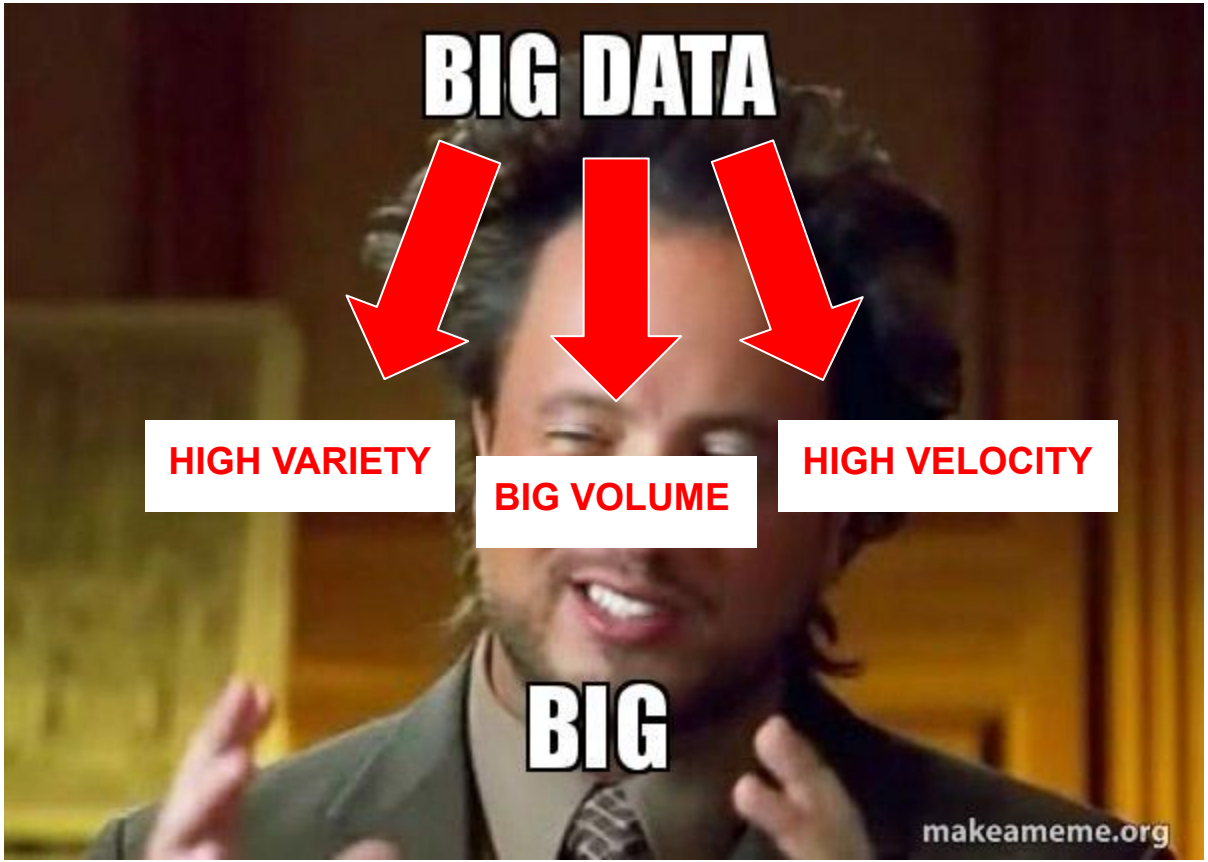
Introduction to Big Data in PySpark

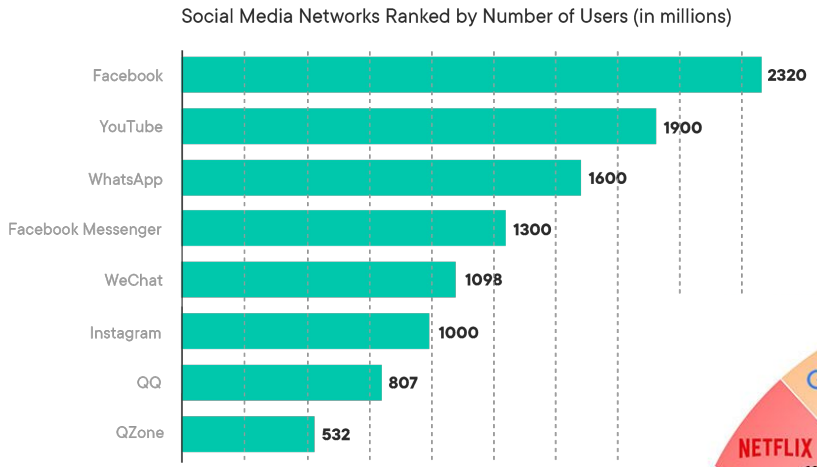
Road Map:

Objectives:

- Introduction to distributed systems for dealing with big data
- Align the relationships between Hadoop and Spark
- Differentiate between Spark RDDs and Spark Dataframes and when each is appropriate
- locate and explore the Spark.ML documentation
- code along to understand similarities between Pyspark and Python (Pandas/Sklearn)

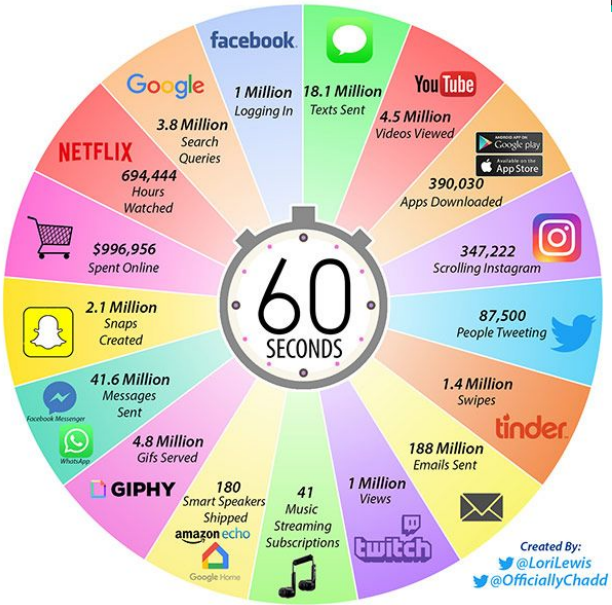

BIG DATA
BIG
makeameme.org


BIG DATA
HIGH VARIETY
BIG VOLUME
HIGH VELOCITY
BIG
makeameme.org



BIG VOLUME

HIGH VELOCITY



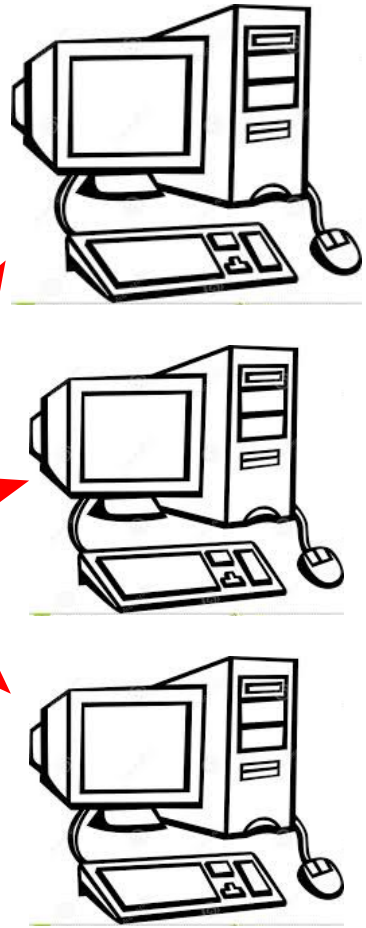
HIGH VARIETY

Where in the Process are we?

1. Business Question
2. Data Science Question(s)
3. Data Acquisition/Cleaning
4. EDA
5. Feature Selection and Engineering
6. Modeling
7. Model Evaluation
8. Addressing Business Questions

