BigData – Spark – Processing

discover Spark Core & Sql

Arnaud Nauwynck

Outline

Example RAW to LAKE transformations

Explanation step-by-step

Dataset

Parallel Distribution

Reminder: Spark RAW to LAKE samples

Typical RAW to LAKE as Spark Java code

Reminder

read

```
spark.read
    .format(« csv »)
    .option(«schema », « col1 type1, … colN typeN »)
    .load(« hdfs://raw/team/domain/table/date=2022-10-12 »)
```

transform

```
    .as(Encoder.bean(Bean.Class)
    .map(bean ->  transformBean(bean) )
    .toDF()
```

write

```
    .repartition(2, « col1 »)
    .sortWithinPartition(« col1, col2, col3 »)
    .write
    .format(« parquet »)
    .save(« hdfs://lake/team/domain/table/date=2022-10-22 »);
```

Reminder

Typical RAW to LAKE processing with Spark as SQL code

write

transform

read

transform

read

write

```
INSERT OVERWRITE
       lake_team_domain.table
SELECT /* +REPARTITION(col1, 2) */
    col1, col2,
    udf_func1(col3, col4)  as col3,
    udf_func2(col4, col5)  as col4,
    ..
FROM
       raw_team_domain.table
JOIN
       lake_anotherTeam_domain.anotherTable x ON x.ID=id
WHERE date='2022-10-22' AND ..
SORT BY col1, col2, col3     -- idem sortWithinPartition
```

Reminder

Example of LAKE Aggregation

```
INSERT OVERWRITE
      lake_team_domain.table
SELECT * FROM (
    SELECT * FROM table1 WHERE ..
  UNION
    SELECT * FROM table2 WHERE ..
  UNION
    SELECT * FROM table3 WHERE ..
  UNION
    SELECT * FROM table4 WHERE ..
)
SORT BY col1, col2, col3     -- idem sortWithinPartition
```

Example of « latest value » cristalisation analytical query « over(partition by) »

Reminder

```
INSERT OVERWRITE
       lake_team_domain.table
SELECT
   col1,col2,…. colN    -- idem * EXCEPT rank   (cf issue SPARK-33164)
FROM (
   SELECT *,
      RANK() OVER (PARTITION BY id ORDER BY update_time DESC) as rank
   FROM lake_team_domain.event_table
)
WHERE rank=1
SORT BY col1, col2, col3     -- idem sortWithinPartition
```

Step-by-Step explained

Reminder

Typical RAW to LAKE as Spark Java code

read
Step **1/4**

```
spark.read
    .format(« csv »)
    .option(«schema », « col1 type1, … colN typeN »)
    .load(« hdfs://raw/team/domain/table/date=2022-10-12 »)
```

transform
Step **2/4**

```
    .as(Encoder.bean(Bean.Class)
    .map(bean ->  transformBean(bean) )
    .toDF()
```

Step **3/4**

```
    .repartition(3, « col1 »)
    .sortWithinPartition(« col1, col2, col3 »)
```

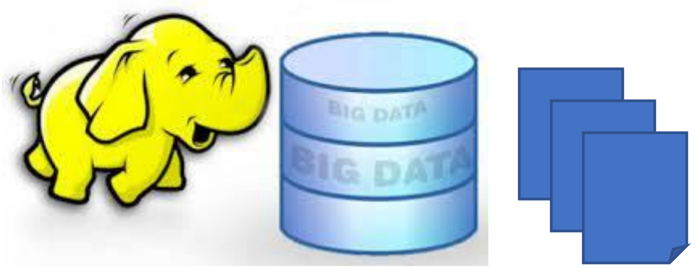
write
Step **4/4**

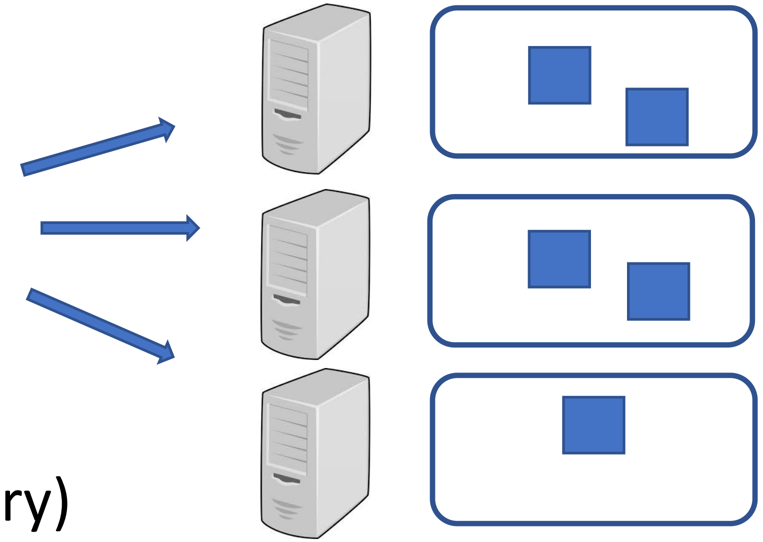
```
    .write
    .format(« parquet »)
    .save(« hdfs://lake/team/domain/table/date=2022-10-22 »);
```

RAW to LAKE – Step 1/4: read to Dataset

read

```
Dataset<Row> ds =
    spark.read
    .format(« csv »)
    .option(«schema », « col1 type1, … colN typeN »)
    .load(« hdfs://raw/team/domain/table/date=2022-10-12 »)
```



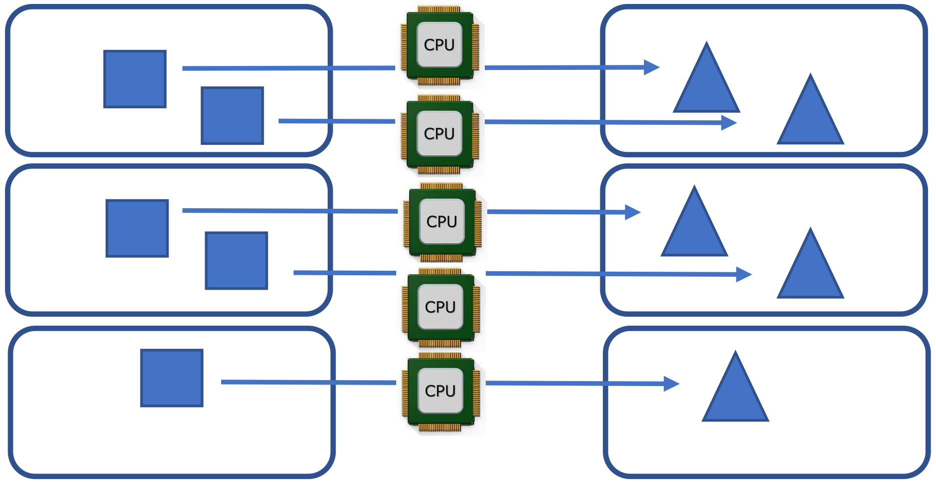


Input = Files
(scanned from Directory)
Distributed Read from Storage

Result =
Distributed Parts in-memory

RAW to LAKE – Step 2/4 : Transform Dataset

transform Dataset<Row> ds2 = ds.map(row -> transformData(row))

