GroupBy Operation

GroupBy Operation

- Categorizing a dataset and applying a function to each group, whether an aggregation or transformation is referred as GroupBy operation and Aggregation.

GroupBy Operation

- Categorizing a dataset and applying a function to each group, whether an aggregation or transformation is referred as GroupBy operation and Aggregation.

How GroupBy operation works?

- It works based on the ‘split-apply-combine ’ formula basically applied in R language.

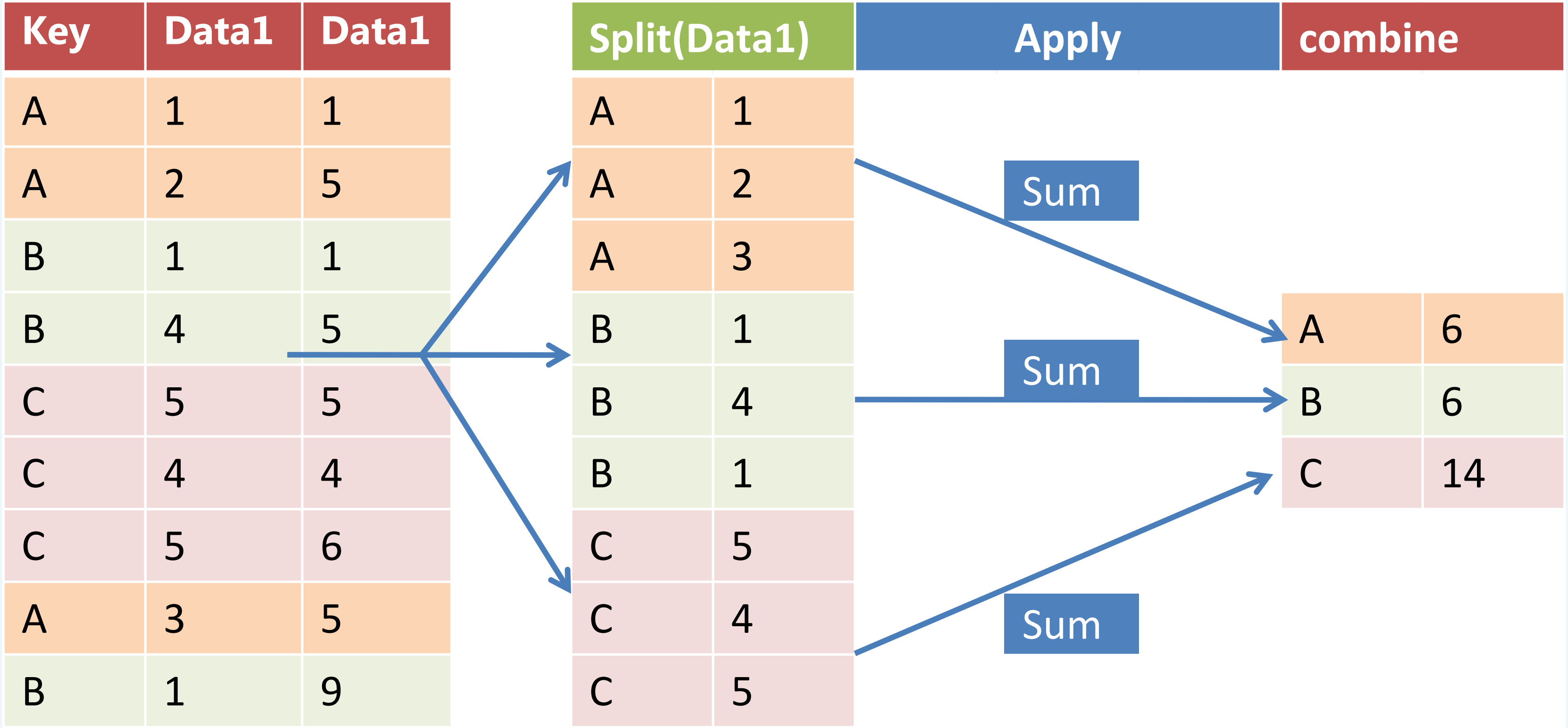
The mechanism uses the Pandas Object: `pandas.groupby`

- By default groupby groups on axis=0(usually along rows); we can use axis=1 (along columns) to perform on any other axis.

GroupBy Operation

How GroupBy operation works?

- It works based on the ‘split-apply-combine ’ formula basically applied in R language.



Key	Data1	Data1
A	1	1
A	2	5
B	1	1
B	4	5
C	5	5
C	4	4
C	5	6
A	3	5
B	1	9

Split(Data1)	
A	1
A	2
A	3
B	1
B	4
B	1
C	5
C	4
C	5

Apply

Sum

Sum

Sum

combine	
A	6
B	6
C	14

GroupBy Operation

How to iterate over Groups?

- The Groupby operation can be used to iterate over groups using the Python’s iterating objects like ‘for’ loop.

GroupBy Operation

How to work with Groupby operation?

- We can select column or subset of columns of a DataFrame by passing a list of column names to the groupby operation. It is referred as indexing a groupby object.
- We can pass dictionary as a key (or grouping parameter) to groupby object.
- We can pass list of elements with the length same as that of the DataFrame len object to groupby object.
- We can pass Python function’s as a key for groupby operation.
- We can even use index level to group in groupby operation.

Data Aggregation

What is Data Aggregation?

- Data Aggregation refers to the data transformation that produces the scalar values from any array or array like object. Ex: sum, count, mean, median etc,.

- Pandas provides ‘`aggregate`’ or ‘`agg`’ method for Data Aggregation.

Data Aggregation

Some commonly used Aggregation functions:

✍Though the listed methods are not the final ones. Instead we can use our own aggregate functions by creating our self.

❑ count	Number of non-NA values in the group
❑ sum	Sum of non-NA values
❑ mean	Mean of non-NA values
❑ median	Arithmetic median of non-NA values
❑ std, var	Unbiased (n – 1 denominator) standard deviation
❑ min, max	Minimum and maximum of non-NA values
❑ prod	Product of non-NA values
❑ first, last	First and last non-NA values

Data Aggregation

How to Aggregate when the Data file is too big?

✍The following steps are followed in that situation.

- ❖ When the Data file is too big to compute aggregation; we can select the particular columns using Groupby object.
- ❖ Then use indexing on grouped object with the specific column again
- ❖ Finally pass a single or a list of aggregation function to the ‘agg’ object.

Apply: Method

How exactly the ‘apply’ function works?

It works systematically like:

- ‘split-apply-combine’

But there exist a beautiful mechanism inside the ‘apply’ function.