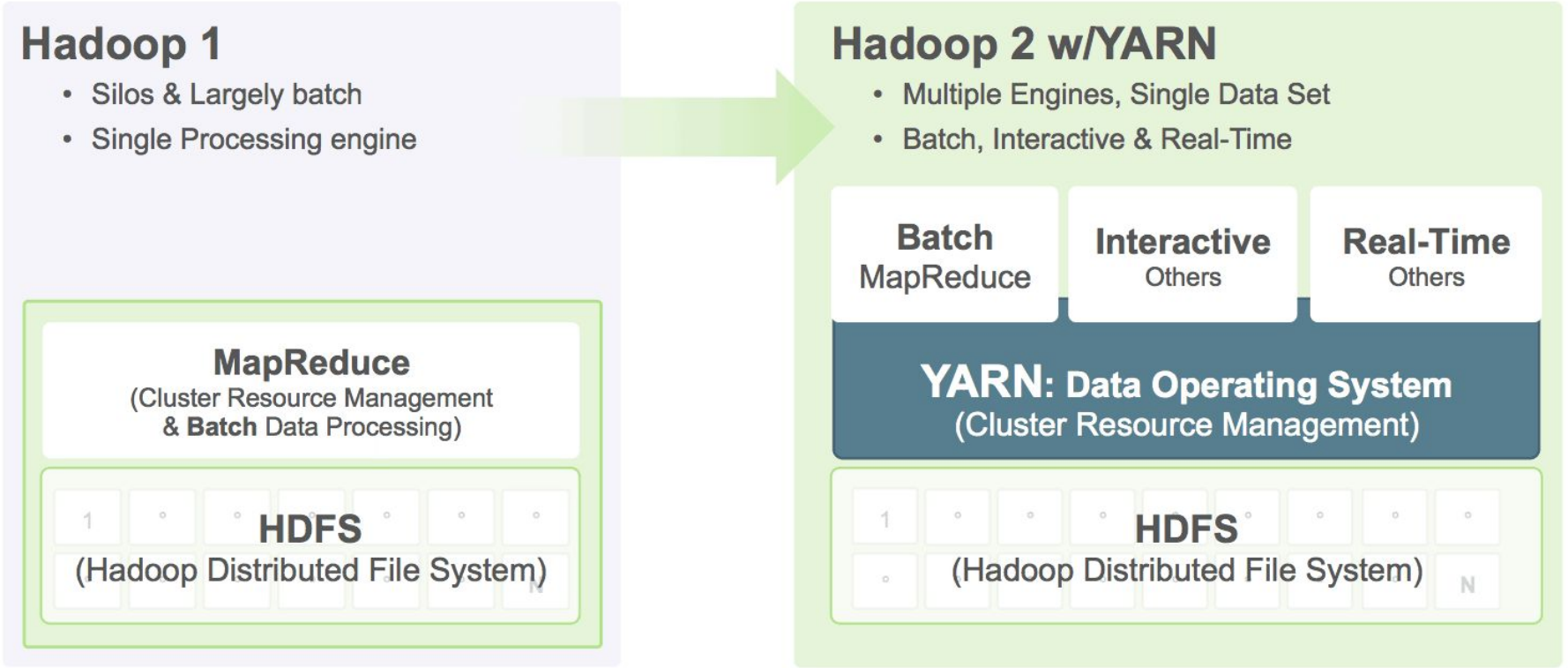
Where in the Process are we?

1. Business Question
2. Data Science Question(s)
3. Data Acquisition/Cleaning
4. EDA
5. Feature Selection and Engineering
6. Modeling
7. Model Evaluation
8. Addressing Business Questions

