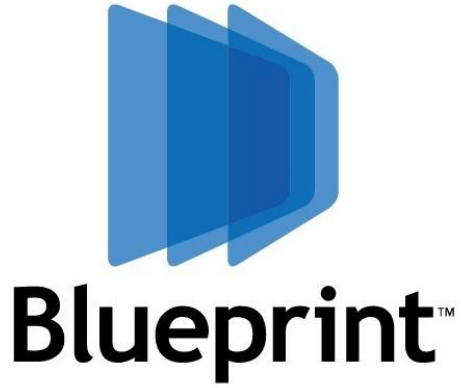
Intro to PySpark Workshop

Garren Staubli
Sr. Data Engineer
@gstaubli



#PySparkWorkshop Resources: garrens.com/pyspark124

Do I know what I'm talking about?

Working with Spark since 2015

- Batch analytics in Spark + Hive, Pig and Hadoop MapReduce
- Real-time big data reporting using Spark/Impala/CDH
- Spark Structured Streaming + ML apps for real-time decision making

50+ answers on stackoverflow for Spark

Main Points

- Apache Spark
- Sample App Walkthrough
- Interactive Azure Jupyter Notebook
- Python-specific Spark advice
- Resources to continue learning

About Apache Spark

Apache Spark™ is a fast and general engine for large-scale data processing.

Lightning-fast cluster computing

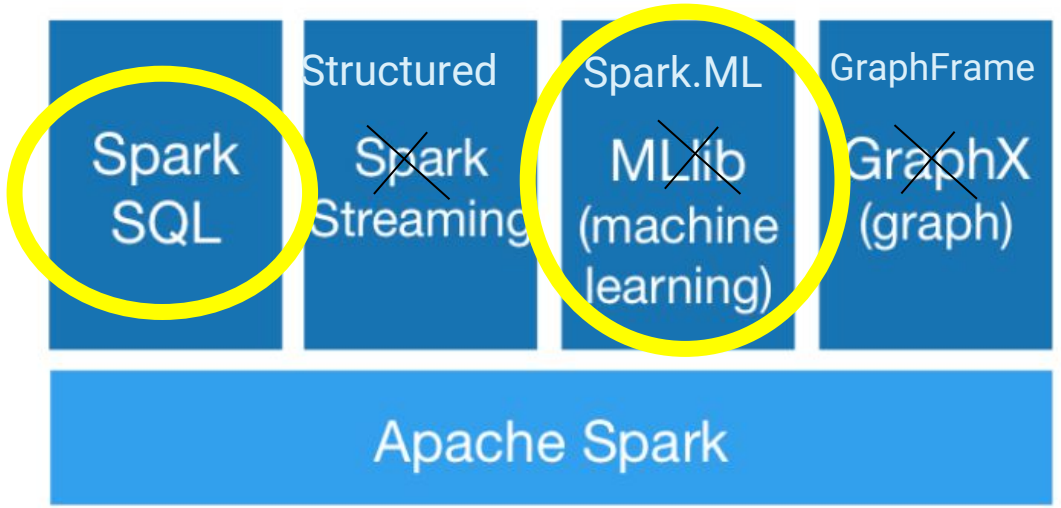
Lazily Evaluated

- Transforms vs Actions

Immutable

Transformations *(lazy)*	Actions
orderBy	show
filter	count
groupBy	take
select	collect
drop	save
join	

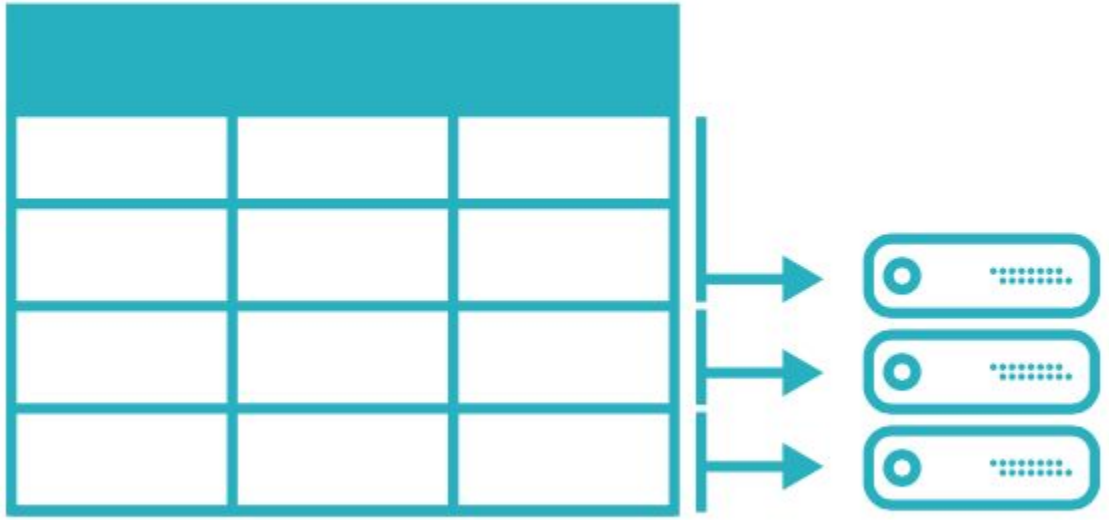
Transformations contribute to a query plan,
but nothing is executed until an action is called



