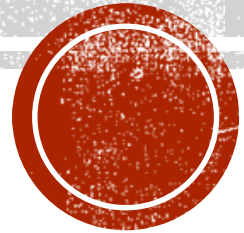
DATAFRAME VS DATASET

.



WHAT IS DATAFRAME AND DATASET

- DataFrame is
  - Immutable Table
  - Collection of Generic Row Objects
  - Not Type Safe
  - Introduced to simplify processing on RDD
- Dataset is
  - Collection of strongly typed objects

UNDERSTAND TYPE SAFETY

Type Of Error	SQL	DataFrame	Dataset
Syntax Error	Runtime	Compile Time	Compile time
Analysis Error	Runtime	Runtime	Compile Time

USE DATASET WHEN

- You Need type Safety

USE DATASET WHEN

- You Need type Safety
- You need SQL + Functional Programming Constructs like map(), mapPartition(), aggregate, reduce etc.

USE DATASET WHEN

- You Need type Safety
- You need SQL + Functional Programming Constructs like map(), mapPartition(), aggregate, reduce etc.
- You need lambda functions

USE DATASET WHEN

- You Need type Safety
- You need SQL + Functional Programming Constructs like map(), mapPartition(), aggregate, reduce etc.
- You need lambda functions
- Want Catalyst Optimization, and Tungsten's efficient code generation

USE DATASET WHEN

- You Need type Safety
- You need SQL + Functional Programming Constructs like map(), mapPartition(), aggregate, reduce etc.
- You need lambda functions
- Want Catalyst Optimization, and Tungsten's efficient code generation
- Want more efficient memory usage using Encoders

USE DATAFRAME WHEN

- You use Python or R

USE DATAFRAME WHEN

- You use Python or R
- Don’t need strong type safety and most of operations will be defined using SQL

USE DATAFRAME WHEN

- You use Python or R
- Don’t need strong type safety and most of operations will be defined using SQL
- No need for lambda expressions