Connector for Spark

Presented by

```
{
   name          : ”Alim S. Gafar",
   title         : ”Developer Advocate",
   company       : ”MongoDB",
   email         : ”alim.gafar@mongodb.com”,

   date          : “October 27, 2017”
}
```

mongoDB

Agenda

Apache Spark

MongoDB

MongoDB Connector for Spark

Working with Spark

Scala example

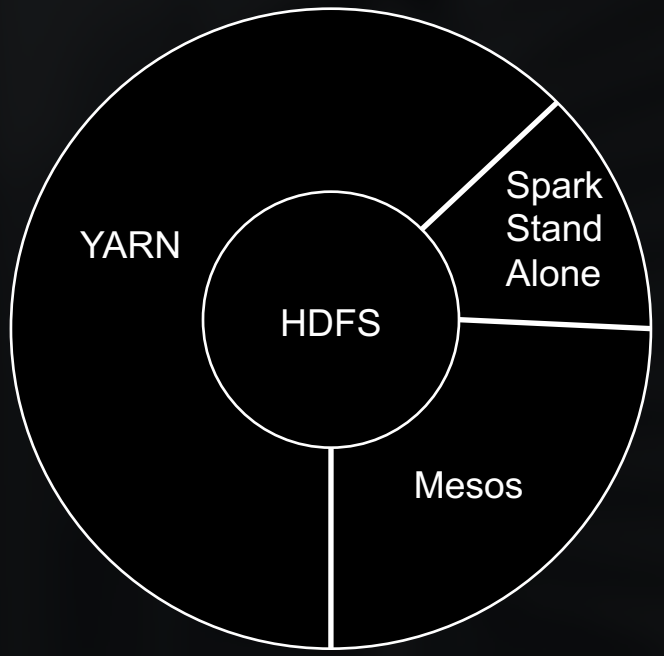
Working with SparkR

Wrap up

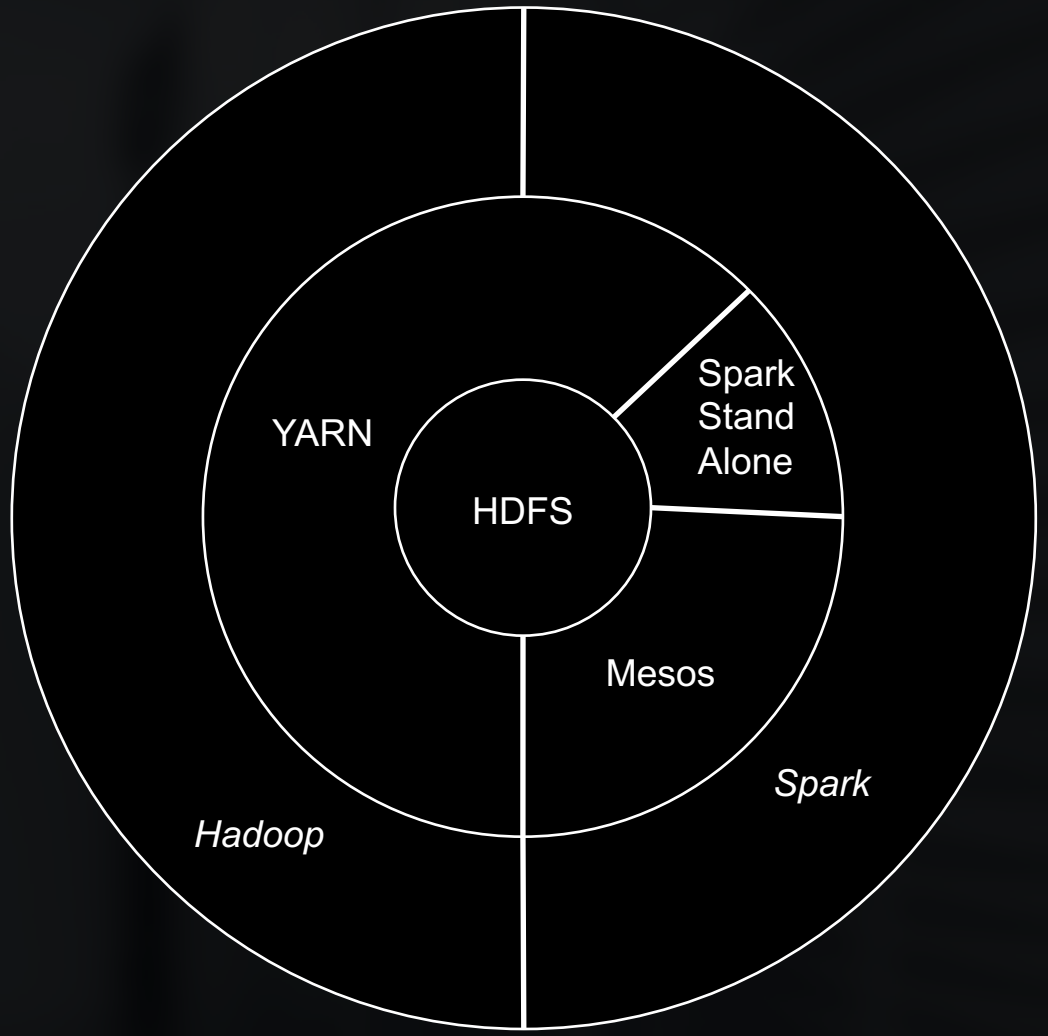

Spark

Distributed Data

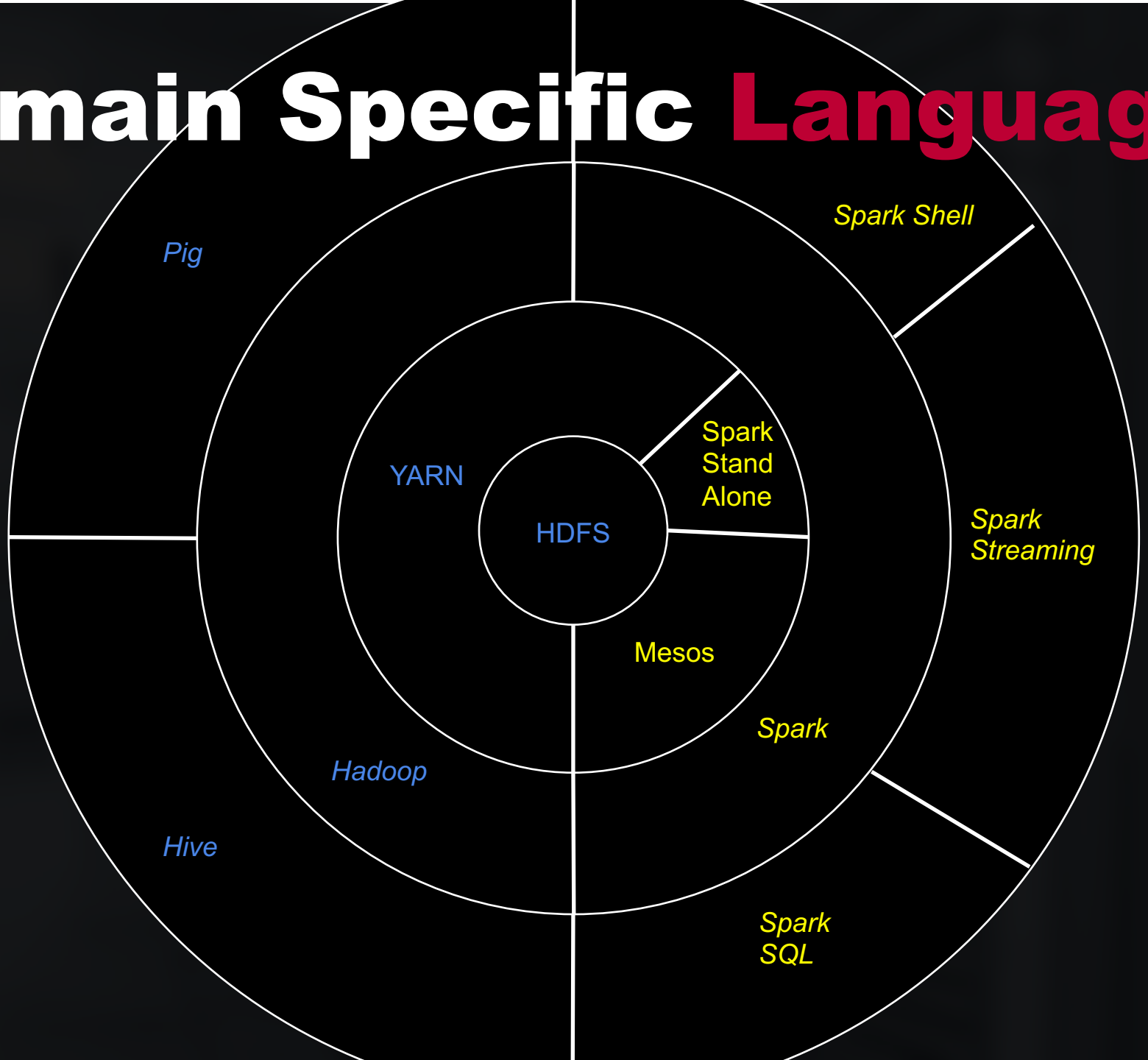
Distributed Resources



Distributed Processing

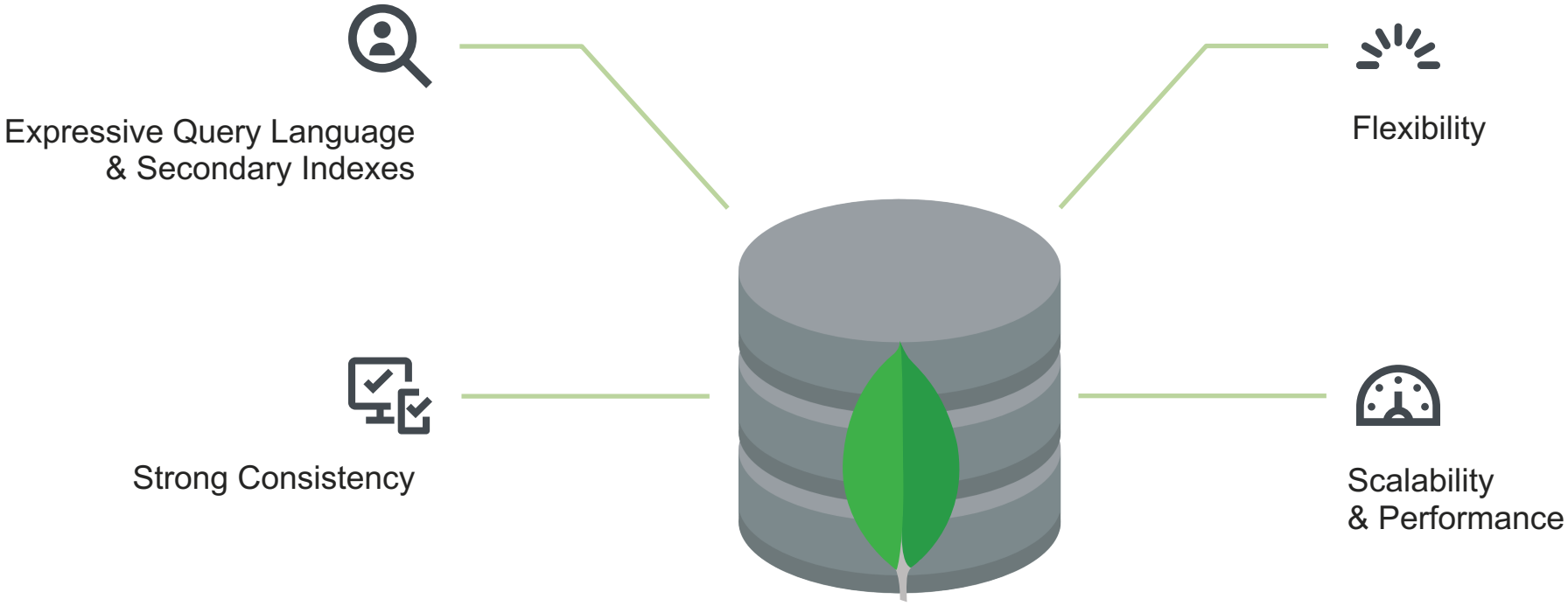


Domain Specific Languages



mongoDB

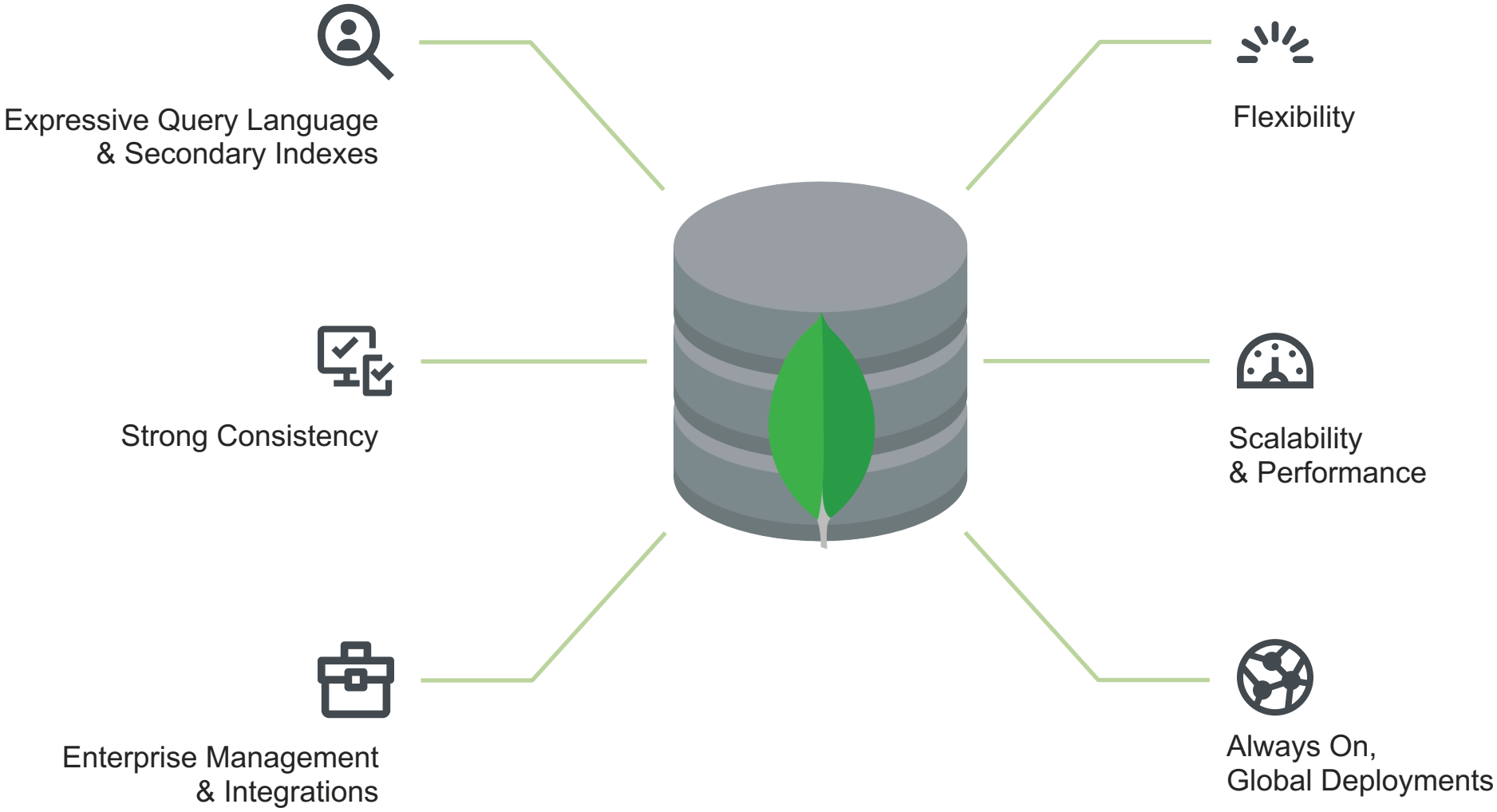
Why Use MongoDB?