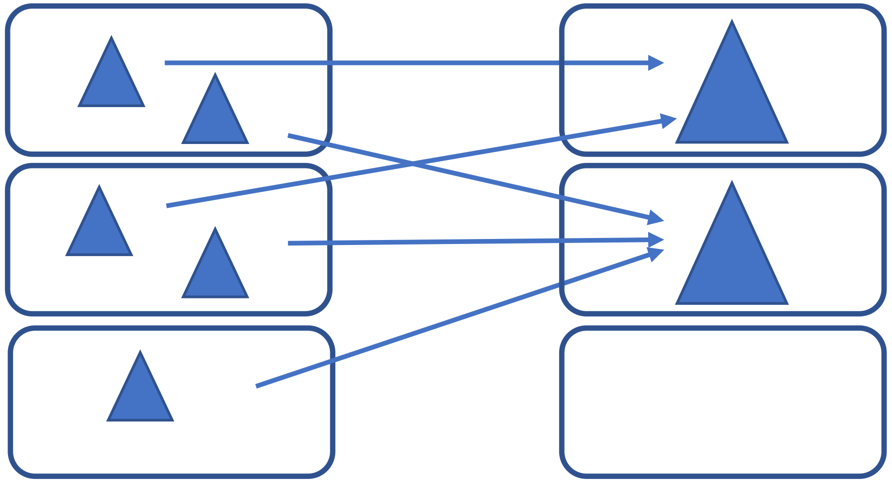


Distributed Processing to compute each new part

RAW to LAKE – Step 3/4 : Repartition Dataset

transform

```
Dataset<Row> ds3 = ds2
   .repartition(2, « col1 »)
   .sortWithinPartition(« col1, col2, col3 »)
```

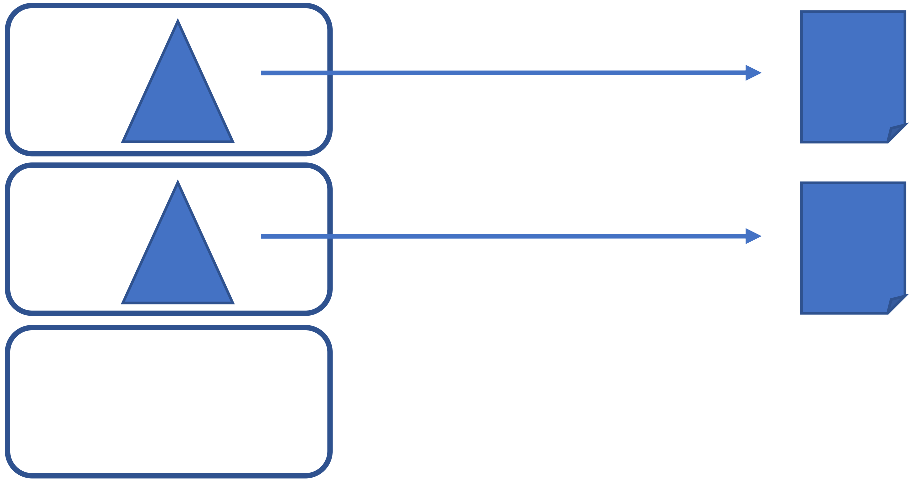


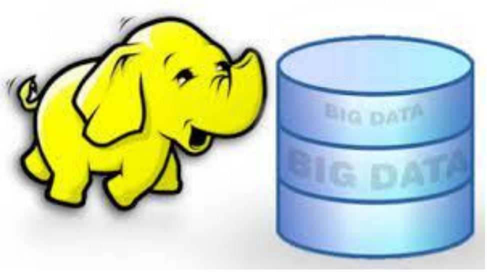
Network Shuffle to distribute / group / sort data

RAW to LAKE – Step 4/4 : Write Dataset

write

```
ds3.write
  .fomat(« parquet »)
  .save(« hdfs://lake/trig/domain/table »)
```



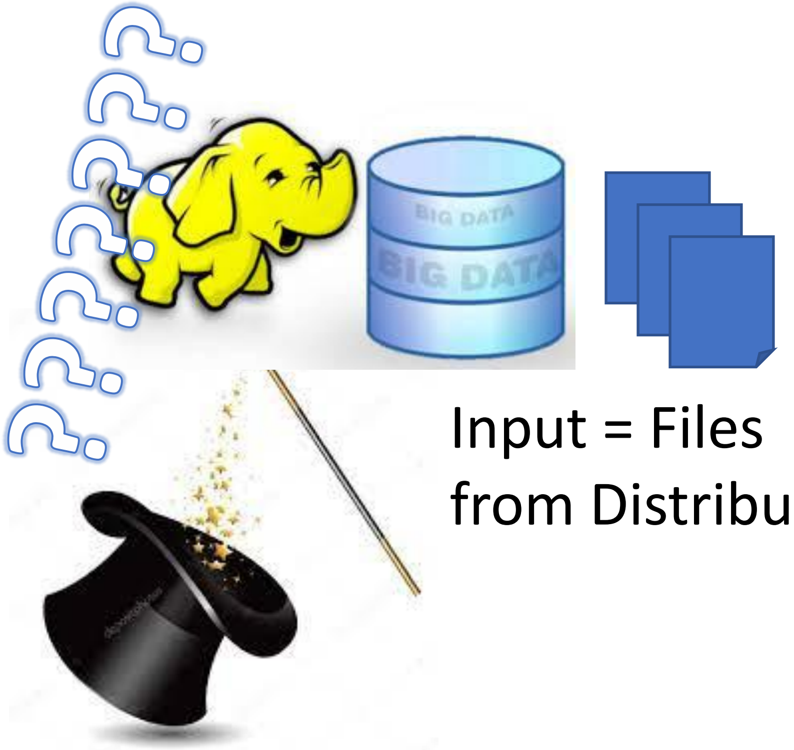


Distributed Write Dataset to Storage

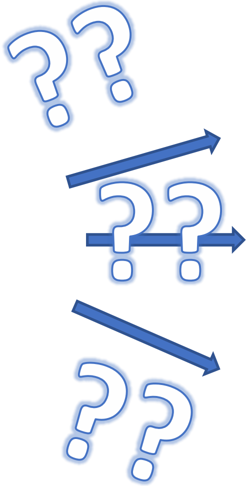
How it works ?

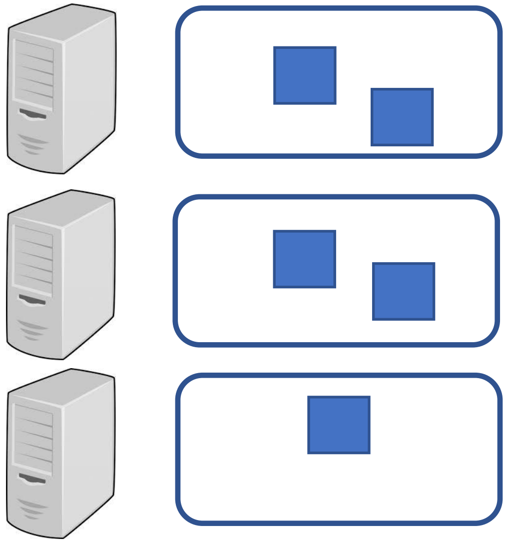
Zooming more ..

RAW to LAKE – Step 1/4: read to Dataset



Input = Files
from Distributed Storage



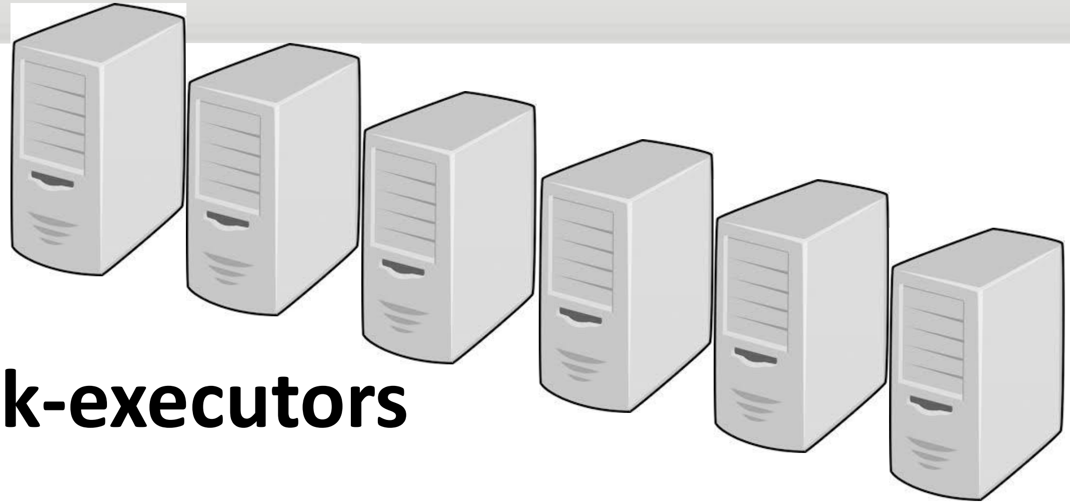


Result = Distributed Parts in-memory

?? **How to Assign** N x Files – P x blocks ⇔ to Q x Executors **??**
+ Retry on Error ?? + Communicate more ??

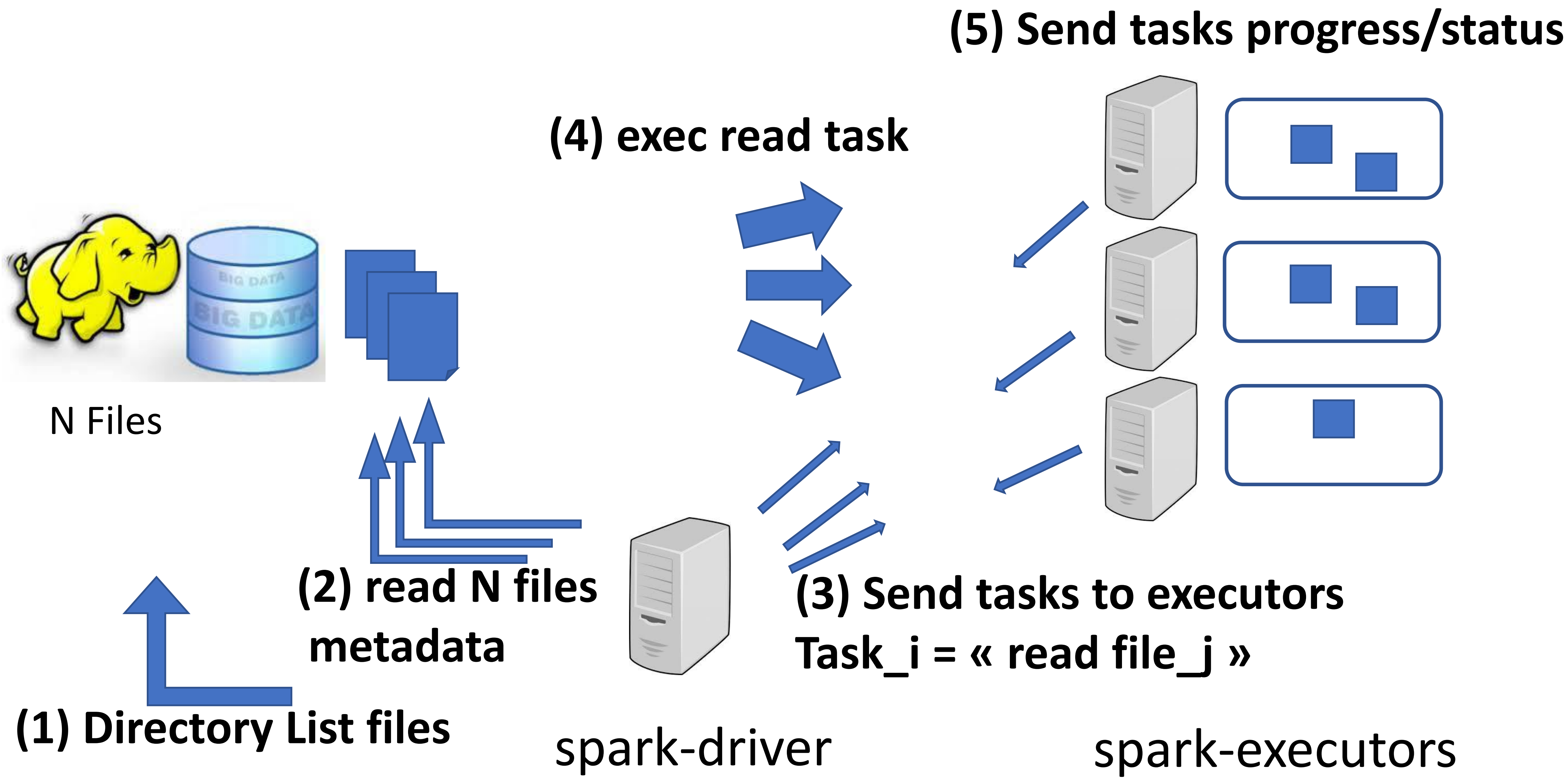
Analogy : How to play music ?
(N musicians without 1 Conductor != 1 Orchestra)

