Improving Type Safety with Datasets

Justin Pihony

@JustinPihony|justin.pihony@blogspot.com

Course Overview

DataFrames

Datasets

Spark Streaming

Optimizing Towards Fast Data

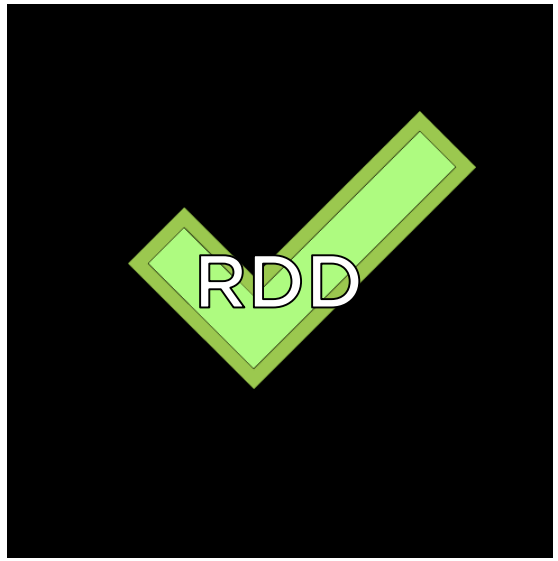
Module Overview

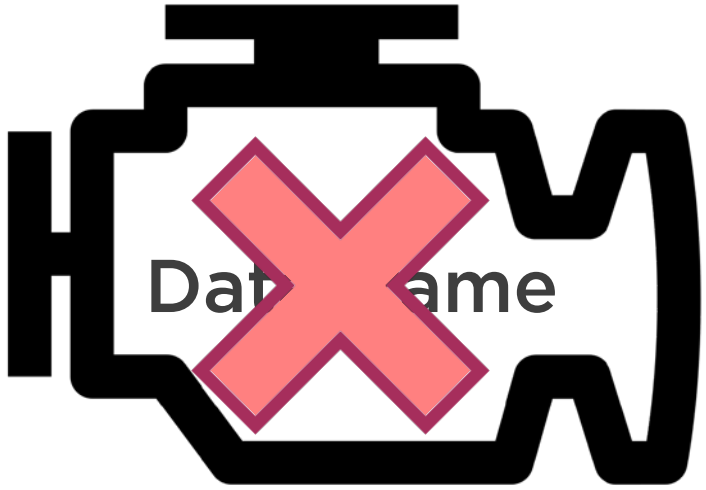
Datasets

- Why
- Embracing type safety
- Encoders
- Datasource expansion

Why Datasets?

Why?





Why?

“A Dataset is a strongly-typed, immutable collection of objects that are mapped to a relational schema.”

https://databricks.com/blog/2016/01/04/introducing-apache-spark-datasets.html

Why?

```
Dataset[MyClass]

DataFrame = Dataset[Row]
```

Why?

