

Lesson 1.2: Spark DataFrames

DISTRIBUTED COMPUTING WITH SPARK SQL

Spark DataFrames



Brooke Wenig
Machine Learning Practice Lead
Databricks

UC DAVIS
Continuing and Professional Education

Slide 2: Welcome Back!

Welcome Back!

Basic Spark architecture: driver and executors

Spark DataFrame

Brief history of Spark

Slide 3: Learning Objective

Learning Objective

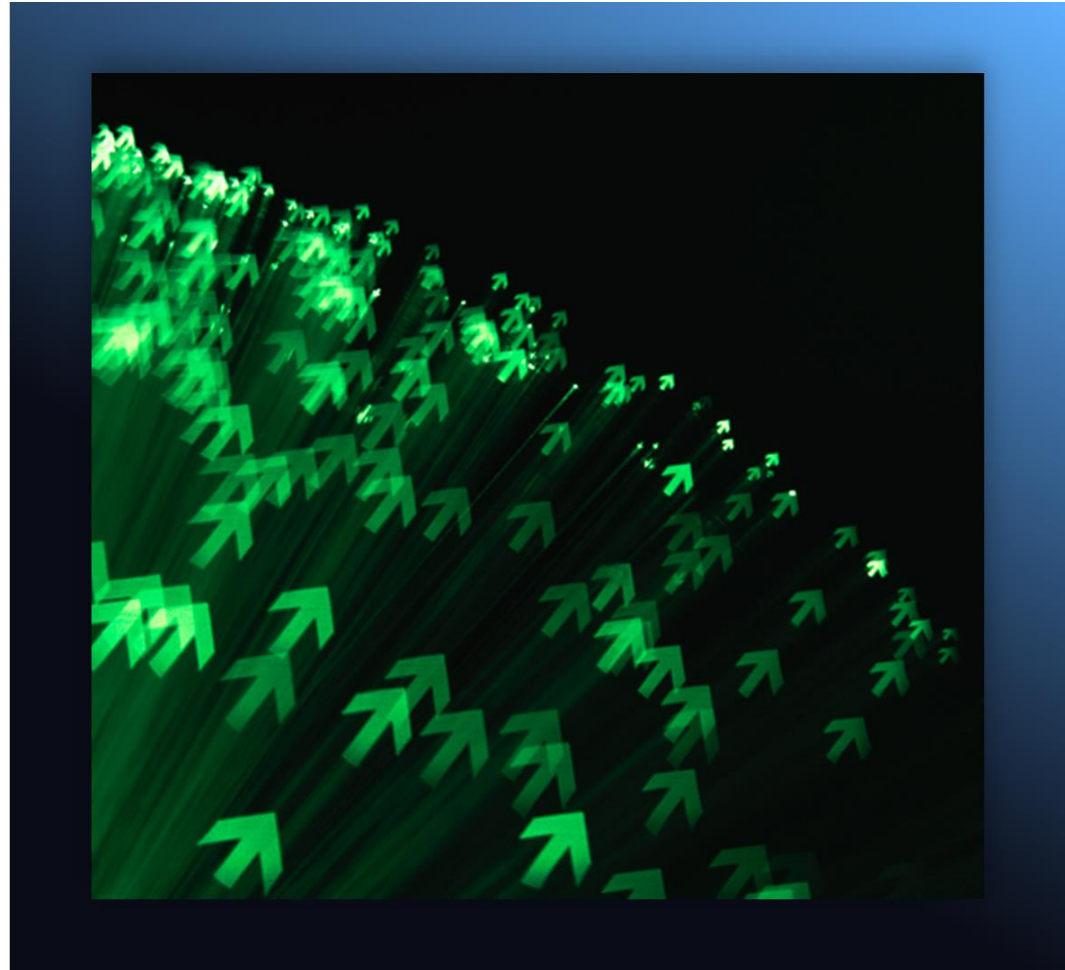
Explain the difference between RDD and DataFrame API within Spark

