Handling Fast Data with Apache Spark SQL and Streaming

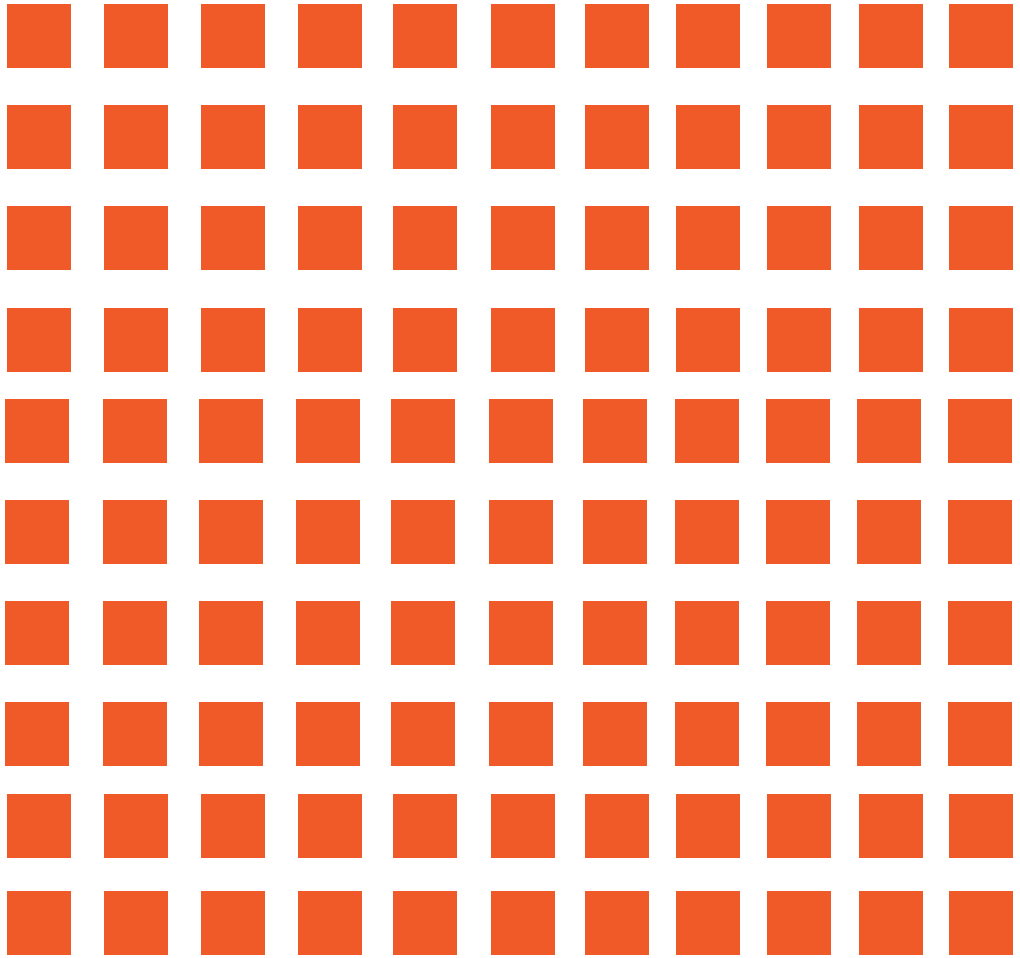
INTRODUCTION

Justin Pihony

@JustinPihony | justin-pihony.blogspot.com

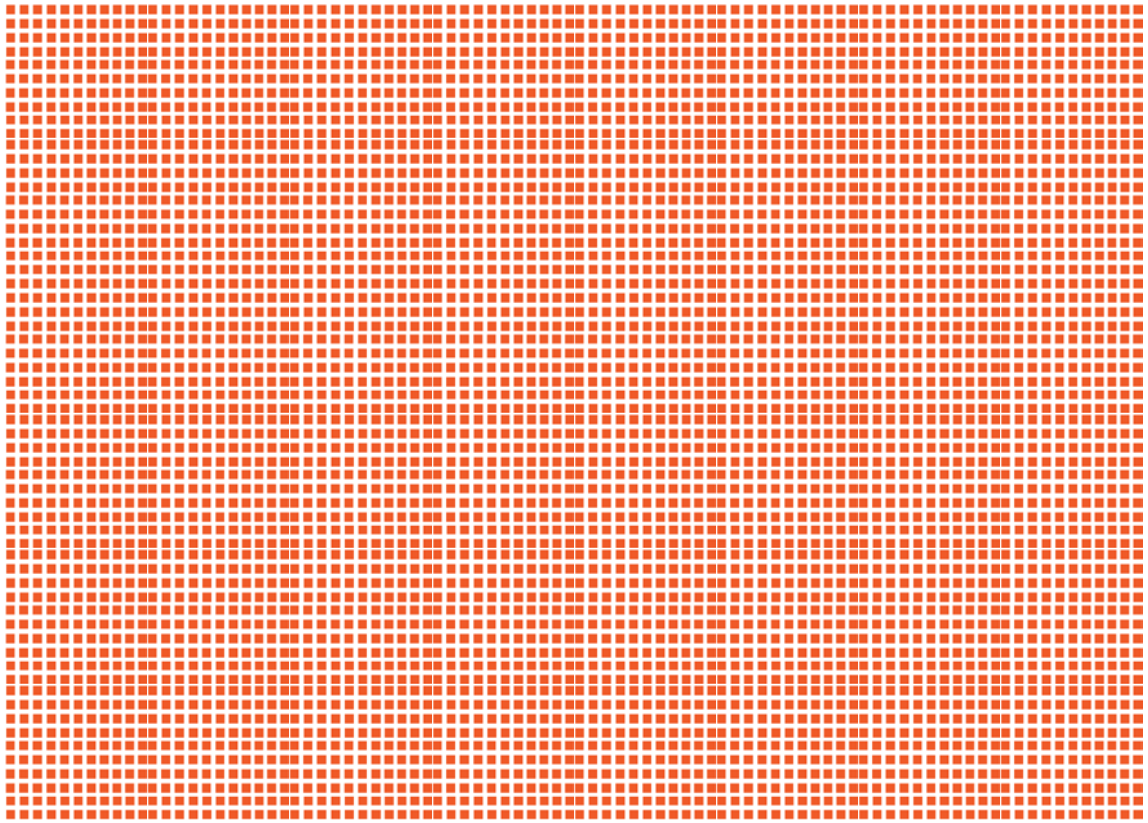
Big Data

Fast Data



Fast Data

Lots of it!



Introduction

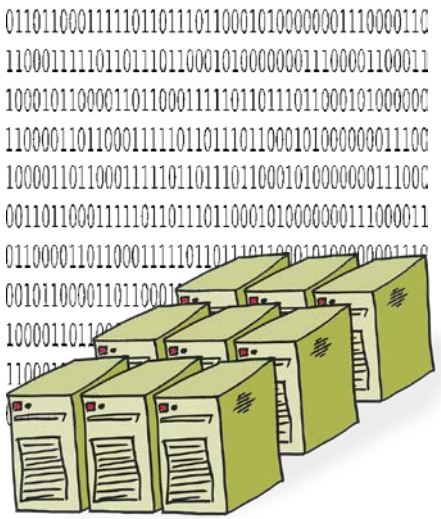
What Is Fast Data?

What to Expect

Spark 2.x

What Is Fast Data?