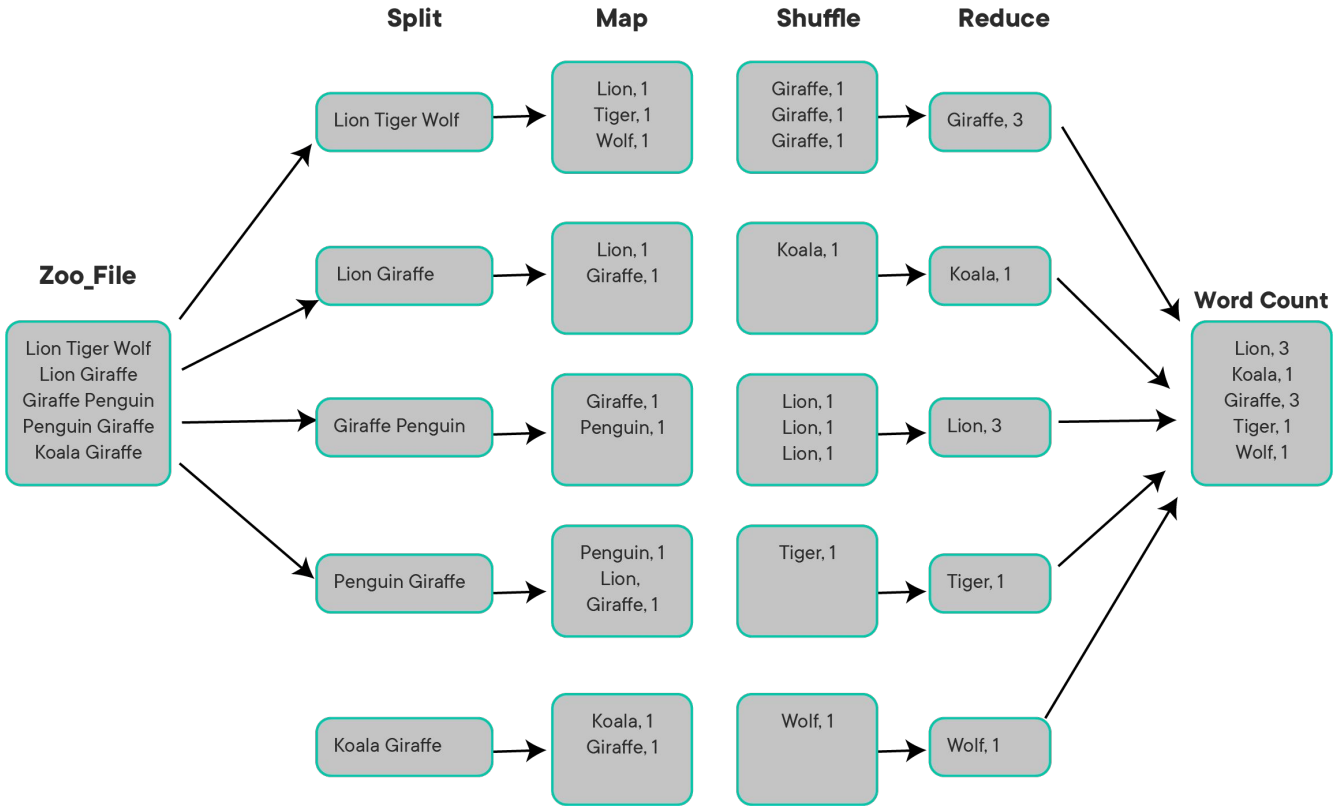


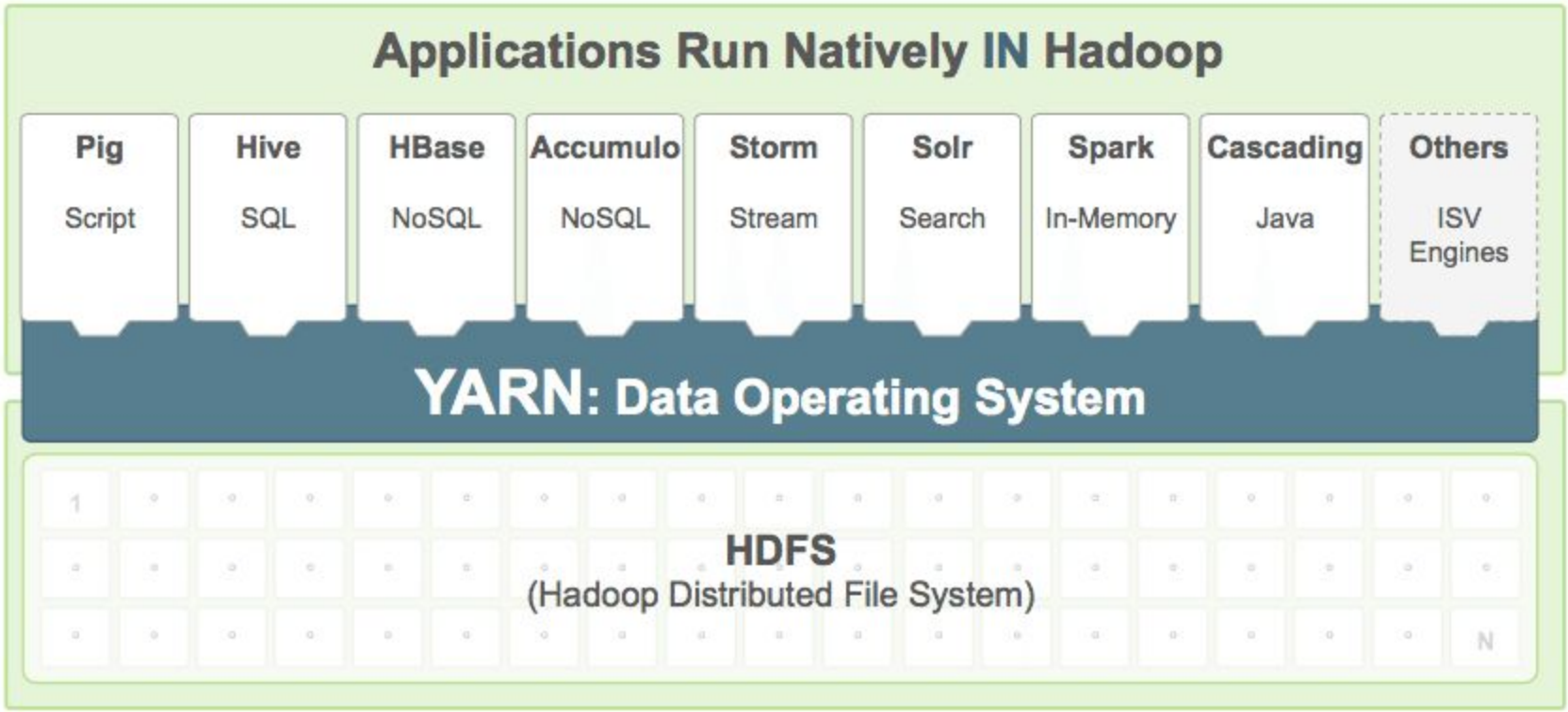

Visualization & Analytics
tableau
Microsoft
looker
R
Computation
APACHE Spark
HIVE
Microsoft Azure
DigitalOcean
hadoop
Storage
amazon S3
APACHE HBASE
Distribution & Data Warehouse
ORACLE EXADATA
Amazon EMR
cassandra
CLOUDERA