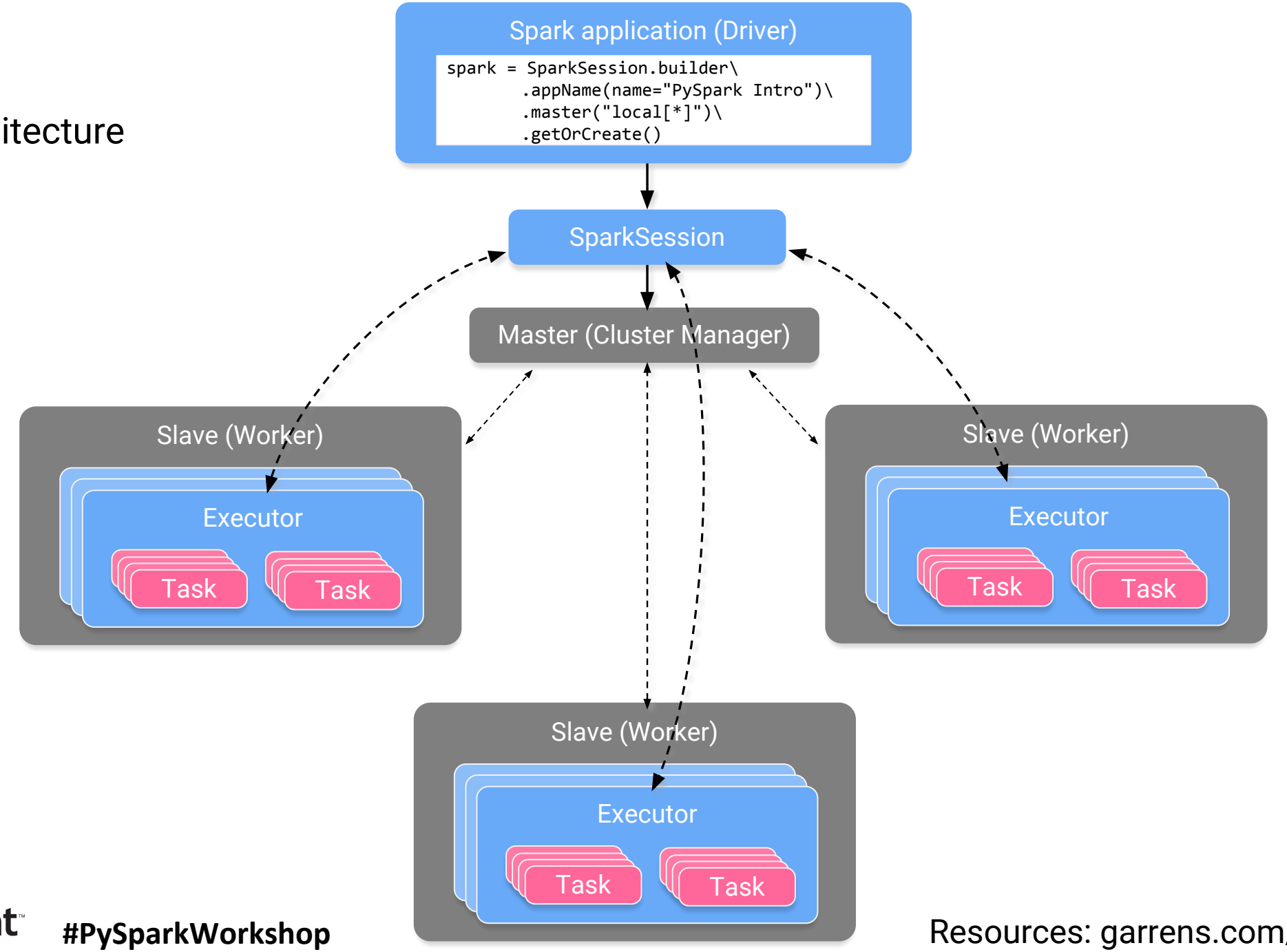
About Apache Spark

Spreadsheet on a single machine

Table or DataFrame partitioned across servers in a data center



detailed architecture



