Input Sources

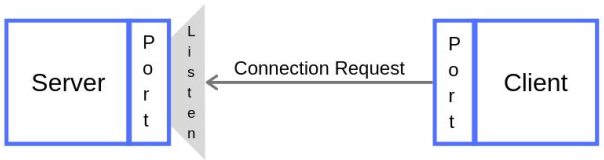
File Source

- Reads files written in a directory as a stream of data
- Files will be processed in the order of file modification time
- latestFirst reverses the order
- Supported file formats include text, CSV, JSON, ORC, Parquet

Kafka

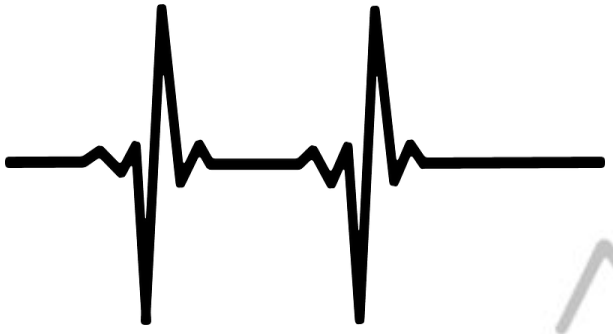
- Only compatible with Kafka broker versions 0.10.0 or higher

Socket Source



- Reads text data from a socket connection
- Listening server socket is at the driver
- Used only for testing as end-to-end fault-tolerance guarantees not provided

Rate Source



- Generates data at the specified number of rows per second
- Each output row contains a timestamp and value
- Used for testing purposes

Input Sources

```
spark = SparkSession. ...

# Read text from socket
socketDF = spark \
    .readStream \
    .format("socket") \
    .option("host", "localhost") \
    .option("port", 9999) \
    .load()

socketDF.isStreaming() # Returns True for DataFrames that have streaming sources

socketDF.printSchema()

# Read all the csv files written atomically in a directory
userSchema = StructType().add("name", "string").add("age", "integer")
csvDF = spark \
    .readStream \
    .option("sep", ";") \
    .schema(userSchema) \
    .csv("/path/to/directory")  # Equivalent to format("csv").load("/path/to/directory")
```

Thank You

Analytics
Vidhya