spark-driver

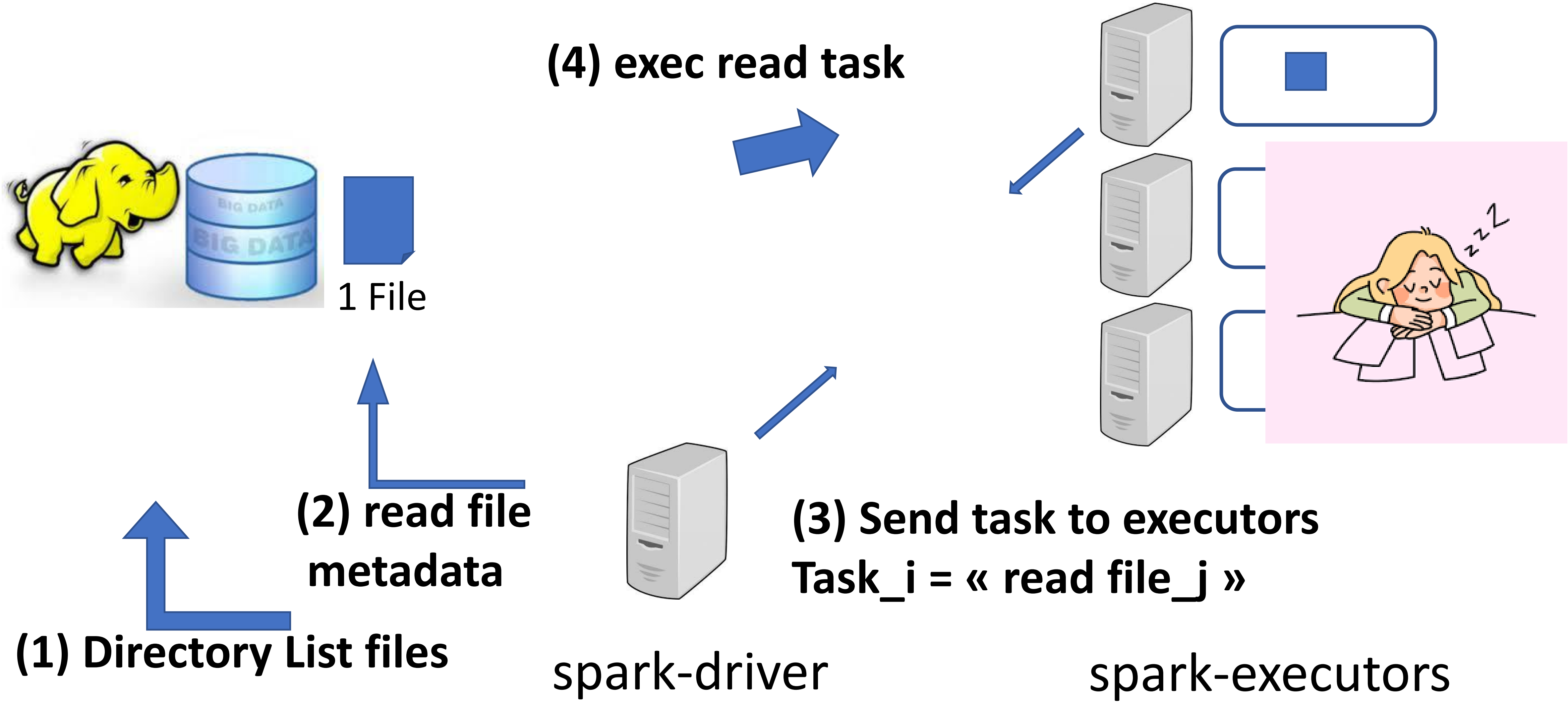


spark-executors

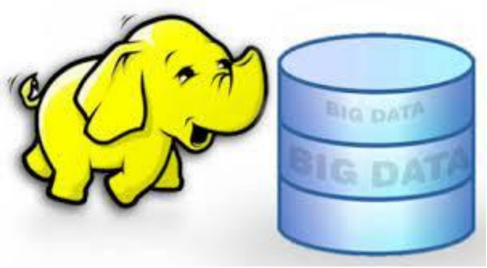
Read N Files – assign Tasks to Executors

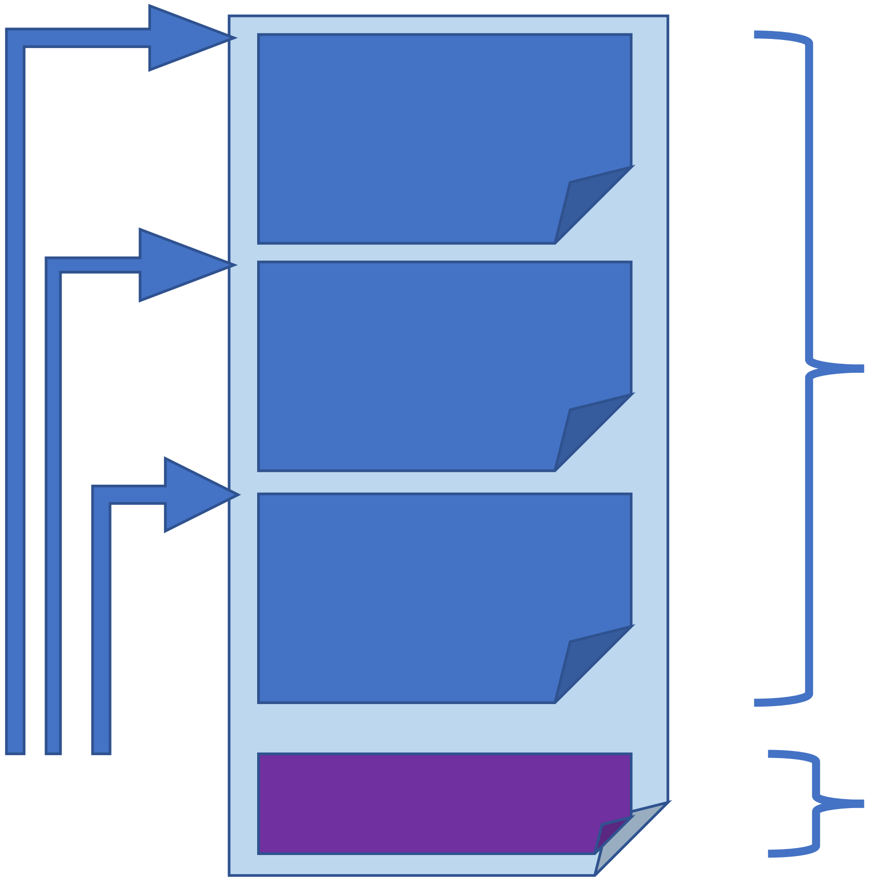


Remark [1/2] on Parallelism
only 1 File -> only 1 Task



Remark [2/2] on Parallelism
Splittable File format (parquet).. Like dir





N x Blocks
(usually 256 Mo)
independent, read at offset

offsets
...Like dir

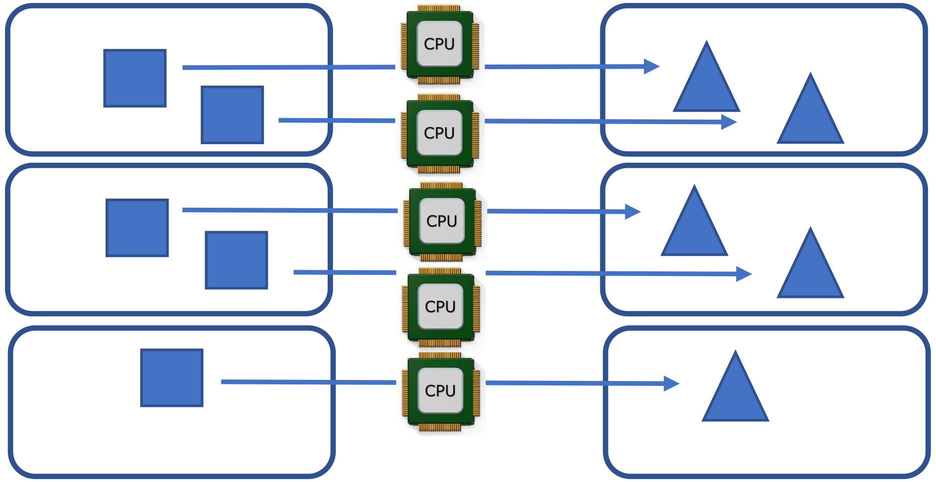
File metadata
= schema + N blocks infos
(offset + stats)

1 splittable file

Zooming RAW to LAKE – Step 2/4 : Transform Dataset

transform

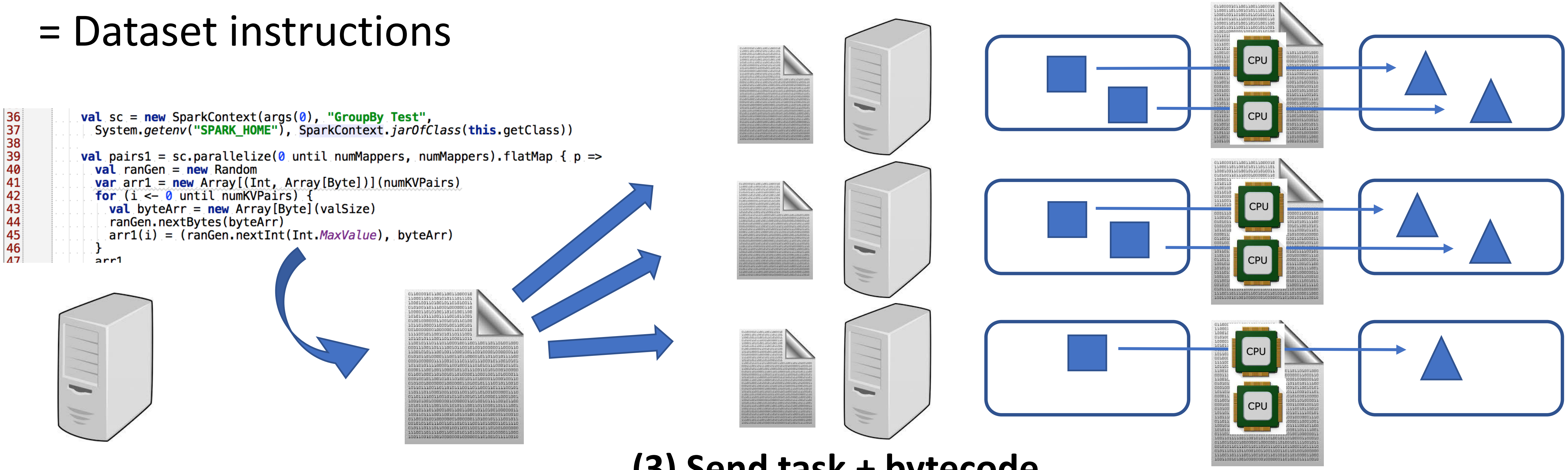
```
Dataset<Row> ds2 = ds.map(row ->  transformData(row) )
```