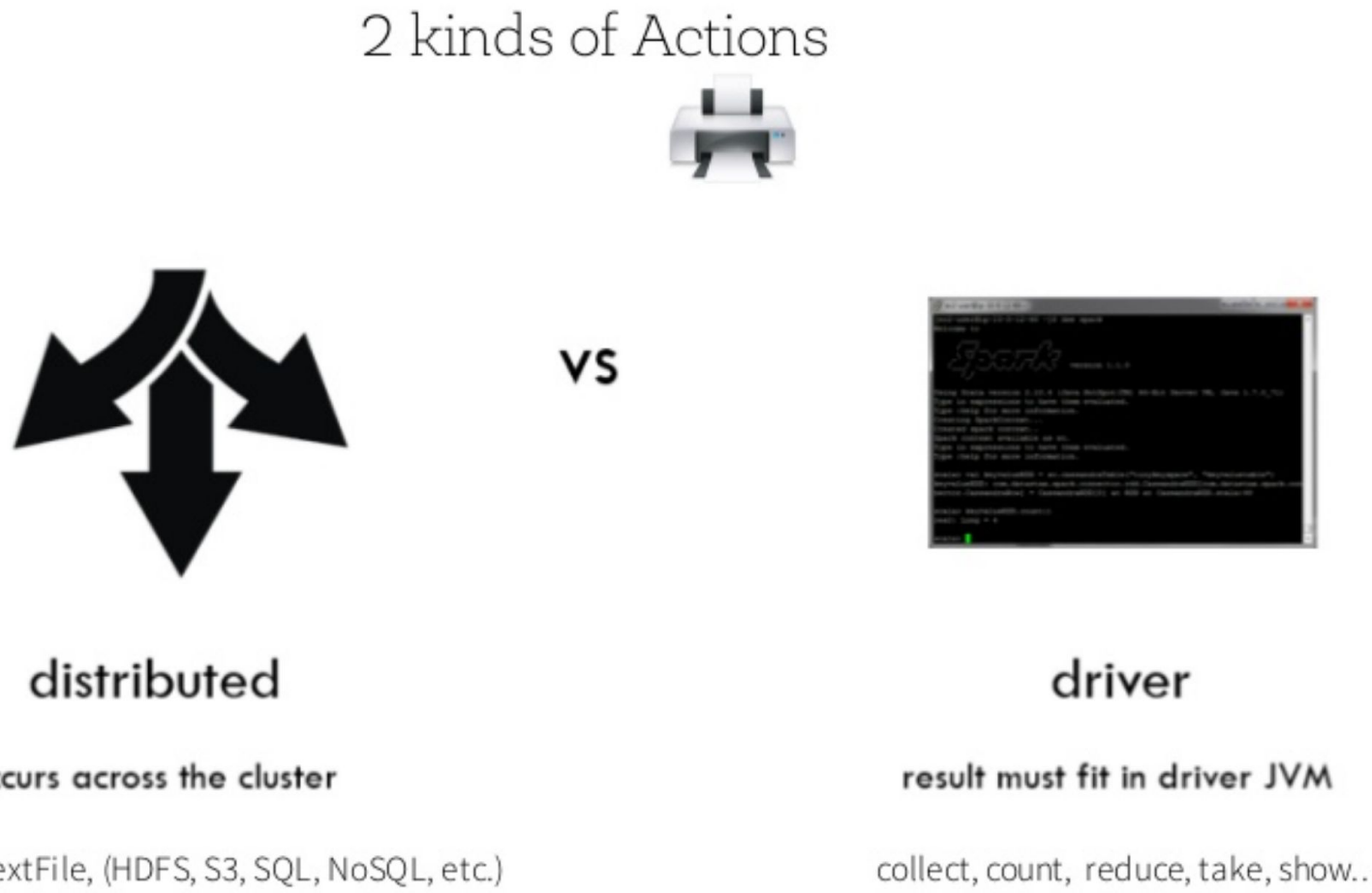
About Apache Spark | Spark SQL

SQL

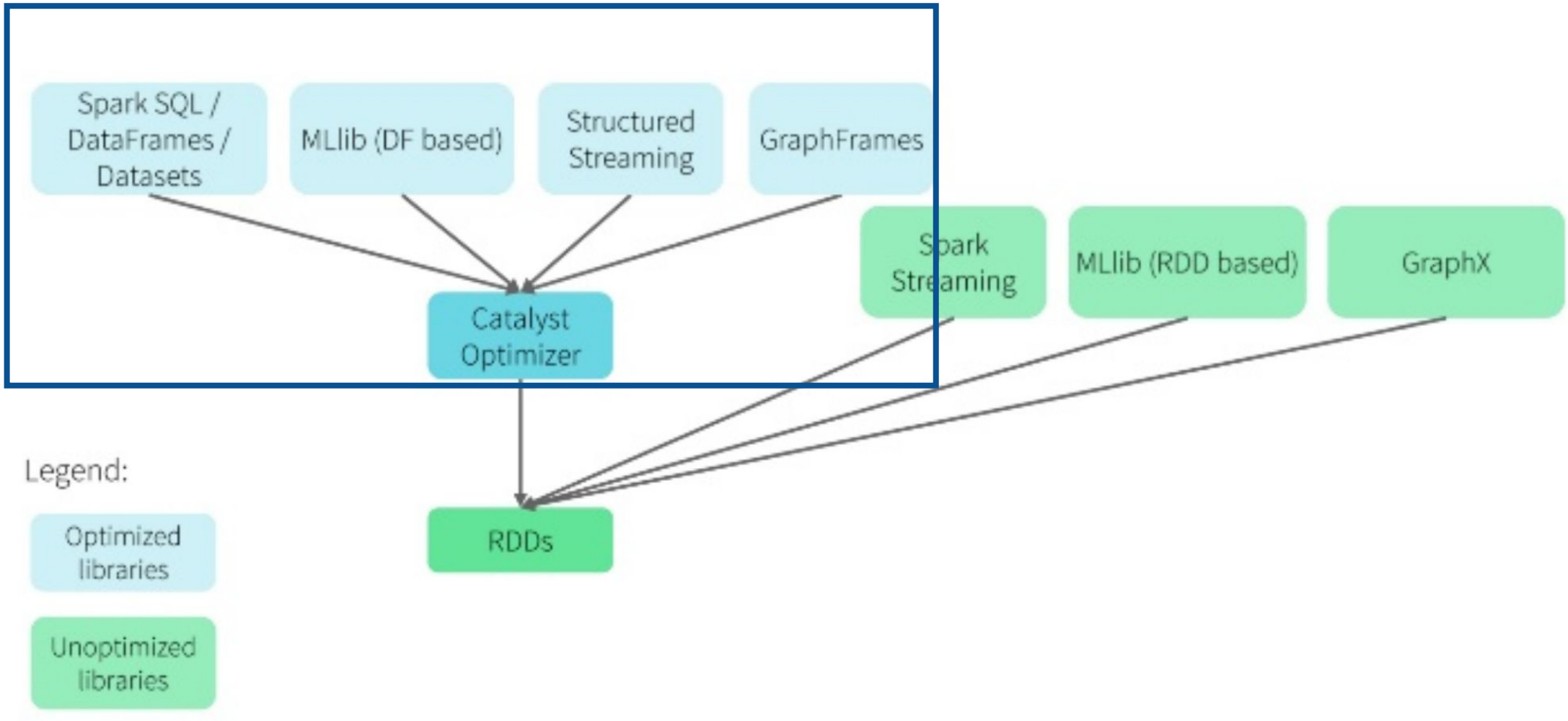
is **not** about SQL

is about **more** than SQL

