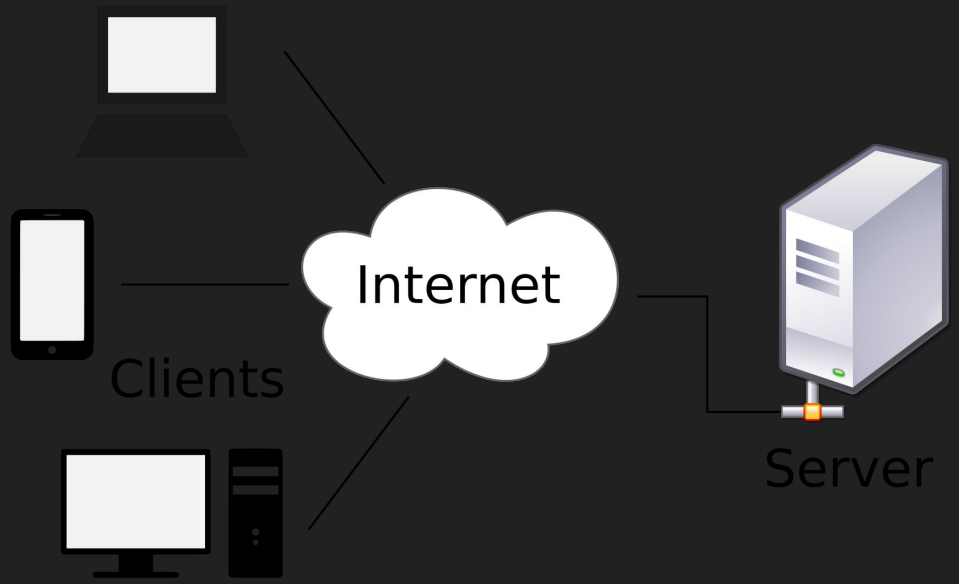
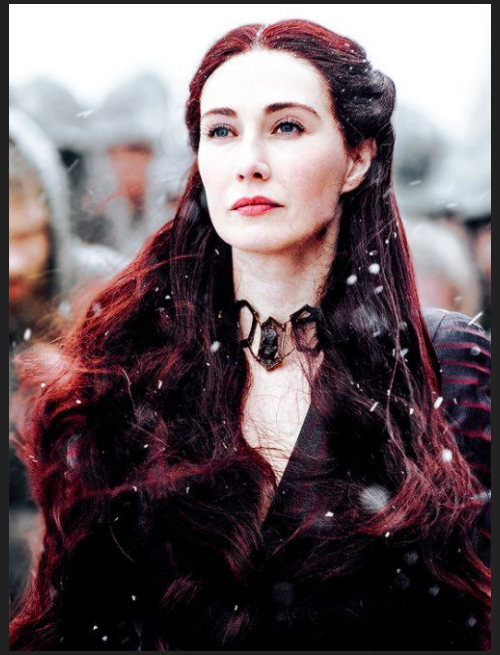
Introduction to Apache Spark

Eda Doko

Single Server

The one true Server





Distributed Computing

~~*The one true Server*~~

Google MapReduce

What is Spark?

Spark builds on MapReduce model.
A framework that enables distributed processing of massive data across clusters.

Key Takeaway: The main feature of Spark: In-memory cluster computing to increase processing speed.
Possible by reducing number of read/write operations to disk. It stores the intermediate processing data in memory.

How does it work?

RDD (Resilient Distributed Datasets) - Lowest level API available to user. An interface to a sequence of data objects that consist of one or more types, located across a variety of machines in a cluster.

Data Frame - a collection of distributed **Row** types. Provide a flexible interface similar to the ones in Pandas or R.

The Dataset - A combination of previous two. Provides typed interface available in RDDs and conveniences of dataframes.

RDD - Fault tolerant abstraction for in-memory cluster computing

Fault Tolerance - To achieve fault tolerance efficiently, RDDs provide a restricted form of shared memory, based on coarse grained transformations rather than fine-grained updates to shared state.

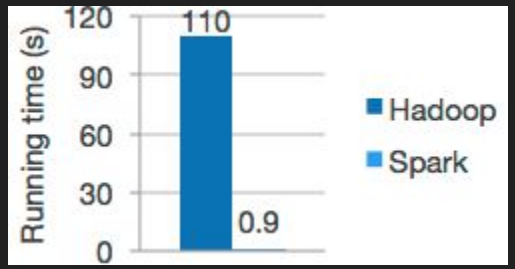
In-Memory: DBMS that primarily relies on main memory, improves performance by an order of magnitude.