Distributed Processing to compute each new part

WholeStageCodeGen

Program
= Dataset instructions



(3) Send task + bytecode to spark-executors

(4) Execute tasks

(1) Generate java code

(RDD Spark sub-class « WholeStageCodeGen$i »)

(2) Compile Bytecode

Advanced Transform …
using Row -> Java ->map()-> Java -> Row

transform

```
ds.as( Encoders.bean(InputBean.class) )
   .toDF()
```

```
class InputBean {
  …
}
```

```
class OutputBean {
  …
}
```

```
OutputBean transformBean(InputBean b) {
    // complex transform in java
    return new OutputBean(… );
}
```

Explained as().map().toDF()

ds.as(Encoders.bean(InputBean.Class))
  .map(bean -> transformBean(bean))
  .toDF()

=

Dataset<**Row**> ds = …

// convert Row->Bean
Dataset<**InputBean**> dsInputBean =
    ds**.as**(Encoder.bean(InputBean.class))

// map
Dataset<**OutputBean**> dsOut =
    dsInputBean**.map**(bean -> transformBean(bean))

// convert OutputBean -> Row
Dataset<**Row**> df = dsOut.**toDF**();

Converting Tabular SQL Row to Java Beans

```
CREATE TABLE MyTable (
  field1 Int,
  field2 String
)
```

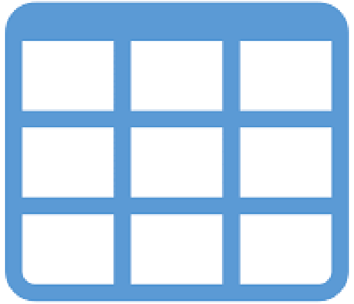
```
public class MyBean {
  public int field1;
  public String field2;
}
```

encoder = Encoders.bean(MyBean.class)

Dataset<Row> df = …

df.as(encoder)

Dataset<MyBean> ds = …

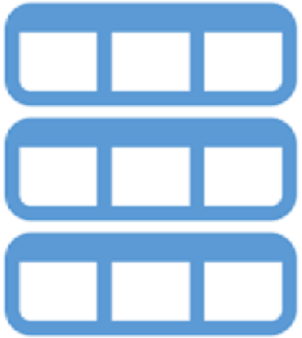


ds.**toDF**()

Object[0]

Object[1]

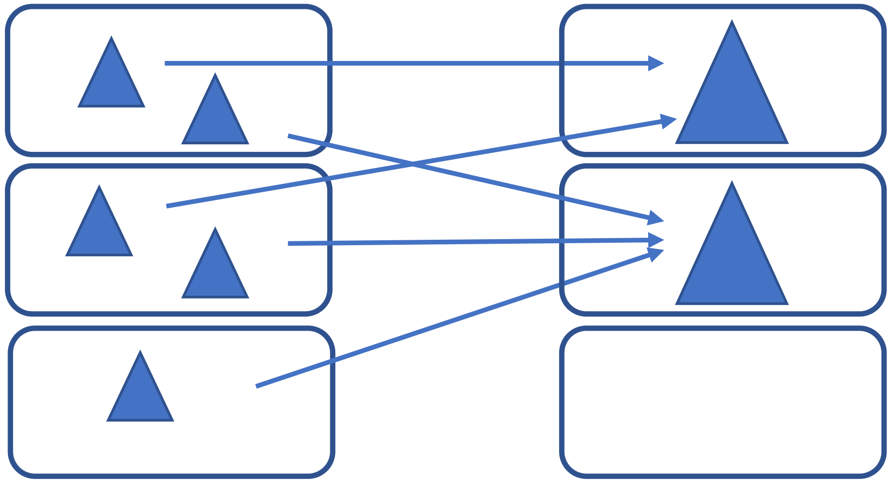
Object[2]



RAW to LAKE – Step 3/4 : Repartition Dataset

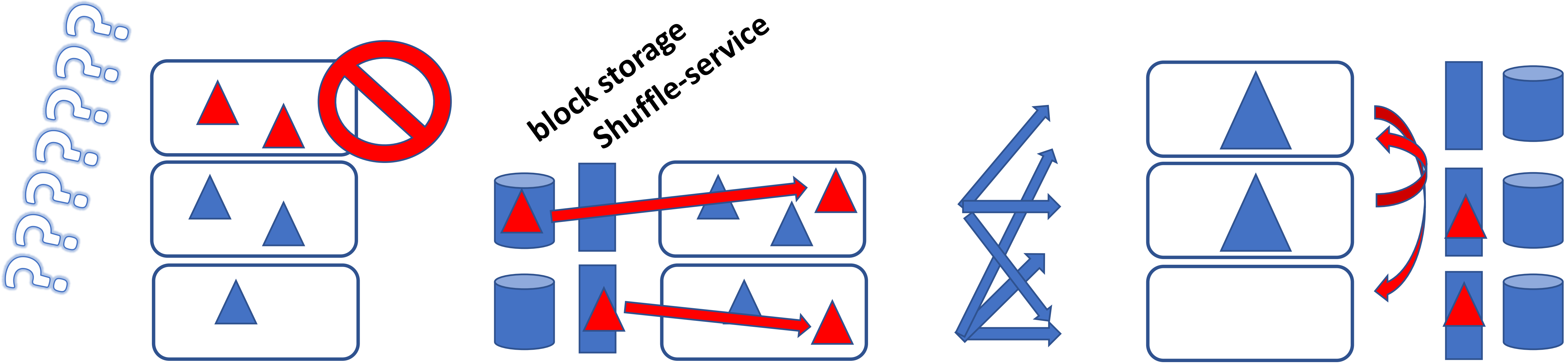
transform

```
Dataset<Row> ds3 = ds2
   .repartition(2, « col1 »)
   .sortWithinPartition(« col1, col2, col3 »)
```



Network Shuffle to distribute / group / sort data

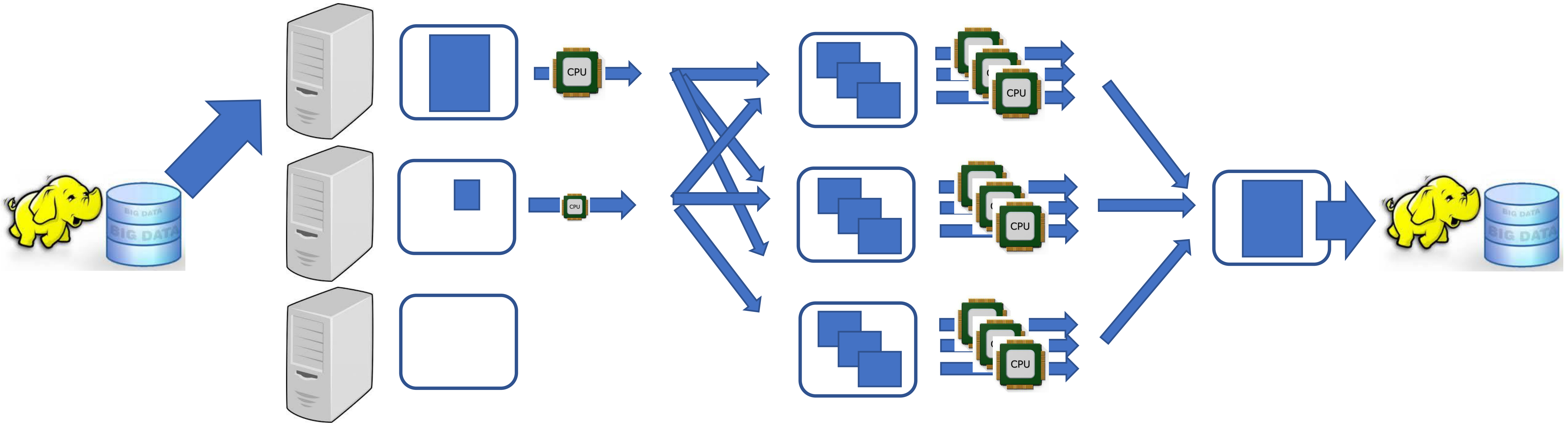
What if a Spark-Executor crash / Node stop ?



