WELCOME TO APACHE SPARK

WHOAMI

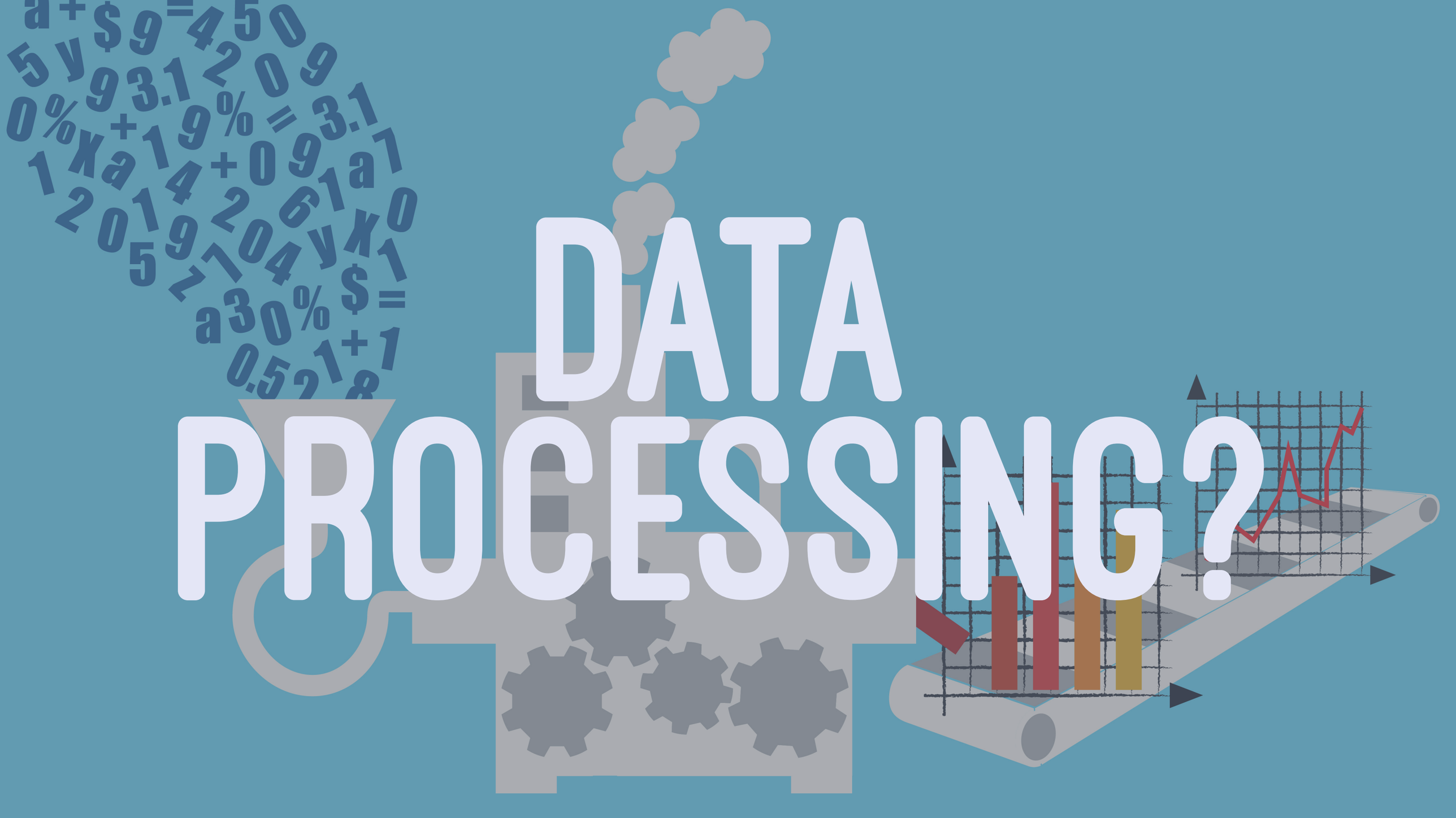
- RUBEN BERENGUEL (@BERENGUEL)
- PHD IN MATHEMATICS
- (BIG) DATA CONSULTANT
- SENIOR CRAFTER IN PYTHON, GO AND SCALA
- RIGHT NOW AT AFFECTV

WHOAMI

- CARLOS PEÑA (@CRAFTY_CODER)
  - SENIOR SOFTWARE ENGINEER
    - (BIG) DATA CONSULTANT
- CRAFTING CODE USING SCALA, JAVA AND PYTHON
  - LEARNING ML & DEEP LEARNING

WHAT IS SPARK?

- DISTRIBUTED COMPUTATION FRAMEWORK
  - OPEN SOURCE
  - SQL FRIENDLY
- FOCUS ON DATA PROCESSING


DATA
PROCESSING?

TRANSFORM DATA

EXTRACT INFORMATION

WHY SPARK?

