Usando funções de agregação no *streaming* estruturado do Spark

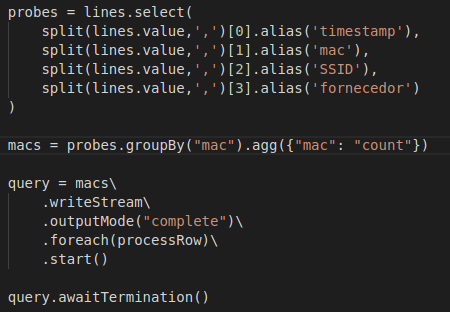
No contexto de *streaming* estruturado do *framework* Apache Spark, agregações são operações que computam vários valores para retornar apenas um. É possível citar diversas funções de agregação, considerando inúmeras aplicações, suportadas por *streaming dataframes* e objetos de dados agrupados (que resultam da utilização do método *groupBy()* em *streaming dataframes*), como:

* *agg()* — método que agrega todo o *dataframe.*
* *avg()* — método que retorna a média dos valores de um grupo.
* *mean()* — método aliás para *avg()*.
* *count()* — método que retorna a quantidade de itens de um grupo.
* *max(x)* — método que retorna o valor máximo da coluna *x* de um grupo.
* *min(x)* — método que retorna o valor mínimo da coluna *x* de um grupo.
* *sum(x)* — método que soma os valores da coluna numérica *x* de um grupo.
* *approx\_count\_distinct(x, rsd = y)* — método que retorna o número aproximado de itens distintos da coluna x, com uma estimativa máxima de erro permitida de y, sem requerer groupBy().

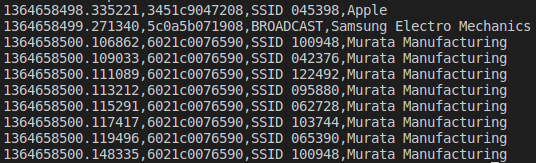
Vale salientar que o *framework* Apache Spark também permite a execução das operações de agregação por *scripts* escritos em SQL após a devida importação dos respectivos métodos necessários.

Para exemplificar as operações citadas, serão utilizados dados processados de pacotes 802.11.

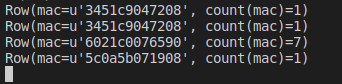
* agg:

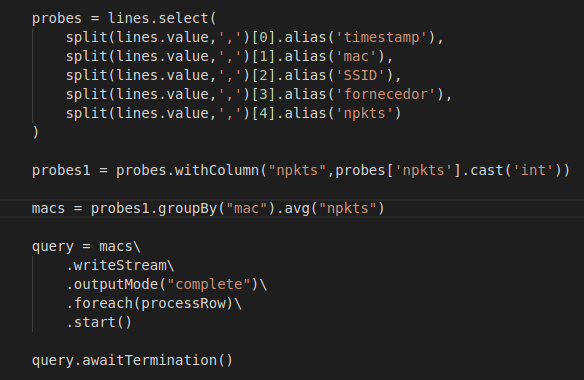


Dados Processados:

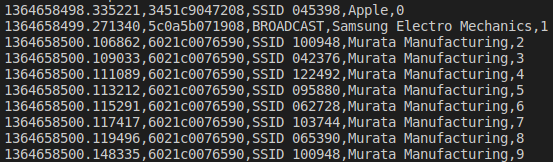


Resposta do spark:

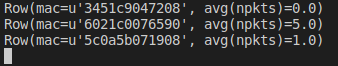


* avg:

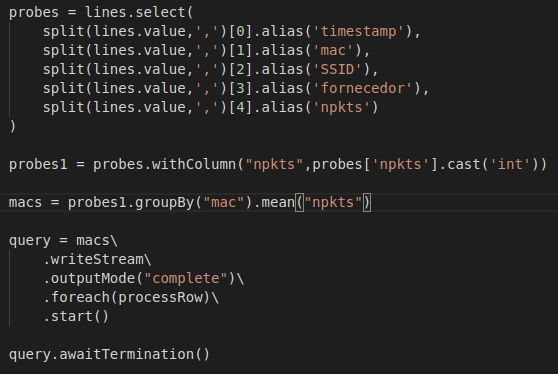
Dados Processados:



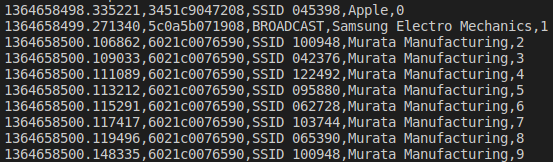
Resposta do spark:



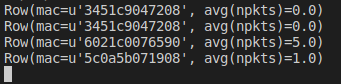
* mean:



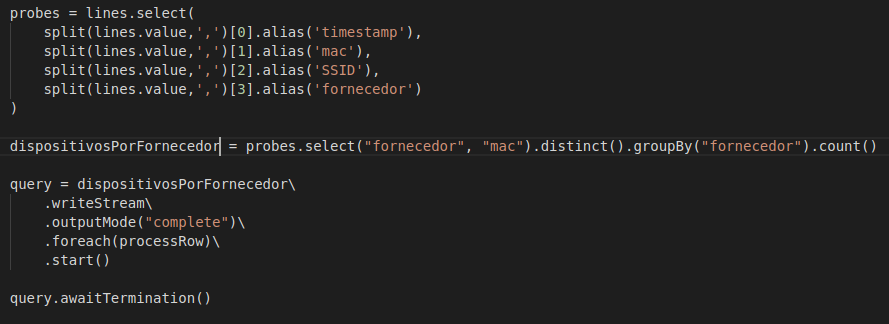
Dados processados:



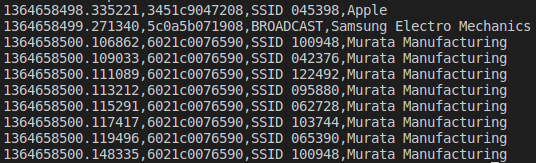
Resposta do spark:



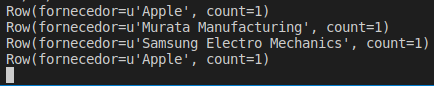
* count:



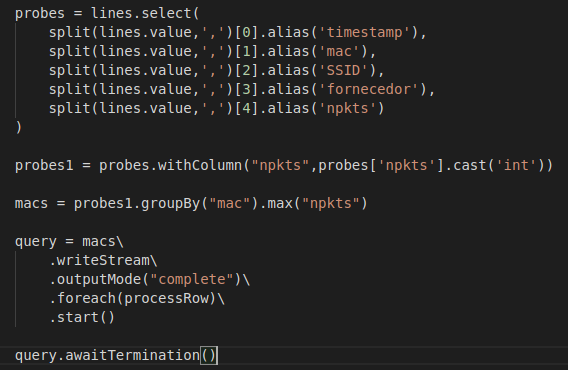
Dados processados:



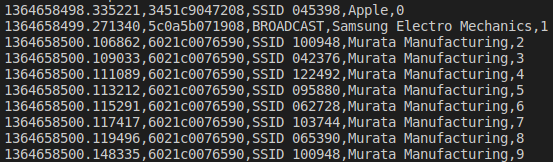
Resposta do spark:



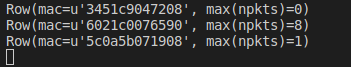
* max:



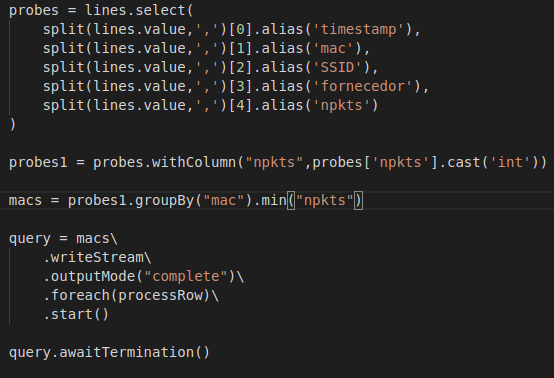
Dados Processados:



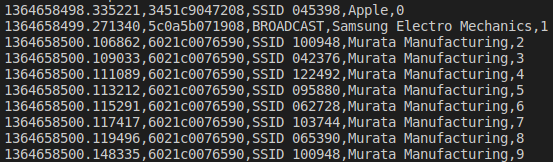
Resposta do spark:



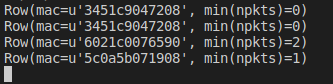
* min:



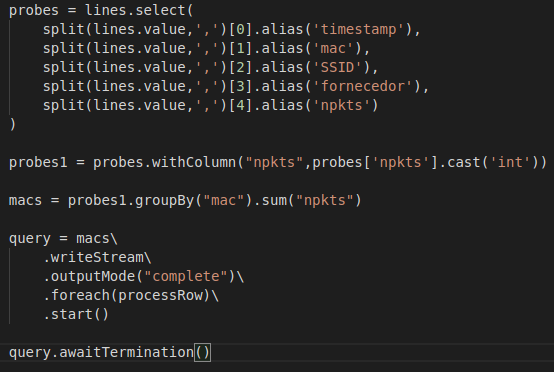
Dados processados:



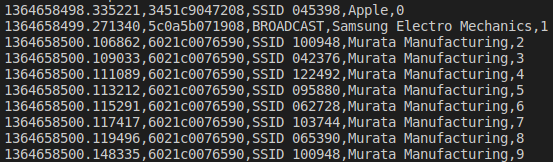
Resposta do spark:



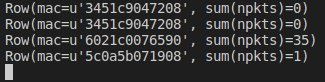
* sum:



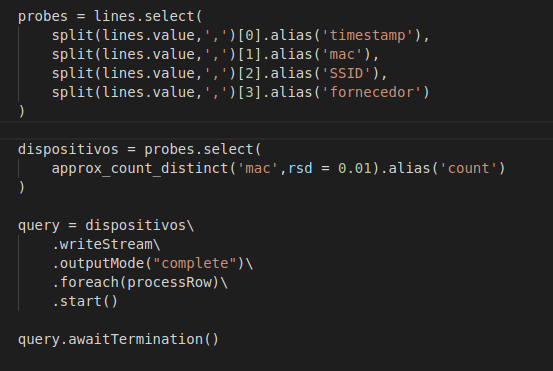
Dados processados:



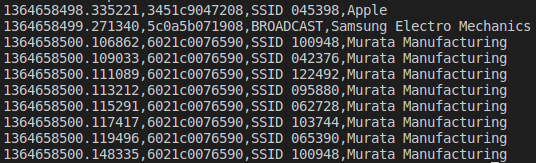
Resposta spark:



* approx\_count\_distinct:



Dados processados:



Resposta do spark:

