CS:199
Applied Cloud Computing

Prof. Robert Brunner

Tyler Kim
Quinn Jarrell

How is research going?

- Rough Draft due tonight at midnight on moodle

Peer Editing

- Each group should fork the repo and push their paper to it. Then open pull request on Github to the main repo
- You will be assigned one groups to peer edit
  - You must edit the assigned group
- You can also suggest edits on other groups which you are not assigned
- Make comments on their pull requests
- You can edit your paper as you go to revise it. Make sure to reply to the comments you've fixed

Streaming Data

- How can you do live analytics?

Streaming Data

- How can you do live analytics?
  - Do mini batches on whatever came in on the last minute/hour/day etc
  - Rerun analytics every day on the entire data set
  - Process the data bit by bit
- Everything ends up being discretized
- Even if it’s a continuous stream, the stream can be broken into sections by time
- Process each discrete time

- Storm was built by Twitter
- It takes in a new piece of data and routes it through a series of operations
  - It’s a graph
  - Never stops processing
- You write functions called ‘bolts’ which take in a tuple of data and returns a tuple
- The same piece of data may be processed multiple times
- Cannot run some operations like sorting anymore

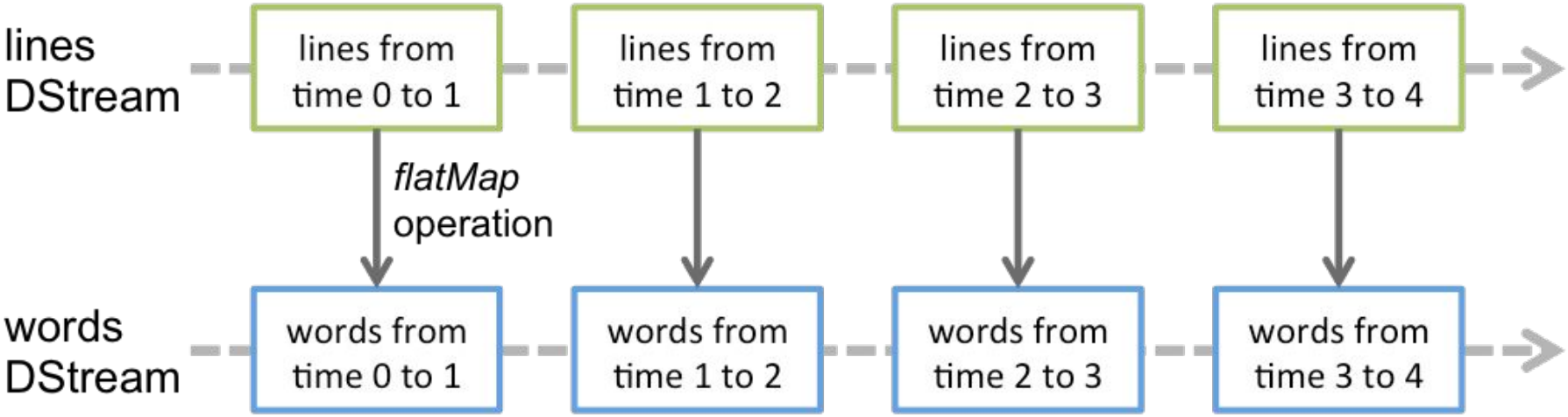
- It’s sort of like Spark
  - Creates a graph of operations
  - Mainly stuff in memory
- So why not use Spark?


Kafka
Flume
HDFS/S3
Kinesis
Twitter
Spark
Streaming
HDFS
Databases
Dashboards

Spark Streaming

- Write almost normal Spark code
- Your Spark code is run on each batch

Spark Streaming

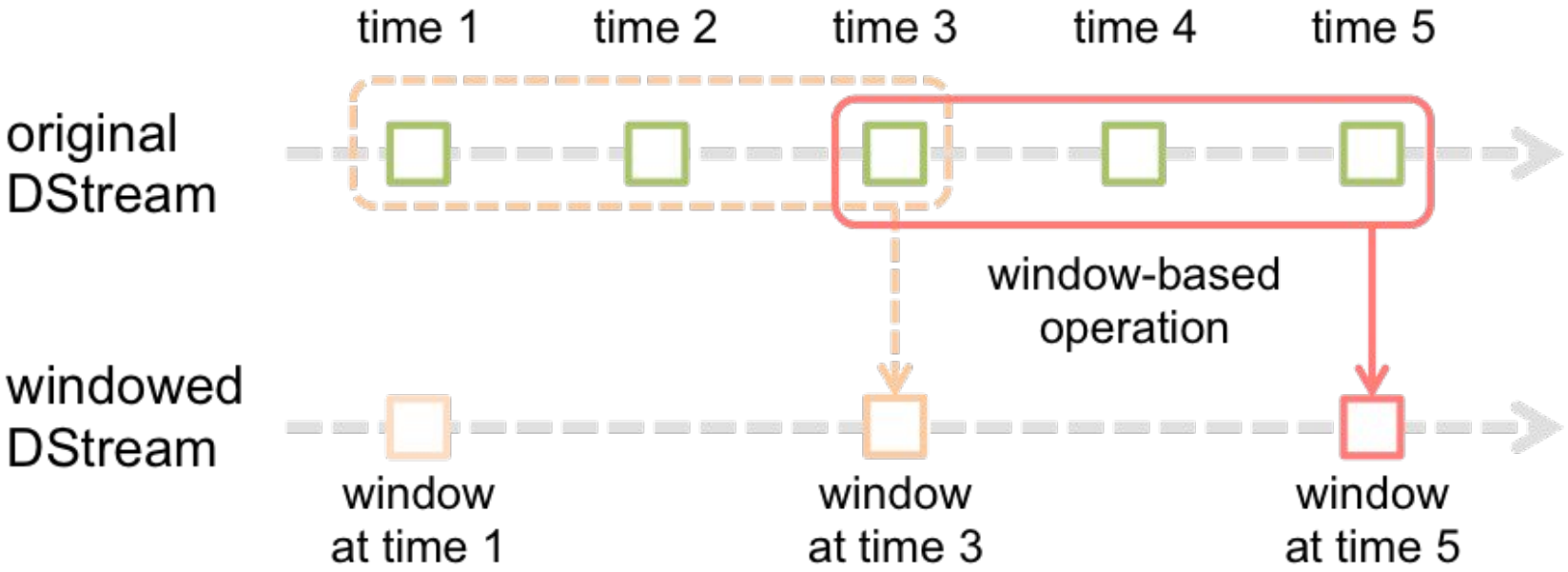


Spark Streaming Differences

- You can use almost all normal Spark operations
- But no sorting or things that do not make sense in a stream
  - Can use the Transform function to sort within a batch

Spark Streaming Windows

- Sometimes we need to operate across mini batches



Spark MLLib and Spark SQL on streaming

- Both have specialized versions for Spark Streaming
- Spark MLLib supports a subset of classifiers like
  - Streaming Linear Regression
  - Streaming KMeans
  - Streaming Logistic Regression
- Also can run other algorithms in large batches
  - Not really live anymore
  - Updates like once a day

