# Spark Streaming

# Python and Spark

Spark Streaming is an extension of the core Spark API that enables scalable, high-throughput, fault-tolerant stream processing of live data streams.

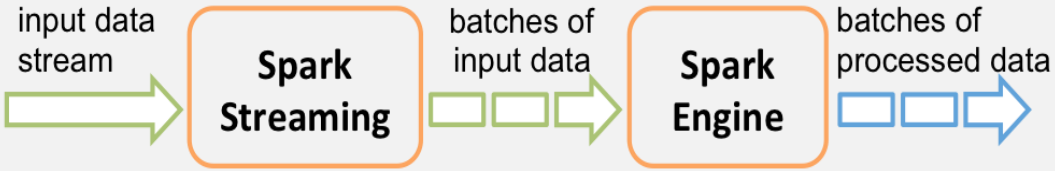
# Python and Spark

Data can be ingested from many sources like Kafka, Flume, Kinesis, or TCP sockets, and can be processed using complex algorithms expressed with high-level functions like map, reduce, join and window.

# Python and Spark

# Python and Spark

Internally, Spark Streaming receives live input data streams and divides the data into batches, which are then processed by the Spark engine to generate the final stream of results in batches.



# Python and Spark

- For this section we will first work through a simple streaming example.
- You will need to simultaneously use jupyter notebook and a terminal for this.
- This is easiest to follow through a local installation using Virtual Box.

# Python and Spark

- The various possible data sources (Kafka, Flume,Kinesis, etc…) can not realistically be shown in a single computer setting.
- If your place of work necessitates use of one of these sources, Spark provides full integration guides.

# Python and Spark

- Keep in mind not every source version is available with the Python API.
- Let's jump to the documentation to show you where you can find additional information on Spark Streaming!

# Spark Streaming Example Code Along