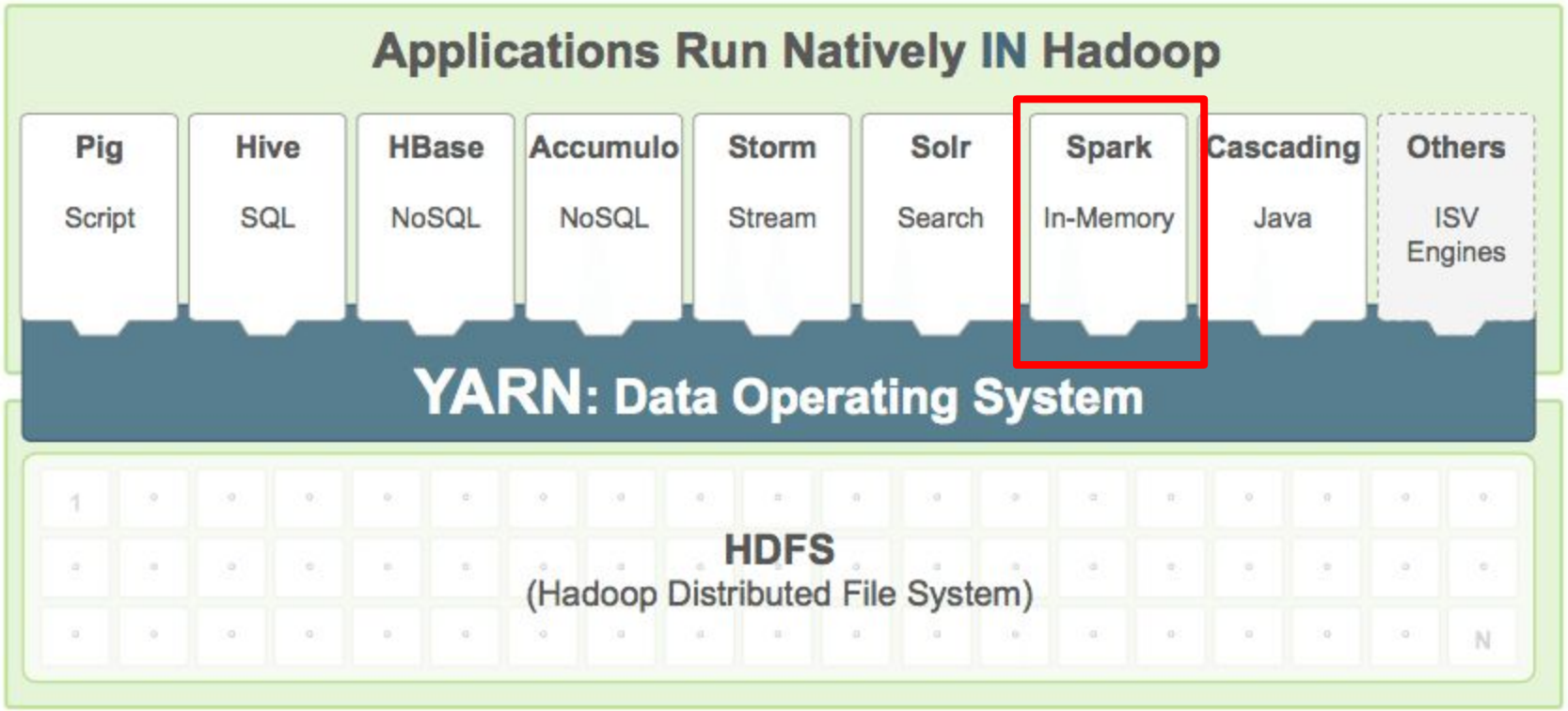

Visualization & Analytics
tableau
Microsoft
looker
R
Computation
APACHE Spark
HIVE
Microsoft Azure
DigitalOcean
hadoop
Storage
amazon S3
APACHE HBASE
Distribution & Data Warehouse
ORACLE EXADATA
Amazon EMR
cassandra
CLOUDERA

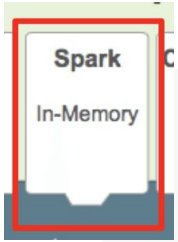

Hadoop 1
• Silos & Largely batch
• Single Processing engine
MapReduce
(Cluster Resource Management
& Batch Data Processing)
HDFS
(Hadoop Distributed File System)
Hadoop 2 w/YARN
• Multiple Engines, Single Data Set
• Batch, Interactive & Real-Time
Batch
MapReduce
Interactive
Others
Real-Time
Others
YARN: Data Operating System
(Cluster Resource Management)
HDFS
(Hadoop Distributed File System)

MapReduce


Split
Map
Shuffle
Reduce
Zoo_File
Lion Tiger Wolf
Lion Giraffe
Giraffe Penguin
Penguin Giraffe
Koala Giraffe
Lion Tiger Wolf
Lion Giraffe
Giraffe Penguin
Penguin Giraffe
Koala Giraffe
Lion, 1
Tiger, 1
Wolf, 1
Lion, 1
Giraffe, 1
Giraffe, 1
Penguin, 1
Penguin, 1
Lion,
Giraffe, 1
Koala, 1
Giraffe, 1
Giraffe, 1
Giraffe, 1
Giraffe, 1
Koala, 1
Lion, 1
Lion, 1
Lion, 1
Tiger, 1
Wolf, 1
Giraffe, 3
Koala, 1
Lion, 3
Tiger, 1
Wolf, 1
Word Count
Lion, 3
Koala, 1
Giraffe, 3
Tiger, 1
Wolf, 1


Applications Run Natively IN Hadoop
Pig
Script
Hive
SQL
HBase
NoSQL
Accumulo
NoSQL
Storm
Stream
Solr
Search
Spark
In-Memory
Cascading
Java
Others
ISV Engines
YARN: Data Operating System
1
HDFS
(Hadoop Distributed File System)
N


Applications Run Natively IN Hadoop
Pig
Script
Hive
SQL
HBase
NoSQL
Accumulo
NoSQL
Storm
Stream
Solr
Search
Spark
In-Memory
Cascading
Java
Others
ISV Engines
YARN: Data Operating System
HDFS
(Hadoop Distributed File System)



Apache Spark Ecosystem

Spark SQL + DataFrames

Streaming

MLlib
Machine Learning

GraphX
Graph Computation

Spark Core API

R

SQL

Python

Scala

Java

Reminder:

Dictionary

Search for a word

API

noun COMPUTING

a set of functions and procedures allowing the creation of applications that access the features or data of an operating system, application, or other service.

Translations, word origin, and more definitions

Spark Data Objects

Distribute collection of JVM objects

Functional Operators (map, filter, etc.)

