R DataFrame

A data frame is used for storing data tables. It is a list of vectors of equal length.

```
> mtcars
                  mpg cyl disp  hp drat   wt ...
Mazda RX4        21.0   6  160 110 3.90 2.62 ...
Mazda RX4 Wag    21.0   6  160 110 3.90 2.88 ...
Datsun 710       22.8   4  108  93 3.85 2.32 ...
                 ............
```

The top line of the table, called the **header**, contains the column names. Each horizontal line afterward denotes **a data row**, which begins with the **name** of the row, and then followed by the actual data. Each data member of a row is called a **cell**.

http://www.r-tutor.com/r-introduction/data-frame

pandas DataFrame (Python)

pandas is the de facto standard (single-node) DataFrame implementation in Python.

The two primary data structures of pandas, **Series** (1-dimensional) and **DataFrame** (2-dimensional)

```
>>> import pandas as pd
>>> pd.Series(data=[1, 2, 3], index=['x', 'y', 'z'])
x    1
y    2
z    3
dtype: int64
>>> pd.DataFrame(data={'col1': [1, 2, 3], 'col2': [4, 5, 6]}, index=['x', 'y', 'z'])
   col1  col2
x     1     4
y     2     5
z     3     6
```

For R users, DataFrame provides everything that R's data.frame provides and much more.

https://pandas.pydata.org/docs/getting_started/overview.html

Distributed pandas DataFrame

Koalas (pandas API on Apache Spark)

The Koalas project makes data scientists more productive when interacting with big data, by implementing the pandas DataFrame API on top of Apache Spark.

Dask

A Dask DataFrame is a large parallel DataFrame composed of many smaller Pandas DataFrames, split along the index. These Pandas DataFrames may live on disk for larger-than-memory computing on a single machine, or on many different machines in a cluster. One Dask DataFrame operation triggers many operations on the constituent Pandas DataFrames.