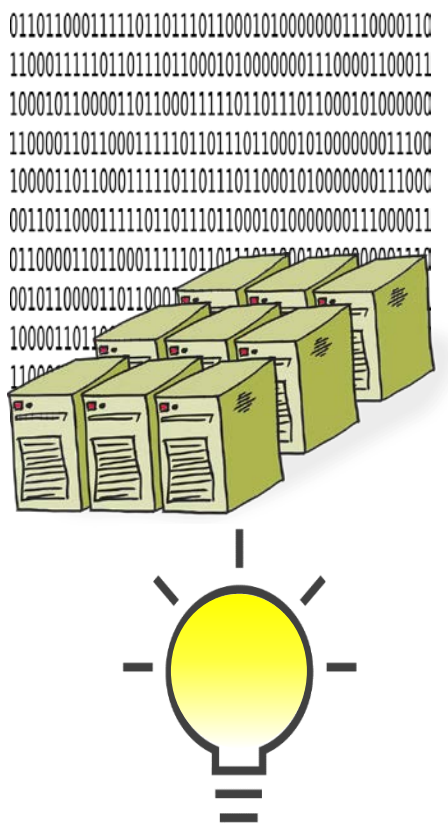
Big Data



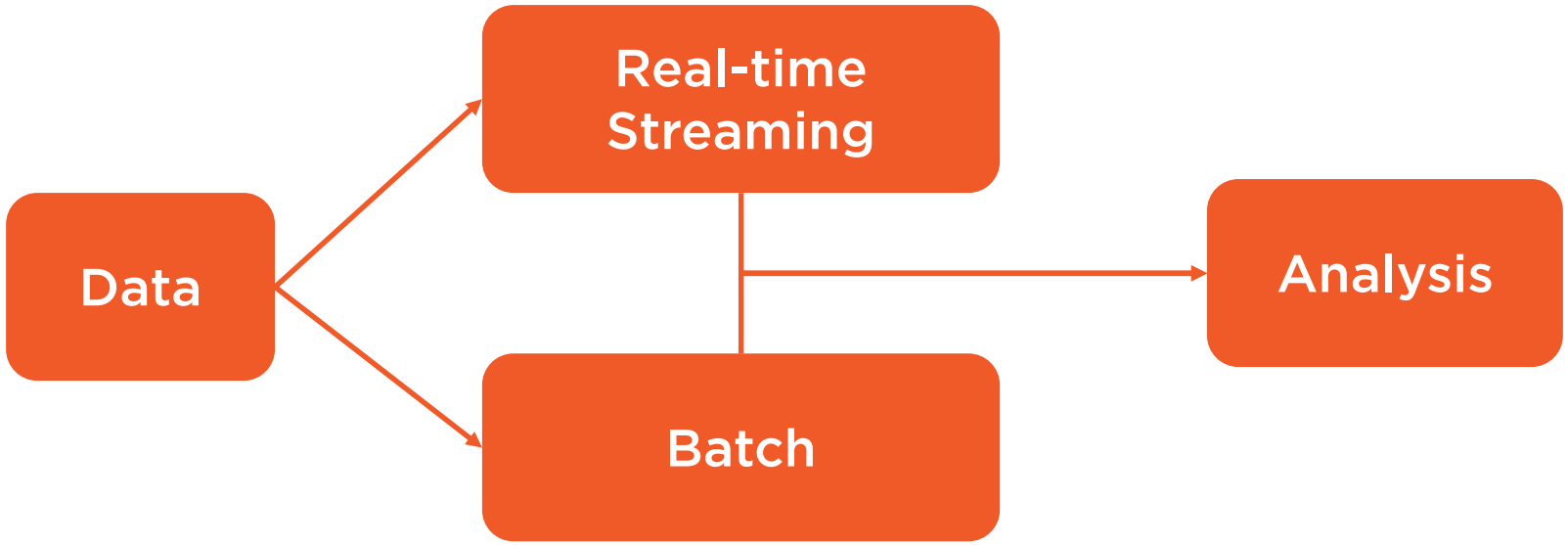
■ ■ ■

Big SLOW Data

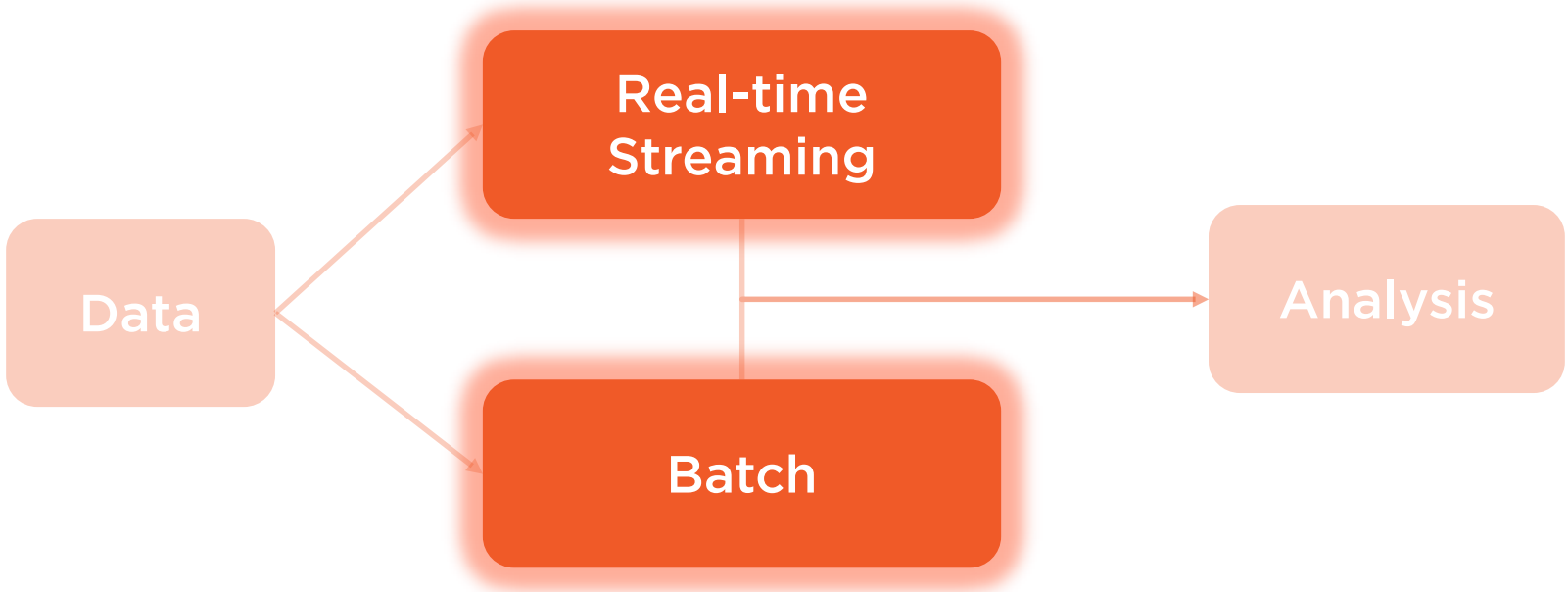
Big FAST Data



Lambda Architecture



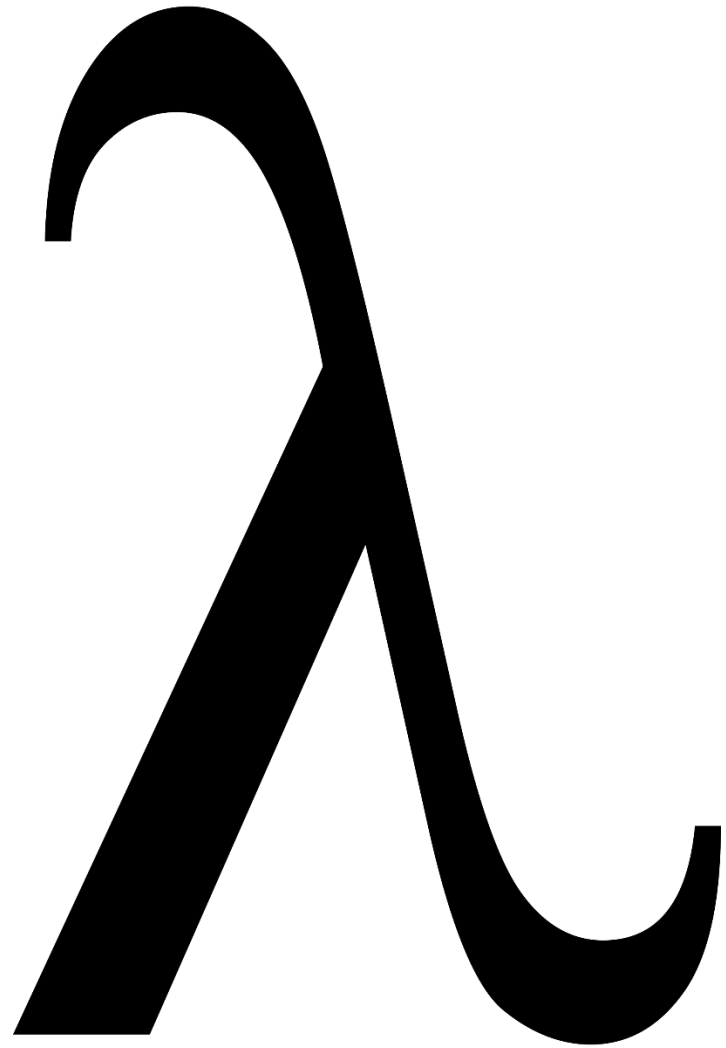
Lambda Architecture

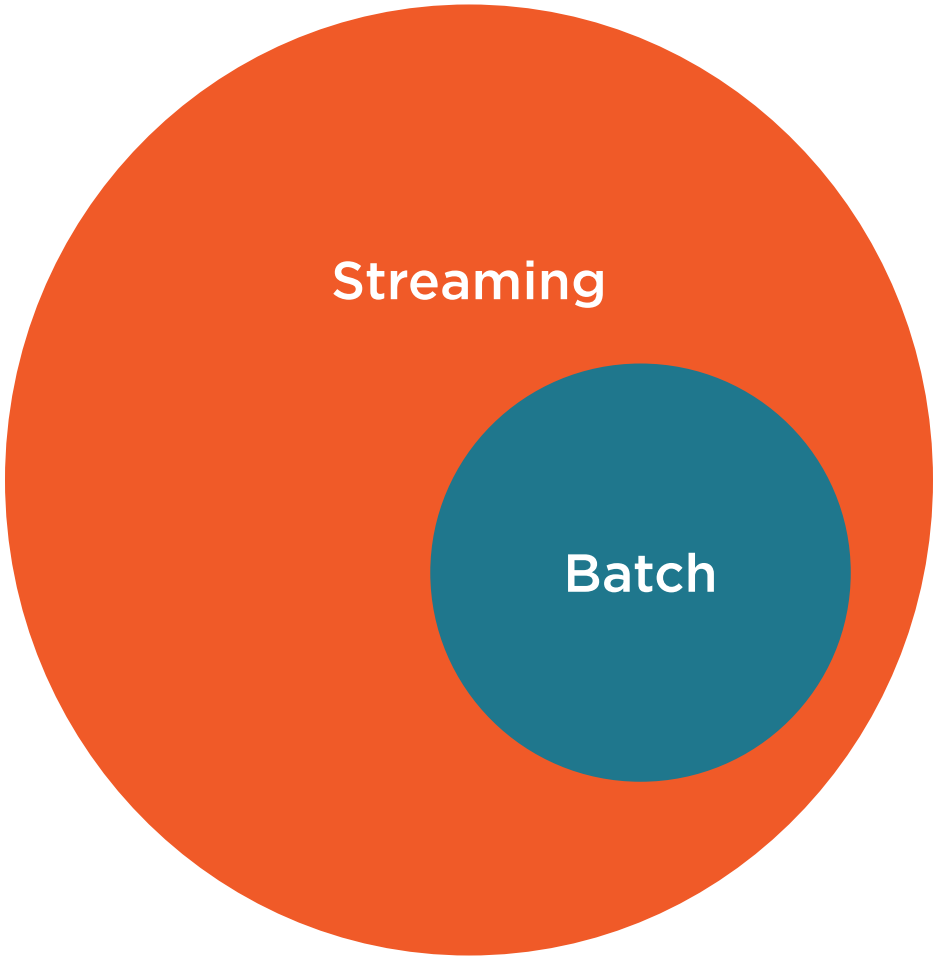


“Why can’t the stream processing system be improved to handle the full problem set in its target domain?”

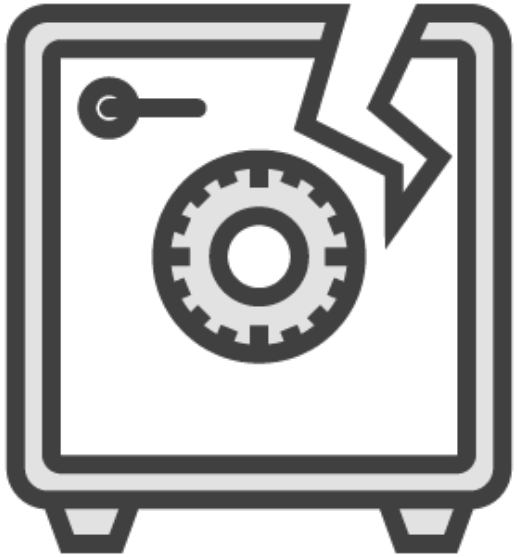
-Jay Kreps

Kappa Architecture



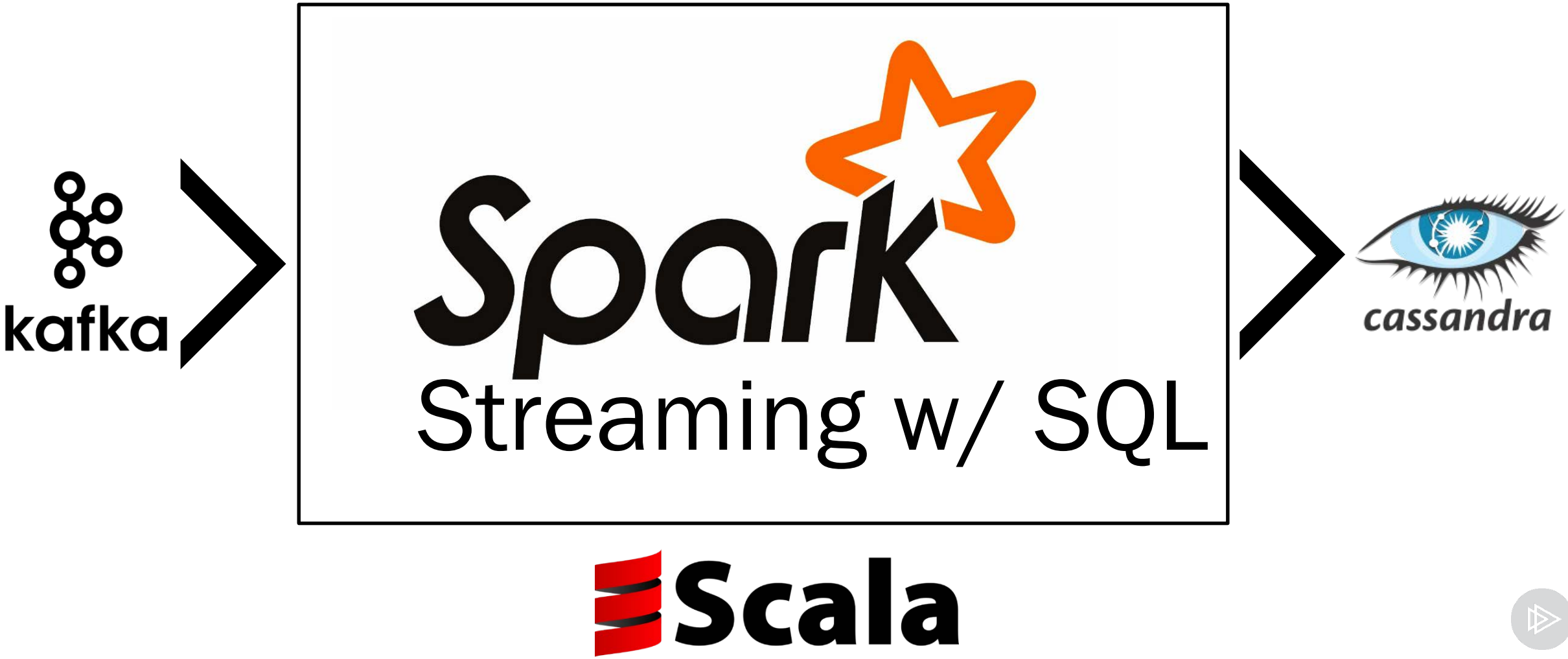

Streaming
Batch

Why Fast Data?



What to Expect

Course Project



Course Project

Course Overview

DataFrames

Datasets

Spark Streaming

Optimizing Towards Fast Data

Spark 2.x

Structured Streaming

"The simplest way to perform streaming analytics is not having to reason about streaming at all"

Tathagata Das

```
val output = df
      .select($“name”, $“age”)
      .where($“age” > 21)
```

◀ Primary Logic

```
val df = spark.read
       .format(“json”)
       .load(“/INPUT/PATH”)

val output = df
       .select($“name”, $“age”)
       .where($“age” > 21)

output.write
       .format(“parquet”)
       .save(“/OUTPUT/PATH”)
```

◀ Input

◀ Primary Logic

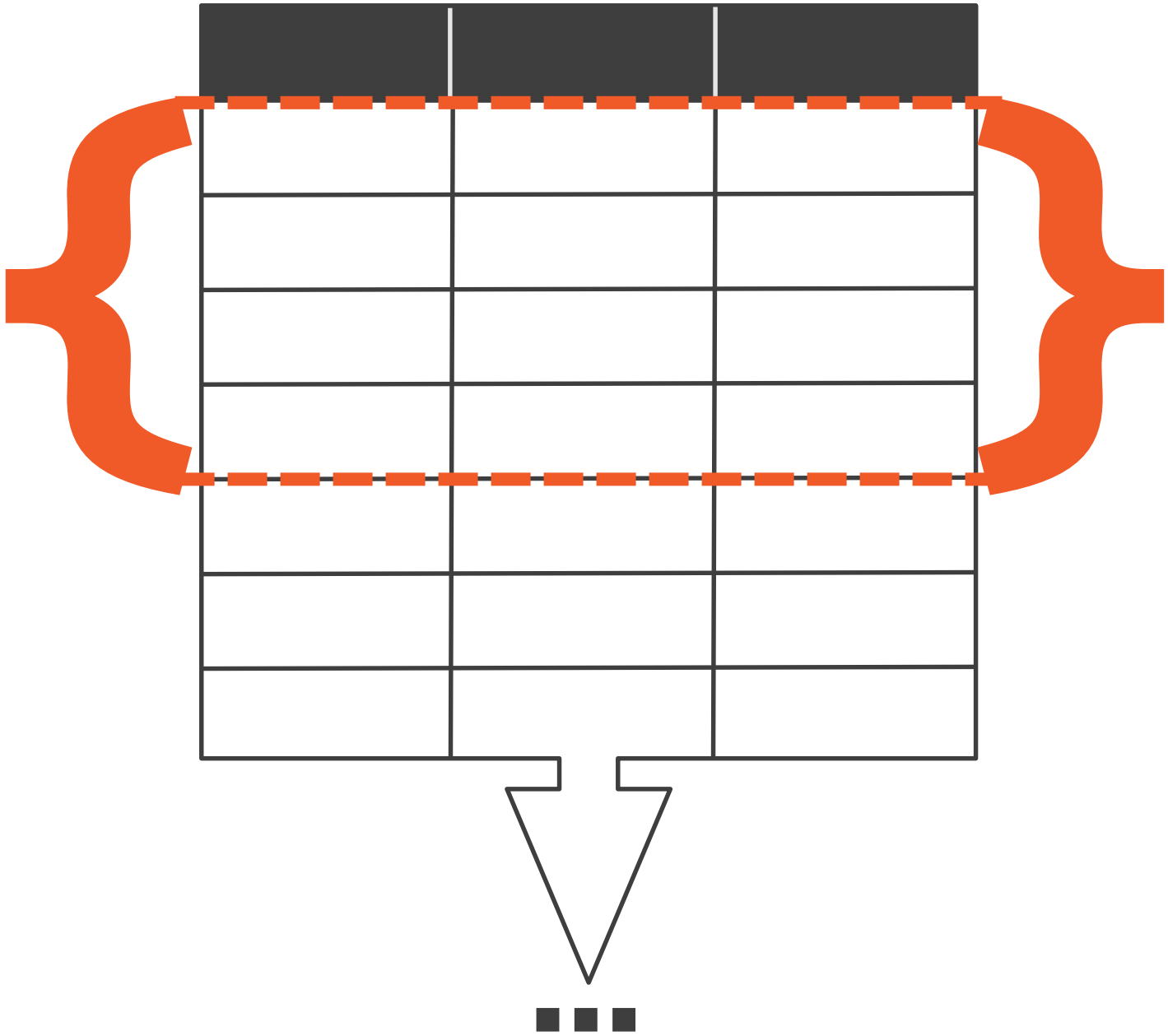
◀ Output

```
val df = spark.readStream
       .format(“json”)
       .load(“/INPUT/PATH”)

val output = df
       .select($“name”, $“age”)
       .where($“age” > 21)

output.writeStream
       .format(“parquet”)
       .start(“/OUTPUT/PATH”)
```

◀ Input

◀ Primary Logic

◀ Output



Streamlined API

```
DataFrame = Dataset[Row]

SQLContext/HiveContext -> SparkSession
                &
          SparkContext
```

Smart == Fast

Tungsten 2.0

Whole-stage code generation (SPARK-12795)

5-10x speedups

Spark as a compiler

Spark 2.x

Expanded SQL

Accumulator simplification

DataFrame focused machine learning

Scala 2.11 as de facto

Resources

- Fast Data
  - Questioning the Lambda Architecture
    - www.oreilly.com/ideas/questioning-the-lambda-architecture
  - Batch is a special case of streaming
    - data-artisans.com/batch-is-a-special-case-of-streaming
  - Which Do We Need More: Big Data or Fast Data?
    - www.entrepreneur.com/article/243123
  - 2016 State of Fast Data & Streaming Applications
    - www.opsclarity.com/wp-content/uploads/2016/06/2016FastDataSurvey.pdf
  - Fast Data: Big Data Evolved/Fast Data Architectures For Streaming Applications
    - info.lightbend.com/COLL-20XX-Fast-Data-Big-Data-Evolved-WP_LP
    - info.lightbend.com/COLL-20XX-Fast-Data-Architectures-for-Streaming-Apps_LP

Summary

Fast Data

Course Project

Course Overview

Spark 2.x