# Building Data Engineering Pipelines in Python

## Signer

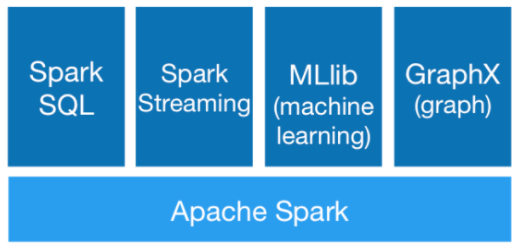
The open-source standard for writing scripts that move data

* Data exchange format: JSON
* Extract and load wit taps an targets
* Communicate over streams:
  + Schema (metadata)
  + State (process meta data)
  + Record (data)

## Creating a data transformation pipeline with PySpark

### PySpark

PySpark is an engine for large-scale data processing. It is built on four libraries. SparkSQL, SparkStreaming, MLlib, GraphX. Those four libraries sit on top of apache Spark.



## Testing your data pipeline

## Managing and orchestrating a workflow