Distribute collection of Row objects

Expression-based operations and UDFs

Logical plans and optimizer

Fast/efficient internal representations

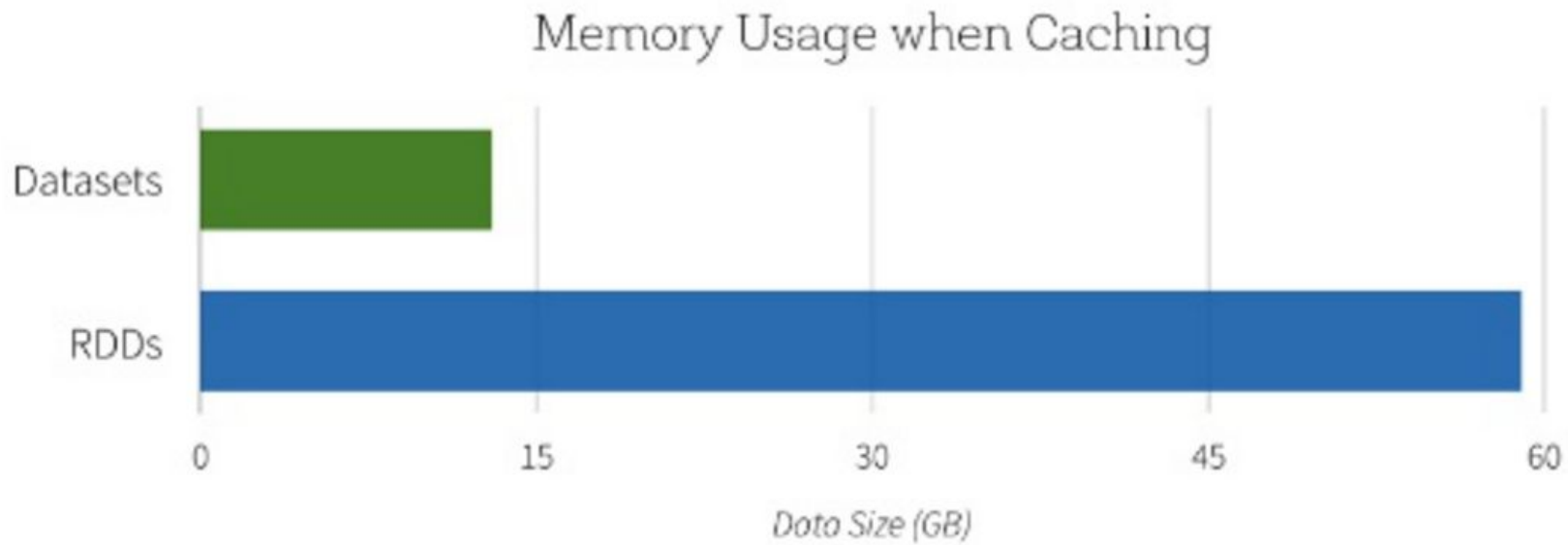
Internally rows, externally JVM objects

Almost the "Best of both worlds": **type safe + fast**

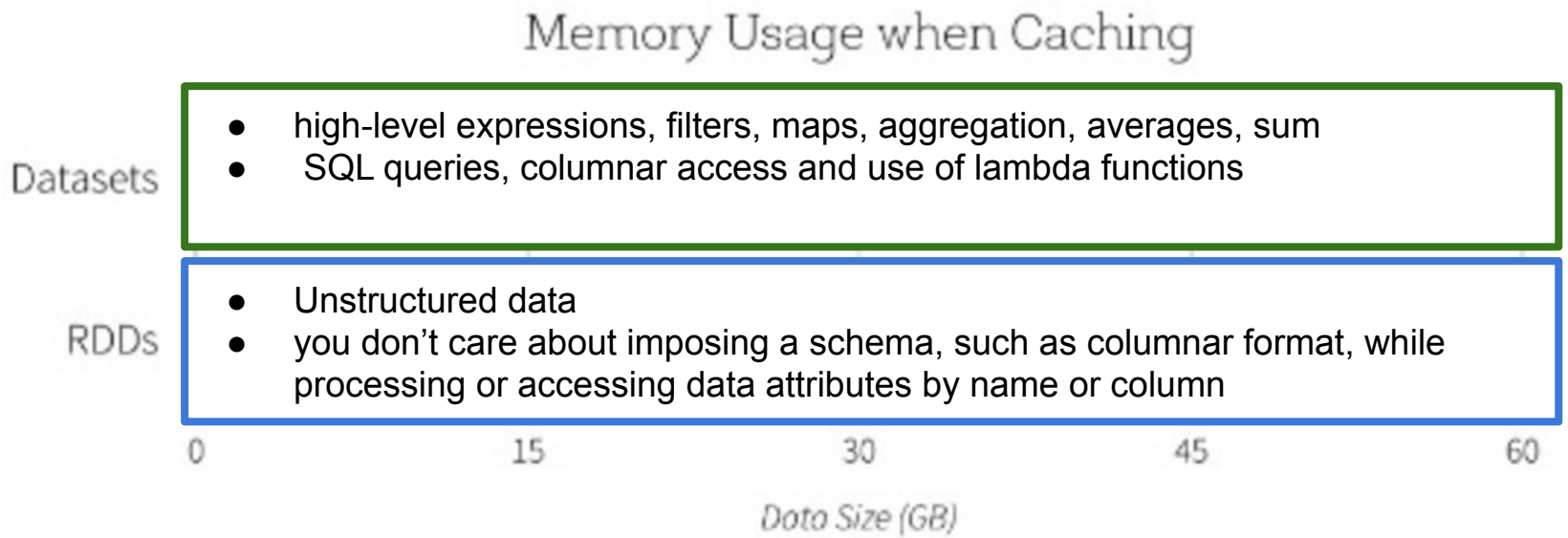
But slower than DF
Not as good for interactive analysis, especially Python