Recompute/recover lost blocks
From mem or disk
...using ShuffleService + storage

after shuffle operation,
=> Replicate (peer-to-peer)
/ Persist Blocks (mem and/or disk)
...using ShuffleService + storage

Example usage:
repartition(N).map(..).repartition(1)



INNEFFICIENT !!!
Undistributed
/ Badly Skewed Data

repartition(N)

EFFICIENT

repartition(1)

Avoid too many small files

Example transformation … in SQL

Reminder

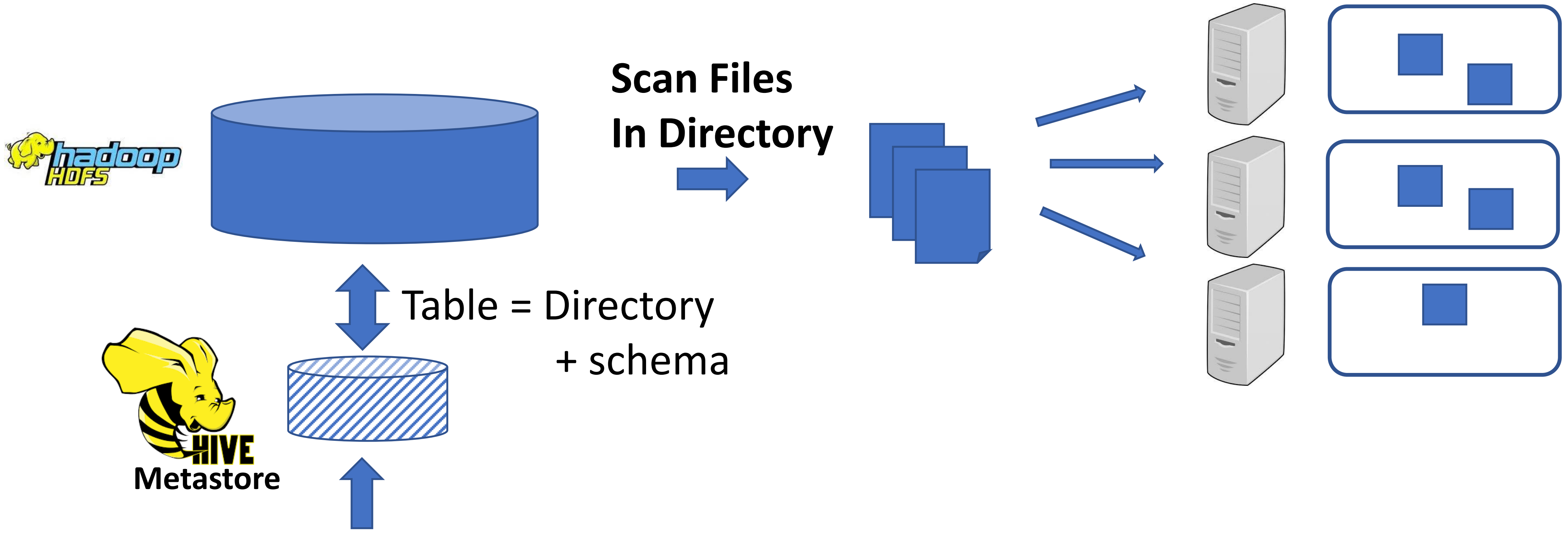
Typical RAW to LAKE as Spark SQL

write	INSERT OVERWRITE lake_team_domain.table
	SELECT /* +REPARTITION(col1, 2) */ col1, col2,
transform	udf_func1(col3, col4) as col3, udf_func2(col4, col5) as col4, ..
read	FROM raw_team_domain.table
transform	JOIN lake_anotherTeam_domain.anotherTable x ON x.ID=id
read	WHERE date='2022-10-22' AND ..
write	SORT BY col1, col2, col3 -- idem sortWithinPartition

Explained … SQL (-> Files) -> Dataset

read

```
SELECT ..
FROM
      raw_team_domain.table
```

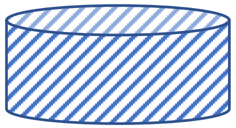


Input = SELECT .. from SQL Table

(Hive) MetaStore

Store ONLY metadatas (DDL + partitions)

Mapping SQL – Dirs+Files



DDL:

CREATE EXTERNAL TABLE **students** (
  socialSecuId: Int,
  firstName string, lastName string,
  birth: Date, …
) PARTITIONED BY (promo: Int)
STORED AS **parquet**
LOCATION 'hdfs://lake/students'

Location Dir + Partitions

« / »   = root hdfs://host/

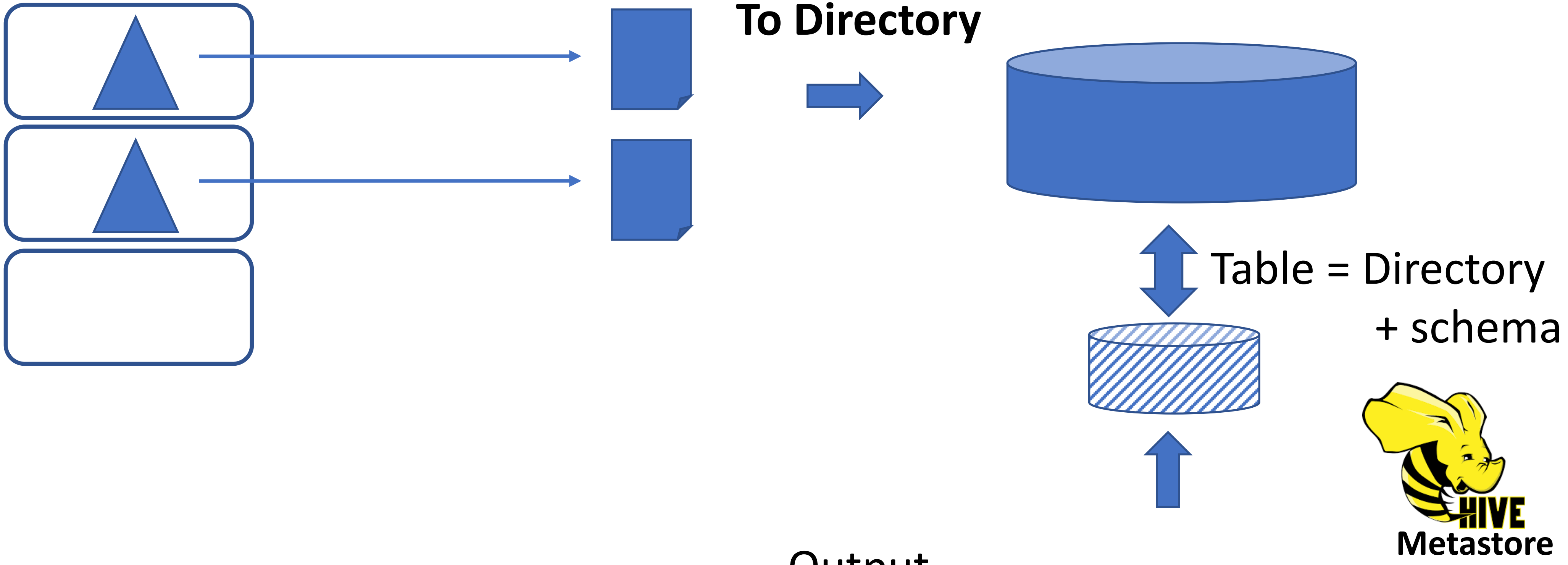
- **« /lake » (dir)**
  - **« /students » (table storage dir)**
    - **« /promo=2021 » (partition dir)**
    - **« /promo=2022 » (partition dir)**
      - **« file1.parquet» (data file)**
      - **« file2.parquet» (data file)**

