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')
```

https://pandas.pydata.org/about/citing.html
https://beam.apache.org/community/logos/

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(..)
```

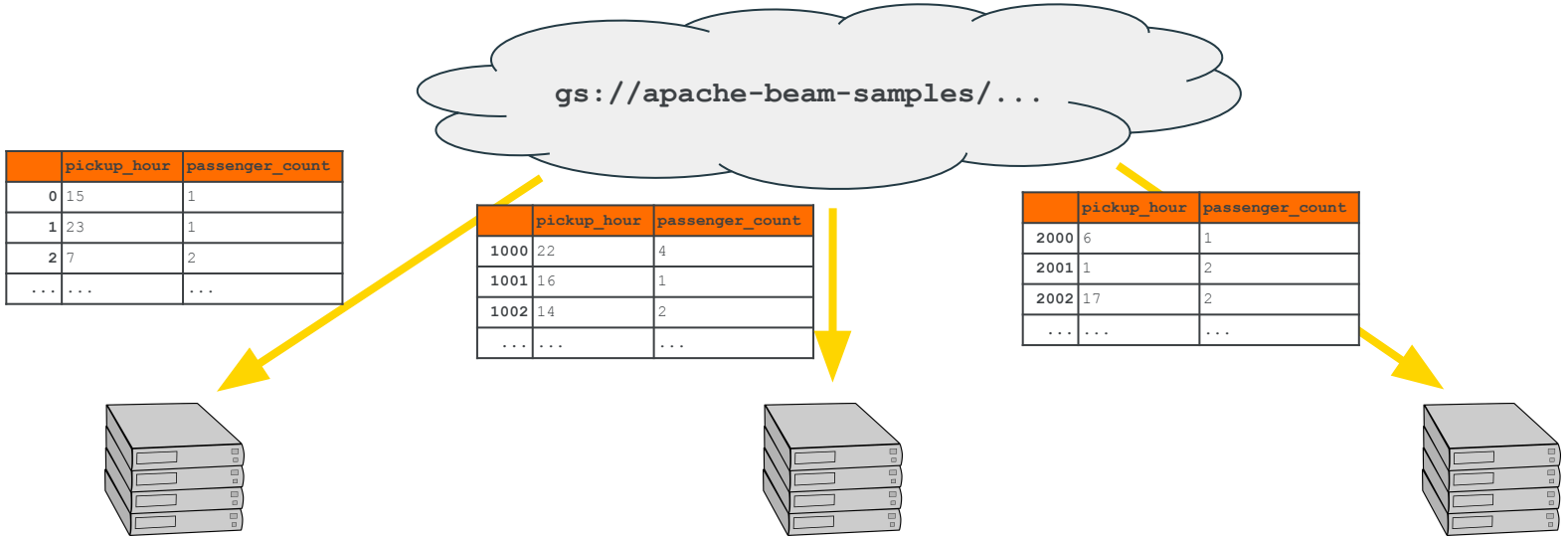

	pickup_hour	passenger_count
...	...	...

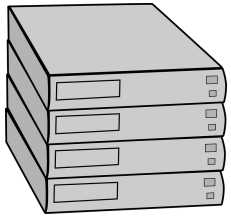
```
bdf.passenger_count.groupby(bdf.pickup_hour).sum()
```


	passenger_count
pickup_hour	
...	...

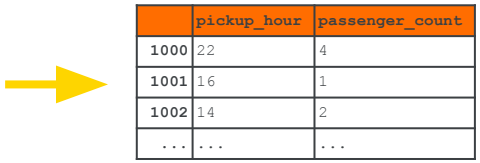
```
p.run()
```


Then what??


gs://apache-beam-samples/...
pickup_hour
passenger_count
0 15 1
1 23 1
2 7 2
... ...
1000 22 4
1001 16 1
1002 14 2
... ...
2000 6 1
2001 1 2
2002 17 2
... ...

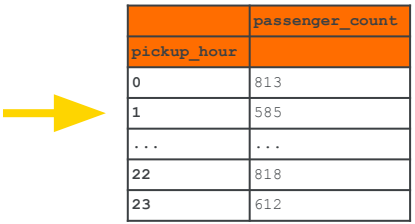


```
df = pd.read_csv(..)
```



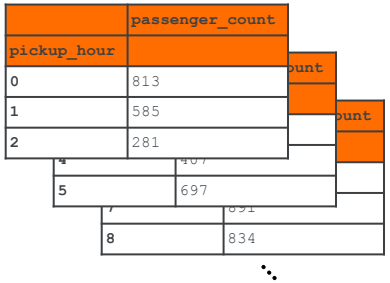
	pickup_hour	passenger_count
1000	22	4
1001	16	1
1002	14	2
...	...	...

