MTY Data Science & Eng Meetup

PySpark - Israel Zúñiga de la Mora
israel@alturin.com

Sobre que si vamos a hablar: □□□

- Intro a Apache Spark
- Ventajas
- Componentes de Apache Spark
- RDD (Resilent Distributed Datasets)
  - Cargar datos
  - Inspeccion
  - Transformacion
- MLlib
  - Clasificacion
- Aplicaciones simples con Apache Spark

Sobre que no vamos a hablar: □□

- RDD: intermedio - avanzado
  - PairRDD
  - UDF (User Defined Functions)
- MLlib: intermedio - avanzado
- Monitoreo
  - Spark Execution Model
  - Spark Web UI
  - Spark Applications:
    - Debugging
    - Tuning
- Apache Spark Data Pipelines
- Apache Spark Streaming
- Apache Spark GraphX

Sobre que no vamos a hablar (II): □□

- DevOps:
  - UNIX/Linux
  - Docker
  - Deployments
  - Cloud Computing
  - Data Center Basics
- Mesos
- Kubernetes
- NLP
- DataViz

Big Data en 10 segundos

- Hadoop (MapReduce)
  - Mahout
- HDFS (GFS)
- Spark
  - MLlib
  - Graphx
- ZooKeeper
- Flume, Scribe
- HBase (Google BigTable)
- Hive, Pig
  - Dremel, Presto, Spark SQL, Impala, Redshift
- Oozie
- Sqoop
- Lucene

Skills para triunfar con Apache Spark

Skills para triunfar con Apache Spark

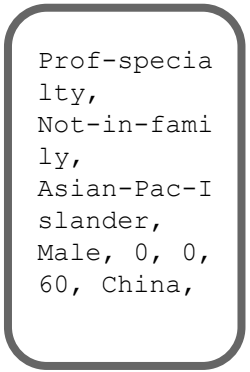
Requeridos:

- Linux. Básico a intermedio:
  - Habilidad de usar un editor de texto (vi)
  - CLI: mv, cp, ssh, grep, cd, useradd, ls
- Principios de Software Dev (SDLC)
- Acceso por SSH y HTTP a la instancia/cluster de Spark

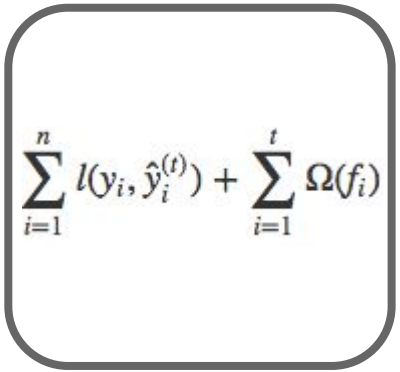
Recomendados:

- Programación funcional
- Scala/Python
- SQL basico
- Big Data (Data Pipeline, roles, glosario)

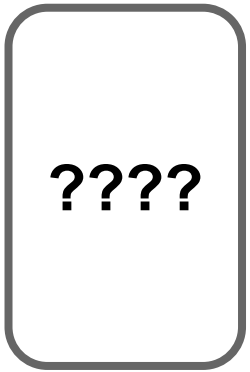
The process of building a machine learning product