NoSQL

mongoDB

MongoDB Nexus Architecture

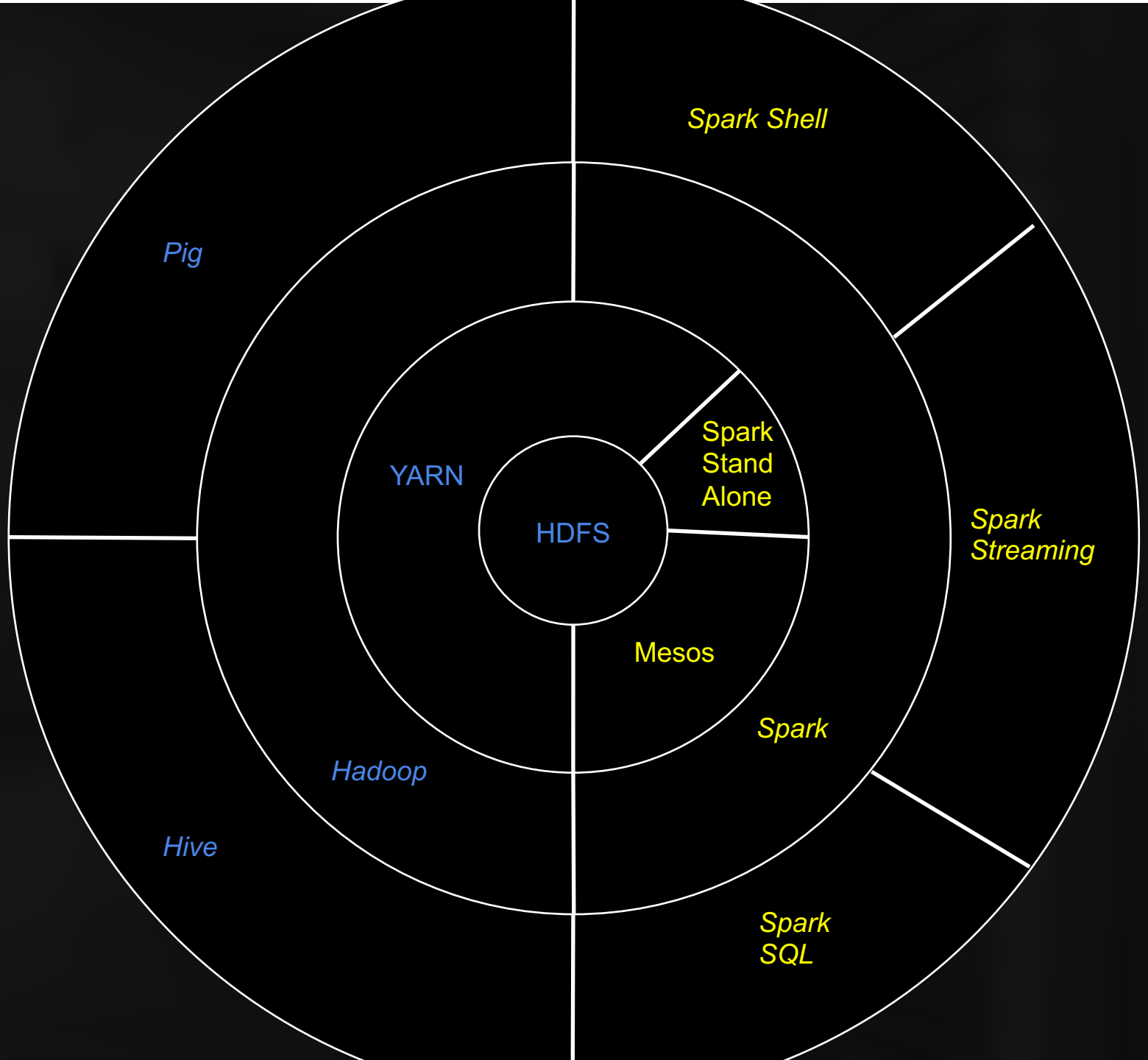


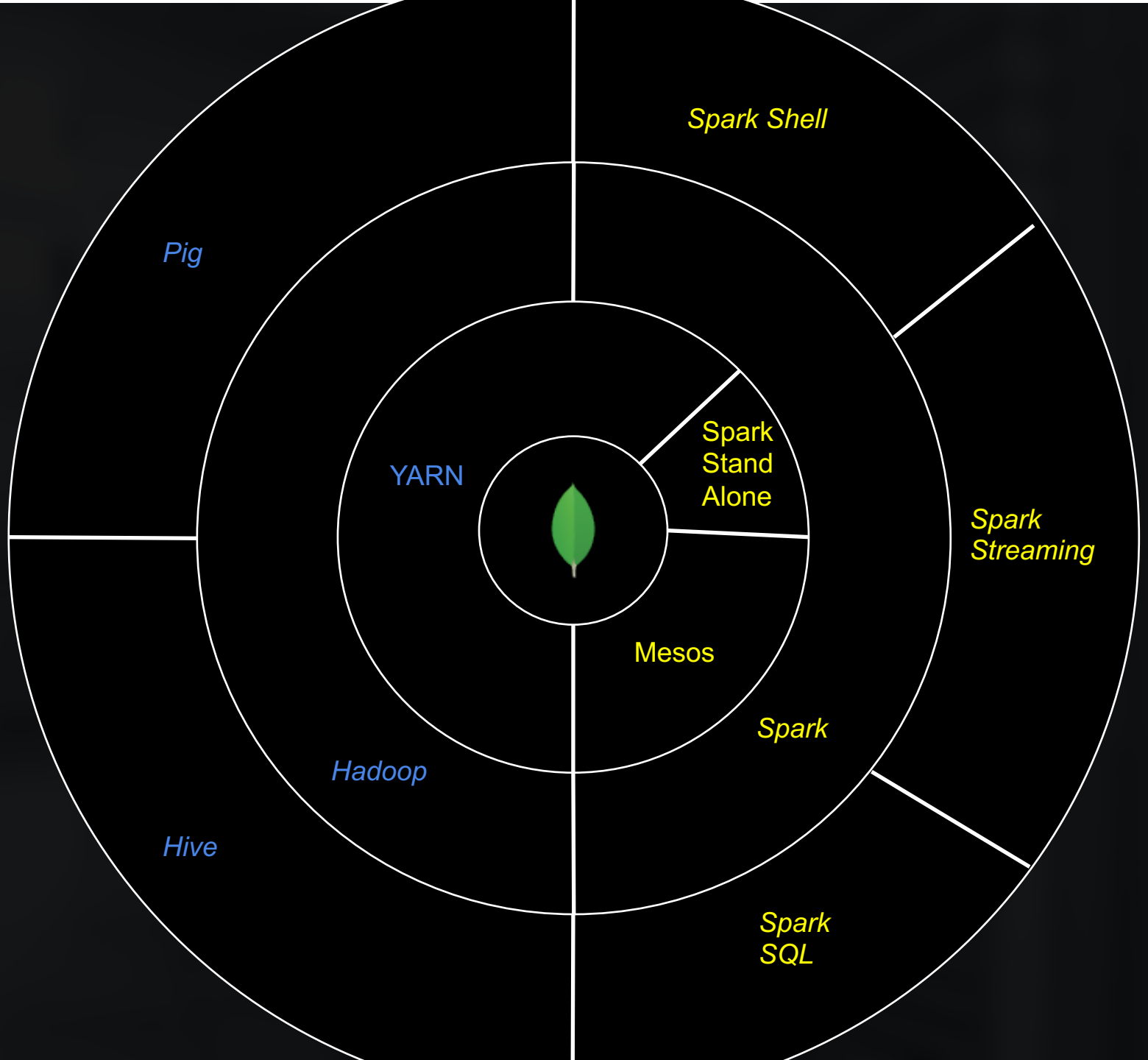
Which MongoDB?

