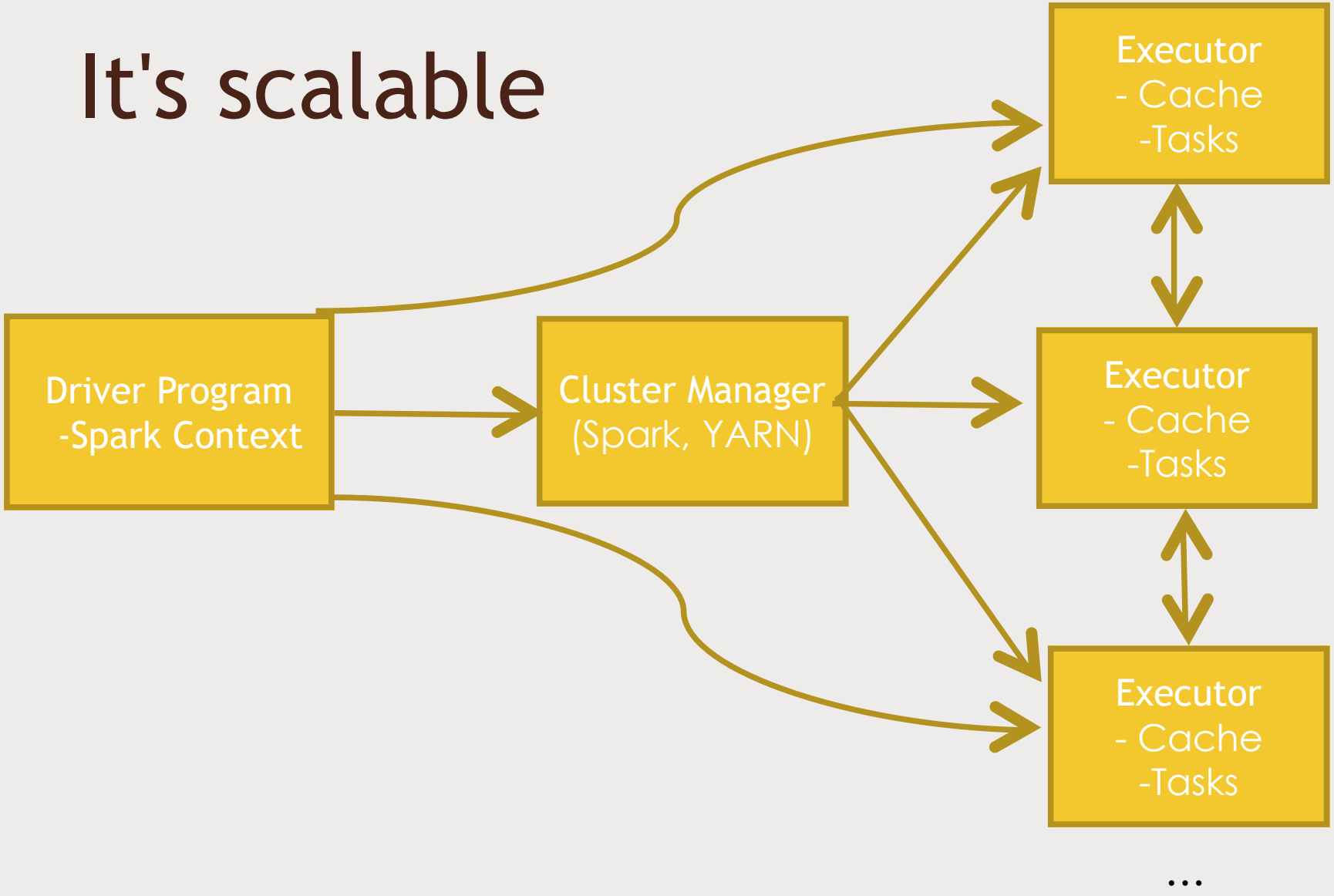
INTRODUCTION TO SPARK

Frank Kane

What is spark?