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

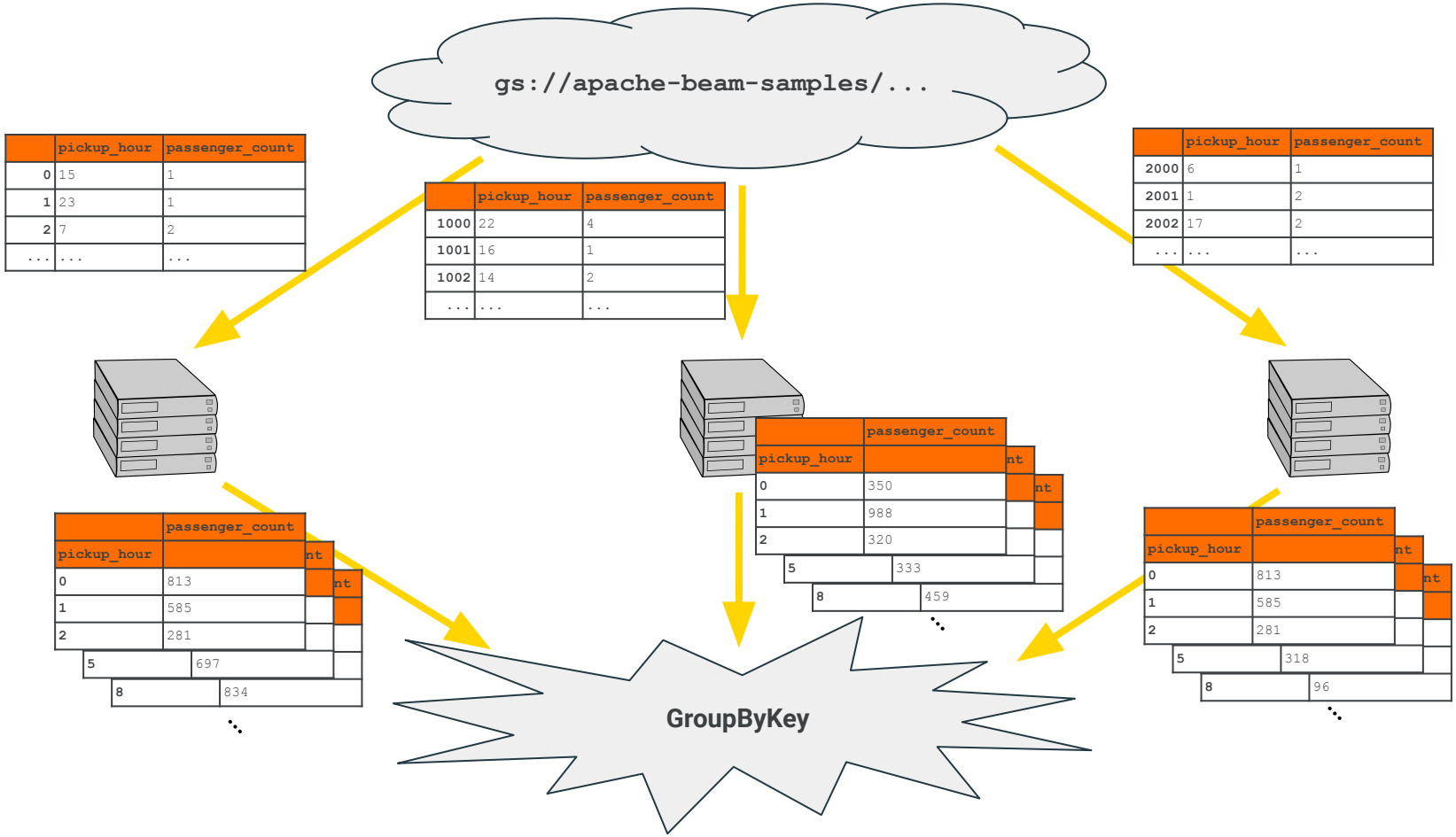


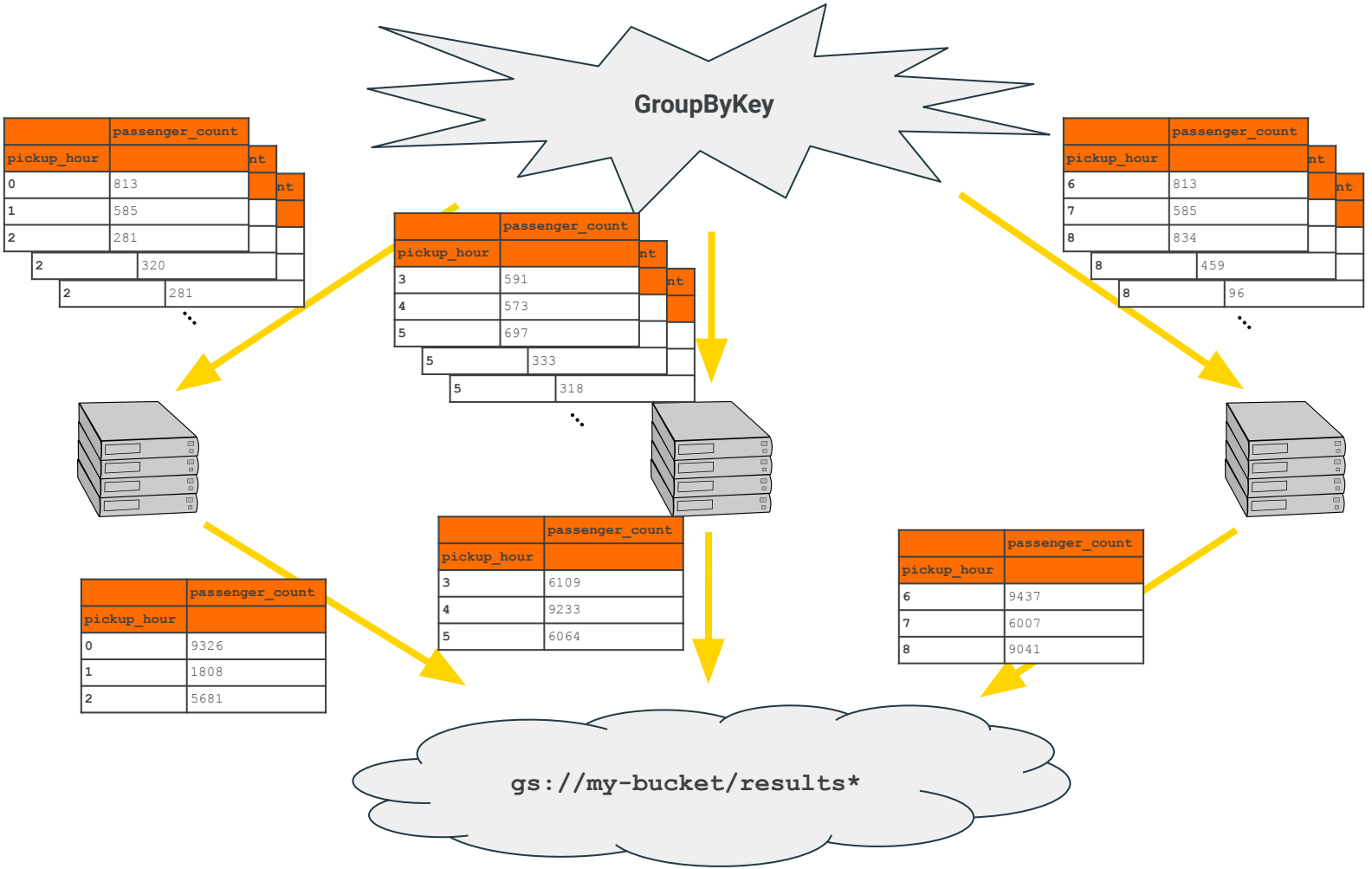
pickup_hour	passenger_count
0	813
1	585
...	...
22	818
23	612

Partition the intermediate result



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


gs://apache-beam-samples/...
GroupByKey


GroupByKey
passenger_count
pickup_hour
0 813
1 585
2 281
2 320
2 281
3 591
4 573
5 697
5 333
5 318
6 813
7 585
8 834
8 459
8 96
passenger_count
pickup_hour
0 9326
1 1808
2 5681
passenger_count
pickup_hour
3 6109
4 9233
5 6064
passenger_count
pickup_hour
6 9437
7 6007
8 9041
gs://my-bucket/results*

Key Takeaways

- Unlike pandas DataFrames, computation on Beam DataFrames is **distributed**.
- Beam DataFrames use **efficient pandas code** on the workers, just inserting **GroupByKeys** 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!