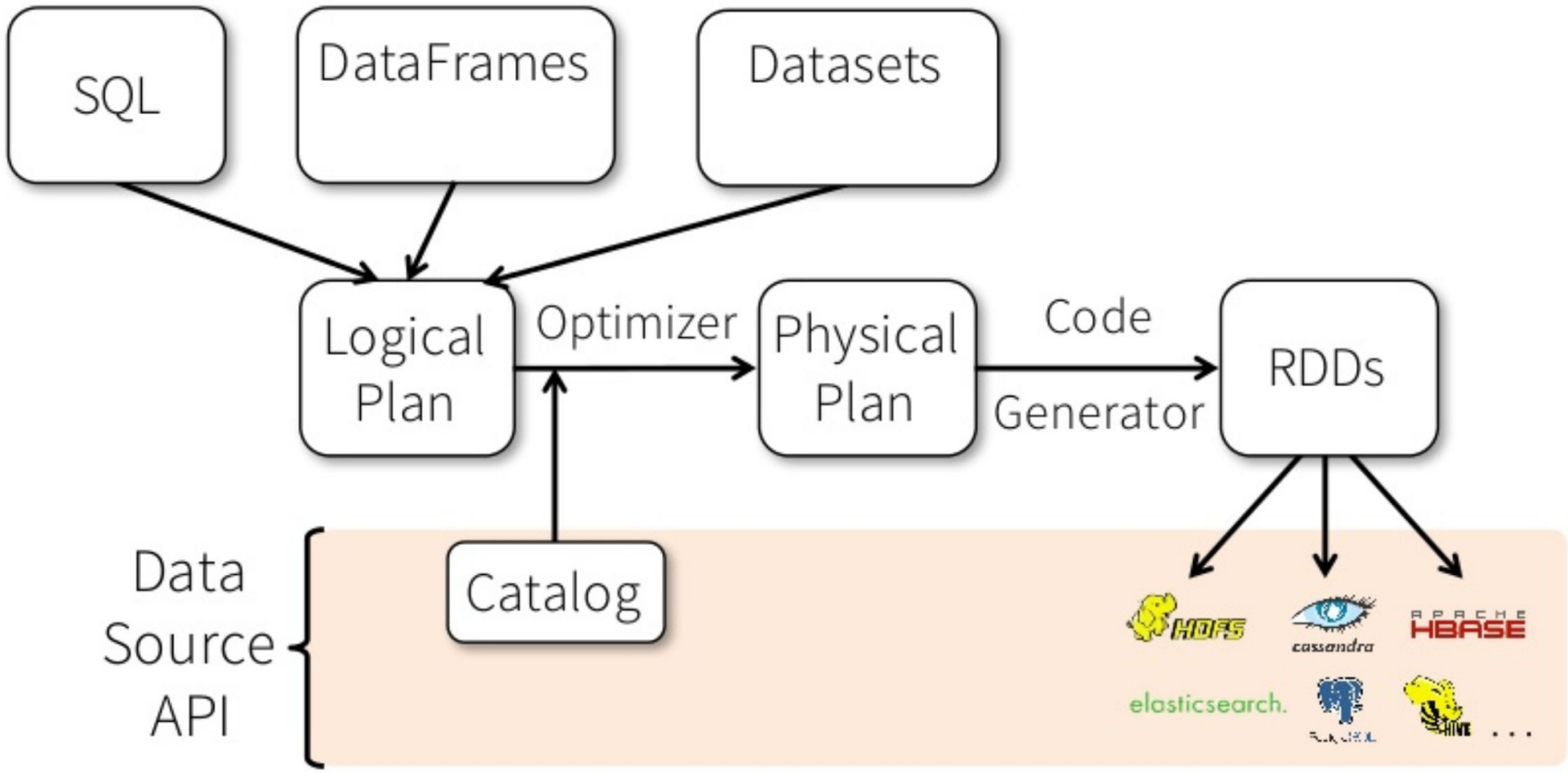
About Apache Spark | 2 Kinds of Actions



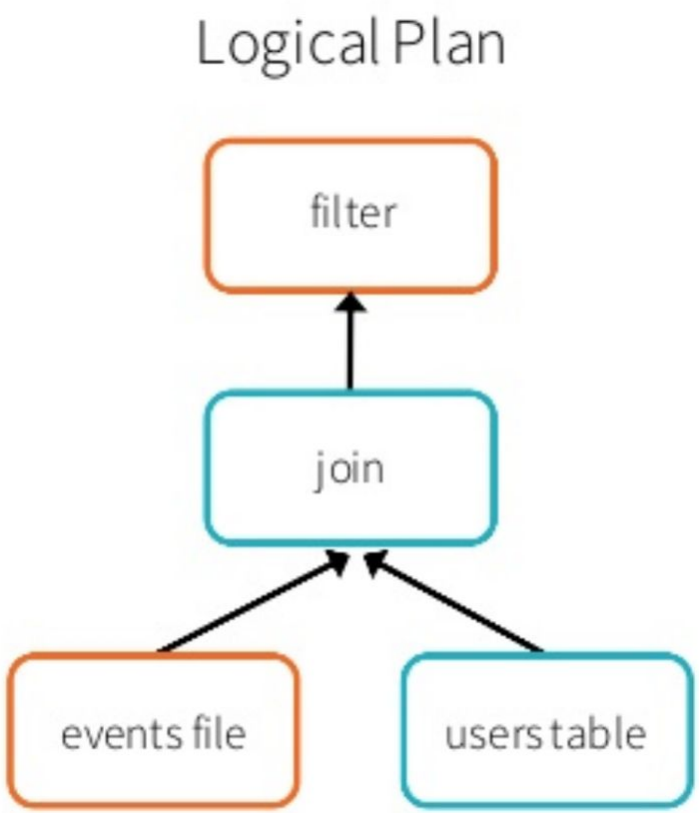
About Apache Spark | Modern vs Legacy

