1) np.array() : This fn is used to create a Numpy array from a python list or tuple. Converts i/p data into an ndarray object.

2) np.arange() : Returns an array with evenly spaced value within a specified range.

3) np.reshape() : Used to change the shape of the array without changing its data.

4) np.mean() : np.mean() computes arithmetic mean along a specified axis or the whole array. Calculates avg of elements in the array.

5) np.max() : Returns max value from an array along specified axis or whole array.

1) df.head() : returns first n rows of a dataframe

2) df.tail(): returns last n rows of a dataframe

3) df.describe(): generates descriptive stats of numerical columns in a dataframe

4) df.sort_values(): used to sort dataframes by one or more columns.

5) df.groupby(): used to group rows of a dataframe based on one or more columns.

6) df.drop(): Remove rows/columns, axis.

1) getdate(): returns current system date & time.

2) dateadd(): adds or subtracts a specified interval

3) datediff(): Calculates diff between two dates.

4) datepart(): returns specified part of a date, such as date, month & year.

5) convert(): converts a given value to a specified data type such as converting date to a different format.

6) Unix_timestamp(): returns the unix epoch seconds from a given date.

RDD API

↳ Represents immutable distributed collection of objects that can be processed in parallel across clusters.

↳ RDD API provides a set of ops.

i) map - applies a fn to each element of RDD & returns a new RDD of results. (Transform)

ii) filter. — operation used to selectively retain elements based on a condition.

iii) reduceByKey - performs reduce operation based on a key value pair of values

iv) join - combines 2 RDDs based on common key. ← Performs inner join

v) collect - retrieves all data from RDD as in-memory ds.

vi) count -

vii) saveAsTextFile. - saves the result of RDD in a txt file