Slide 4: Evolution of Spark

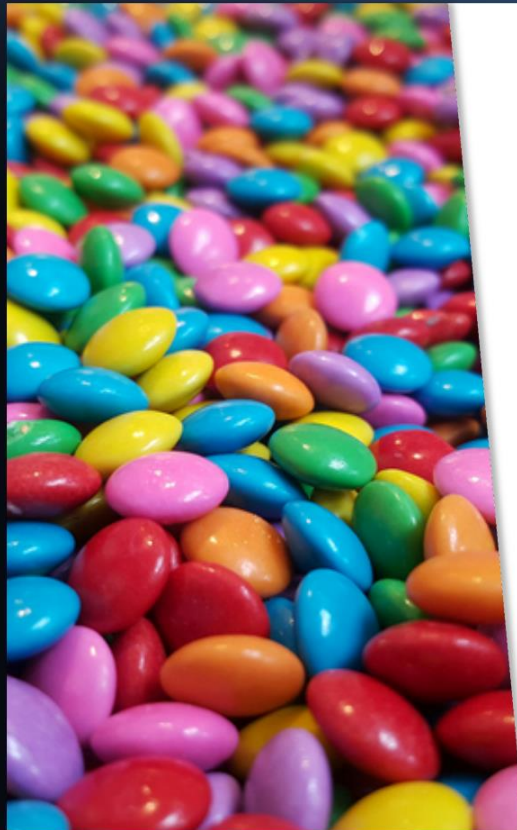
Evolution of Spark

Resilient Distributed Dataset (RDD)

Spark 1.3 – DataFrame
– more functionality
and optimizations



Slide 5: Resilient Distributed Databases (RDD)



Resilient Distributed Datasets (RDD)

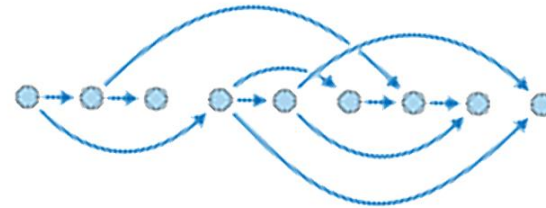
Resilient: fault-tolerant

If you lose a worker,
only recompute work
that worker was
responsible for



Slide 6: Resilient Distributed Databases (RDD)

Directed Acyclic Graph



Series of transformations to apply to your data

However, you cannot change any of the transformations that came before you in this graph

Slide 7: Distributed Dataset

Distributed Dataset

Computed across multiple nodes

Results are aggregated by the driver



Slide 8: Why This Matters

Why This Matters

DataFrame inherits
RDD properties
(resilient + distributed)
plus metadata

Metadata

Number of columns

Data types

```
1 SELECT `Unit ID`, `Call Type`  
2 FROM fireCalls  
3 LIMIT 10
```

► (1) Spark Jobs

Unit ID	Call Type
E08	Medical Incident
M18	Medical Incident
M36	Medical Incident
E12	Structure Fire
M14	Medical Incident
M43	Medical Incident
E10	Alarms

Slide 9: DataFrame

DataFrame

Highly optimized
beyond RDDs

Always use DataFrames
where possible

Spark SQL commands
execute against
DataFrames

```
1 SELECT `Unit ID`, `Call Type`  
2 FROM fireCalls  
3 LIMIT 10
```

► (1) Spark Jobs

Unit ID ▼	Call Type ▼
E08	Medical Incident
M18	Medical Incident
M36	Medical Incident
E12	Structure Fire
M14	Medical Incident
M43	Medical Incident
E10	Alarms

