



Work with pandas-like DataFrames in Apache Beam

Brian Hulette

Software Engineer @ Google, Apache Beam Committer
github.com/theneuralbit



The Beam DataFrame API...

... is a [pandas](#)-compatible API for building Beam pipelines.

... is a **deferred** API.

... is easier to use and more performant than Beam Python.



```
import pandas as pd
```

```
df = pd.read_csv('data.csv')  
agg = df.foo.groupby(df.bar).sum()  
agg.to_csv('result.csv')
```



```
from apache_beam.dataframe.io import read_csv
```

```
with beam.Pipeline() as p:  
    bdf = p | read_csv('gs://all-the/data/*.csv')  
    agg = bdf.foo.groupby(bdf.bar).sum()  
    agg.to_csv('gs://all-the/results/result.csv')
```

Live Demo!

What Happens when I `run()`
a DataFrame pipeline?

Execution with pandas DataFrames

```
df = pd.read_csv(...)
```



	pickup_hour	passenger_count
0	15	1
1	23	1
2	7	2
...

```
df.passenger_count.groupby(df.pickup_hour).sum()
```



	passenger_count
pickup_hour	
0	28630
1	32778
...	...
22	11460
23	14822

Execution with beam DataFrames

```
bdf = p | beam.dataframe.io.read_csv(...)
```



	pickup_hour	passenger_count
...

```
bdf.passenger_count.groupby(bdf.pickup_hour).sum()
```

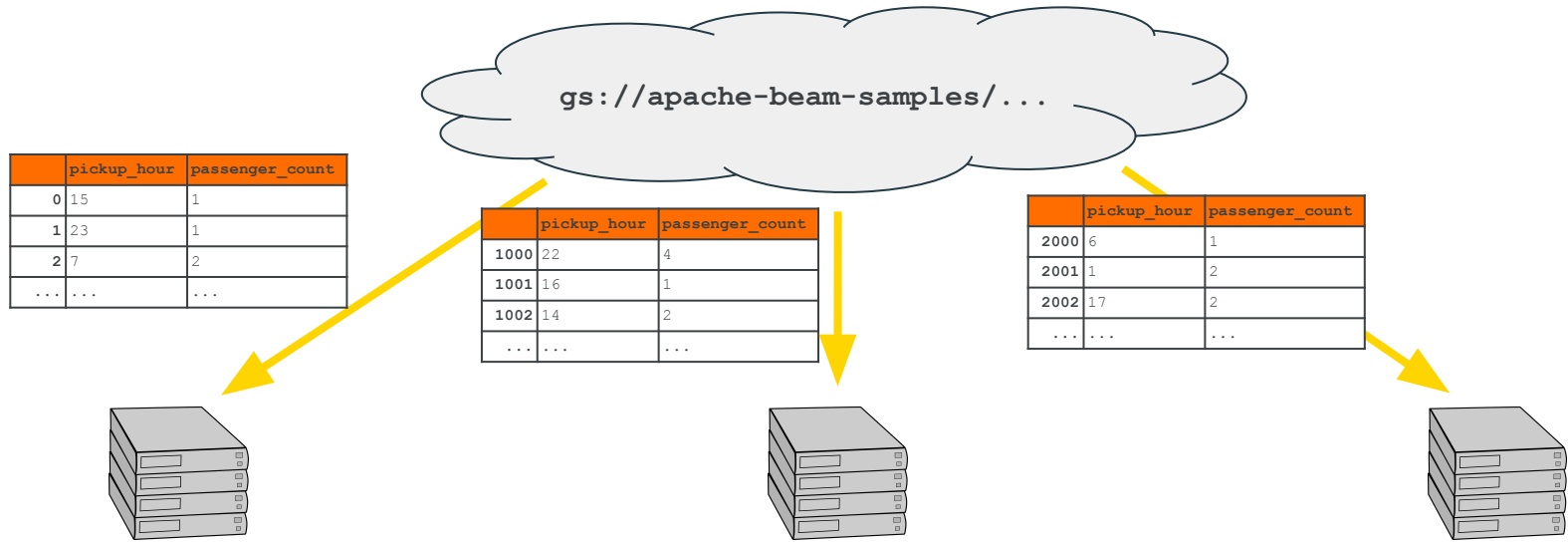


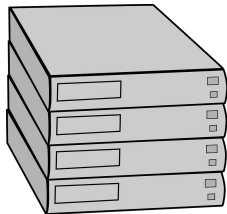
	passenger_count
pickup_hour	
...	...

```
p.run()
```



Then what??





```
df = pd.read_csv(...)
```



	pickup_hour	passenger_count
1000	22	4
1001	16	1
1002	14	2
...

```
df.passenger_count.groupby(df.pickup_hour).sum()
```