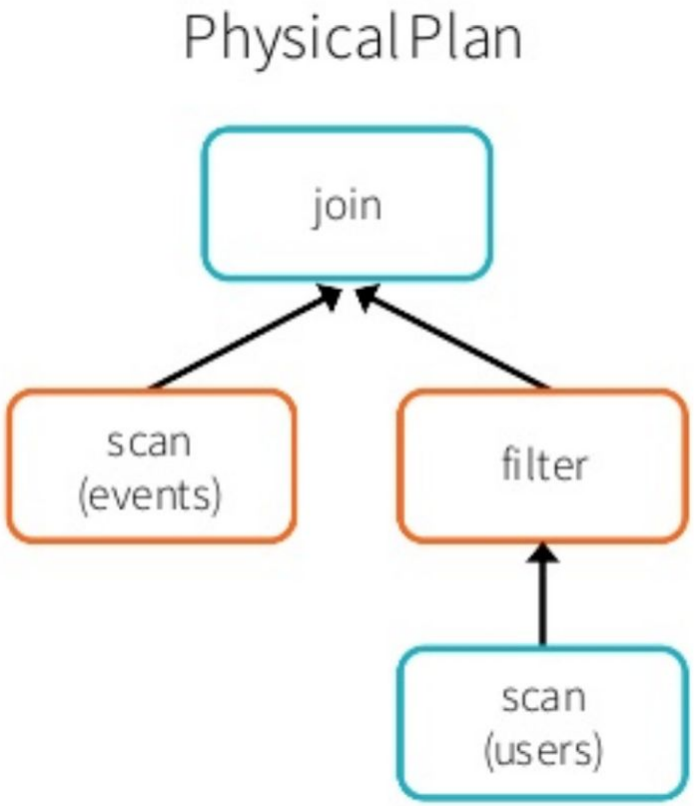


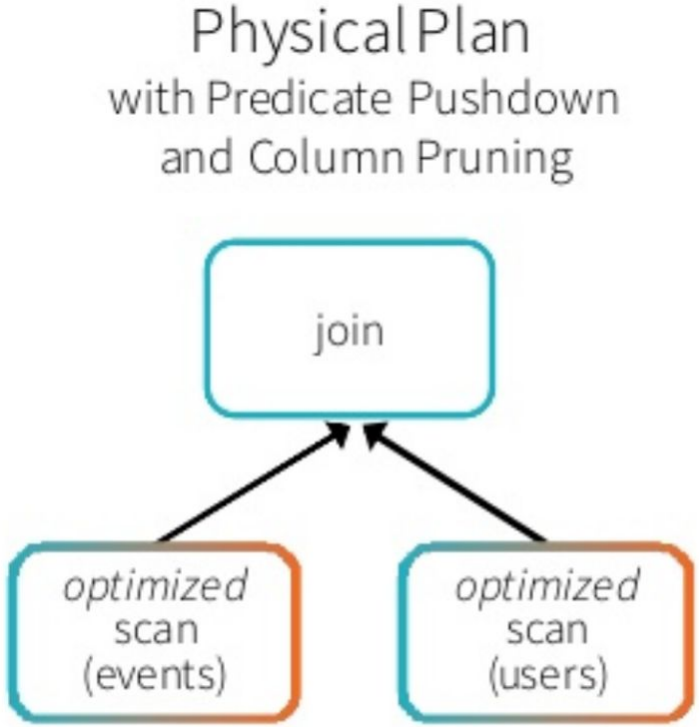
About Apache Spark | Modern Optimization



About Apache Spark | Planning







Walkthrough | Create Spark Session

```
from pyspark.sql import SparkSession
spark = SparkSession.builder\
    .appName(name="PySpark Intro")\
    .master("local[*]")\
    .getOrCreate()
```