SCALES
HORIZONTALLY
AND
VERTICALLY

LETS YOU WRITE
COMPLEX OPERATIONS
EASILY AND SAFELY

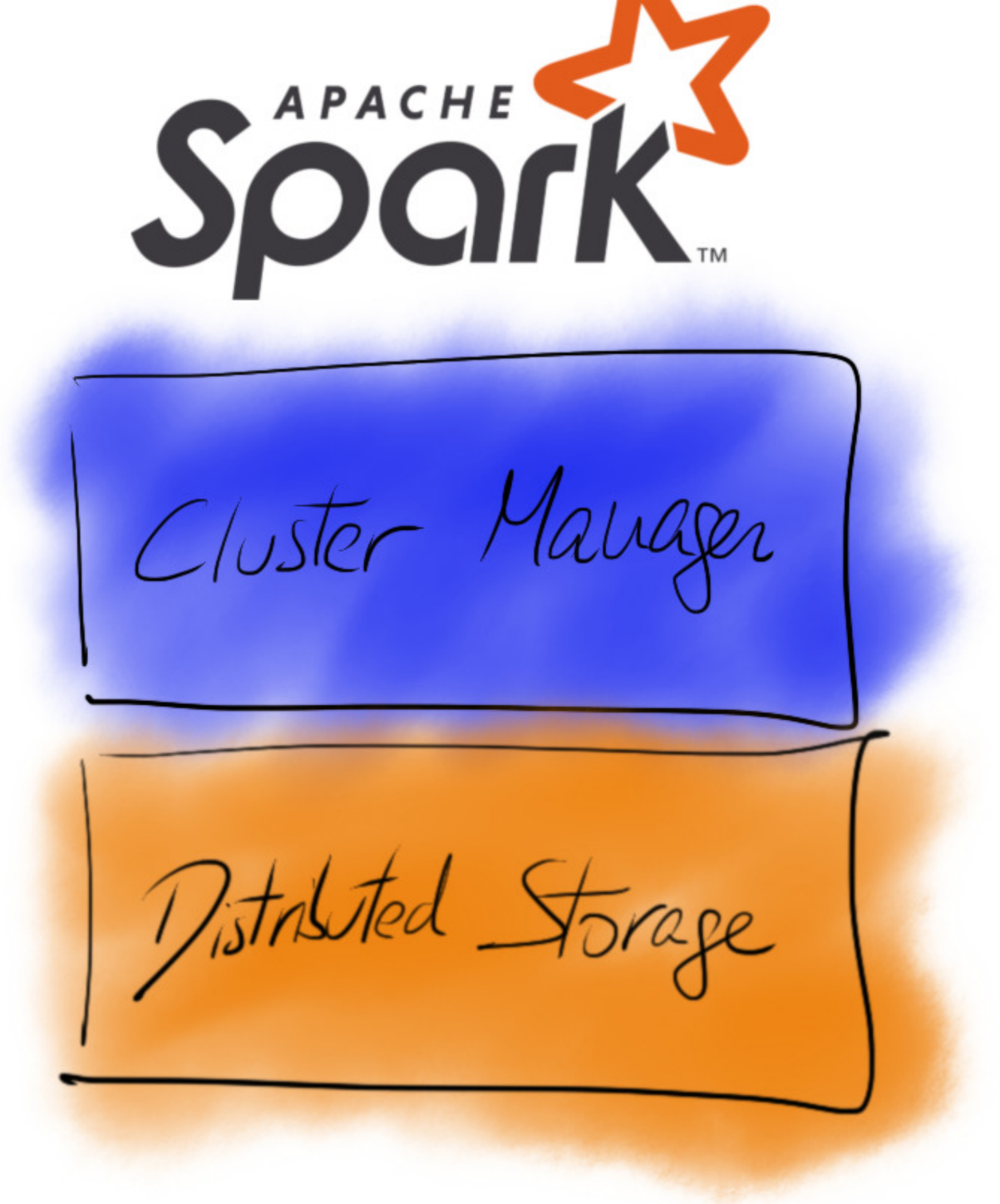

INTERACTIVE

SUPPORTS SCALA. JAVA. PYTHON & R

HOW DOES SPARK WORK?

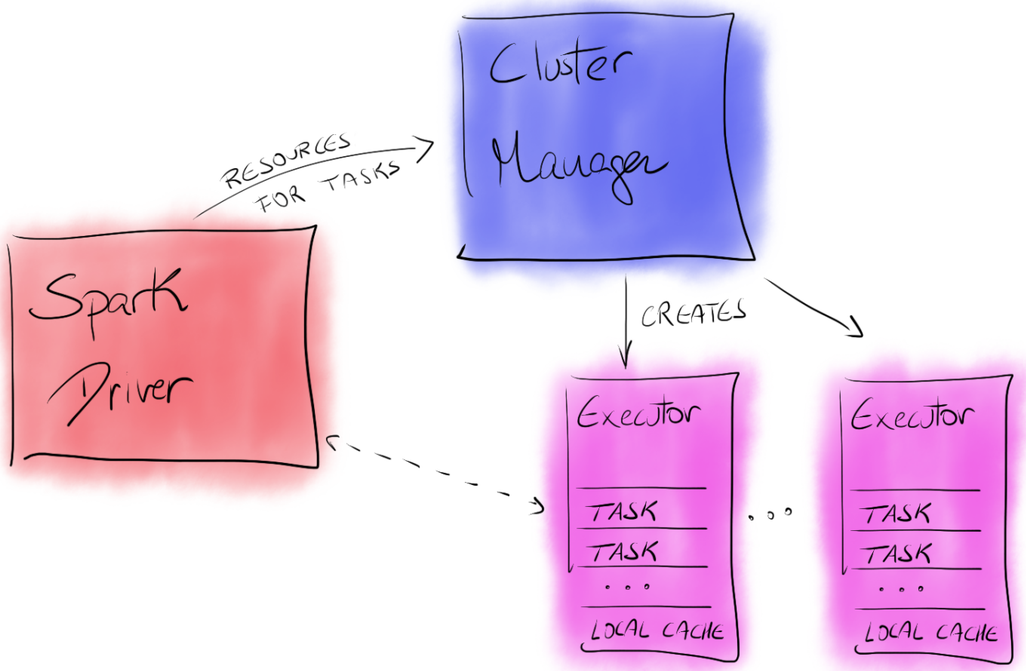
SPARK USUALLY SITS ON TOP OF A CLUSTER MANAGER

AND A DISTRIBUTED STORAGE



A SPARK PROGRAM RUNS IN THE DRIVER

THE DRIVER REQUESTS RESOURCES TO THE CLUSTER MANAGER TO RUN TASKS



TWO KIND OF OPERATIONS:

TRANSFORMATIONS

&

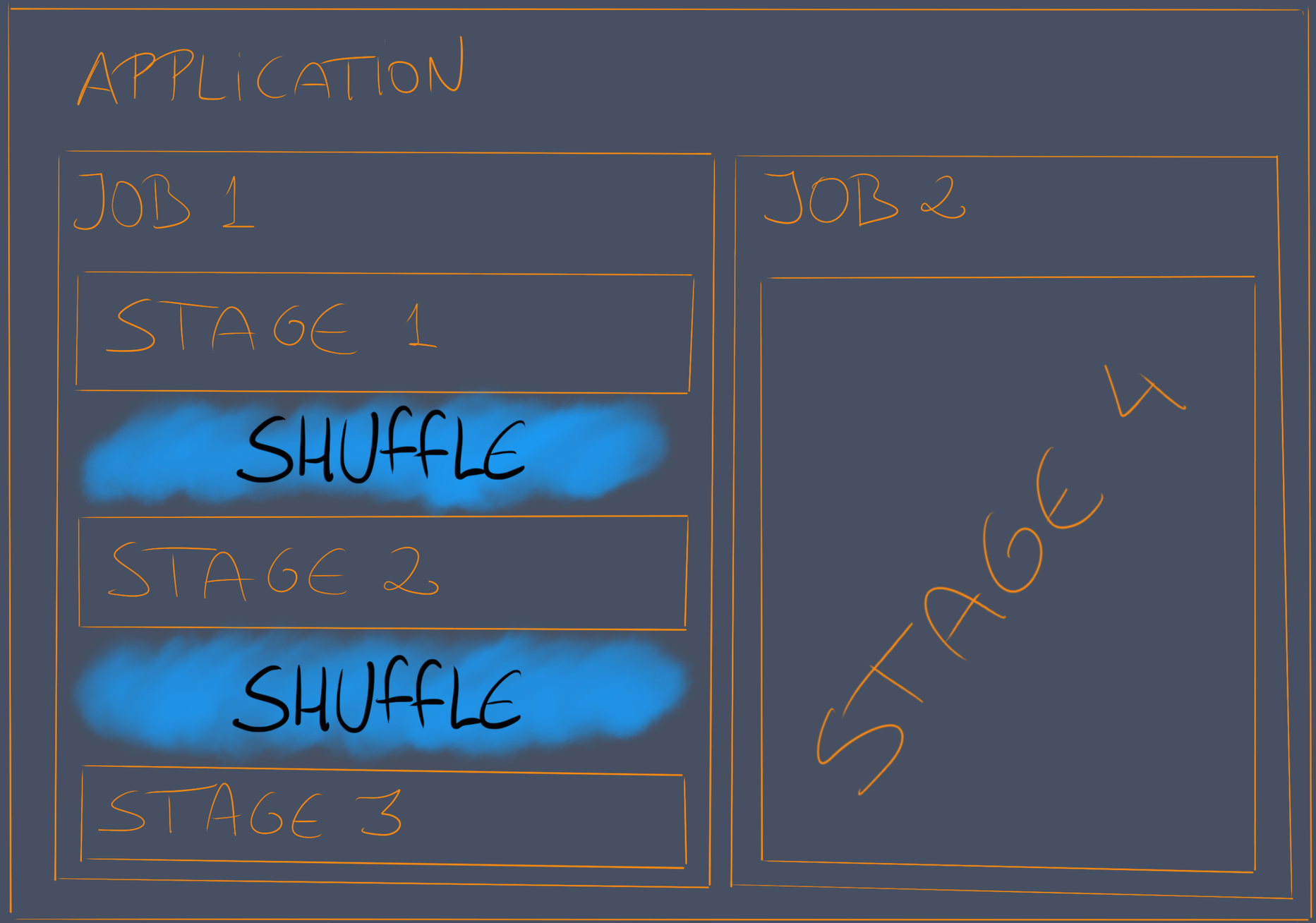
ACTIONS

TRANSFORMATIONS: RESHAPE THE DATA. THEY ARE PART OF STAGES

- FILTER
- MAP
- JOIN

ACTIONS: RETURN A RESULT. THEY DEFINE JOBS

- COUNT
- SHOW
- WRITE


APPLICATION
JOB 1
STAGE 1
SHUFFLE
STAGE 2
SHUFFLE
STAGE 3
JOB 2
STAGE 4

EXAMPLE TIME!

