to each element of the RDD
to return a new RDD of the result.

② `filter` Returns a new RDD containing
only the elements that
satisfy a given predicate.

3) reduce:- Aggregates the elements of the RDD using a specified binary operator

4) flat map:- similar to map, each input can be mapped to zero or more output items

5) distinct:- Return new RDD with unique elements from the original RDD

6) sort by:- Return a new RDD sorted by the specified key

7) group by key:- Groups the values of each key in RDD and return new RDD of key-value pairs

8) reduce by key:- similar to reduce. but operates on key-value pairs so apply the reduction function to the value for each key

Join:- perform an inner join between two RDDs based on these keys

↳ Collect:- Returns all the elements of the RDD, as an array

Spark session object:-

↳ Spark session object is the entry point for working with structured data and executing operations in Spark

↳ include creating data frames, execute SQL queries, and manage config.

create:-

Spark session builder :-

builder:- is used to create a Spark session-builder object, which is responsible for configuring

and creating the 'spark session'.

appName:- sets the name of the spark applications.

Config:- used to set spark configuration options, you can specify various Config options

getOrCreate:- is used to create a new 'spark session' or retrieve an existing one if available.

Read data from CSV
df = spark.read.csv(' ', header, inferSchema)

spark.sql(' ',) → create sql

df_filtered = df.filter(df["age"])

transformation of data frame

df_groupby = df.groupby('gender').agg(' ')