Deploy modes:
Local, standalone, YARN,
Mesos and Kubernetes

SparkSession - in-memory
SparkContext

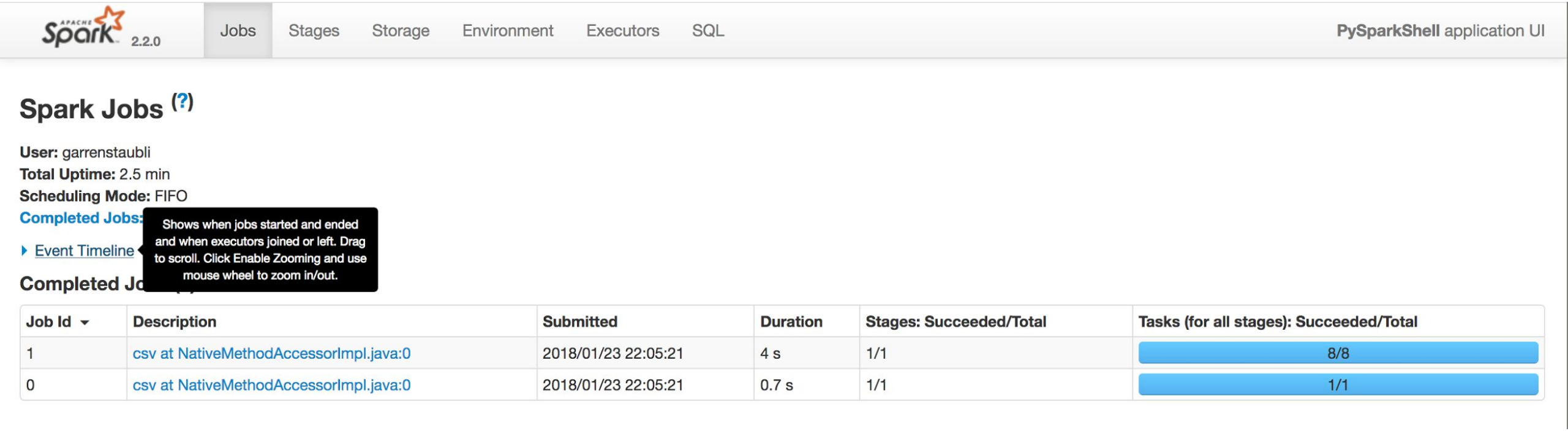
Spark UI
Version
v2.2.1
Master
local[*]
AppName
PySpark Intro

Walkthrough | Read CSV into DataFrame

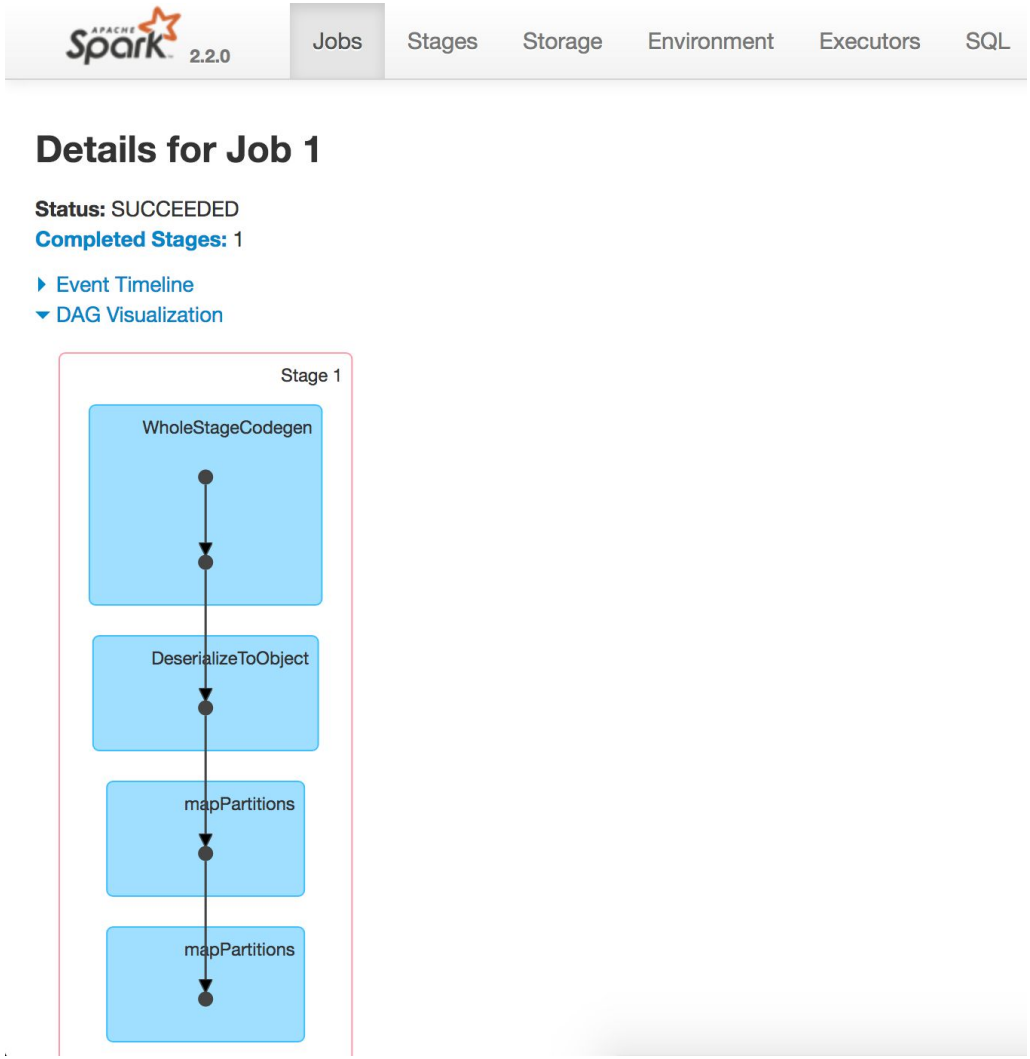
```
green_trips = spark.read\
            .option("header", "true")\
.option("inferSchema", "true")\
.csv("green_tripdata_2017-06.csv")
```