WE WILL USE A KAGGLE DATASET WITH BASKETBALL SHOTS FROM THE 2016 NBA SEASON

W	SHOT_NUMBER	GAME_CLOCK	SHOT_CLOCK	DRIBBLES	TOUCH_TIME	SHOT_DIST	PTS_TYPE	CLOSEST_DEFENDER	CLOSE_DEF_DIST	FGM	PTS	player_name
L	9	1:45	18.4	1	2.2	2.3	2	Withey, Jeff	0.0	0	0	kyrie irving
L	9	4:29	10.0	0	0.7	8.6	2	Jefferson, Al	2.5	0	0	nikola vucevic
L	12	0:38	19.7	0	0.8	24.8	3	Gasol, Marc	6.8	1	3	markieff morris
L	9	1:26	11.8	0	0.7	22.6	3	Bradley, Avery	5.5	0	0	pj tucker
W	5	10:29	9.6	0	-7.5	2.9	2	Roberson, Andre	0.3	0	0	deandre jordan
L	7	6:15	4.8	0	1.8	24.9	3	Iguodala, Andre	3.5	0	0	nicolas batum
L	14	5:44	19.0	5	4.9	24.5	3	Lopez, Brook	2.2	0	0	jamal crawford
W	5	11:14	19.3	6	4.9	15.9	2	Vucevic, Nikola	4.4	1	2	gerald henderson
W	14	5:43	5.9	0	1.5	4.6	2	Ibaka, Serge	1.6	1	2	anthony davis
W	4	2:33	14.7	0	0.8	24.3	3	Thompson, Hollis	5.8	1	3	terrence ross
L	8	4:46	20.2	0	0.8	25.0	3	Ellis, Monta	9.6	0	0	evan fournier
L	1	9:19	7.1	1	1.8	24.7	3	Williams, Mo	5.7	1	3	ty lawson
W	3	9:01	13.7	3	5.7	7.1	2	Len, Alex	2.7	0	0	pau gasol
L	1	3:35	15.0	0	1.0	24.8	3	Jerebko, Jonas	6.5	0	0	derrick williams
L	8	3:11	9.1	6	4.4	14.5	2	Temple, Garrett	2.9	1	2	evan fournier
L	9	5:20	14.2	3	2.6	6.9	2	Teague, Jeff	3.6	0	0	jj redick
W	9	7:28	15.6	9	6.6	20.1	2	Rose, Derrick	4.2	0	0	jeff teague
L	11	2:22	13.0	0	3.1	23.5	2	Daniels, Troy	9.6	1	2	ben mclemore
W	2	8:08	16.9	6	5.4	4.2	2	Calderon, Jose	2.0	1	2	tyreke evans
L	2	8:09	20.9	0	0.6	4.1	2	Korver, Kyle	1.4	1	2	dwayne wade
W	9	4:24	12.9	11	11.7	24.5	3	Smart, Marcus	6.5	0	0	john wall
L	8	7:55	11.8	1	1.6	6.1	2	Early, Cleanthony	2.7	1	2	james johnson
L	2	0:48	14.1	0	0.7	4.2	2	Udrih, Beno	2.1	0	0	shabazz muhammad
L	5	10:01	19.5	4	4.3	5.9	2	Patterson, Patrick	2.6	0	0	kenneth faried
L	12	2:00	11.0	5	5.1	3.1	2	Holiday, Jrue	3.1	1	2	eric bledsoe
W	10	6:17	18.1	0	0.8	19.5	2	Hollins, Ryan	5.5	1	2	marreese speights
W	6	11:36	19.2	5	6.2	25.1	3	Bledsoe, Eric	8.5	1	3	chris paul
L	3	9:05	6.6	2	3.9	12.7	2	Bennett, Anthony	2.5	1	2	robert sacre
L	9	11:26	12.3	0	4.1	2.5	2	Asik, Omer	2.6	1	2	tim duncan
W	7	3:55	6.8	0	0.0	2.5	2	Lowry, Kyle	0.5	1	2	kevin love

INFORMATION WE WANT TO EXTRACT

- BEST SCORERS
- AND THEIR SHOOTING RANGE

SELECT

#	DIST	PTS	PLAYER
0	6.9	0	CARLOS BOOZER
1	6.2	2	STEPHEN CURRY
2	4.5	0	ANTHONY DAVIS
3	19.1	3	DIRK NOWITZKI
4	20.3	2	KOBE BRYANT

FILTER (1)

#	DIST	PTS	PLAYER
0	6.9	0	CARLOS BOOZER
1	6.2	2	STEPHEN CURRY
2	4.5	0	ANTHONY DAVIS
3	19.1	3	DIRK NOWITZKI
4	20.3	2	KOBE BRYANT

FILTER (2)

DIST	PTS	PLAYER
6.2	2	STEPHEN CURRY
19.1	3	DIRK NOWITZKI
20.3	2	KOBE BRYANT

GROUP BY PLAYER

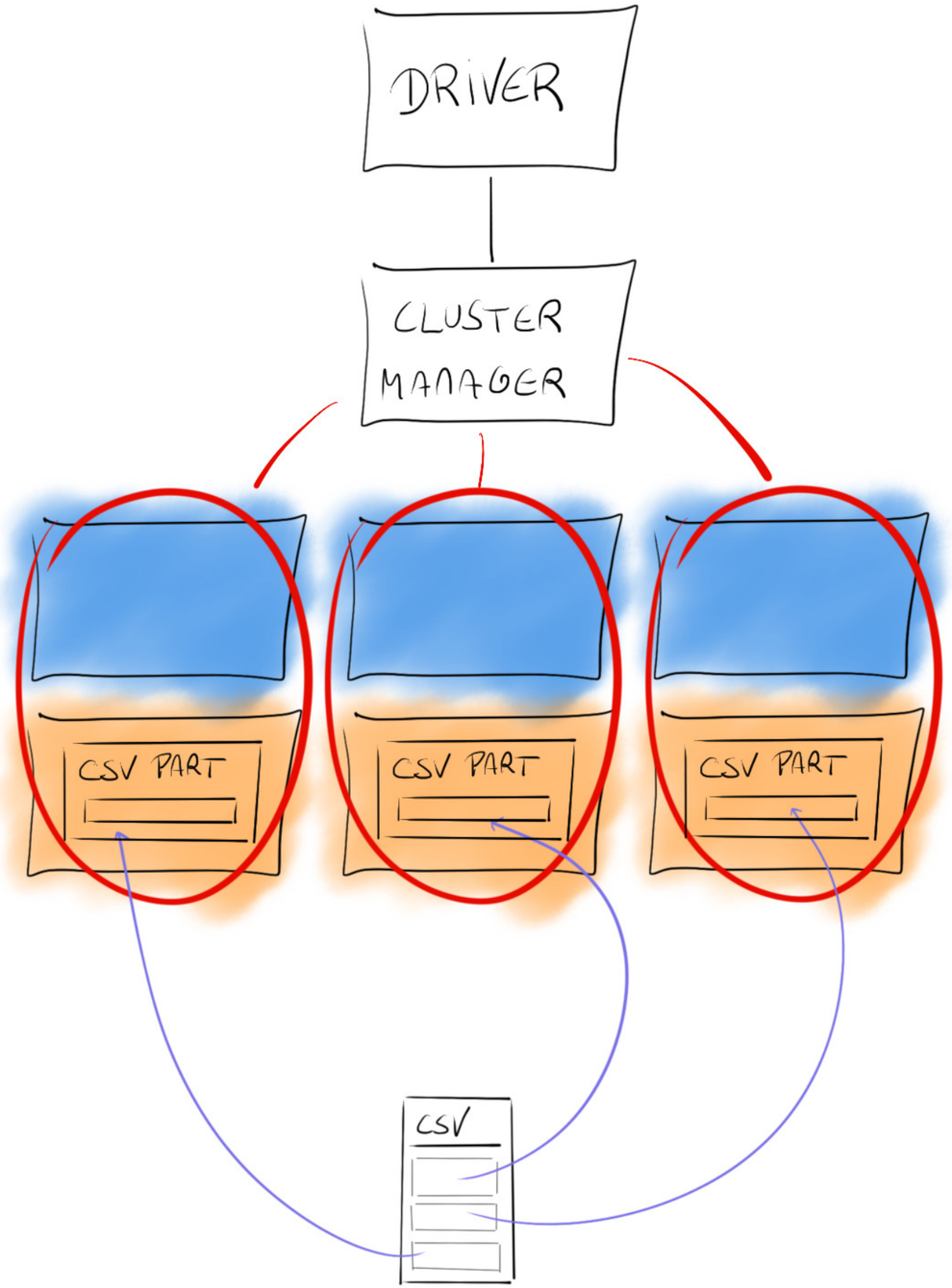
DIST (GROUP)	PTS (GROUP)	PLAYER
[6.2, 3.1,...]	[2, 2, ...]	STEPHEN CURRY
[19.1, 35,...]	[3, 2, ...]	DIRK NOWITZKI
[20.3, 0,...]	[2 2, ...]	KOBE BRYANT

AGGREGATE & SORT

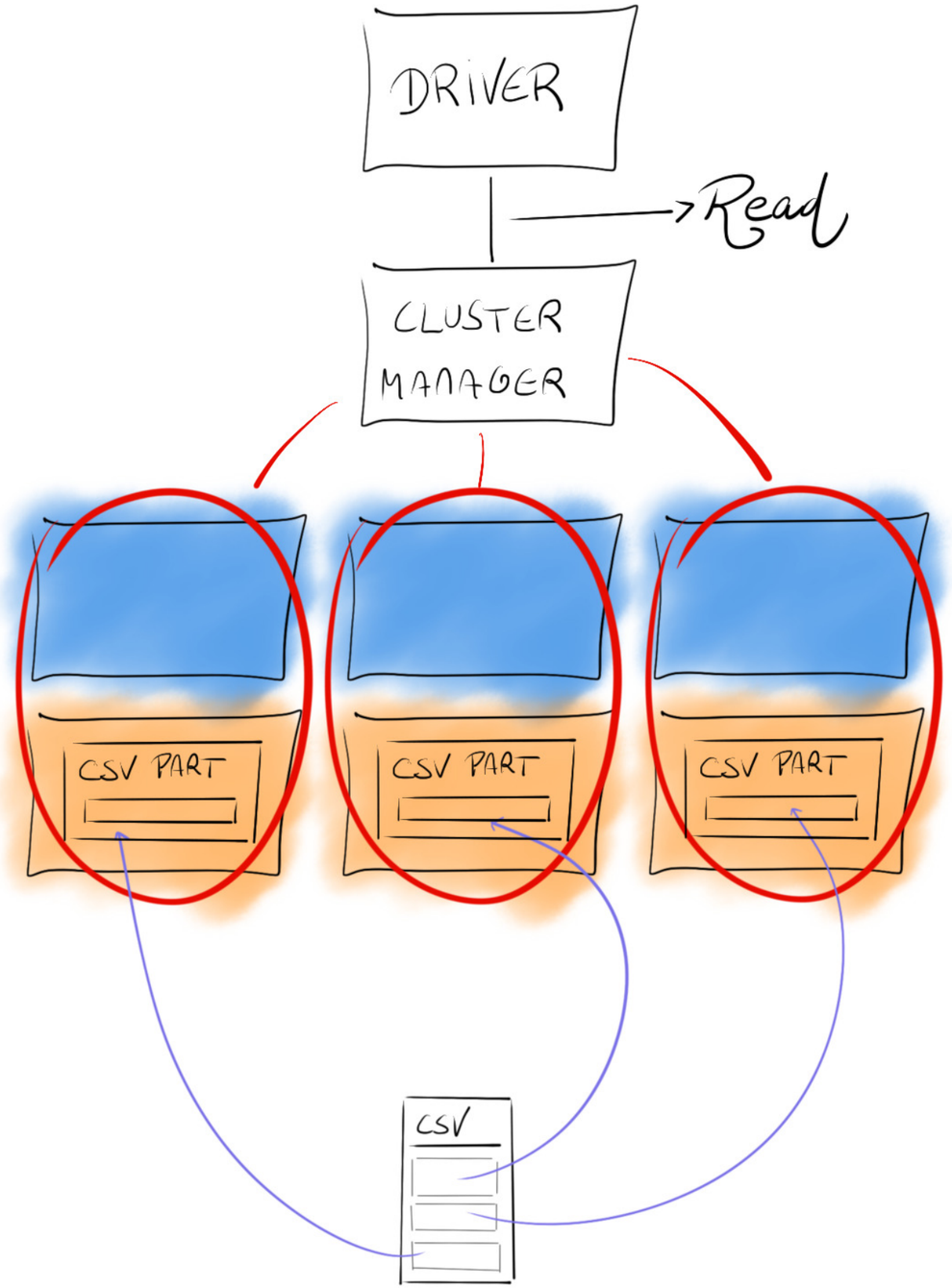
AVG(DIST)	SUM(PTS) 🔽	COUNT(1)	PLAYER
12.1	999	450	STEPHEN CURRY
42.0	998	442	DIRK NOWITZKI
12.5	997	300	KOBE BRYANT

```
spark.read
  .option("header", true)
  .csv("shot_logs.csv")
  .filter('PTS !== 0)
  .select(
      'player_name,
      'SHOT_DIST,
      'PTS)
  .groupBy('player_name)
  .agg(
       avg('SHOT_DIST),
       count(1),
       sum('PTS))
  .sort(desc("sum(PTS)"))
  .show
```