Spark MLLib and Spark SQL on streaming

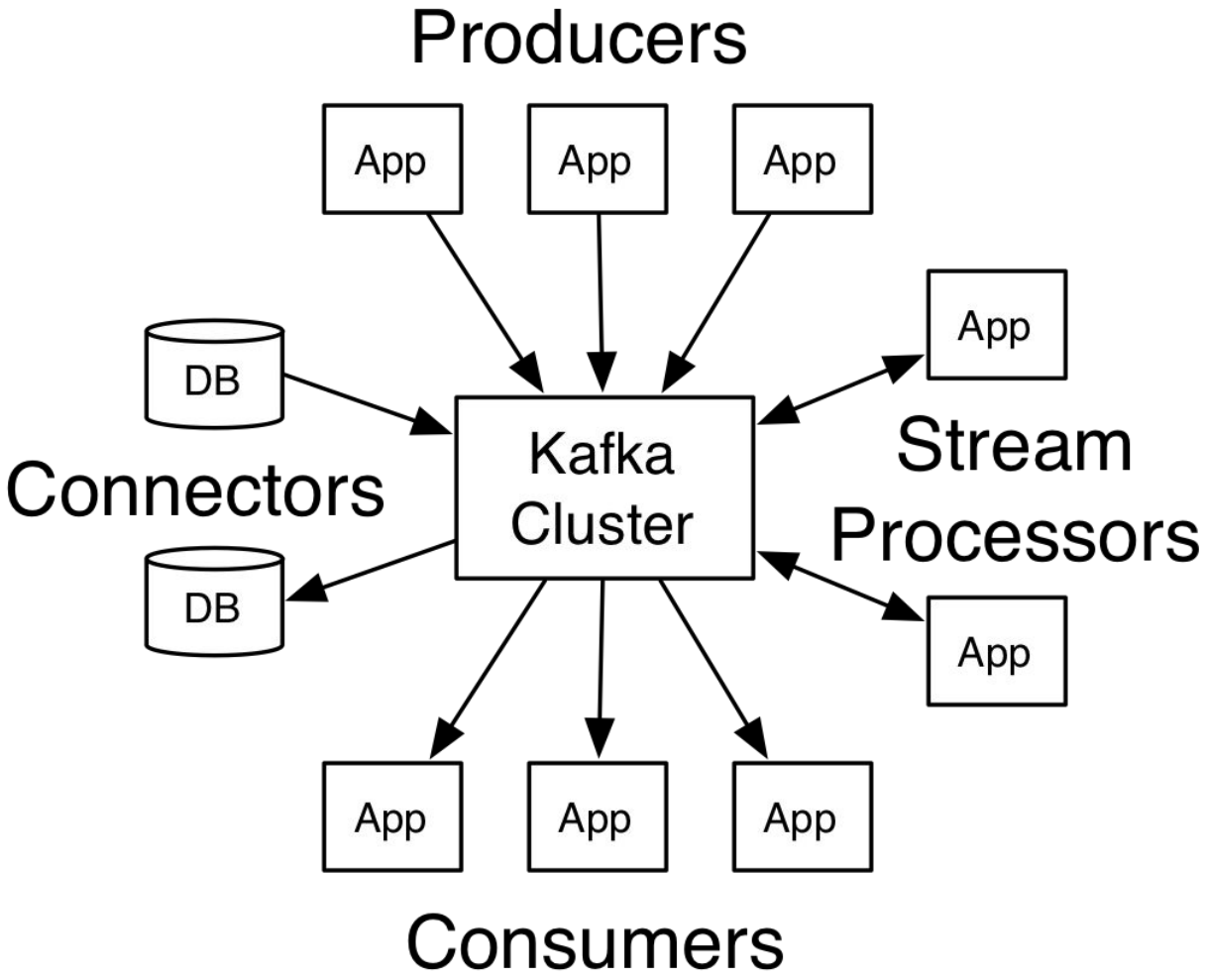
- You can make non streaming algorithms streaming through offline training
- Train on whole dataset through normal Spark MLLib
- Predict on live stream by using the transform function

```
trainingData = data.load()
Model = model.train(trainingData)
Stream = socketStream(localhost, 9999)
Cleaned = stream.map(cleaner)
Predictions = cleaned.transform(lambda rdd: model.predict(rdd))
predictions.pprint() # Output or send it to another stream/file/db
```

Sources of Streams

- File based
  - Runs Spark Streaming every time a new file appears in a directory
- Network based
  - Can run when a line or text is received over a port
  - Can pull from AWS S3 or from a distributed log like Kafka

- Handles streams and everything about streams
- Sort of like a centralized database for streams
- Groups objects by ‘topic’
  - An object has a key, value and a timestamp
- Programs can subscribe to a topic to resume the stream
- Plays a similar role to HDFS for streams


Producers
App
App
App
App
DB
Connectors
DB
Kafka
Cluster
Stream
Processors
App
App
App
App
App
Consumers