databricks

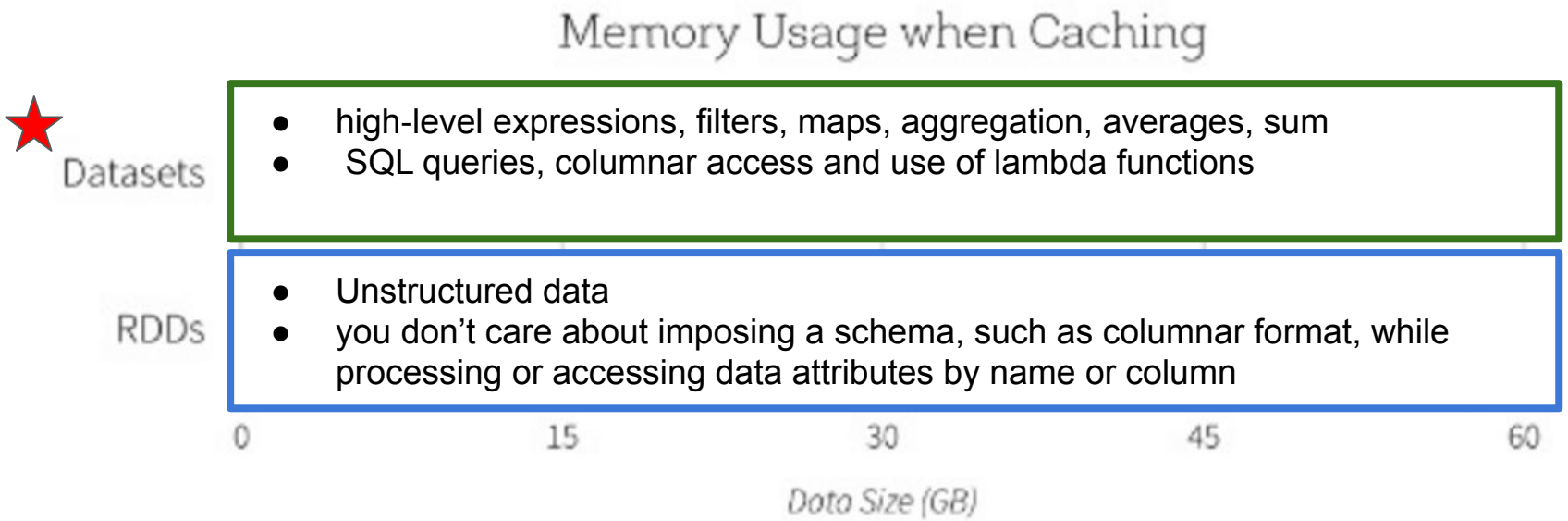
Space Efficiency



Space Efficiency



Space Efficiency



- high-level expressions, filters, maps, aggregation, averages, sum
- SQL queries, columnar access and use of lambda functions

- Unstructured data
- you don’t care about imposing a schema, such as columnar format, while processing or accessing data attributes by name or column

DataFrame or RDD?

1. You are grabbing live tweets about the the 2020 Political Election.

2. You have an RDD of data that you wish to use to build a predictive model. Should you leave it as an RDD or transform it to a DataFrame?

3. You want to access audio and video stored on your HDFS

Lets Code