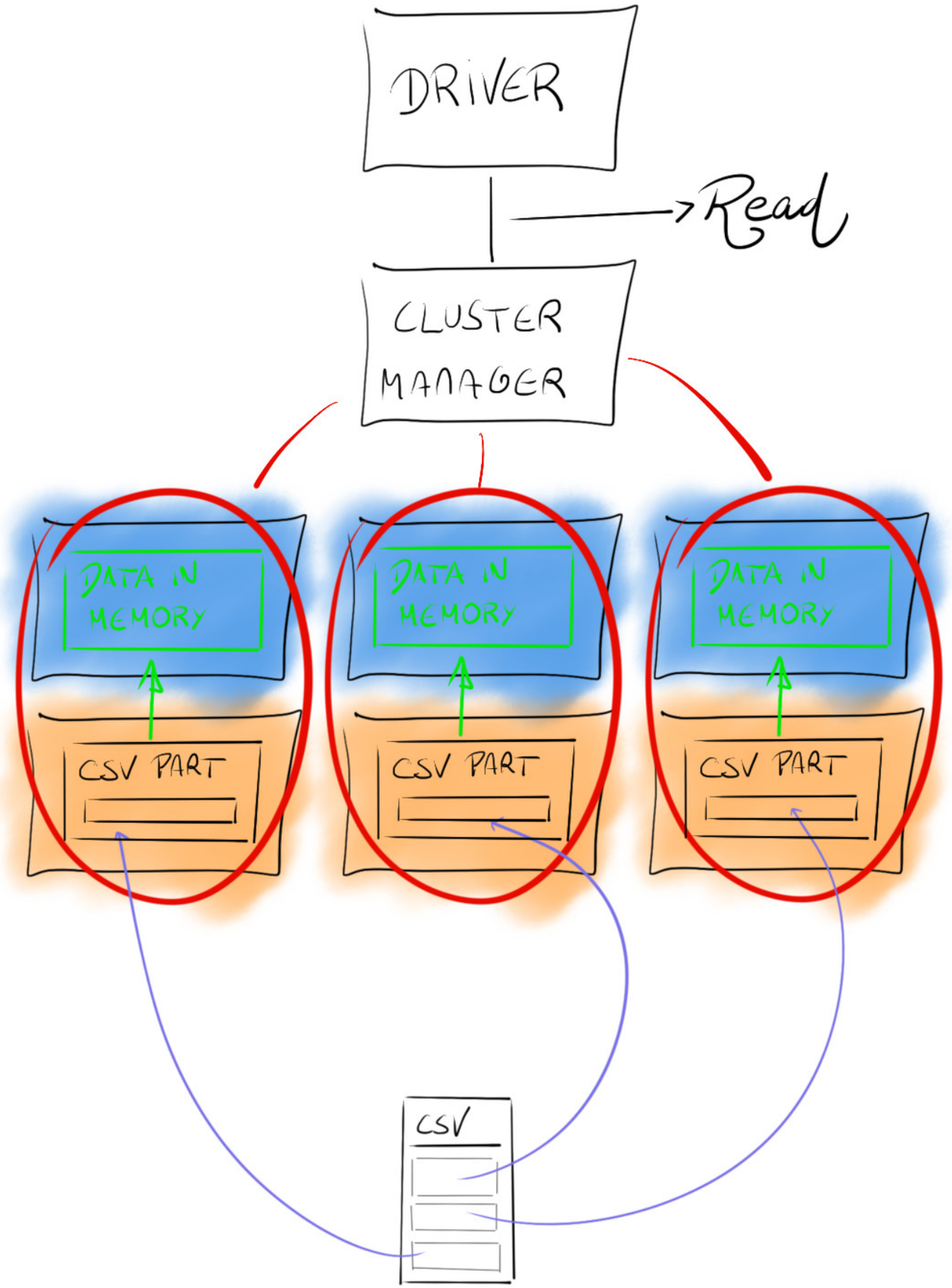
```
spark.read
   .option("header", true)
   .csv("shot_logs.csv")
   .filter('PTS !== 0)
   .select(
        'player_name,
        'SHOT_DIST,
        'PTS)
   .groupBy('player_name)
   .agg(
         avg('SHOT_DIST),
         count(1),
         sum('PTS))
   .sort(desc("sum(PTS)"))
   .show
```



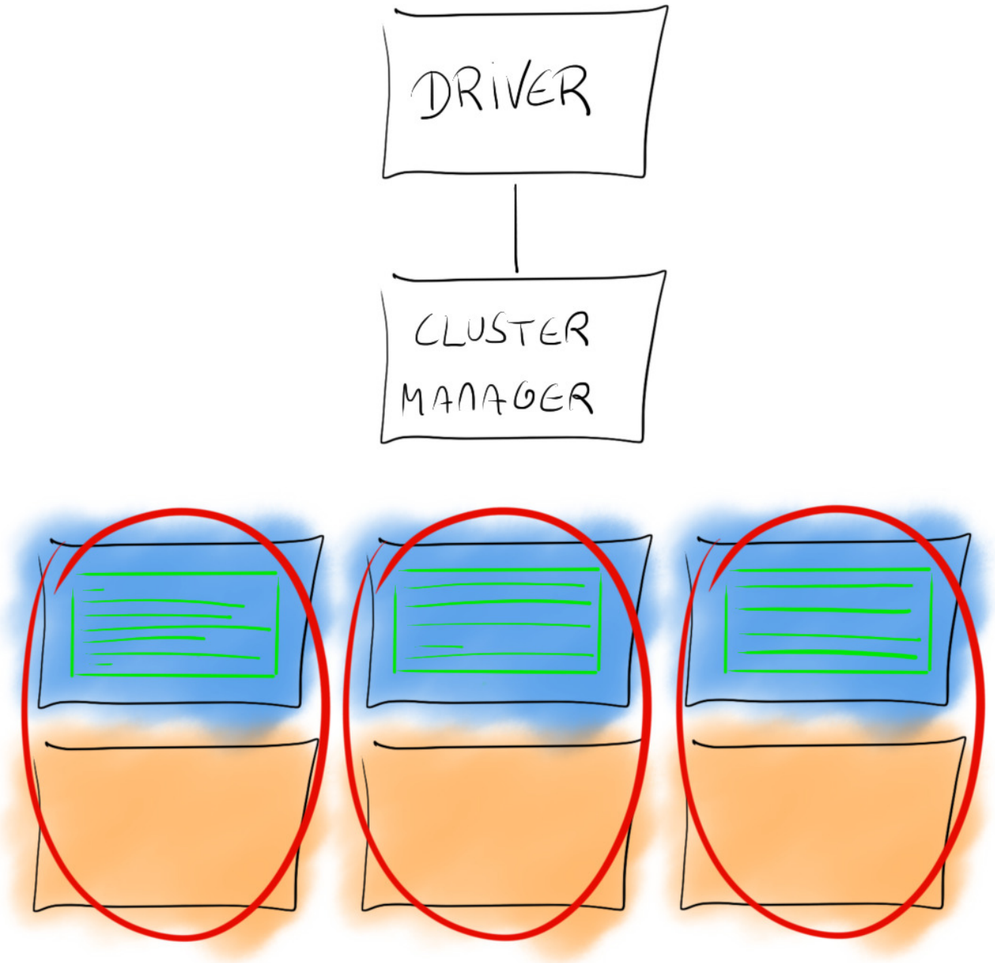
```
spark.read
   .option("header", true)
   .csv("shot_logs.csv")
   .filter('PTS !== 0)
   .select(
        'player_name,
        'SHOT_DIST,
        'PTS)
   .groupBy('player_name)
   .agg(
        avg('SHOT_DIST),
        count(1),
        sum('PTS))
   .sort(desc("sum(PTS)"))
   .show
```



