



Summary





Unified Platform

Spark SQL

Spark Streaming PySpark SparkR MLlib spark.ml

GraphX

Spark Core

Standalone Scheduler

YARN

Mesos





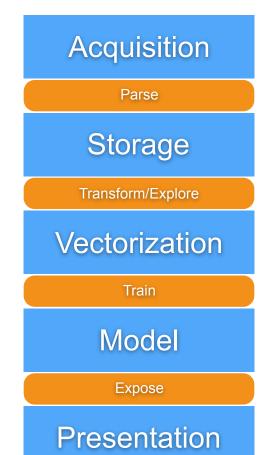
Data Pipeline Locally At Scale Acquisition PySpark requests Parse BeautifulSoup4 Storage **HDFS** pymongo Dataframes/Spark SQL pandas Transform/Explore Vectorization MLlib/spark.ml scikit-learn/NLTK Model model.save() Expose pickle Presentation **Spark Streaming** Flask

©2016 Pearson, Inc.

At Scale



Data Pipeline



Locally

requests

BeautifulSoup4

pymongo

pandas

scikit-learn/NLTK

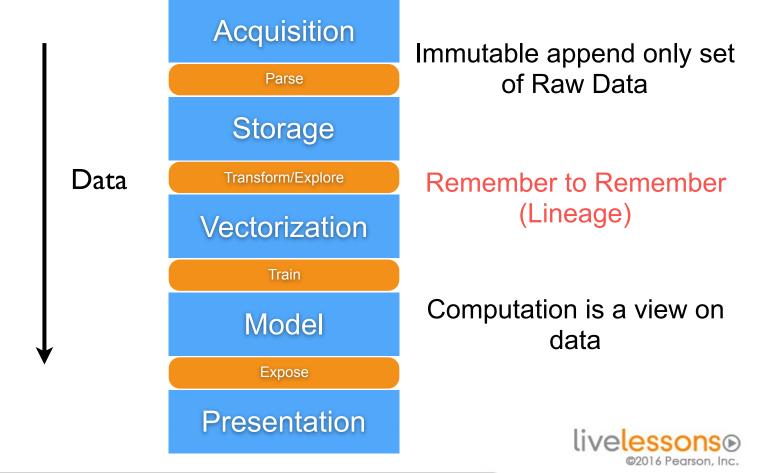
pickle

Flask





Data Pipeline





Remember: How to Scale







Next Up: Keep Learning!