Slide 10: What DataFrames Are Not

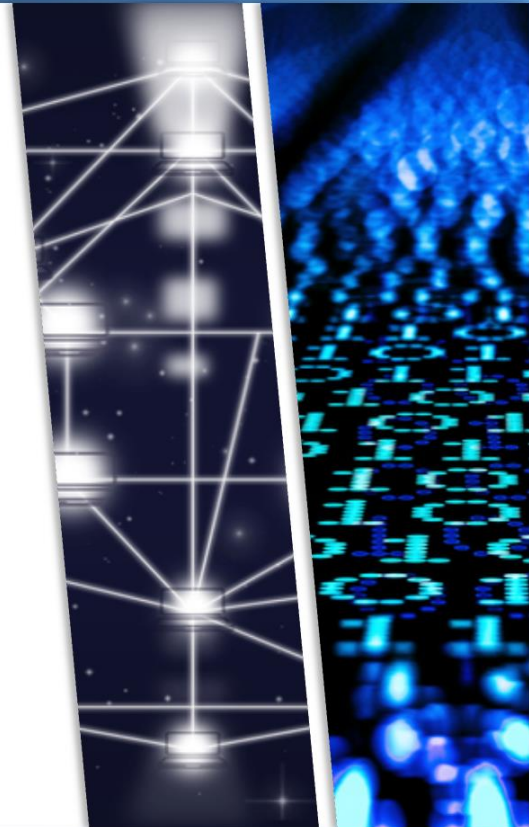
What DataFrames Are Not

Spark is not a database

It's a compute engine that can read
from databases

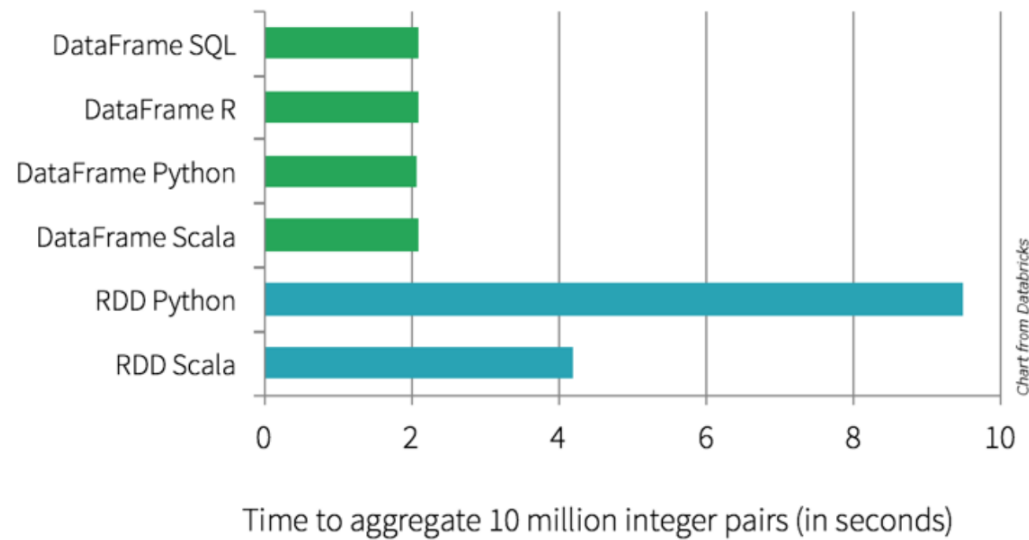
Data is ephemeral

DataFrames are not SQL tables,
Excel files, etc.