Cluster Computing: connected computers that work together

Why use Spark? Features!

1. Speed:

Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.



Logistic Regression in Hadoop and Spark

Why use Spark? Features!

2. Simple

Supports different languages by providing high level API (Scala, Java, Python, R)

```
> from pyspark.sql.functions import *
  titanic = sqlContext.read.format('com.databricks.spark.csv').options(header='true', inferSchema='true').load('/databricks-datasets/Rdatasets/data-001/csv/COUNT/titanic.csv').cache()
  titanic.select("*").where(col("survived") == "no").show()
```

▸ (3) Spark Jobs

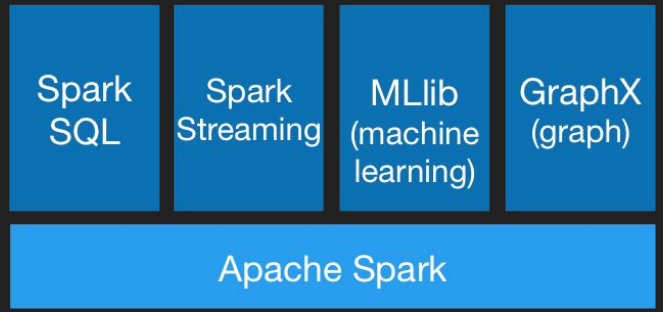
```
+---+---------+------+---+--------+
|   |    class|   age|sex|survived|
+---+---------+------+---+--------+
| 58|1st class|adults|man|      no|
| 59|1st class|adults|man|      no|
| 60|1st class|adults|man|      no|
| 61|1st class|adults|man|      no|
| 62|1st class|adults|man|      no|
| 63|1st class|adults|man|      no|
| 64|1st class|adults|man|      no|
| 65|1st class|adults|man|      no|
| 66|1st class|adults|man|      no|
| 67|1st class|adults|man|      no|
| 68|1st class|adults|man|      no|
| 69|1st class|adults|man|      no|
| 70|1st class|adults|man|      no|
| 71|1st class|adults|man|      no|
| 72|1st class|adults|man|      no|
| 73|1st class|adults|man|      no|
| 74|1st class|adults|man|      no|
| 75|1st class|adults|man|      no|
| 76|1st class|adults|man|      no|
| 77|1st class|adults|man|      no|
+---+---------+------+---+--------+
only showing top 20 rows
```

Why use Spark? Features!

3. Advanced Analytics

Supports and can combine the following libraries into same application.

Spark SQL: for Sql and unstructured data processing
Spark Streaming: Stream processing of live data streams
MLlib: machine learning algorithms
GraphX: Graph Processing



Why use Spark? Features!

3. Runs Everywhere

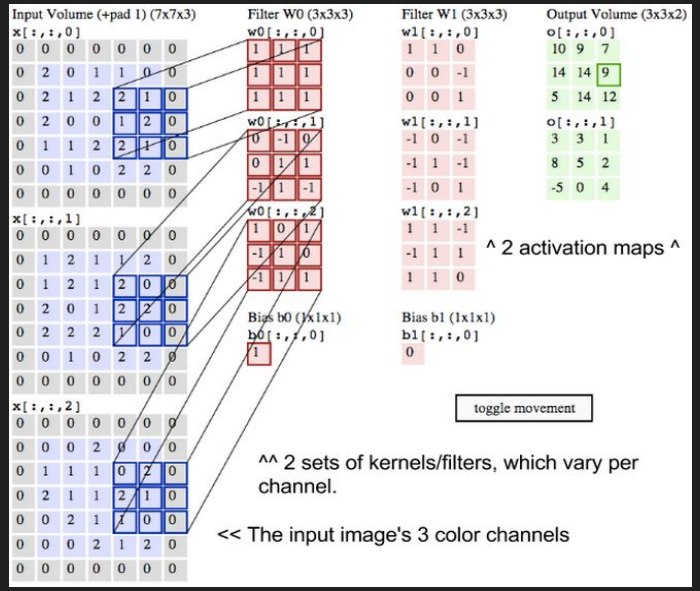
Spark runs on Hadoop, Mesos, standalone, or in the cloud. It can access diverse data sources including HDFS, Cassandra, HBase, and S3.