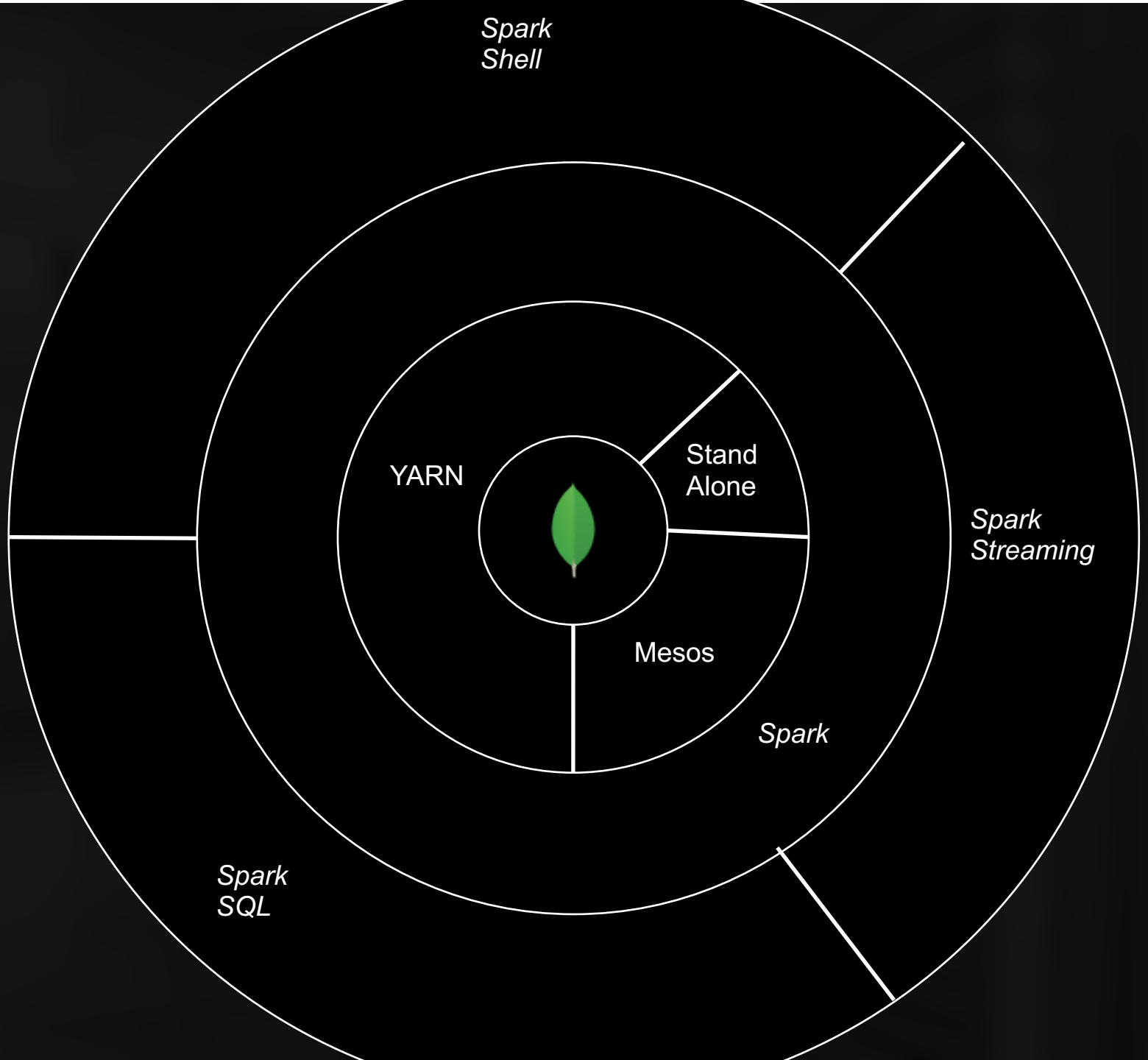
Community

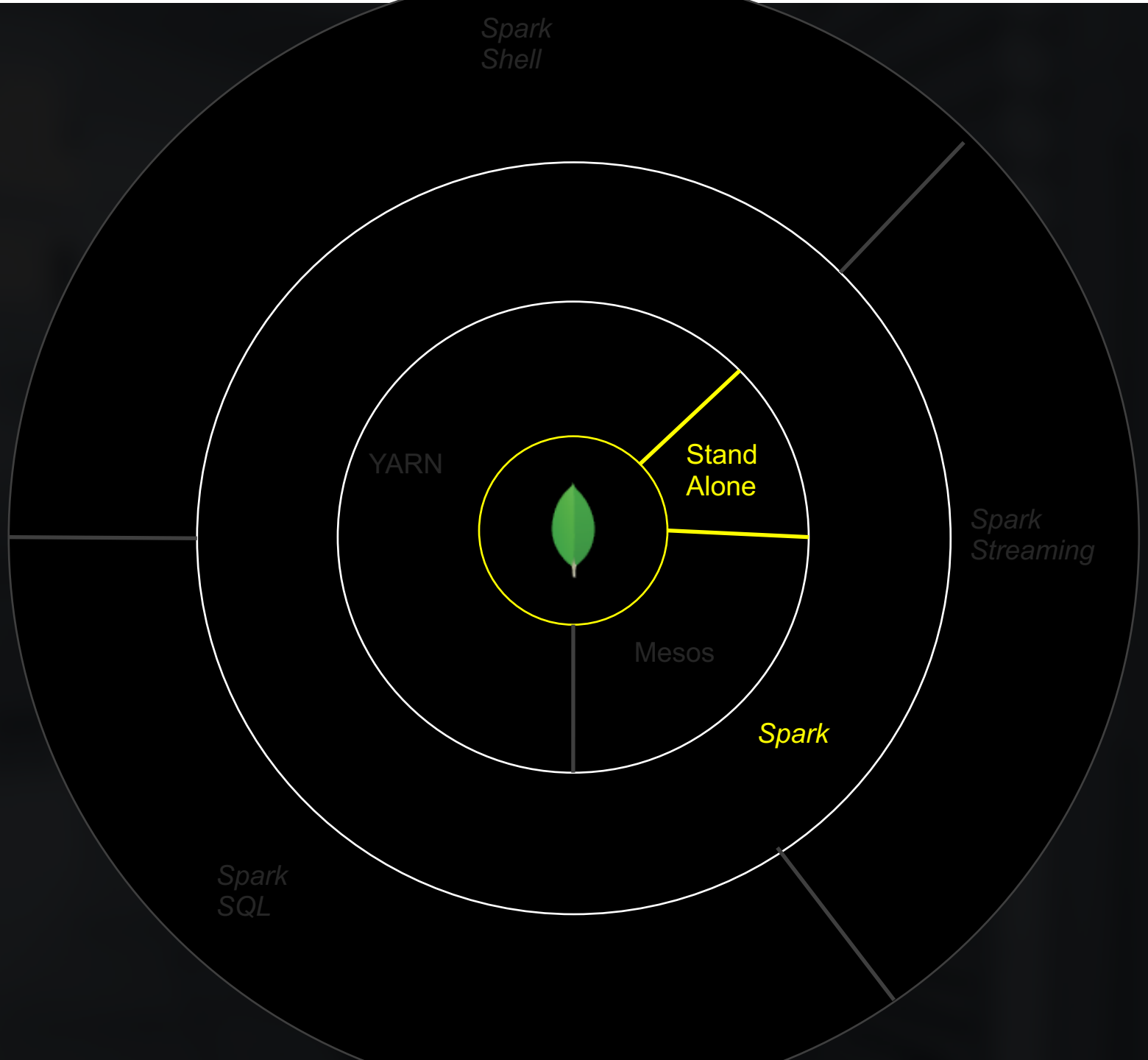
Commercial

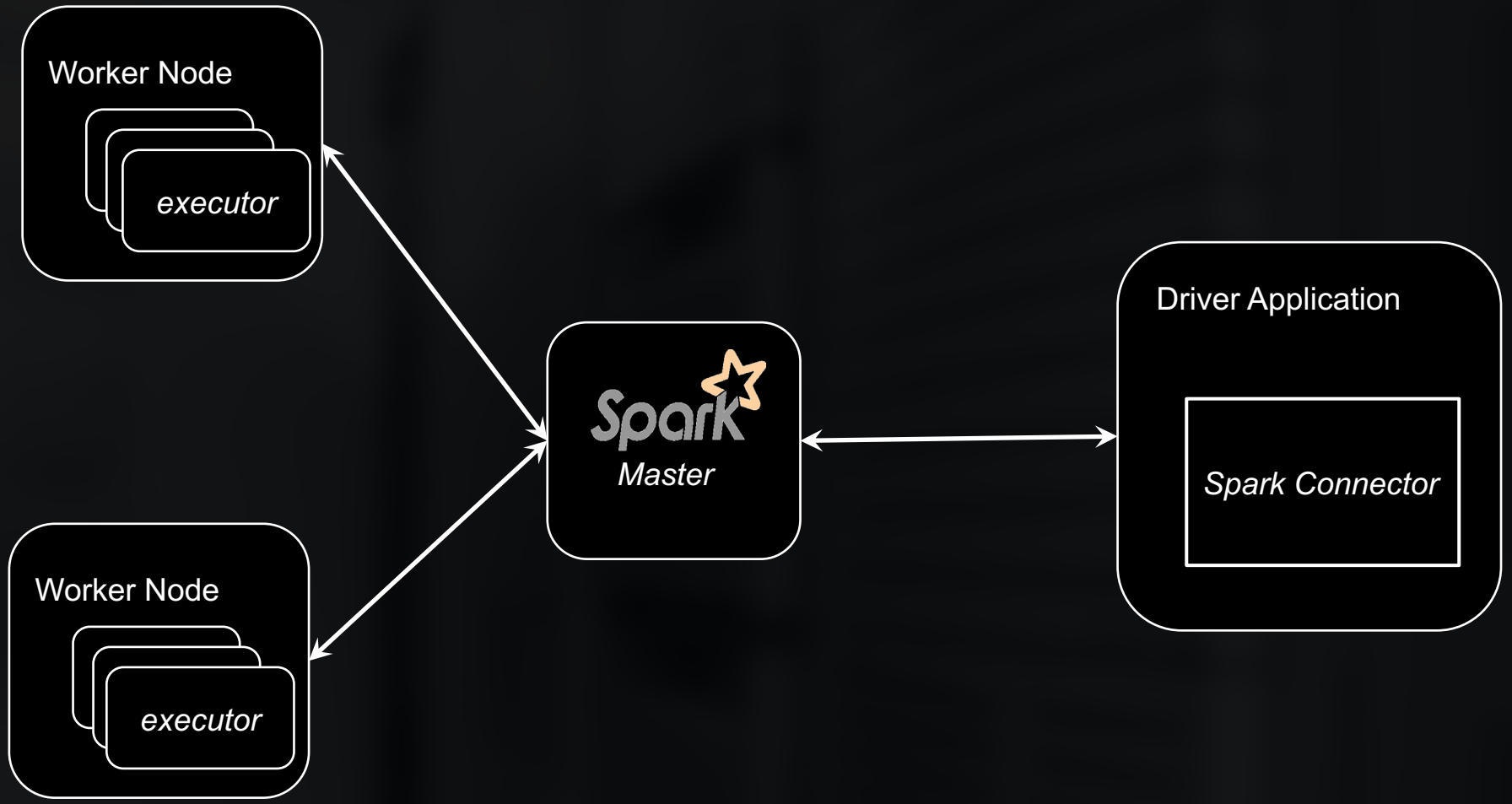
Atlas


Spark Shell
Pig
Spark
Stand
Alone
YARN
HDFS
Spark
Streaming
Mesos
Spark
Hadoop
Hive
Spark
SQL


Spark Shell
Pig
Spark
Stand
Alone
YARN
Spark
Streaming
Mesos
Spark
Hadoop
Hive
Spark
SQL


Spark
Shell
Spark
Streaming
Spark
SQL
Spark
YARN
Stand
Alone
Mesos


Spark
Shell
Spark
Streaming
Spark
SQL
YARN
Mesos
Stand
Alone
Spark


Worker Node
executor
Worker Node
executor
Spark
Master
Driver Application
Spark Connector

Parallelize

Parallelize

Parallelize

Parallelize

Parellelize Transform

Parellelize Transform

Parellelize Transform

Parellelize Transform

Transformations

filter(func)
union(func)
intersection(set)
distinct(n)
map(function)

Parellelize Transform Transform

Parellelize Transform Transform

Parellelize Transform Transform

Parellelize Transform Transform

Parellelize Transform Transform Action

Parellelize Transform Transform Action

Parellelize Transform Transform Action

Parellelize Transform Transform Action

Actions

collect()
count()
first()
take(n)
reduce(function)

Parellelize Transform Transform Action Result

Parellelize Transform Transform Action Result

Parellelize Transform Transform Action Result

Parallelize Transform Transform Action Result

Lineage

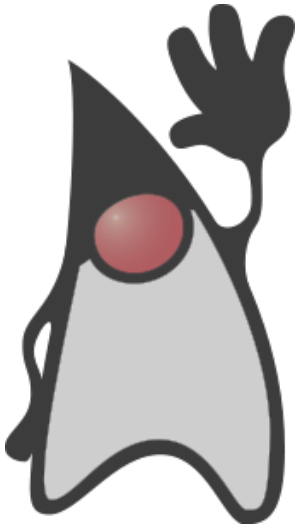
Parellelize Transform Transform Action Result