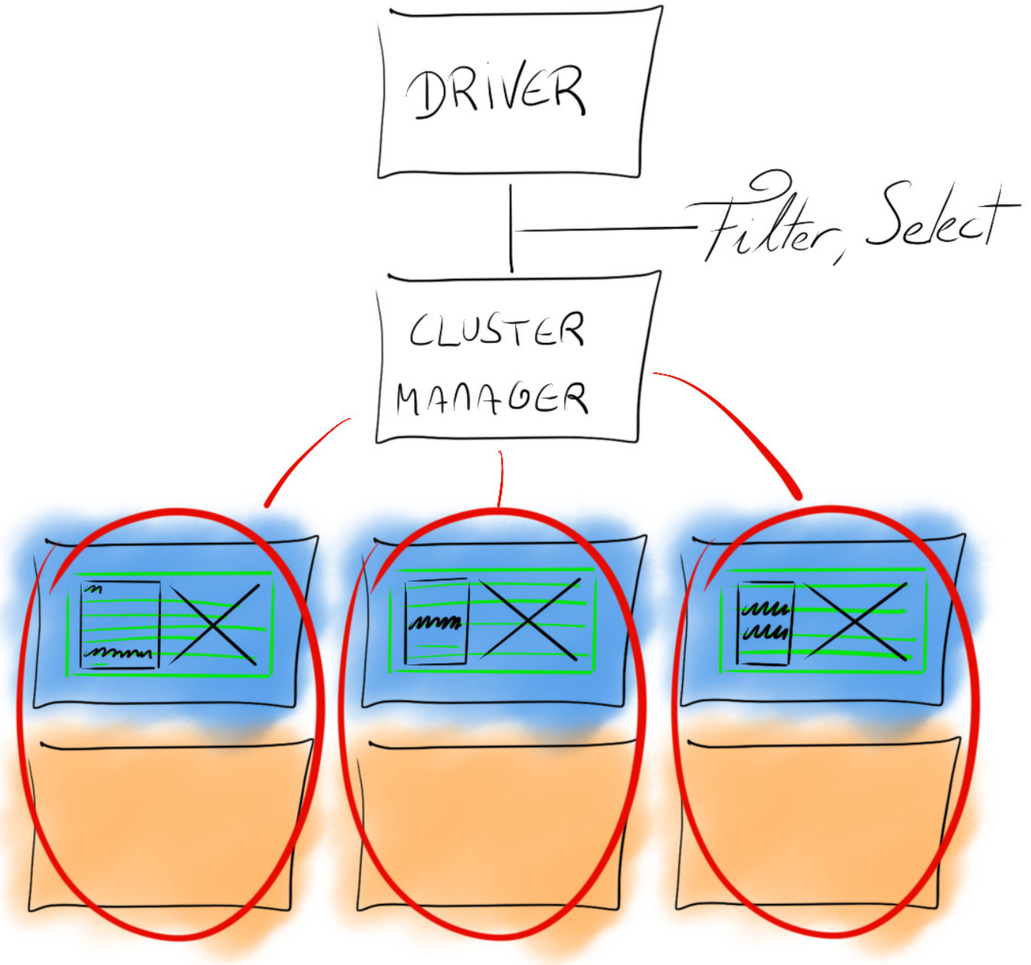
```
spark.read
   .option("header", true)
   .csv("shot_logs.csv")
   .filter('PTS !== 0)
   .select(
        'player_name,
        'SHOT_DIST,
        'PTS)
   .groupBy('player_name)
   .agg(
        avg('SHOT_DIST),
        count(1),
        sum('PTS))
   .sort(desc("sum(PTS)"))
   .show
```



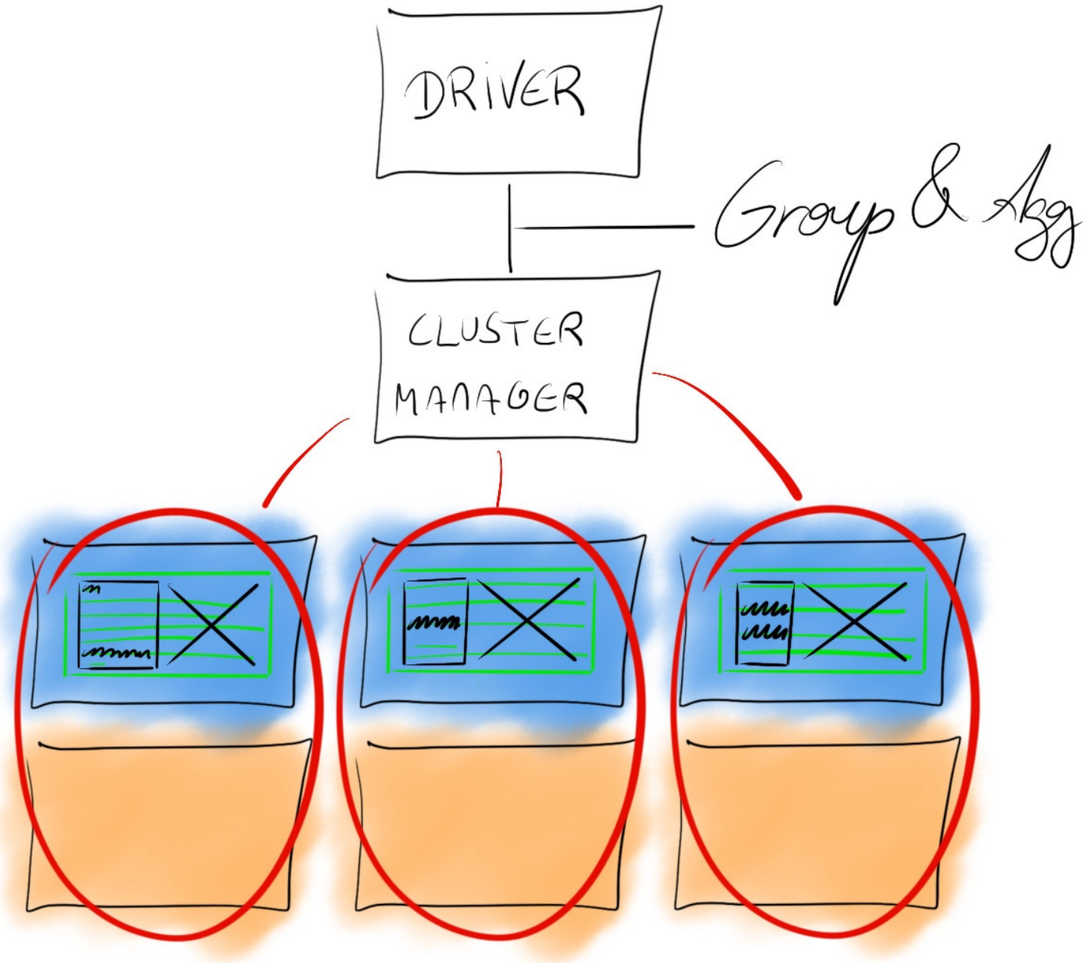
```
spark.read
  .option("header", true)
  .csv("shot_logs.csv")
  .filter('PTS !== 0)
  .select(
      'player_name,
      'SHOT_DIST,
      'PTS)
  .groupBy('player_name)
  .agg(
       avg('SHOT_DIST),
       count(1),
       sum('PTS))
  .sort(desc("sum(PTS)"))
  .show
```



