Lesson Reflection

Summary

In this lesson, we explored alternatives to Pandas DataFrames for working with tabular data across parallel computing systems. The core takeaways included:

Key Points

- Pandas builds on top of Numpy ndarrays, which offer speed through vectorization
- PySpark enables distributed, parallel processing on clusters, ideal for big data
- Dask allows larger-than-memory computing on a single machine
- Polars is an emerging new standard for DataFrames

Reflection Questions

- What are some limitations you've encountered when working with Pandas? When might an alternative be better suited?
- How could PySpark or Dask help you work with larger datasets than possible in Pandas alone?
- What infrastructure and configuration is needed to leverage PySpark or Dask effectively?
- What are tradeoffs between ease of use with Pandas vs scalability with PySpark/Dask?
- How might you leverage all 3 tools together on a larger project?

Challenges

- Load a 1GB CSV into Pandas. How does performance compare to Dask?
- Set up PySpark locally and compare reading 1M rows to Pandas
- Combine output from Dask ingest to PySpark machine learning pipeline
- Use Pandas to analyze random sample output from PySpark cluster
- Distributed apply: Calculate mean of 100M rows in PySpark, output to Pandas

Code Examples

Here is code to generate a large Pandas DataFrame with random fruit prices, and calculate statistics on prices partitioned by fruit type:

```python
import pandas as pd
import numpy as np

# Create small dummy DataFrame
np.random.seed(1)
fruits = ['apple', 'banana', 'strawberry', 'kiwi']
N = 1000
df = pd.DataFrame({
    'fruit': np.random.choice(fruits, N),
    'price': np.random.uniform(1, 10, N)
})

# Function to process DataFrame
def get_prices_by_fruit(df):
    return df.groupby('fruit')['price'].agg(['count', 'mean', 'min', 'max'])

prices = get_prices_by_fruit(df)
print(prices)
```

```
            count      mean       min       max
fruit
apple         265  5.610179  1.047159  9.986685
banana        241  5.625379  1.071396  9.934308
kiwi          237  5.420487  1.006879  9.941144
strawberry    257  5.358464  1.028016  9.922119
```