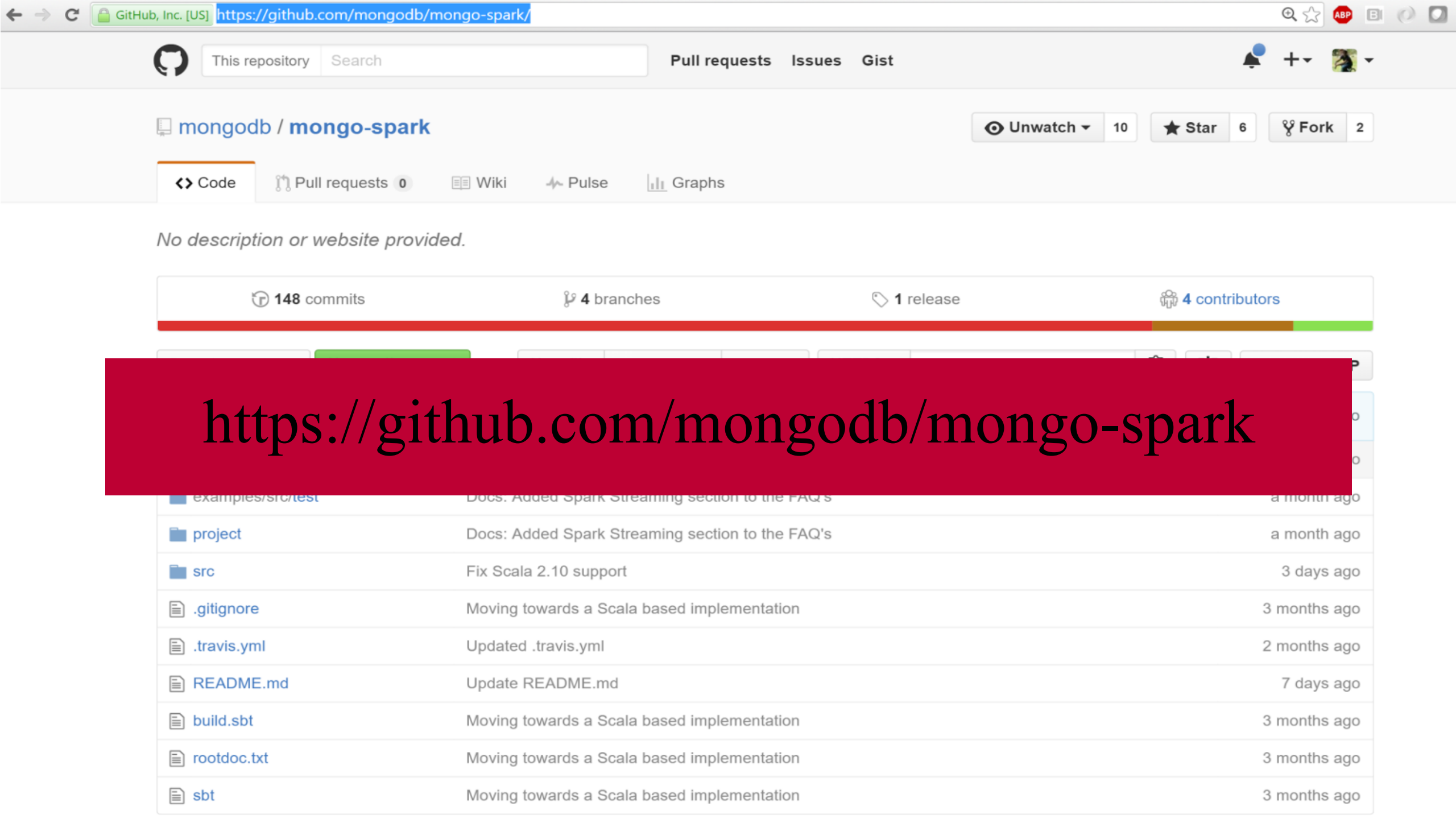
Parellelize Transform Transform Action Result

Parellelize Transform Transform Action Result

Parellelize Transform Transform Action Result

Driver Types




GitHub, Inc. [US] https://github.com/mongodb/mongo-spark/
This repository Search
Pull requests Issues Gist
mongodb / mongo-spark
Unwatch 10 Star 6 Fork 2
Code Pull requests 0 Wiki Pulse Graphs
No description or website provided.
148 commits 4 branches 1 release 4 contributors
https://github.com/mongodb/mongo-spark
project Docs: Added Spark Streaming section to the FAQ's a month ago
src Fix Scala 2.10 support 3 days ago
.gitignore Moving towards a Scala based implementation 3 months ago
.travis.yml Updated .travis.yml 2 months ago
README.md Update README.md 7 days ago
build.sbt Moving towards a Scala based implementation 3 months ago
rootdoc.txt Moving towards a Scala based implementation 3 months ago
sbt Moving towards a Scala based implementation 3 months ago

Spark Overview

Apache Spark is a fast and general-purpose cluster computing system. It provides high-level APIs in Java, Scala, Python and R, and an optimized engine that supports general execution graphs. It also supports a rich set of higher-level tools including Spark SQL for SQL and structured data processing, MLlib for machine learning, GraphX for graph processing, and Spark Streaming.

Downloading

Get Spark from the downloads page of the project website. This documentation is for Spark version 1.6.1. Spark uses Hadoop's client libraries for HDFS and YARN. Downloads are pre-packaged for a handful of popular Hadoop versions. Users can also download a "Hadoop free" binary and run Spark with any Hadoop version by augmenting Spark's classpath.

http://spark.apache.org/docs/latest/

Running the Examples and Shell

Spark comes with several sample programs. Scala, Java, Python and R examples are in the `examples/src/main` directory. To run one of the Java or Scala sample programs, use `bin/run-example <class> [params]` in the top-level Spark directory. (Behind the scenes, this invokes the more general `spark-submit` script for launching applications). For example,

```
./bin/run-example SparkPi 10
```

You can also run Spark interactively through a modified version of the Scala shell. This is a great way to learn the framework.

```
./bin/spark-shell --master local[2]
```

The `--master` option specifies the master URL for a distributed cluster, or `local` to run locally with one thread, or `local[N]` to run locally with N

MovieLens

GroupLens Research has collected and made available rating data sets from the MovieLens web site (http://movielens.org). The data sets were collected over various periods of time, depending on the size of the set. Before using these data sets, please review their README files for the usage licenses and other details.

Help our research lab: Please take a short survey about the MovieLens datasets

MovieLens 100K Dataset

Stable benchmark dataset. 100,000 ratings from 1000 users on 1700 movies. Released 4/1998.

- README.txt
- ml-100k.zip (size: 5 MB, checksum)
- Index of unzipped files

Permalink: http://grouplens.org/datasets/movielens/100k/

MovieLens 1M Dataset

Stable benchmark dataset. 1 million ratings from 6000 users on 4000 movies. Released 2/2003.

Datasets

The New MongoDB Connector for Apache Spark In Action: Building a Movie Recommendation Engine

< View all blog posts

Sam Weaver

June 28, 2016

Featured

3.2 Events Spark Community

Using the Connector

```
{
  "_id" : ObjectId("578be1fe1fe699f2deb80807"),
  "user_id" : 196,
  "movie_id" : 242,
  "rating" : 3,
  "timestamp" : 881250949
}
```

```
./bin/spark-shell \
 --conf \
    "spark.mongodb.input.uri=mongodb://127.0.0.1/movies.movie_ratings" \
 --conf \
   "spark.mongodb.output.uri=mongodb://127.0.0.1/movies.user_recommendations" \
 --packages org.mongodb.spark:mongo-spark-connector_2.10:1.0.0
```

```
./bin/spark-shell \
 --conf \
    "spark.mongodb.input.uri=mongodb://127.0.0.1/movies.movie_ratings" \
 --conf \
   "spark.mongodb.output.uri=mongodb://127.0.0.1/movies.user_recommendations" \
 --packages org.mongodb.spark:mongo-spark-connector_2.10:1.0.0
```

```
./bin/spark-shell \
 --conf \
    "spark.mongodb.input.uri=mongodb://127.0.0.1/movies.movie_ratings" \
 --conf \
   "spark.mongodb.output.uri=mongodb://127.0.0.1/movies.user_recommendations" \
 --packages org.mongodb.spark:mongo-spark-connector_2.10:1.0.0
```

```
./bin/spark-shell \
 --conf \
    "spark.mongodb.input.uri=mongodb://127.0.0.1/movies.movie_ratings" \
 --conf \
   "spark.mongodb.output.uri=mongodb://127.0.0.1/movies.user_recommendations" \
 --packages org.mongodb.spark:mongo-spark-connector_2.10:1.0.0
```