- "A fast and general engine for large-scale data processing"

It's scalable



It's fast

- "Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk."
- DAG Engine (directed acyclic graph) optimizes workflows

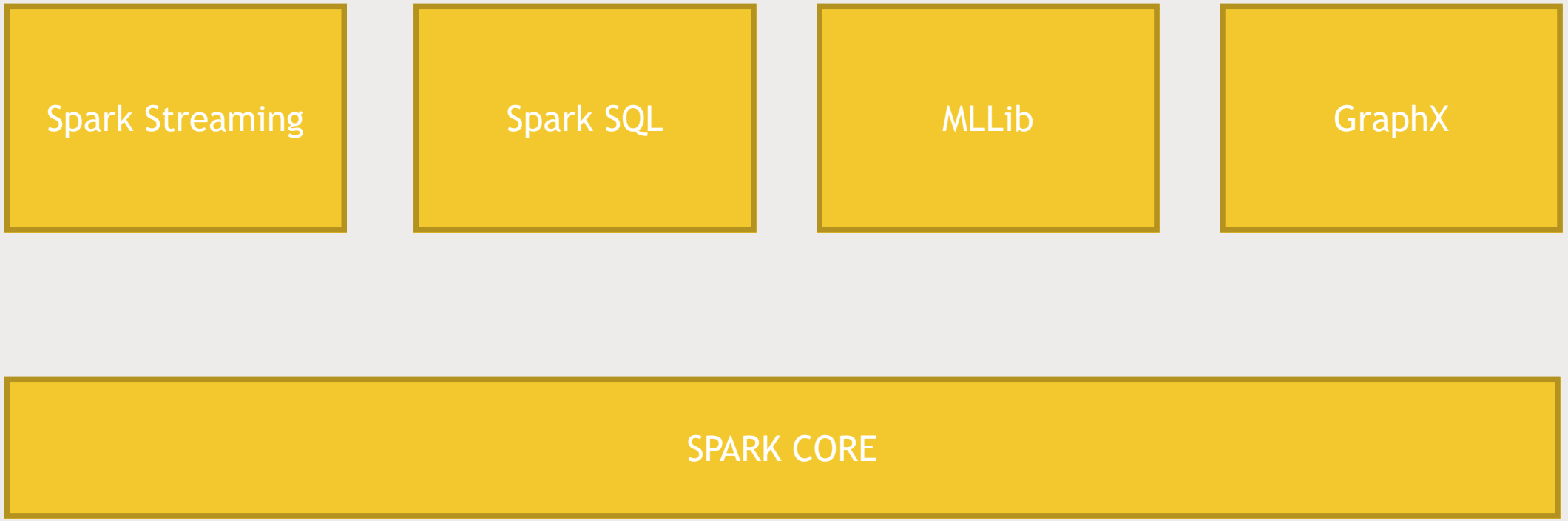
It's hot

- Amazon
- Ebay: log analysis and aggregation
- NASA JPL: Deep Space Network
- Groupon
- TripAdviser
- Yahoo
- Many others: https://cwiki.apache.org/confluence/display/SPARK/Powered+By+Spark

It's not that hard

- Code in Python, Java, or Scala
- Built around one main concept: the Resilient Distributed Dataset (RDD)

Components of spark



Let's Use Python

- Why Python?
  - *It's a lot simpler, and this is just an overview.*
  - *Don't need to compile anything, deal with JAR's, dependencies, etc*
- But...
  - *Spark itself is written in Scala*
  - *Scala's functional programming model is a good fit for distributed processing*
  - *Gives you fast performance (Scala compiles to Java bytecode)*
  - *Less code & boilerplate stuff than Java*
  - *Python is slow in comparison*

FEAR NOT

- Scala code in Spark looks a LOT like Python code.

Python code to square numbers in a data set:

```
nums = sc.parallelize([1, 2, 3, 4])
squared = nums.map(lambda x: x * x).collect()
```

Scala code to square numbers in a data set:

```
val nums = sc.parallelize(List(1, 2, 3, 4))
val squared = nums.map(x => x * x).collect()
```