Data set

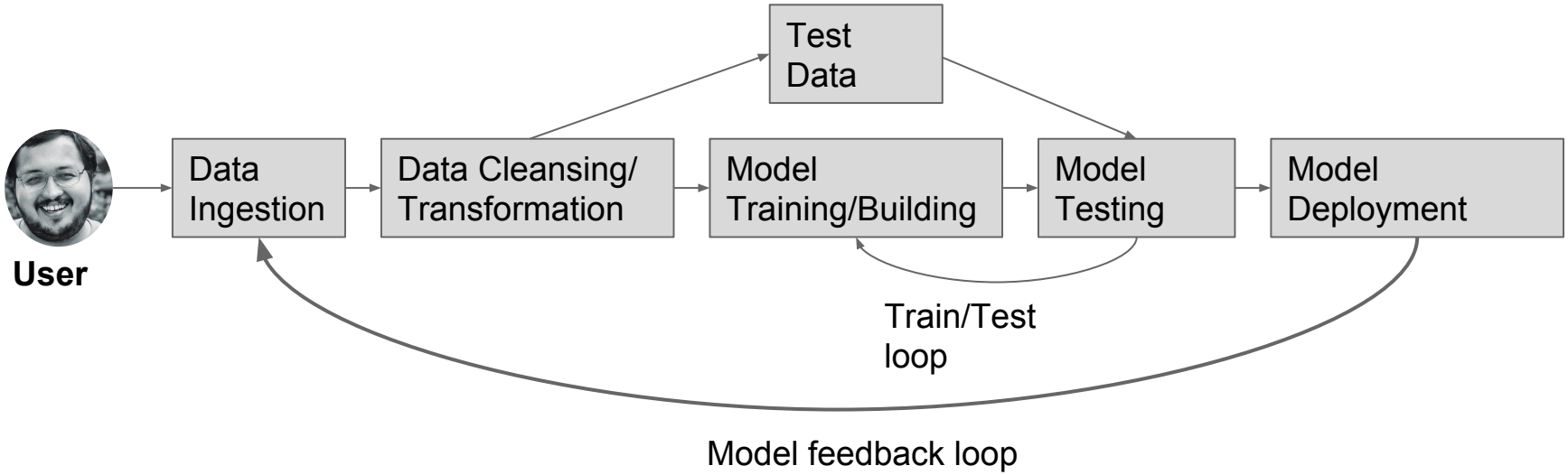


Models

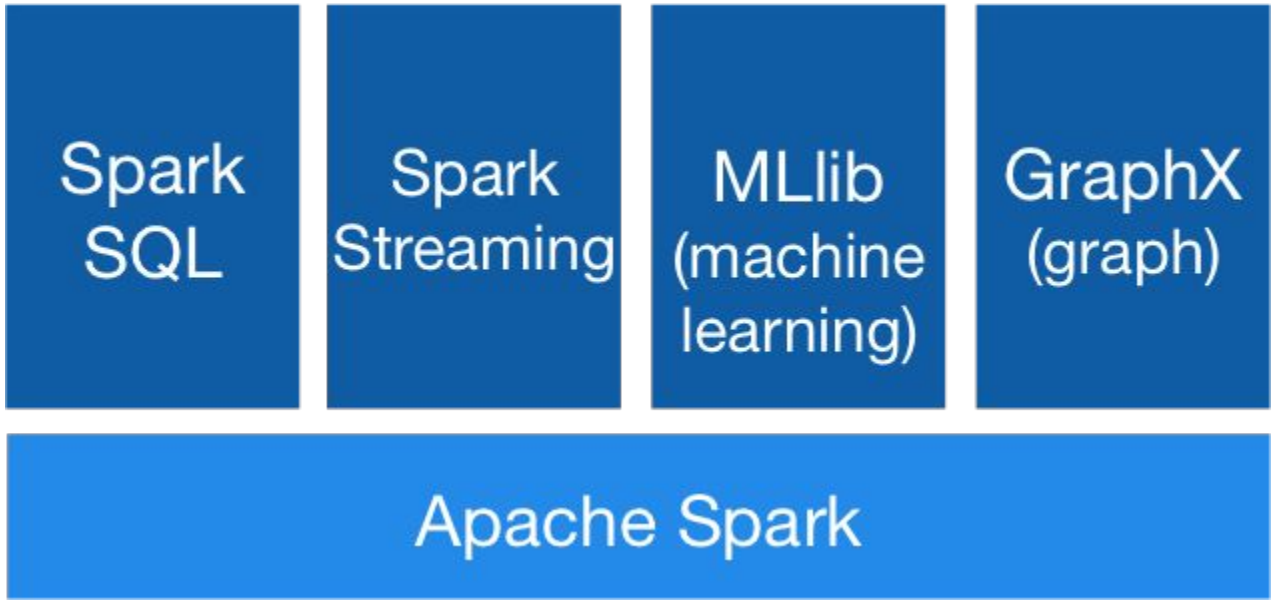


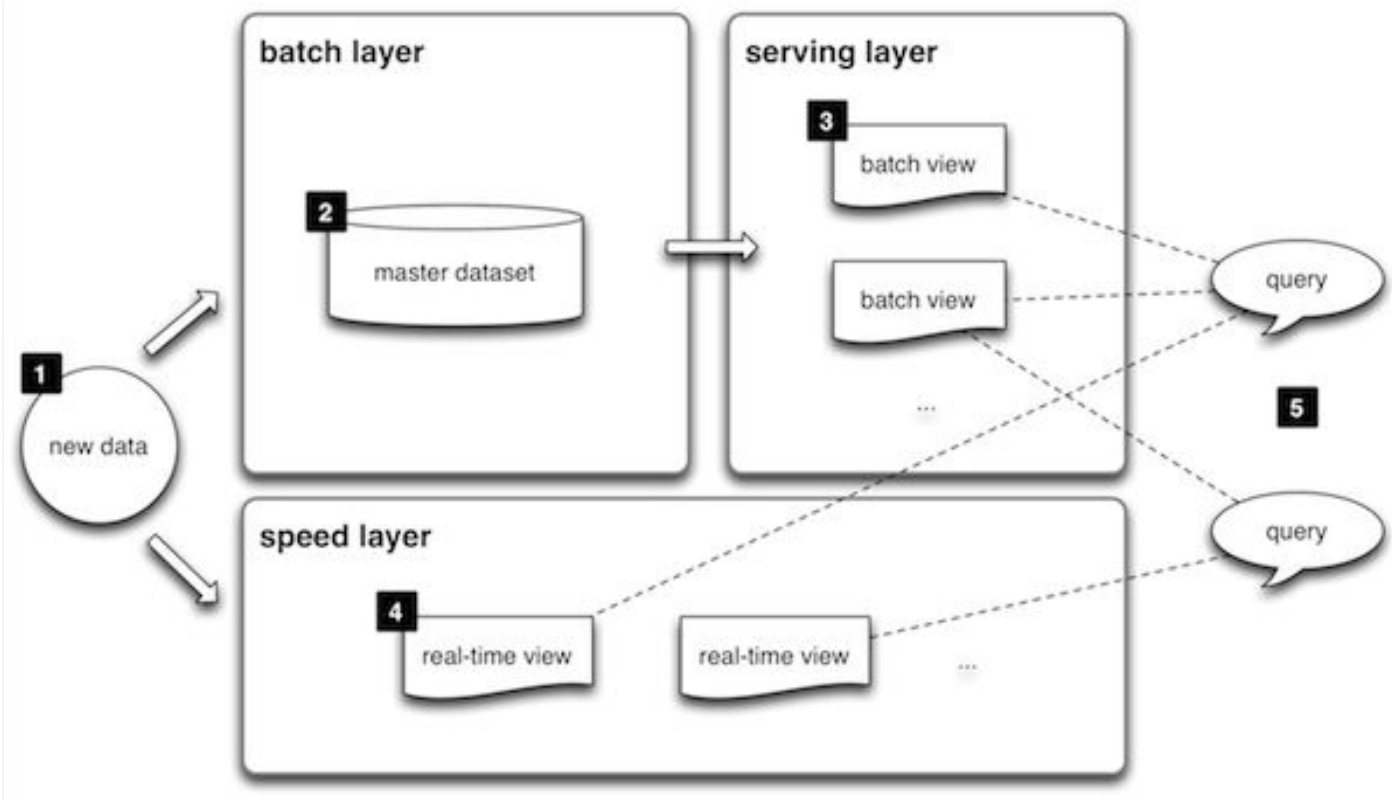
Prediction

Typical Machine Learning Workflow



Apache Spark - spark.apache.org





Lambda Architecture - http://lambda-architecture.net/

Tasks:

- Spark Internals - https://github.com/JerryLead/SparkInternals
- PySpark Pictures
- Fast PySpark
- CI/CD
- PANCAKE STACK