Slide 11: DataFrame > RDD

DataFrame > RDD

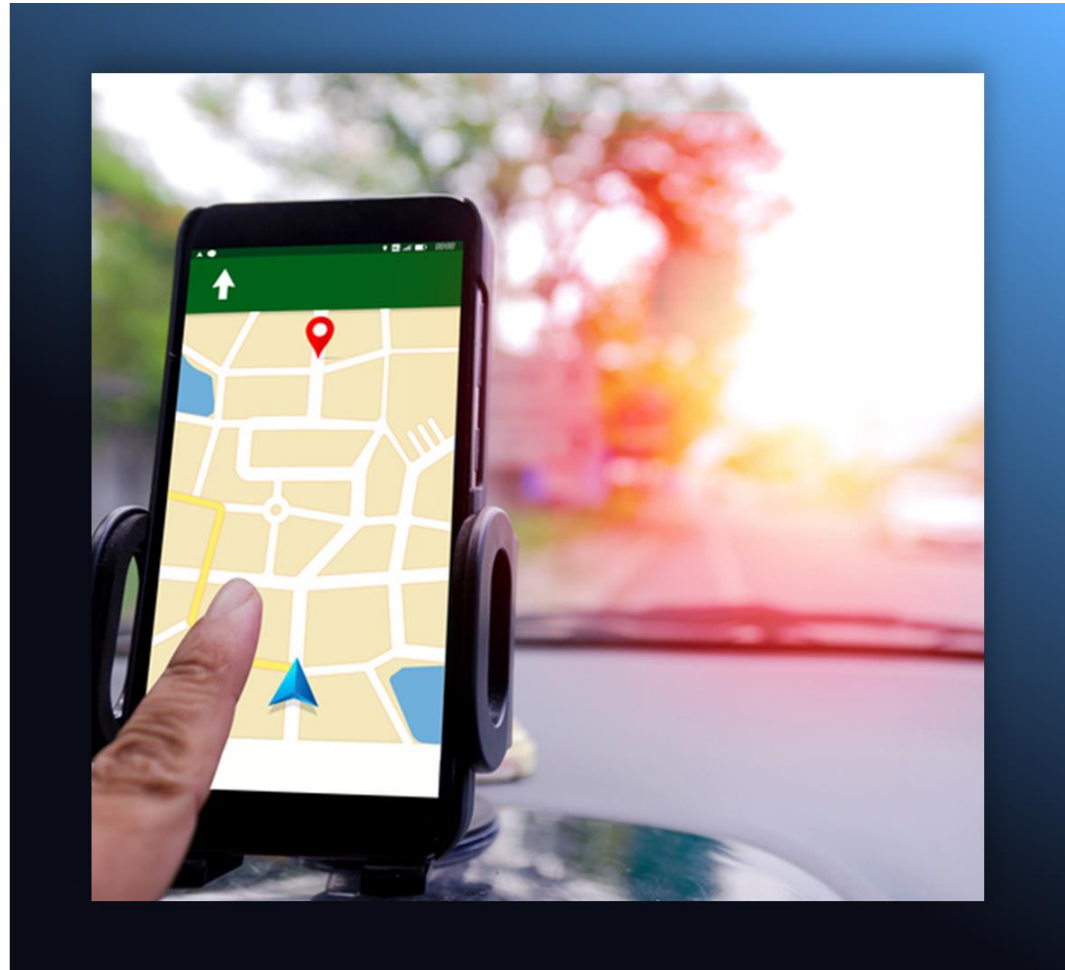


Slide 12: Spark DataFrame Execution

Spark DataFrame Execution: Catalyst

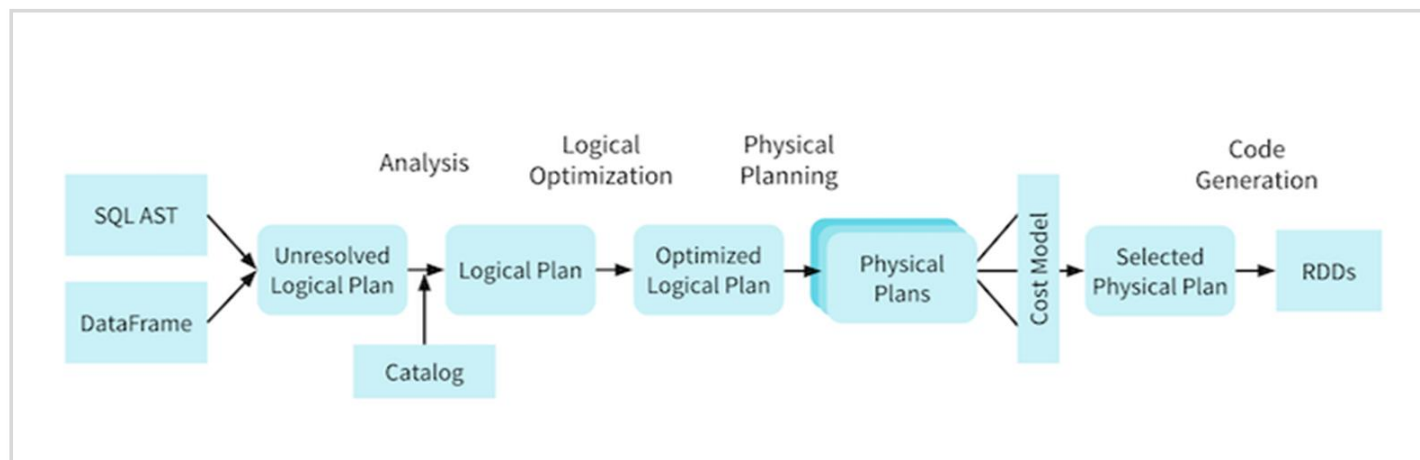
Google Maps Analogy:

Specify **what** you want to do, not how you want to do it



Spark DataFrame Execution

1. Unresolved logical plan before look-up in data catalog
2. Then Catalyst resolves them and creates a logical plan



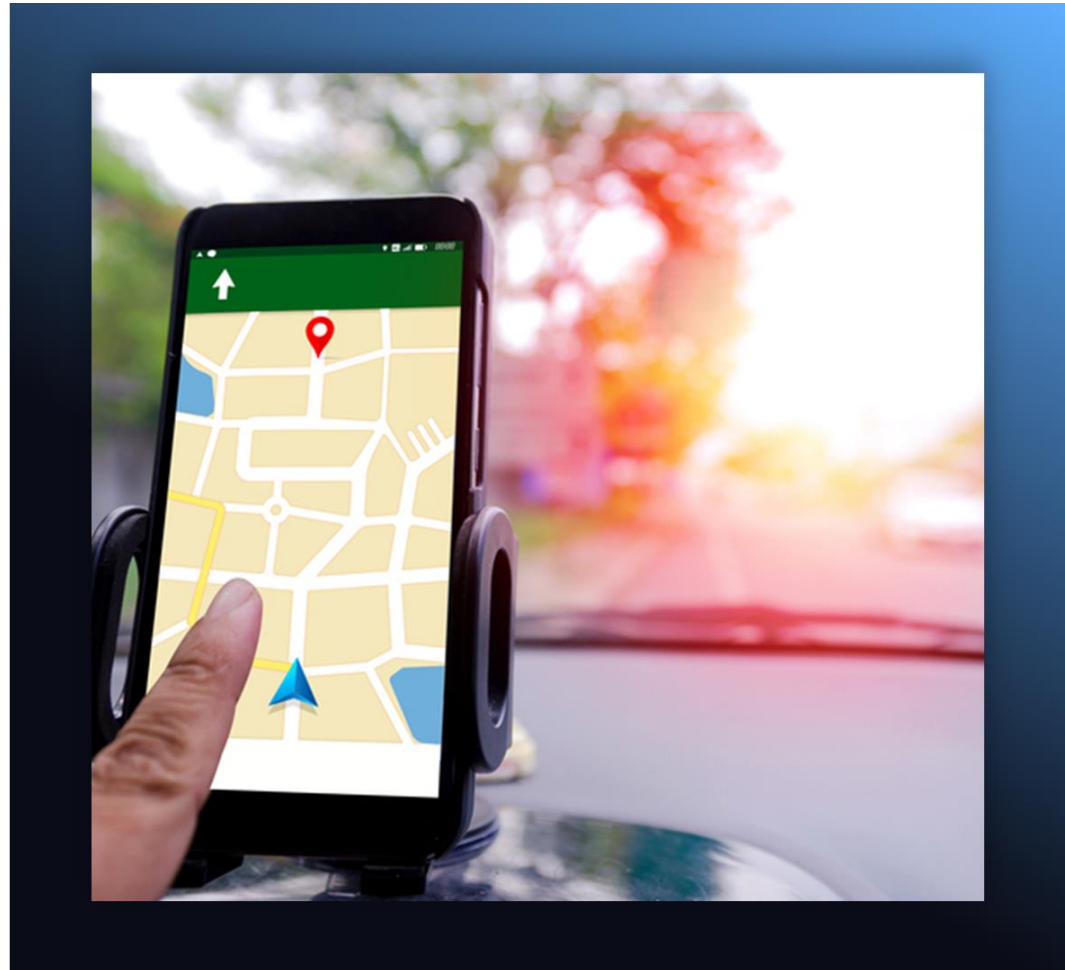
Slide 14: Spark DataFrame Execution

Spark DataFrame Execution

Maps Analogy:

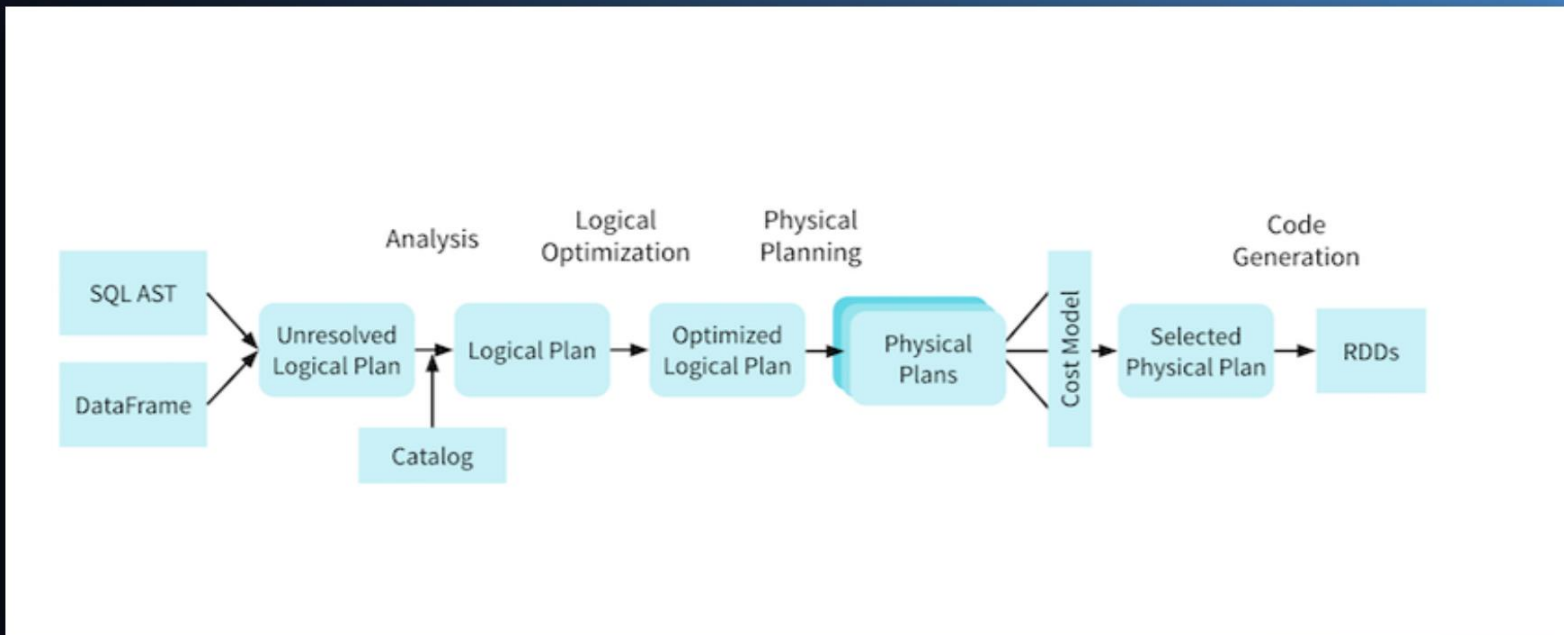
There are many ways
to get from point A to
point B

What you want to do, not
how you want to do it



Slide 15: Using Catalyst in Spark SQL

Using Catalyst in Spark SQL



Slide 16: Coming Up



Coming Up

Look at Tungsten which illustrates why
DataFrames are more performant than
RDDs