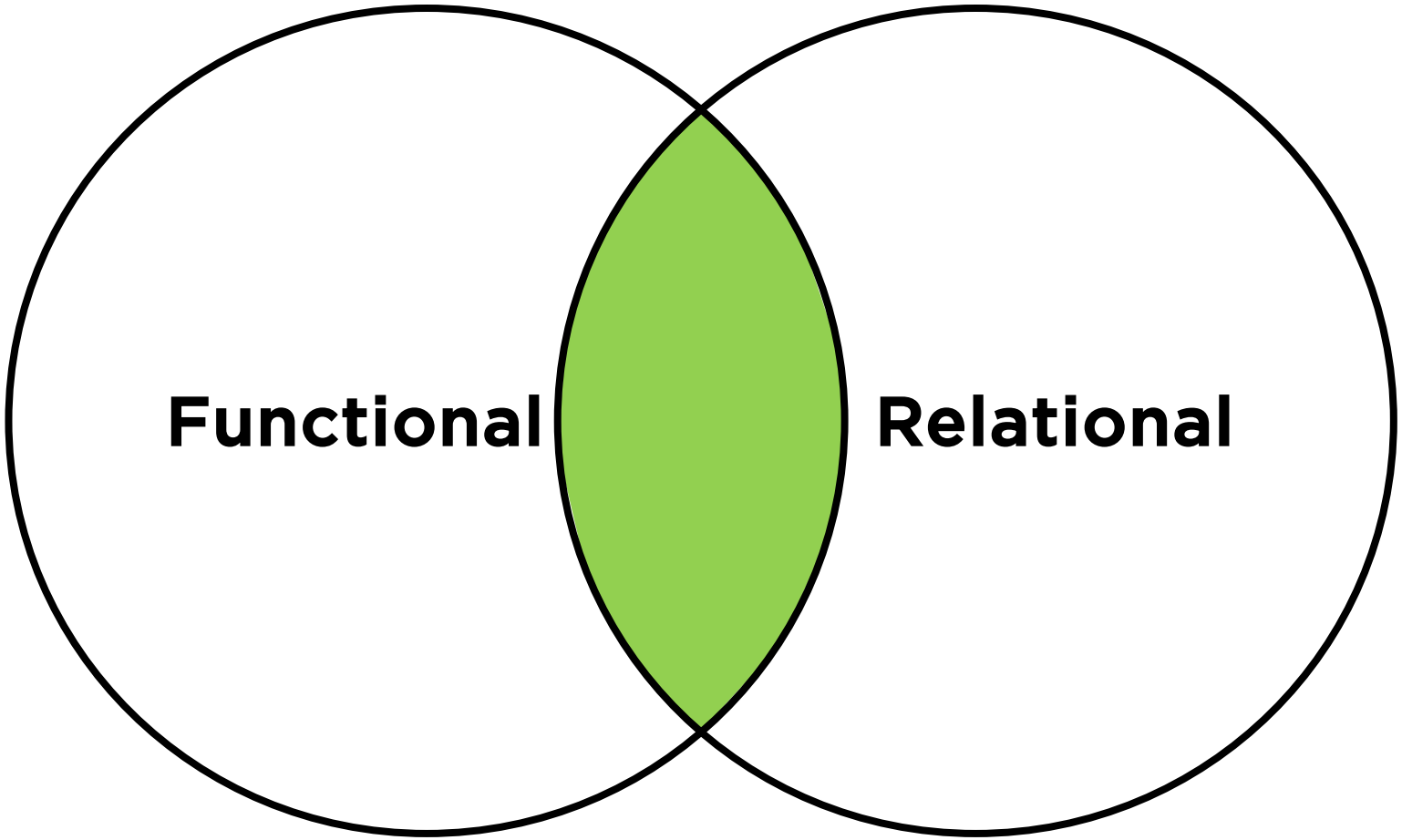
```
import org.apache.spark.sql._
import org.apache.spark.sql.types._
val schema = StructType(List(
                StructField("test", BooleanType, true)))
val rdd = spark.sparkContext.parallelize(
            List(Row(0), Row(true), Row("stuff")))
val df = spark.createDataFrame(rdd, schema)
df.collect
```

java.lang.ClassCastException

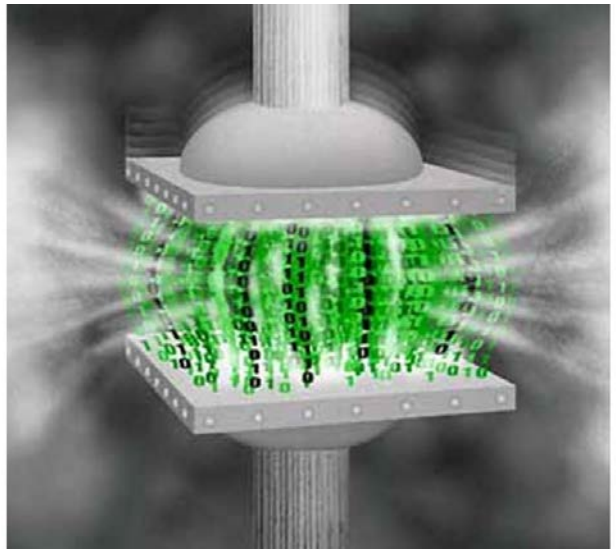
Why?

```
import org.apache.spark.sql._
import org.apache.spark.sql.types._
val schema = StructType(List(
                StructField("test", BooleanType, true)))
val rdd = spark.sparkContext.parallelize(
            List(Row(0), Row(true), Row("stuff")))
val ds = spark.createDataset(rdd, schema)
ds.collect
```

Why?



Why?

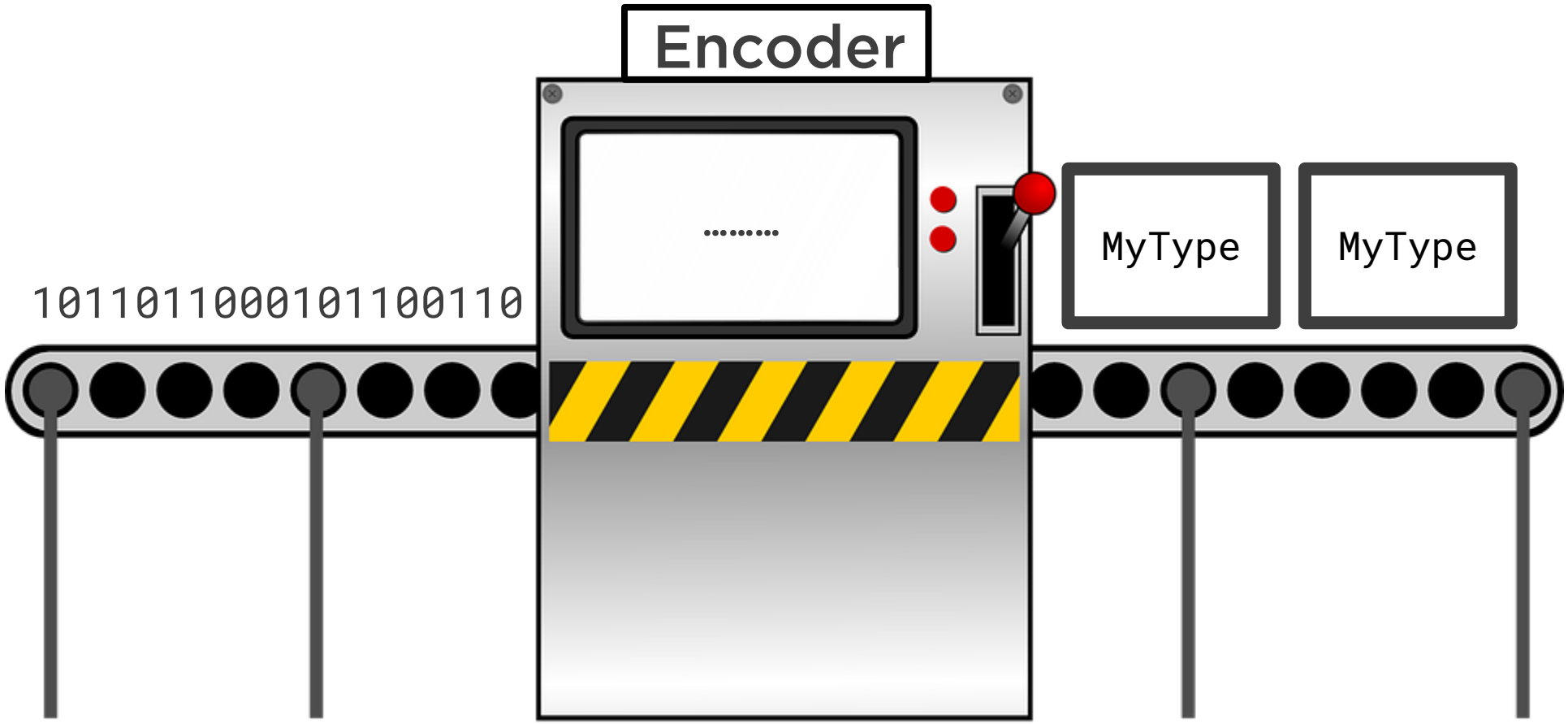


DataFrames Are Datasets

All About Types

- **Actions**
  - `collect`, `take`, `foreach`, **...**
  - `toLocalIterator` **(2.0)**
- `toDF` **and** `toDS`
- **Typed vs Untyped Transformations**

Encoders


Encoder
.....
MyType
MyType
10110110001011001l0

“...the serialized data is already in the Tungsten binary format, which means that many operations can be done in-place, without needing to materialize an object at all”

https://databricks.com/blog/2016/01/04/introducing-apache-spark-datasets.html

Encoders

Scala

```
  import spark.implicits._
```

Java

```
org.apache.spark.sql.Encoders.BOOLEAN
                              .DATE
                              .DOUBLE
                              .INT
                              .TIMESTAMP
                              .STRING
                              .bean[T]
                              …
```

[Key_1, NewData]

...
[Key_1, Data]
...

Cassandra for Developers
https://www.pluralsight.com/courses/cassandra-developers

...
[Key_1, **NewData**]
...

Cassandra for Developers
https://www.pluralsight.com/courses/cassandra-developers

Data Sources

Native

spark-packages

JDBC

Native

JDBC

com.databricks.spark.csv

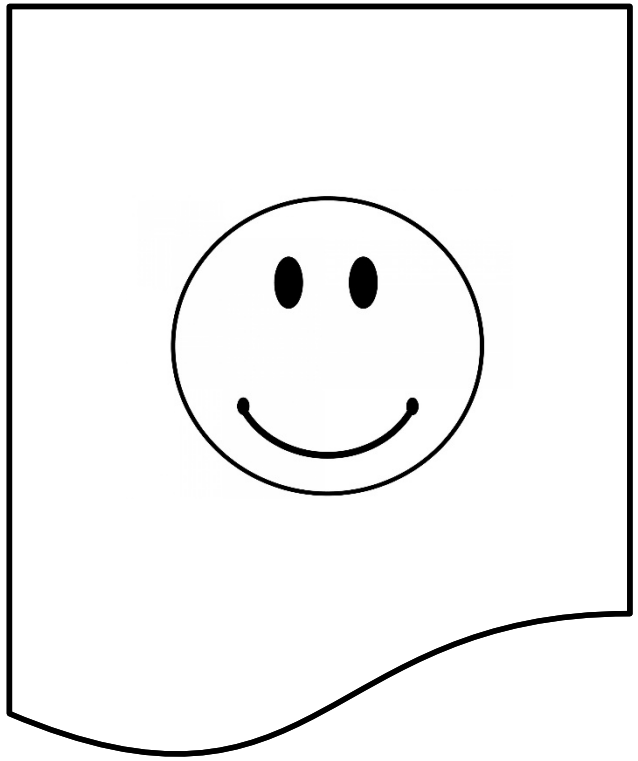
spark-packages


```
wholeFile", true)
```

```
lingRatio", (0…1.0])
```

JDBC

--driver-class-path PATH/TO/JAR



JDBC

MS SQL

MySQL

PostgreSQL

...

Data Sources

http://bit.ly/2u3frhN

```
sources.RelationProvider

..sources.BaseRelation
```

Resources

- Introducing Apache Spark Datasets: Databricks
  - databricks.com/blog/2016/01/04/introducing-apache-spark-datasets
- A Tale of Three Apache Spark APIs: RDDs, DataFrames, and Datasets: Databricks
  - databricks.com/blog/2016/07/14/a-tale-of-three-apache-spark-apis-rdds-dataframes-and-datasets
- Mastering Apache Spark 2 - Encoders: Jacek Laskowski
  - jaceklaskowski.gitbooks.io/mastering-apache-spark/content/spark-sql-Encoder
- Structuring Apache Spark: Michael Armbrust
  - youtube.com/watch?v=1a4pgYzeFwE
- Cassandra
  - app.pluralsight.com/library/courses/cassandra-developers
  - Connector
    - github.com/datastax/spark-cassandra-connector
      - /blob/master/doc
        - /14_data_frames.md
        - /15_python.md
        - /7_java_api.md
    - datastax.com/dev/blog/accessing-cassandra-from-spark-in-java

Summary

Datasets for the win!

Encoders

Beyond native datasources