Forces eager evaluation;
default is **false**

Walkthrough | Behind the Scenes: UI



Walkthrough | Behind the Scenes: UI



Walkthrough | DataFrame Schema

```
green_trips.printSchema()
```

Eagerly evaluated (inferSchema = true)

```
root
 |-- VendorID: integer (nullable = true)
 |-- lpep_pickup_datetime: timestamp (nullable = true)
 |-- lpep_dropoff_datetime: timestamp (nullable = true)
 |-- store_and_fwd_flag: string (nullable = true)
 |-- RatecodeID: integer (nullable = true)
 |-- PULocationID: integer (nullable = true)
 |-- DOLocationID: integer (nullable = true)
 |-- passenger_count: integer (nullable = true)
 |-- trip_distance: double (nullable = true)
 |-- fare_amount: double (nullable = true)
 |-- extra: double (nullable = true)
 |-- mta_tax: double (nullable = true)
 |-- tip_amount: double (nullable = true)
 |-- tolls_amount: double (nullable = true)
 |-- ehail_fee: string (nullable = true)
 |-- improvement_surcharge: double (nullable = true)
 |-- total_amount: double (nullable = true)
 |-- payment_type: integer (nullable = true)
 |-- trip_type: integer (nullable = true)
```

Lazily evaluated (inferSchema = false)

```
root
 |-- VendorID: string (nullable = true)
 |-- lpep_pickup_datetime: string (nullable = true)
 |-- lpep_dropoff_datetime: string (nullable = true)
 |-- store_and_fwd_flag: string (nullable = true)
 |-- RatecodeID: string (nullable = true)
 |-- PULocationID: string (nullable = true)
 |-- DOLocationID: string (nullable = true)
 |-- passenger_count: string (nullable = true)
 |-- trip_distance: string (nullable = true)
 |-- fare_amount: string (nullable = true)
 |-- extra: string (nullable = true)
 |-- mta_tax: string (nullable = true)
 |-- tip_amount: string (nullable = true)
 |-- tolls_amount: string (nullable = true)
 |-- ehail_fee: string (nullable = true)
 |-- improvement_surcharge: string (nullable = true)
 |-- total_amount: string (nullable = true)
 |-- payment_type: string (nullable = true)
 |-- trip_type: string (nullable = true)
```

Blueprint™

CONSULTING SERVICES

You guessed it…
We're hiring!

Seattle Business 100 WASHINGTON'S BEST COMPANIES TO WORK FOR
- 2014
- 2015
- 2016

100 FASTEST-GROWING PRIVATE COMPANIES
- 2015 #1
- 2016 #1

BEST WASHINGTON'S BEST WORKPLACES
- 2016

PUGET SOUND BUSINESS JOURNAL 50 EASTSIDE FASTEST-GROWING PRIVATE COMPANIES
- 2014
- 2015 #1
- 2016 #1
- 2017 #4

Inc. 500
- 2015 #373
- 2016 #166
- 2017 #161

Resources: garrens.com/pyspark124