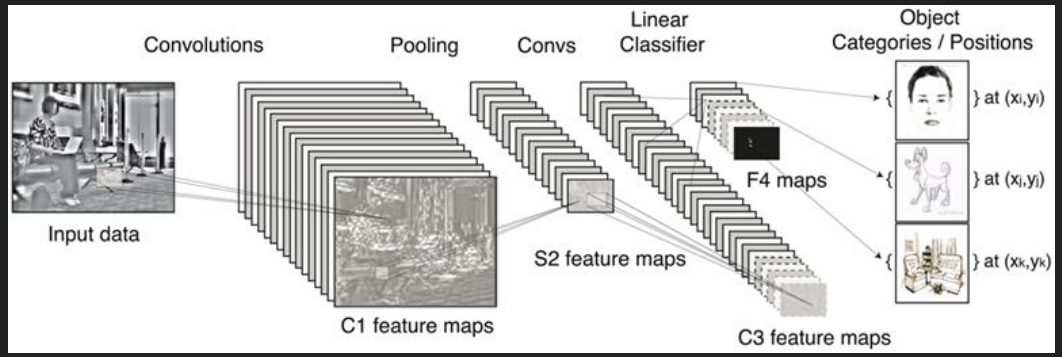
What can you do with it?

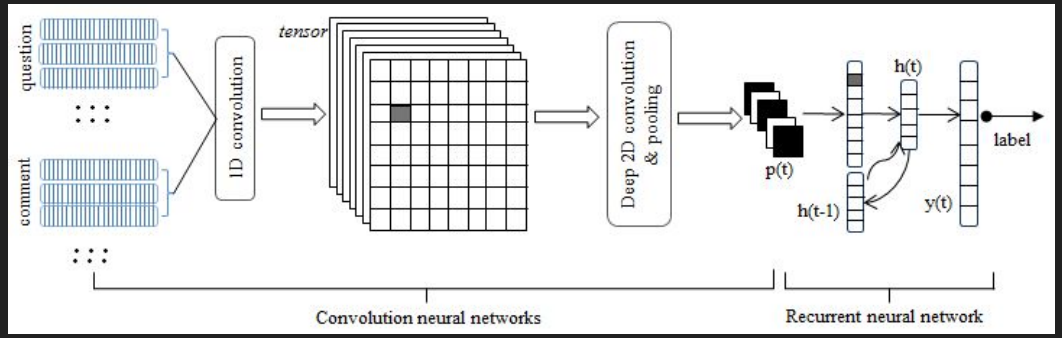
Machine Learning!

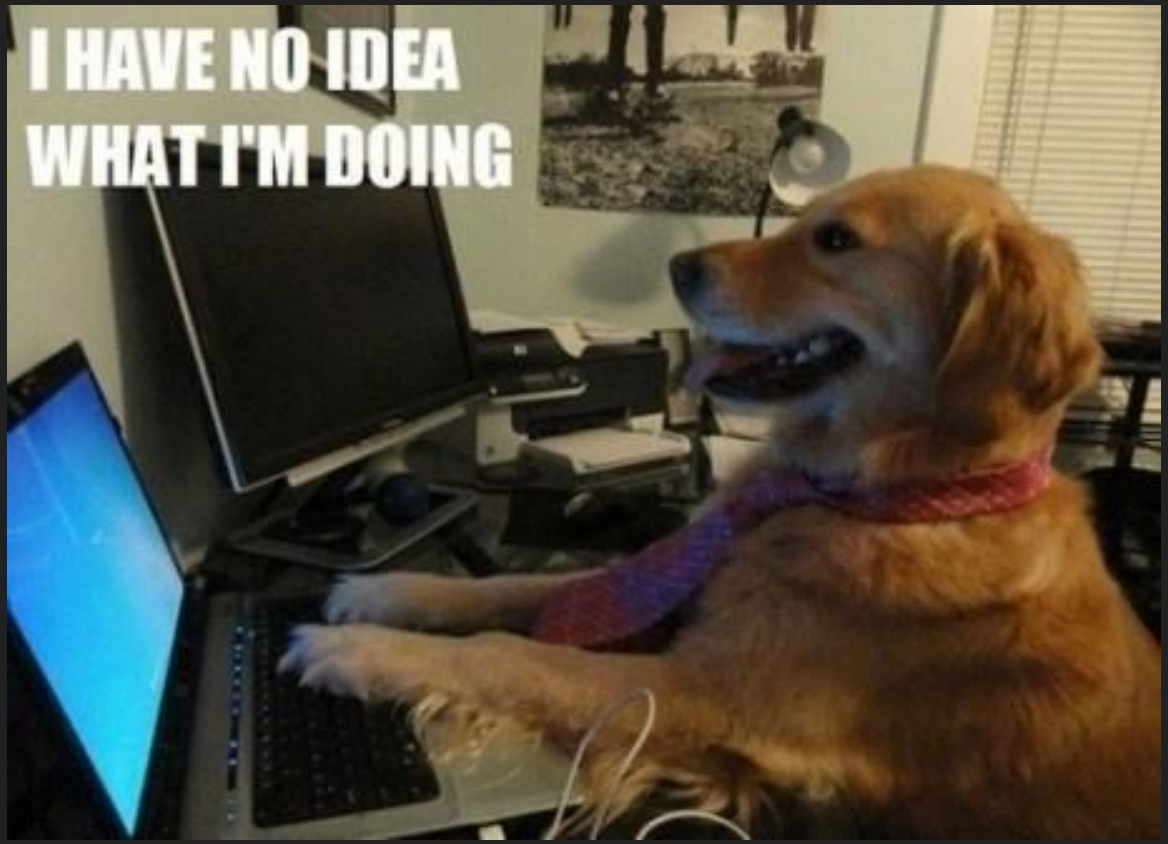
What people think Machine Learning is like...