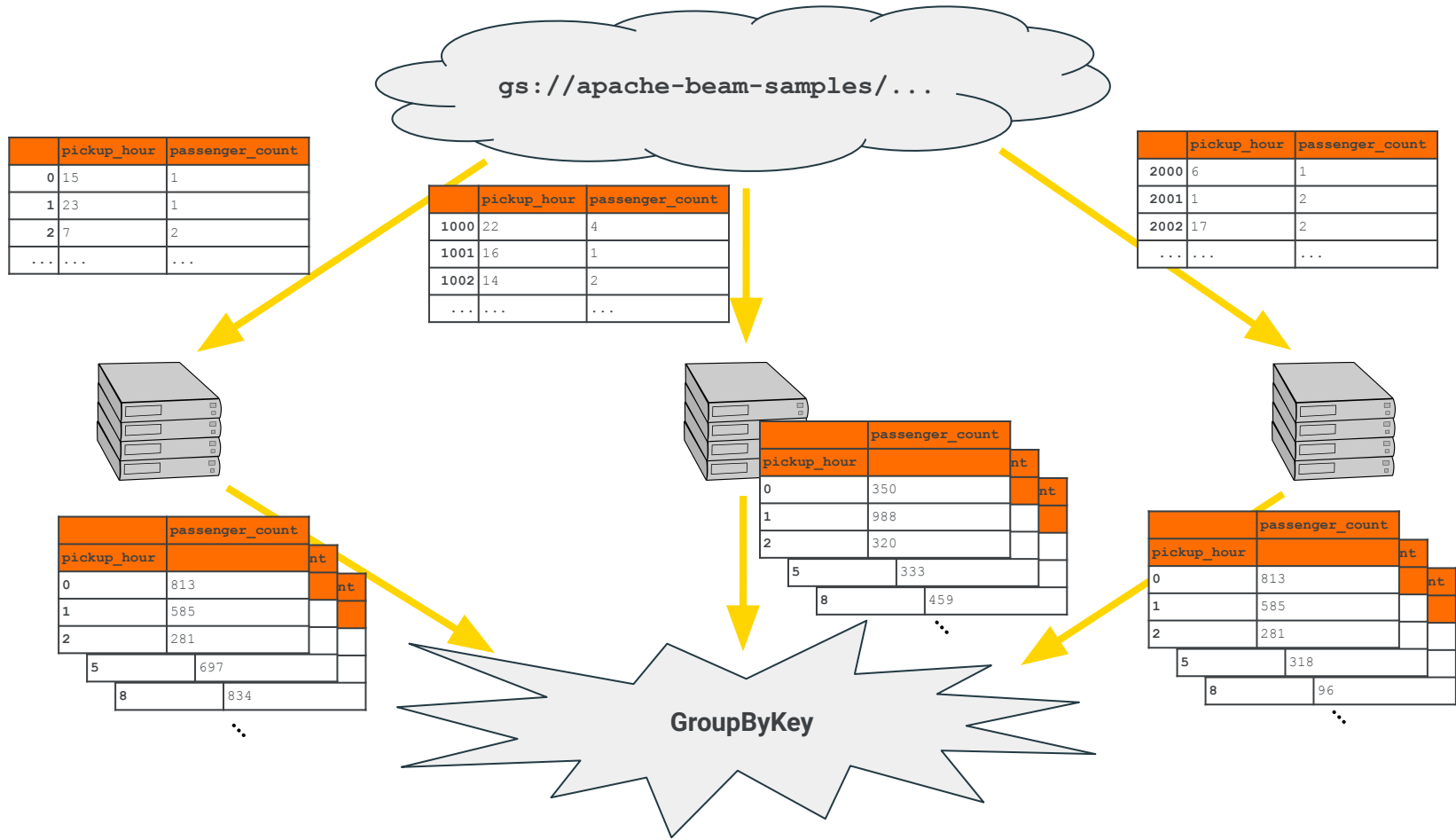


	passenger_count
pickup_hour	
0	813
1	585
...	...
22	818
23	612

Partition the intermediate result



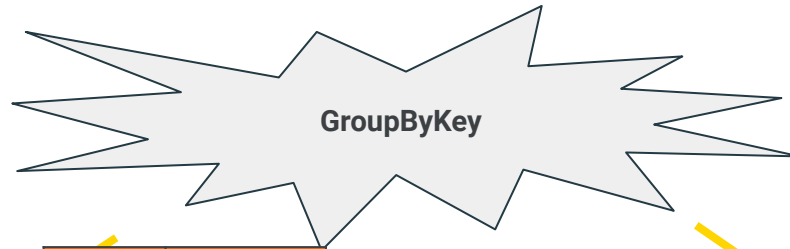
	passenger_count	
pickup_hour		
0	813	unt
1	585	unt
2	281	unt
...
5	697	unt
...
8	834	unt
...



passenger_count			
pickup_hour		nt	
0	813		nt
1	585		
2	281		
2		320	
2		281	
...			



passenger_count	
pickup_hour	
0	9326
1	1808
2	5681

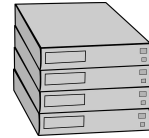


passenger_count			
pickup_hour		nt	
3	591		nt
4	573		
5	697		
5		333	
5		318	
...			

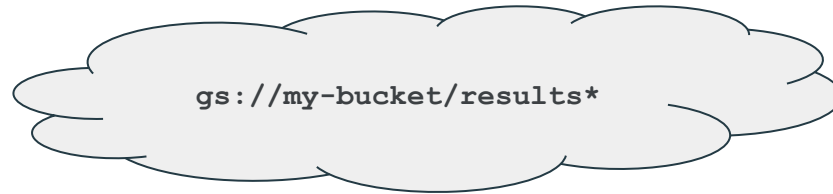


passenger_count	
pickup_hour	
3	6109
4	9233
5	6064

passenger_count			
pickup_hour		nt	
6	813		nt
7	585		
8	834		
8		459	
8		96	
...			



passenger_count	
pickup_hour	
6	9437
7	6007
8	9041



Key Takeaways

- Unlike pandas DataFrames, computation on Beam DataFrames is **distributed**.
- Beam DataFrames use **efficient pandas code** on the workers, just inserting **GroupByKey** where necessary.

Want to learn more?

- [Beam DataFrame API Documentation](#)
- [Design doc](#)
- Other talks with more detail:
 - [Simpler Python Pipelines - Beam Summit 2020](#)
 - [Scaling up pandas with the Beam DataFrame API - Beam Summit 2022](#)
(coming soon)

Thank you!