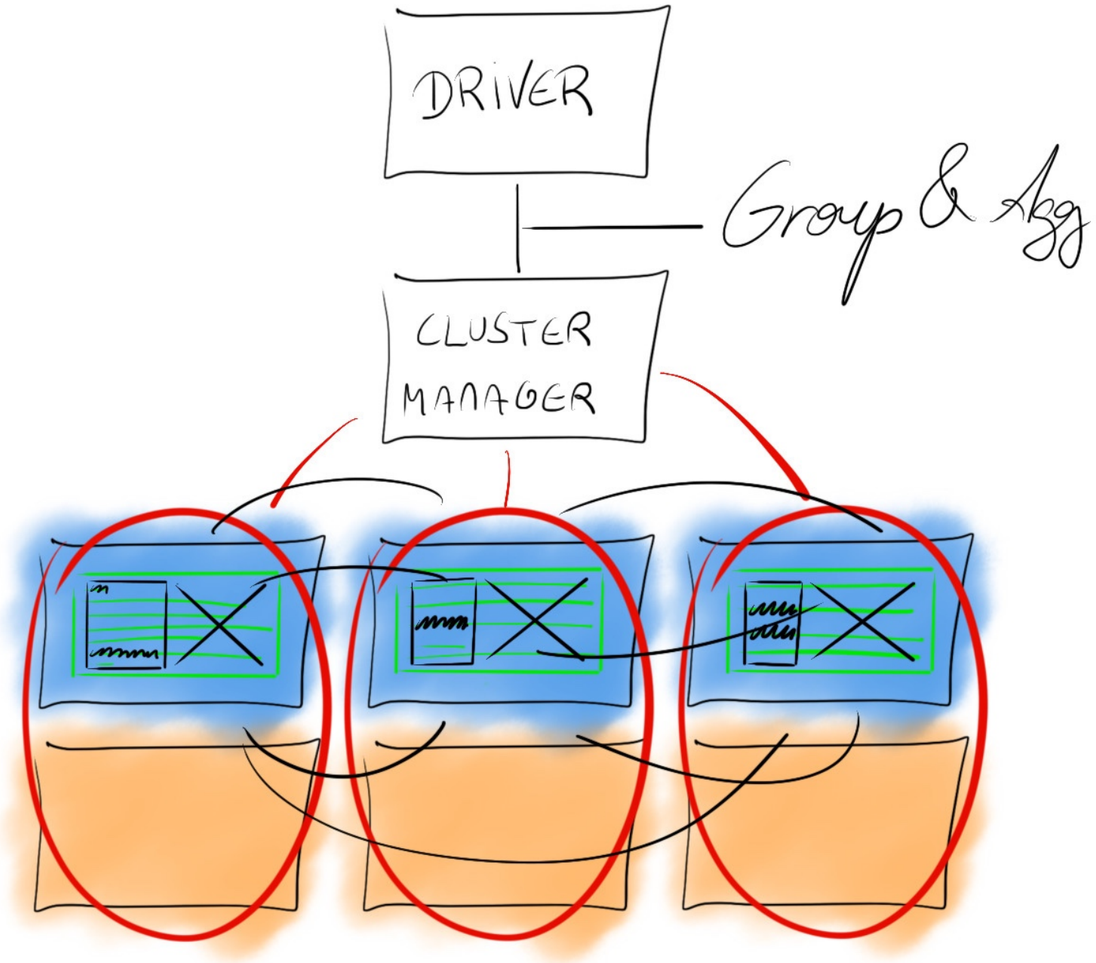
```
spark.read
  .option("header", true)
  .csv("shot_logs.csv")
  .filter('PTS !== 0)
  .select(
      'player_name,
      'SHOT_DIST,
      'PTS)
  .groupBy('player_name)
  .agg(
       avg('SHOT_DIST),
       count(1),
       sum('PTS))
  .sort(desc("sum(PTS)"))
  .show
```



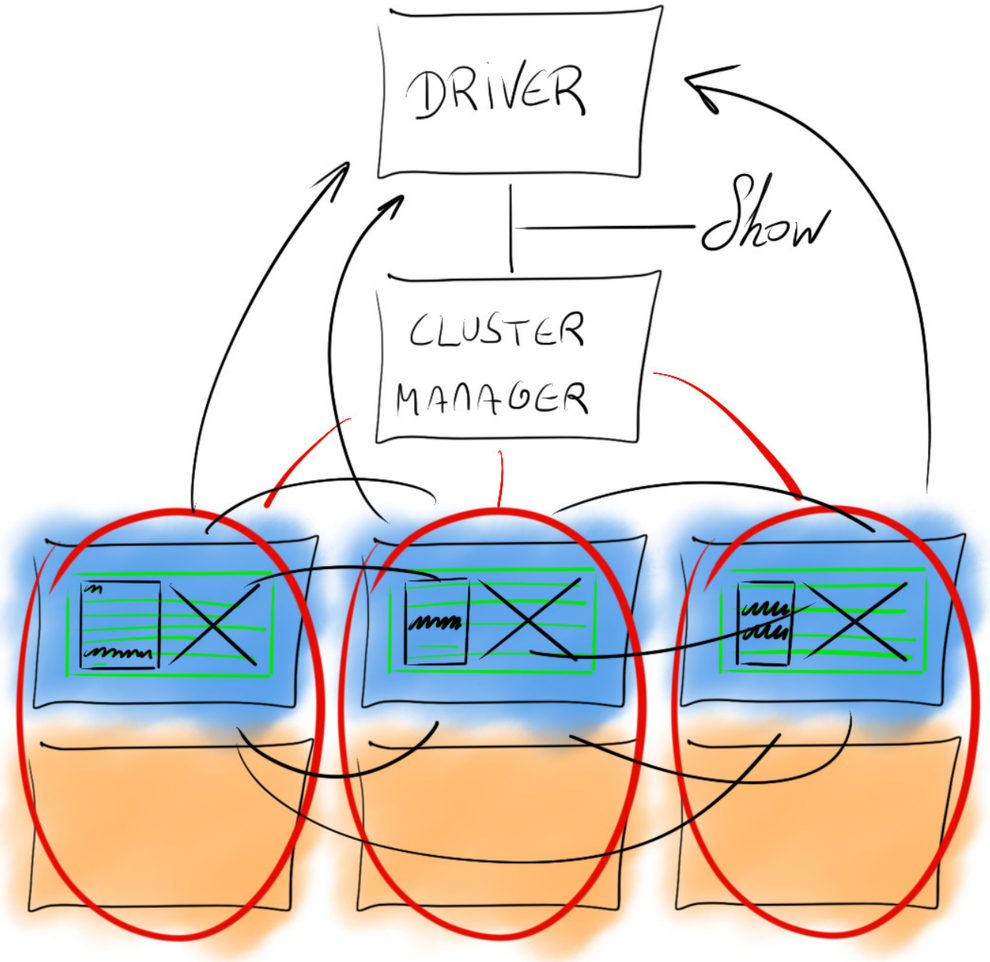
```
spark.read
   .option("header", true)
   .csv("shot_logs.csv")
   .filter('PTS !== 0)
   .select(
        'player_name,
        'SHOT_DIST,
        'PTS)
   .groupBy('player_name)
   .agg(
        avg('SHOT_DIST),
        count(1),
        sum('PTS))
   .sort(desc("sum(PTS)"))
   .show
```