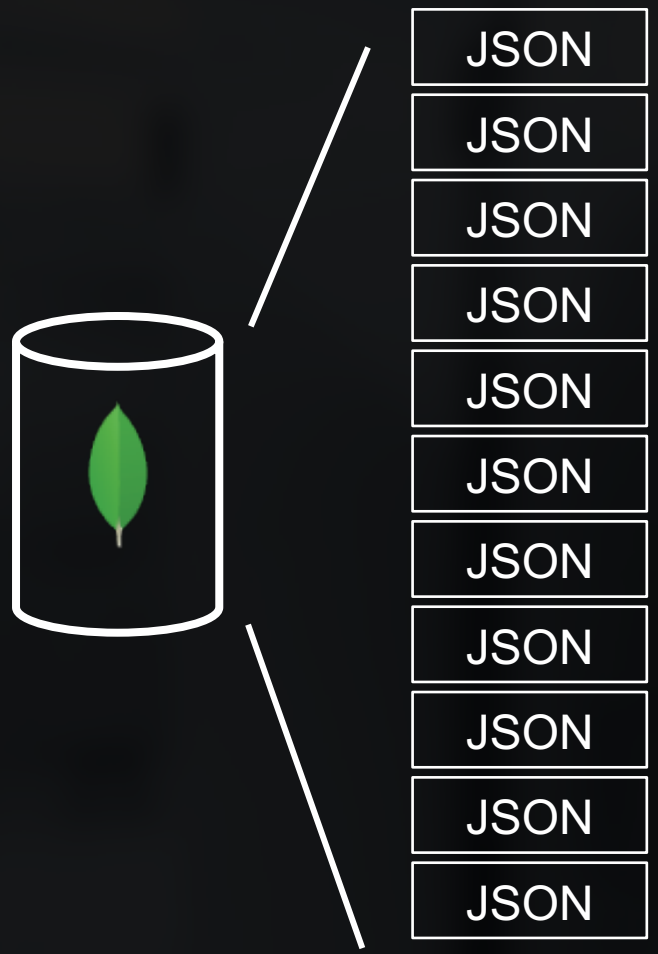
```
import com.mongodb.spark._
import com.mongodb.spark.rdd.MongoRDD
import org.bson.Document

val rdd = sc.loadFromMongoDB()
for( doc <- rdd.take( 10 ) ) println( doc )
```

SparkR

Advanced Activities

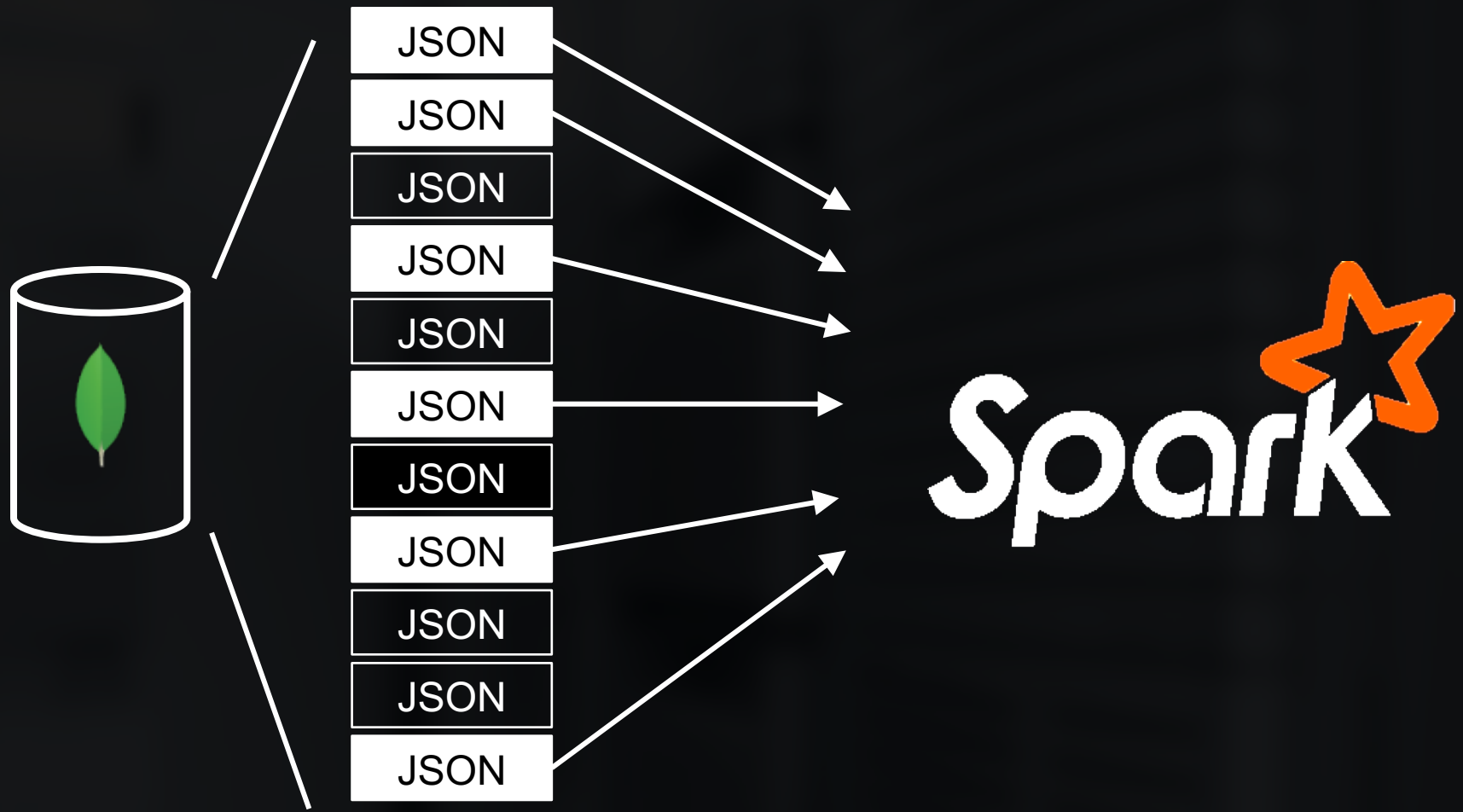
Aggregation Filters

```
$match  |  $project | $group
```



