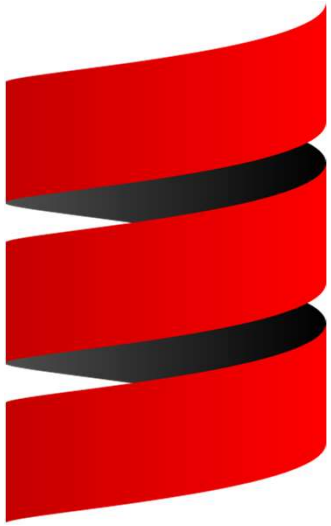
Understanding a Typed API: Datasets Works with Scala, Not Python

Xavier Morera

HELPING DEVELOPERS UNDERSTAND SEARCH & BIG DATA

@xmorera www.xaviermorera.com

Scala



General purpose programing language

- Support for functional programming
- Object oriented

Aimed to address some criticisms of Java

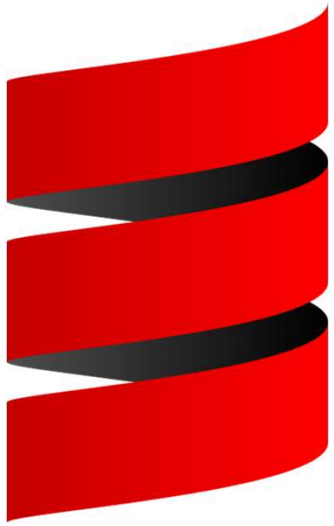
Statically typed

Runs on JVM

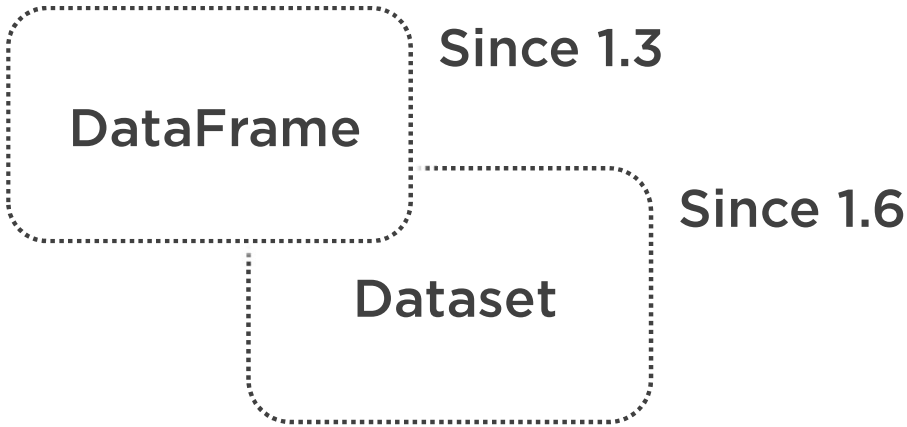
- Interoperability with Java

What's Built in Scala?

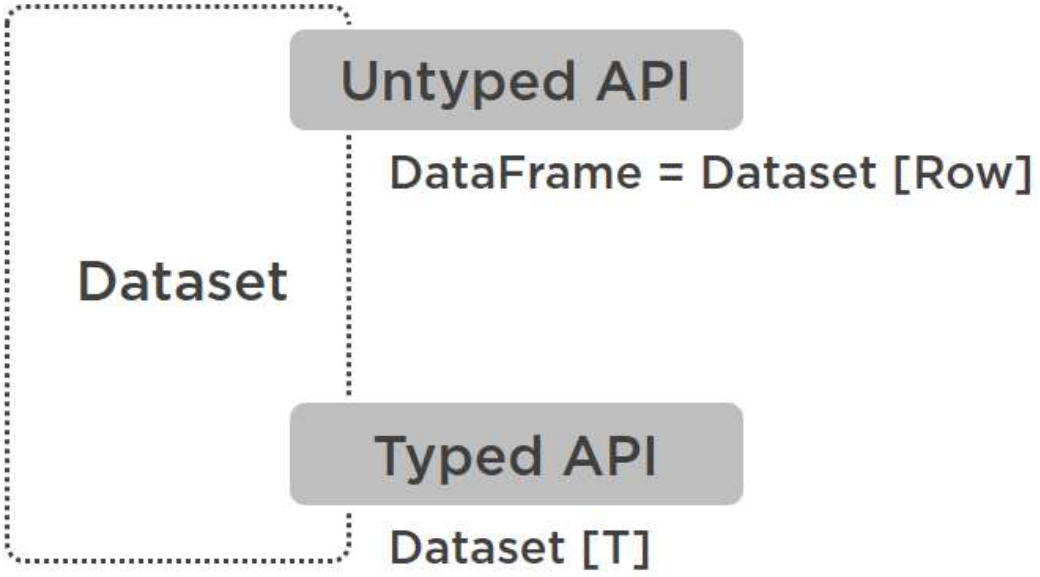
Scala is a Whole Different Course



Higher Level API

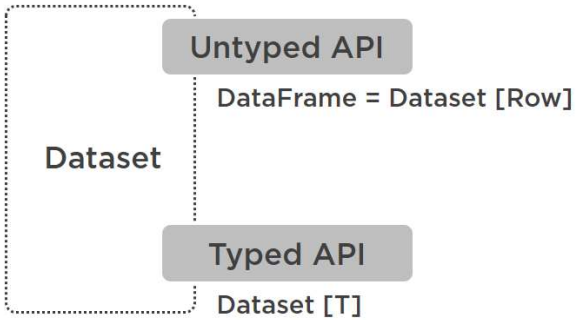


Higher Level



Since 2.0

Dataset



Collection of strongly typed objects

- Distributed and resilient

Schema

Typed columns

- Typed transformations
- Errors at compile time

Case classes or primitive types

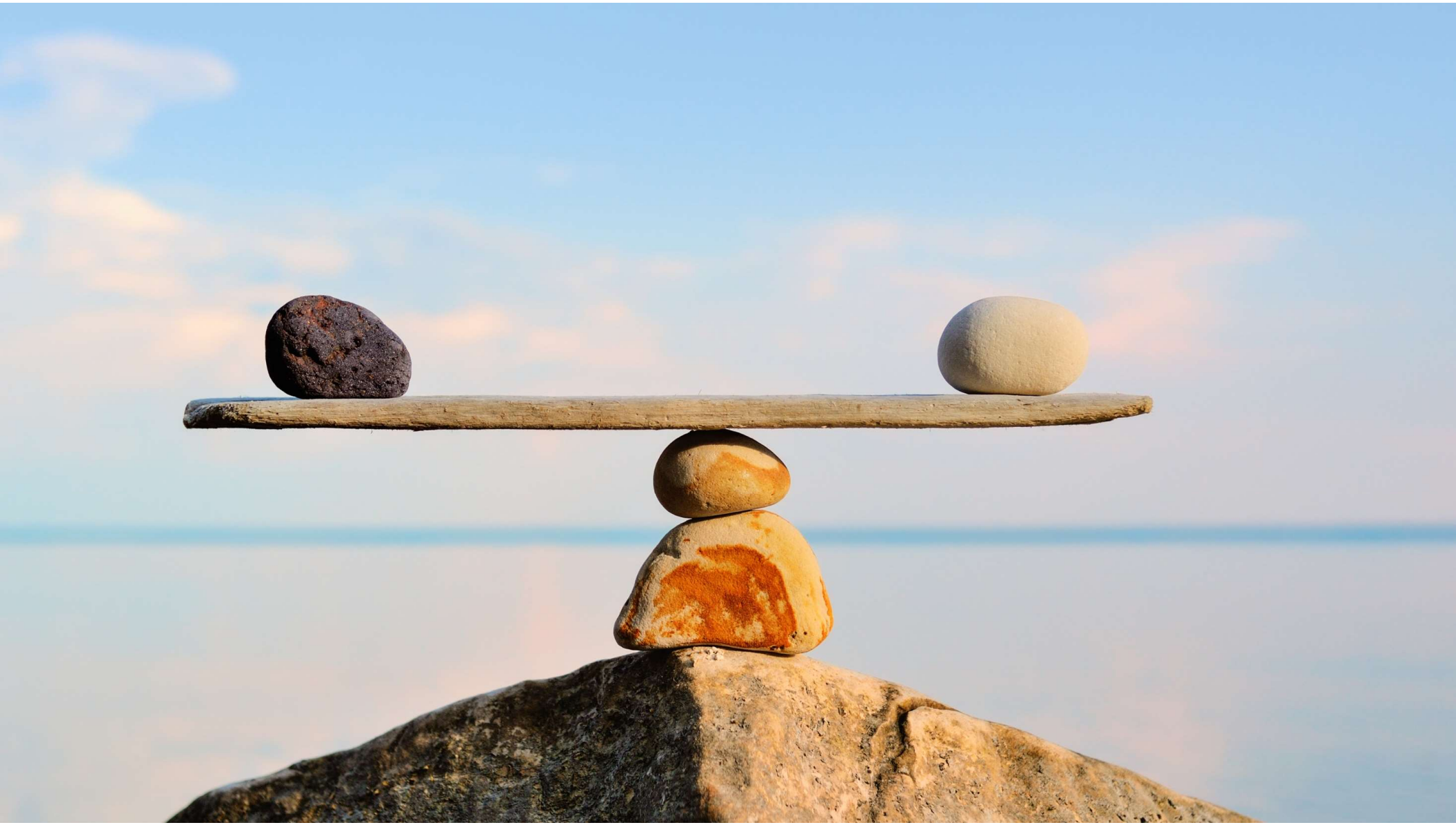
```
val commentsDF =
spark.read.parquet('/user/cloudera/stackexchange/comments_parquet')

case class Comments(Id: Int, PostId: Int, Score: Int, CreationDate:
java.sql.Timestamp, UserId: Int)

val commentsDS = comments_parquetDF.as[Comments]

commentsDS

commentsDS.show(3)
```

Datasets

Several ways of creating Datasets, let's load some data

Define a case class and use as[]

We have a Dataset!



Takeaway

Higher Level API

DataFrames

- Dataset[Row]

Haven't used Datasets

- Why?
- Python is not statically typed
- Scala is strongly typed

Takeaway

Spark implemented in Scala

Defined a Dataset

- Case class
- Used as[ ]

Dataset API out of scope

- Check the library for more