Dataset -> INSERT SQL Table (-> Files)

write

INSERT OVERWRITE
lake_team_domain.table

Write Files
(append/overwrite)
To Directory



Output

SQL: INSERT INTO/OVERWRITE Table

More Java <-> Sql Interactions

Executing SQL from Java


```
for( int i = 0; i < 10; i++) {
   String sql = « SELECT * from db.table » + i;

   Dataset[Row] ds = spark.sql(sql);


   ..
}
```

NO imperative in SQL (cf PL/Sql extensions)
=> OK in java code : if, for(), …

Java DataSet as SQL View

SQL

Dataset[Row] ds = ..

ds.**createTemporaryView**(« **myview1** »)

spark.sql(« SELECT * FROM **myview1** »)

Java

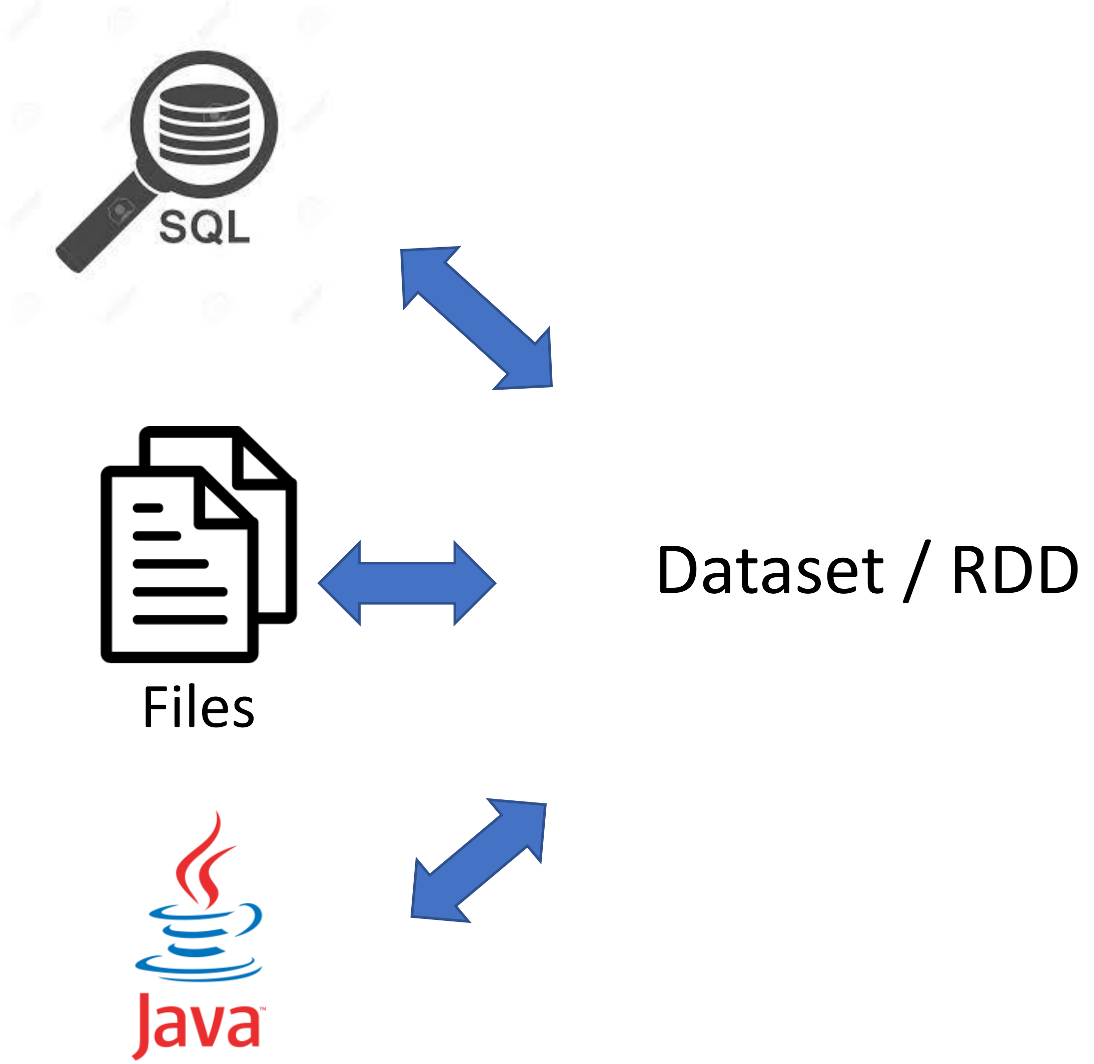
Calling Java from SQL : User-Defined Function

transform

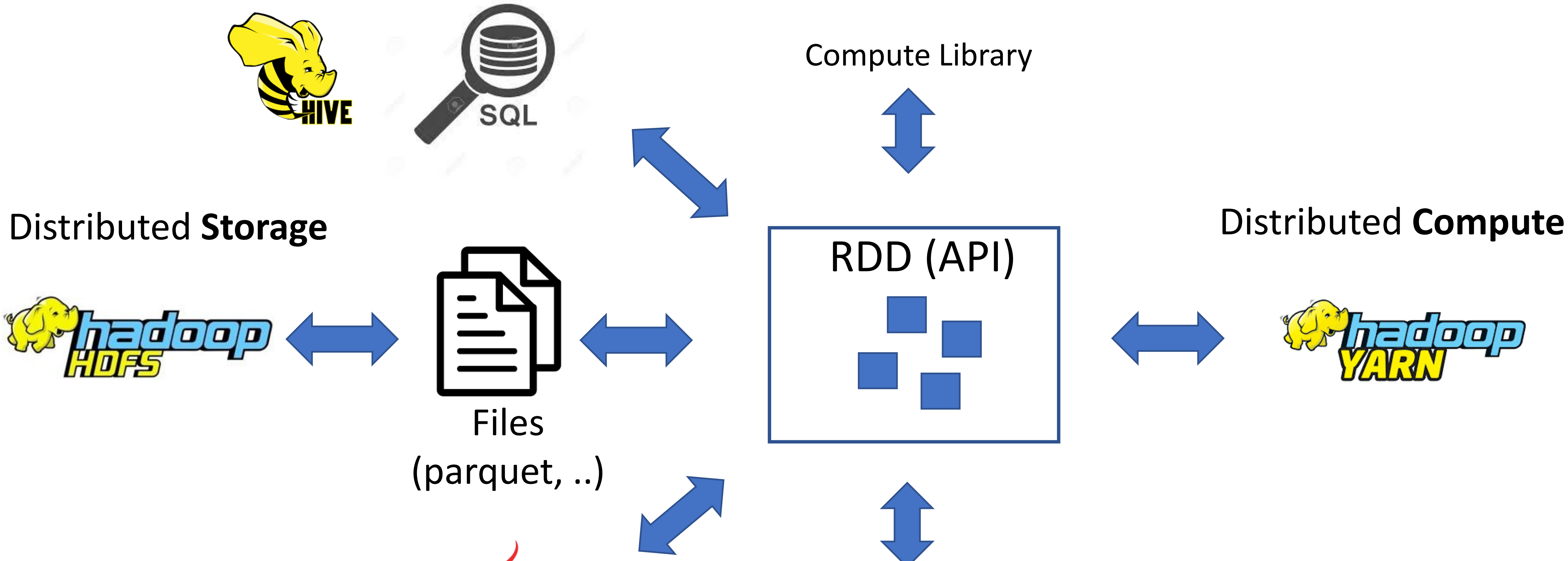
```
SELECT ..
   udf_func1(col3, col4)  as col3,
   udf_func2(col4, col5)  as col4,
```


```
int func1(int x, int y) { return x+y; }

spark.udf().register(« udf_func1",
     (UDF2<Integer,Integer, Integer>) ::func1,
     DataTypes.IntegerType);
```