Spark

mongoDB

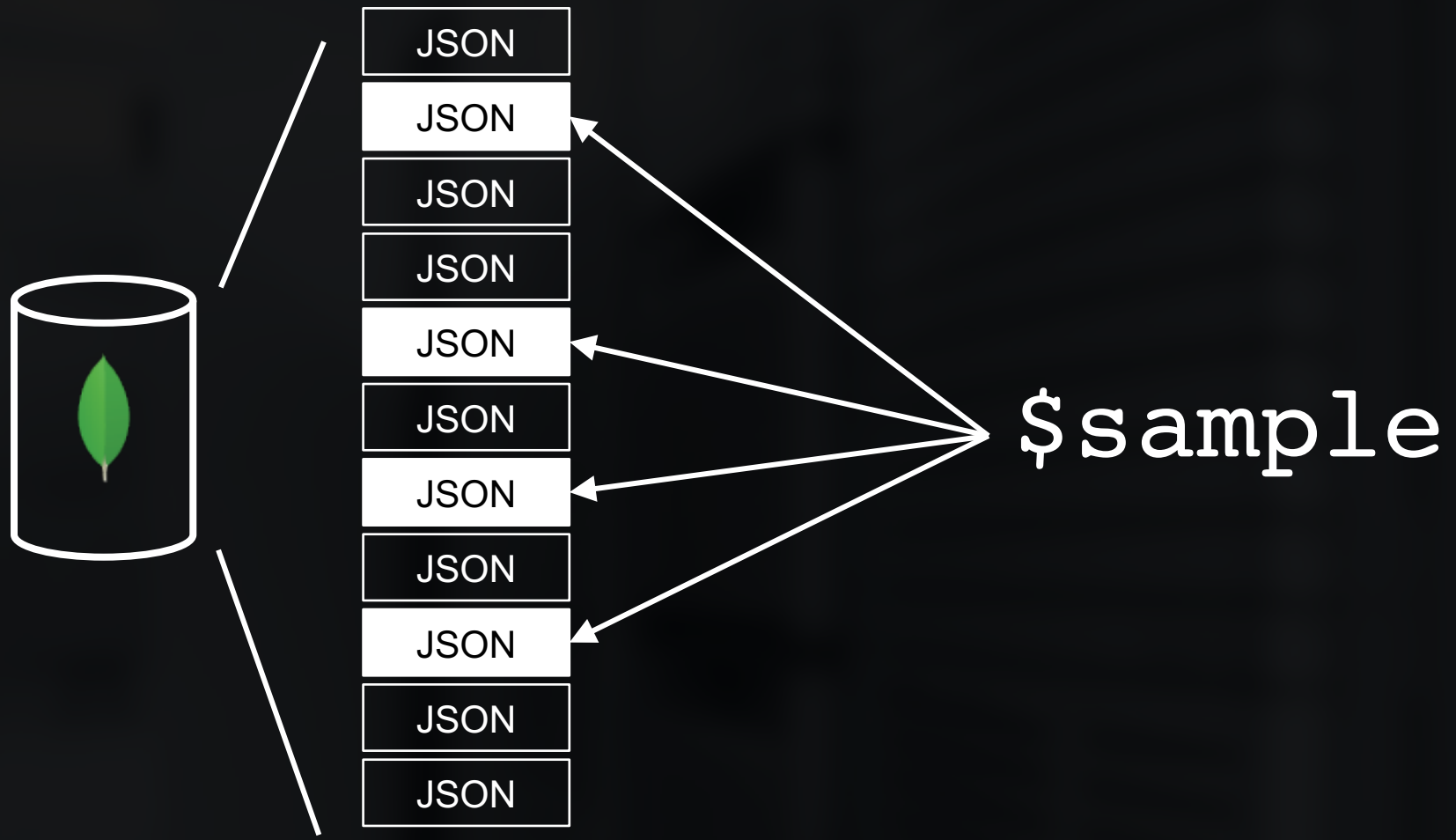

JSON
JSON
JSON
JSON
JSON
JSON
JSON
JSON
JSON
JSON
JSON
Spark

```
val aggRdd =
    rdd.withPipeline(
        Seq(
            Document.parse(
                "{ $match: { Country: \"USA\" } }"
            )
        )
    )
```

Spark SQL + Dataframes

RDD + Schema = Dataframe

```
15/12/02 11:10:22 INFO DAGScheduler: Job 7 finished: show at <console>:36, took 0.047074 s
+--------------------+-------+--------------------+--------+--------+-------+--------------------+-----------+
|                 _id|country|                crew|    date|duration|minutes|             purpose|    vehicle|
+--------------------+-------+--------------------+--------+--------+-------+--------------------+-----------+
|[565f1dd8d4c6d9a9...|    USA|            Ed White|  6/3/65|    0:36|   36.0|First U.S. EVA. U...|  Gemini IV|
|[565f1dd8d4c6d9a9...|    USA|       Eugene Cernan|  6/5/66|    2:07|  127.0|Inadequate restra...|Gemini IX-A|
|[565f1dd8d4c6d9a9...|    USA|        Mike Collins| 7/19/66|    0:50|   50.0|Standup EVA.  UV ...|   Gemini X|
|[565f1dd8d4c6d9a9...|    USA|        Mike Collins| 7/20/66|    0:39|   39.0|Retrieved MMOD ex...|   Gemini X|
|[565f1dd8d4c6d9a9...|    USA|      Richard Gordon| 9/13/66|    0:44|   44.0|Attached tether b...|  Gemini XI|
|[565f1dd8d4c6d9a9...|    USA|      Richard Gordon| 9/14/66|    2:10|  130.0|Standup EVA.  Too...|  Gemini XI|
|[565f1dd8d4c6d9a9...|    USA|         Buzz Aldrin|11/12/66|    2:29|  149.0|Standup EVA.  Sci...| Gemini XII|
|[565f1dd8d4c6d9a9...|    USA|         Buzz Aldrin|11/13/66|    2:06|  126.0|Attached tether b...| Gemini XII|
|[565f1dd8d4c6d9a9...|    USA|         Buzz Aldrin|11/14/66|    0:55|   55.0|Standup EVA.  Jet...| Gemini XII|
|[565f1dd8d4c6d9a9...|    USA|         David Scott|  3/6/69|    0:47|   47.0|Standup EVA from ...|   Apollo 9|
|[565f1dd8d4c6d9a9...|    USA|    Russ Schweickart|  3/6/69|    0:51|   51.0|Lunar module base...|   Apollo 9|
|[565f1dd8d4c6d9a9...|    USA|Neil Armstrong   ...| 7/20/69|    2:32|  152.0|First to walk on ...|  Apollo 11|
|[565f1dd8d4c6d9a9...|    USA|Neil Armstrong   ...| 7/20/69|    0:05|    5.0|Jettison suit bac...|  Apollo 11|
|[565f1dd8d4c6d9a9...|    USA|Allen Bean       ...|11/19/69|    3:39|  219.0|Collected 75.6 lb...|  Apollo 12|
|[565f1dd8d4c6d9a9...|    USA|Allen Bean       ...|11/20/69|    3:48|  228.0|Retrieved parts o...|  Apollo 12|
|[565f1dd8d4c6d9a9...|    USA|Allen Bean       ...|11/20/69|    0:05|    5.0|Jettison suit bac...|  Apollo 12|
|[565f1dd8d4c6d9a9...|    USA|Ed Mitchell     ...|  2/5/71|    4:48|  288.0|Collected 94.4 lb...|  Apollo 14|
|[565f1dd8d4c6d9a9...|    USA|Ed Mitchell     ...|  2/6/71|    4:34|  274.0|Sought but did no...|  Apollo 14|
|[565f1dd8d4c6d9a9...|    USA|Ed Mitchell     ...|  2/6/71|    0:05|    5.0|Jettison suit bac...|  Apollo 14|
|[565f1dd8d4c6d9a9...|    USA|         David Scott| 7/30/71|    0:33|   33.0|Standup EVA to sc...|  Apollo 15|
+--------------------+-------+--------------------+--------+--------+-------+--------------------+-----------+


scala> df.printSchema()
root
 |-- _id: struct (nullable = true)
 |    |-- $oid: string (nullable = true)
 |-- country: string (nullable = true)
 |-- crew: string (nullable = true)
 |-- date: string (nullable = true)
 |-- duration: string (nullable = true)
 |-- minutes: double (nullable = true)
 |-- purpose: string (nullable = true)
 |-- vehicle: string (nullable = true)
```


JSON
JSON
JSON
JSON
JSON
JSON
JSON
JSON
JSON
JSON
JSON
$sample

Resources

DOCS OPEN SOURCE COMPANY

Sign in

MongoDB University

CERTIFICATION ONLINE COURSES TRAINING

https://university.mongodb.com/courses/M233/about

Next Session

Start: 17 May 2016 at 17:00 UTC

End: 16 May 2017 at 17:00 UTC

CONTACT

```
{
   name          : "Daniel M Farrell",
   title         : "Solution Architect",
   location      : "Somewhere, CO",
   email         : "daniel.farrell@mongodb.com",
   fon           : "303/808-2225”
}
```