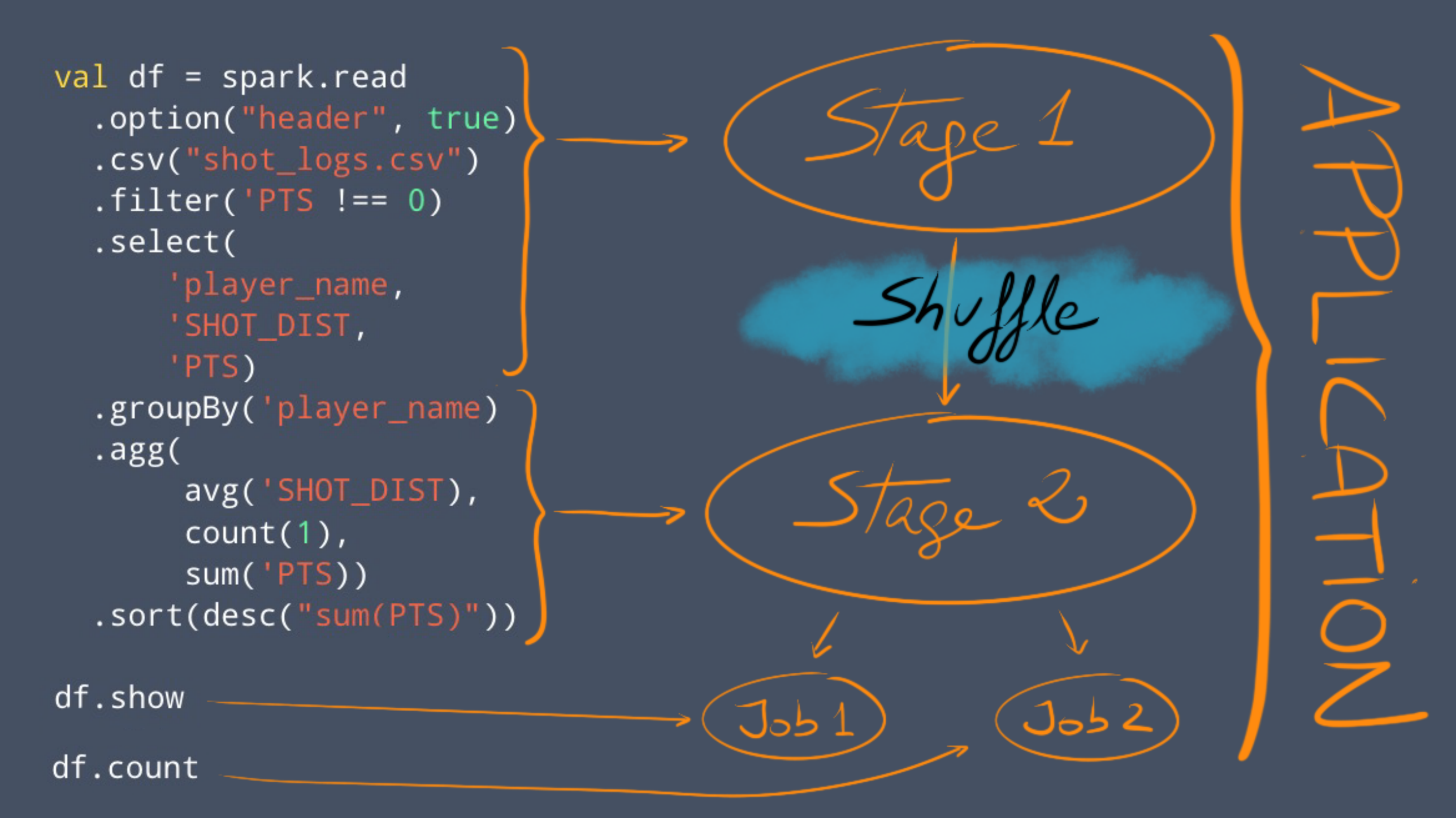


```
spark.read
   .option("header", true)
   .csv("shot_logs.csv")
   .filter('PTS !== 0)
   .select(
        'player_name,
        'SHOT_DIST,
        'PTS)
   .groupBy('player_name)
   .agg(
        avg('SHOT_DIST),
        count(1),
        sum('PTS))
   .sort(desc("sum(PTS)"))
   .show
```



```
spark.read
  .option("header", true)
  .csv("shot_logs.csv")
  .filter('PTS !== 0)
  .select(
      'player_name,
      'SHOT_DIST,
      'PTS)
  .groupBy('player_name)
  .agg(
      avg('SHOT_DIST),
      count(1),
      sum('PTS))
  .sort(desc("sum(PTS)"))
  .show
```




val df = spark.read
.option("header", true)
.csv("shot_logs.csv")
.filter('PTS !== 0)
.select(
'player_name,
'SHOT_DIST,
'PTS)
.groupBy('player_name)
.agg(
avg('SHOT_DIST),
count(1),
sum('PTS))
.sort(desc("sum(PTS)"))
df.show
df.count
Stage 1
Shuffle
Stage 2
Job 1
Job 2
APPLICATION

```
spark.read
  .option("header", true)
  .csv("/Users/ruben/Downloads/shot_logs.csv")
  .filter('PTS =!= 0)
  .select('player_name, 'SHOT_DIST.cast(DoubleType).alias("SHOT_DIST"), 'PTS.cast(IntegerType).alias("PTS"))
  .groupBy('player_name).agg(round(avg('SHOT_DIST), 2), count('SHOT_DIST), sum('PTS))
  .sort(desc("sum(PTS)"))
  .show
```

```
+----------------+----------------------+---------------+--------+
|     player_name|round(avg(SHOT_DIST), 2)|count(SHOT_DIST)|sum(PTS)|
+----------------+----------------------+---------------+--------+
|   stephen curry|                 15.48|            470|    1130|
|    james harden|                 13.14|            474|    1103|
|   klay thompson|                 15.73|            449|    1075|
|    lebron james|                 11.29|            478|    1041|
|      mnta ellis|                  12.6|            473|    1018|
|    kyrie irving|                 13.02|            439|     998|
|  damian lillard|                 13.76|            426|     995|
|lamarcus aldridge|                  12.3|            473|     971|
|  nikola vucevic|                  8.64|            480|     962|
|      chris paul|                 15.64|            425|     947|
|   anthony davis|                  8.78|            457|     915|
|   blake griffin|                 10.23|            447|     902|
|russell westbrook|                  10.8|            422|     889|
|  gordon hayward|                 13.35|            389|     875|
| wesley matthews|                 16.91|            336|     845|
|    tyreke evans|                   8.3|            396|     842|
|        rudy gay|                 11.74|            392|     841|
|       john wall|                 12.03|            392|     831|
|  brandon knight|                 15.14|            355|     829|
|  dirk nowtizski|                 16.52|            377|     828|
+----------------+----------------------+---------------+--------+
```

THERE IS NO SILVER BULLET

RESOURCES

- SPARK DOCUMENTATION
- HIGH PERFORMANCE SPARK BY HOLDEN KARAU
- MASTERING APACHE SPARK 2.3 BY JACEK LASKOWSKI
- SPARK'S GITHUB
- BECOME A CONTRIBUTOR

QUESTIONS?


THANKS!