What it's really like...



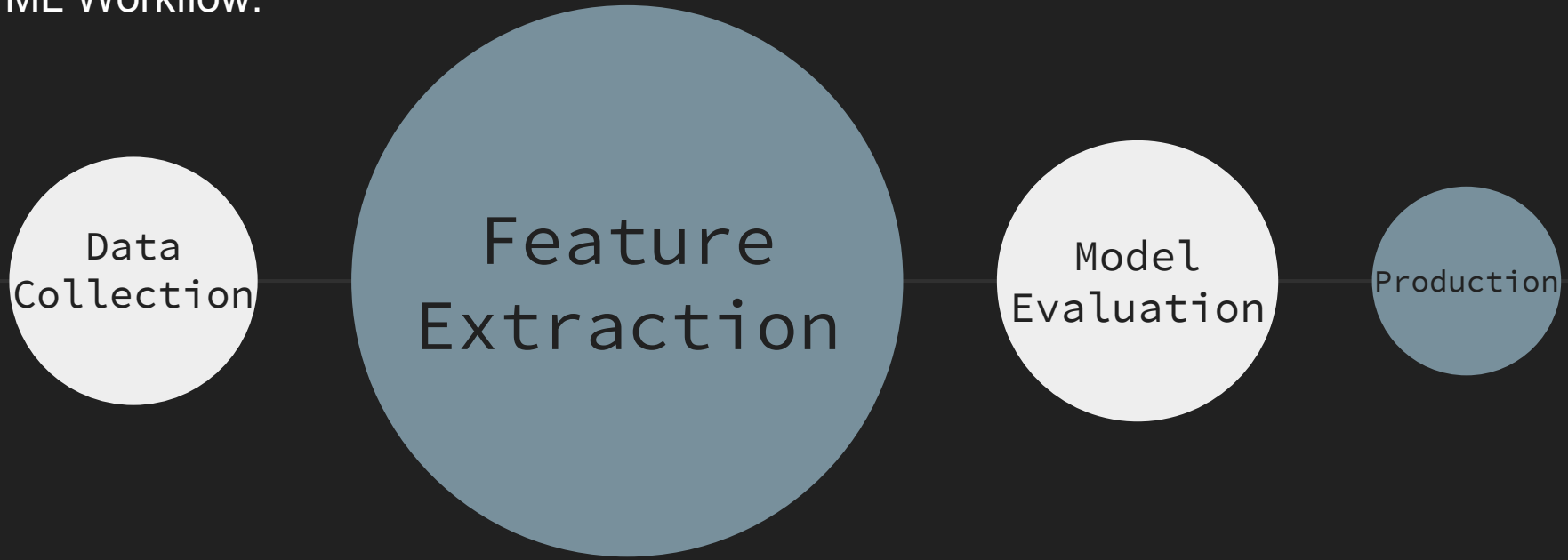





I HAVE NO IDEA
WHAT I'M DOING

But seriously...

ML Workflow:



Spark has tools to support the ML workflow

1. Feature Preparation - RDDs, Spark SQL
2. Model training - MLlib
3. Model Evaluation - MLlib
4. Production use - model.predict()

(all operate on RDD's)

Databricks

Founded by the creators of Spark.

Offers a hosted service, Databricks cloud. Spark on EC2 with notebooks, dashboards, scheduled jobs.

Demo

Resources

http://spark.apache.org/

https://cs.stanford.edu/~matei/papers/2012/nsdi_spark.pdf

Spark Summit 2015

https://community.cloud.databricks.com