Spark = Unified Sql-Files-Java



Spark : Unified Engine
(Distributed Storage, Distributed Compute)

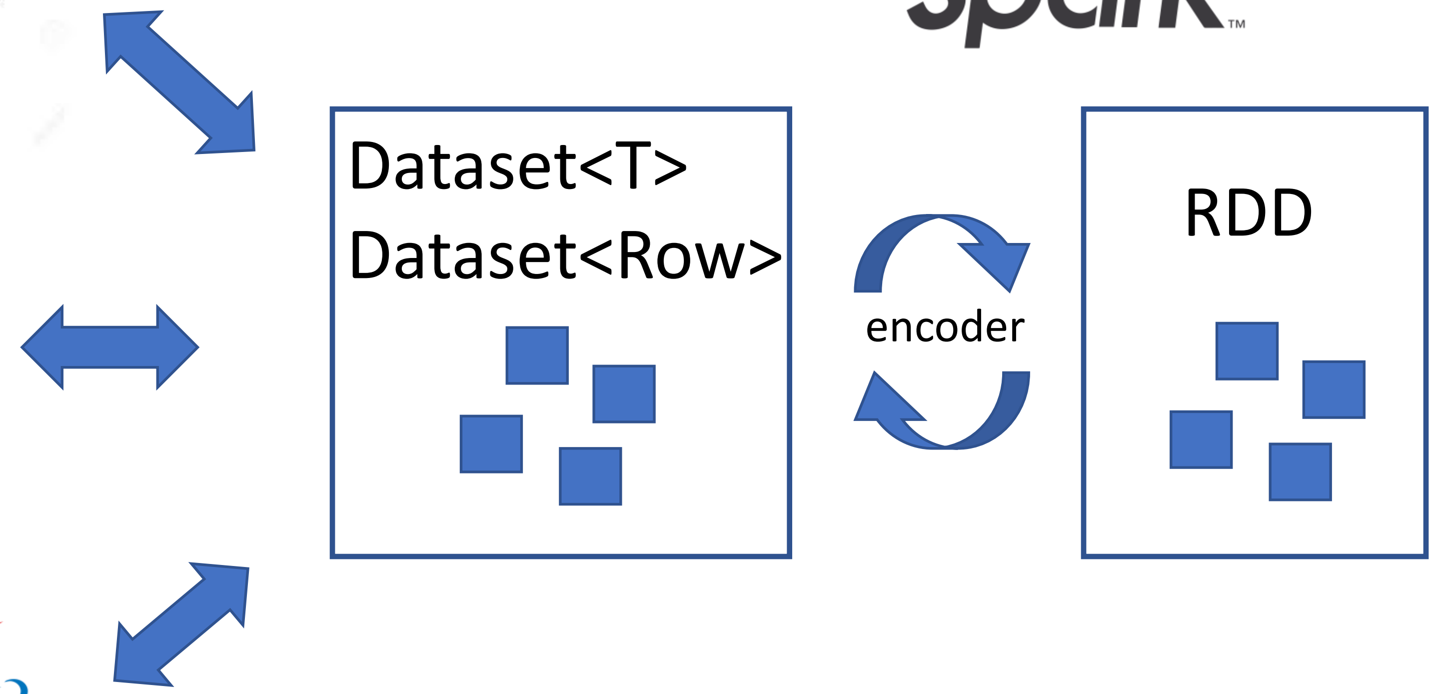


RDD is an abstract class API
=> Can plug any User-Defined implementations

Dataset API .... SQL Extensions

Column API
Expression API
(operator overload)

Lamba /Function
+ runtime **compile to bytecode**
(WholeStageCodeGen)



More Extensions: Hadoop FileSystem API

HDFS implements FileSystem

Distributed Storage API

java.io.File adapter

Java

abstract class **FileSystem {**
..read, write, list,
}

Amazon S3

Azure Data Lake Storage Gen2

ICEBERG

DELTA LAKE

More adapters

....

Spark rely on API
=> Can plug any implementations

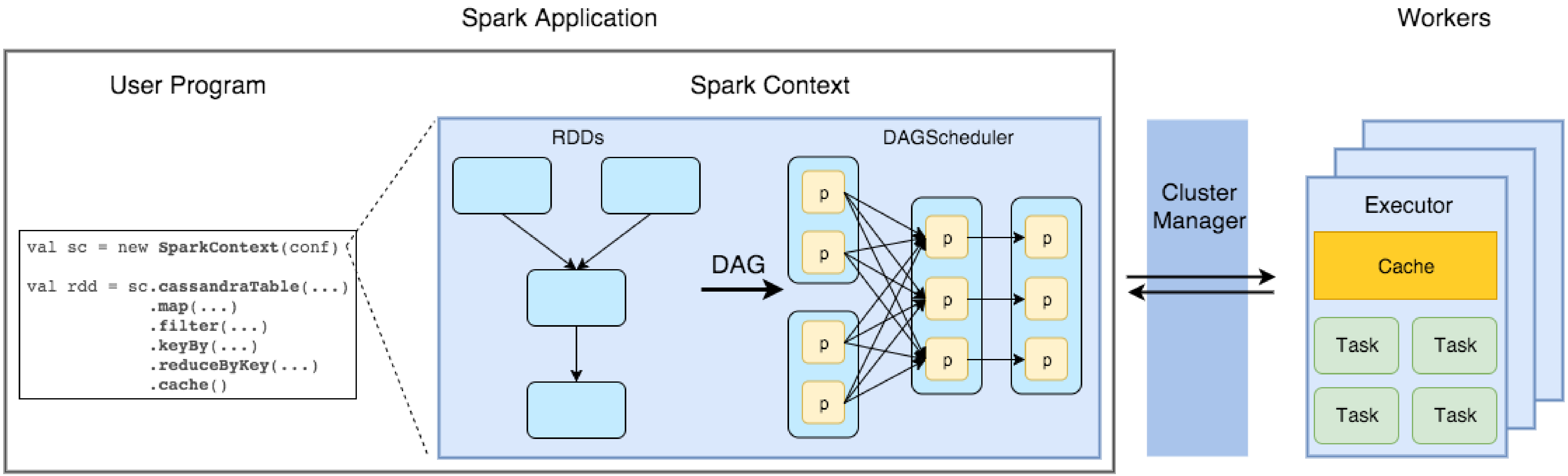
More Extensions: Cluster Scheduler API

Cluster Manager
Scheduler API

TaskScheduler (SPI)

abstract class **TaskScheduler {**
..start,stop,
submitTasks,cancelTasks,
notify Host-Executor-Task changes
}

Standalone Cluster



Spark Application
Workers
User Program
Spark Context
val sc = new SparkContext(conf)
val rdd = sc.cassandraTable(...)
.map(...)
.filter(...)
.keyBy(...)
.reduceByKey(...)
.cache()
RDDs
DAGScheduler
DAG
p
Cluster Manager
Executor
Cache
Task
Task
Task
Task

More « Extensions »
backport API to other Langages (Python, R)

python API

Java API

class DataFrame:

... hundred methods ...

def **method123** (self, x) -> Y
  self._jdf.method(x, self.sparkSession)

```
from py4j.java_gateway import JavaObject
from subprocess import Popen
from pyspark.context import SparkContext
..
pid = Popen(« spark-submit »  ...)
Socket(.. )
```

py4j

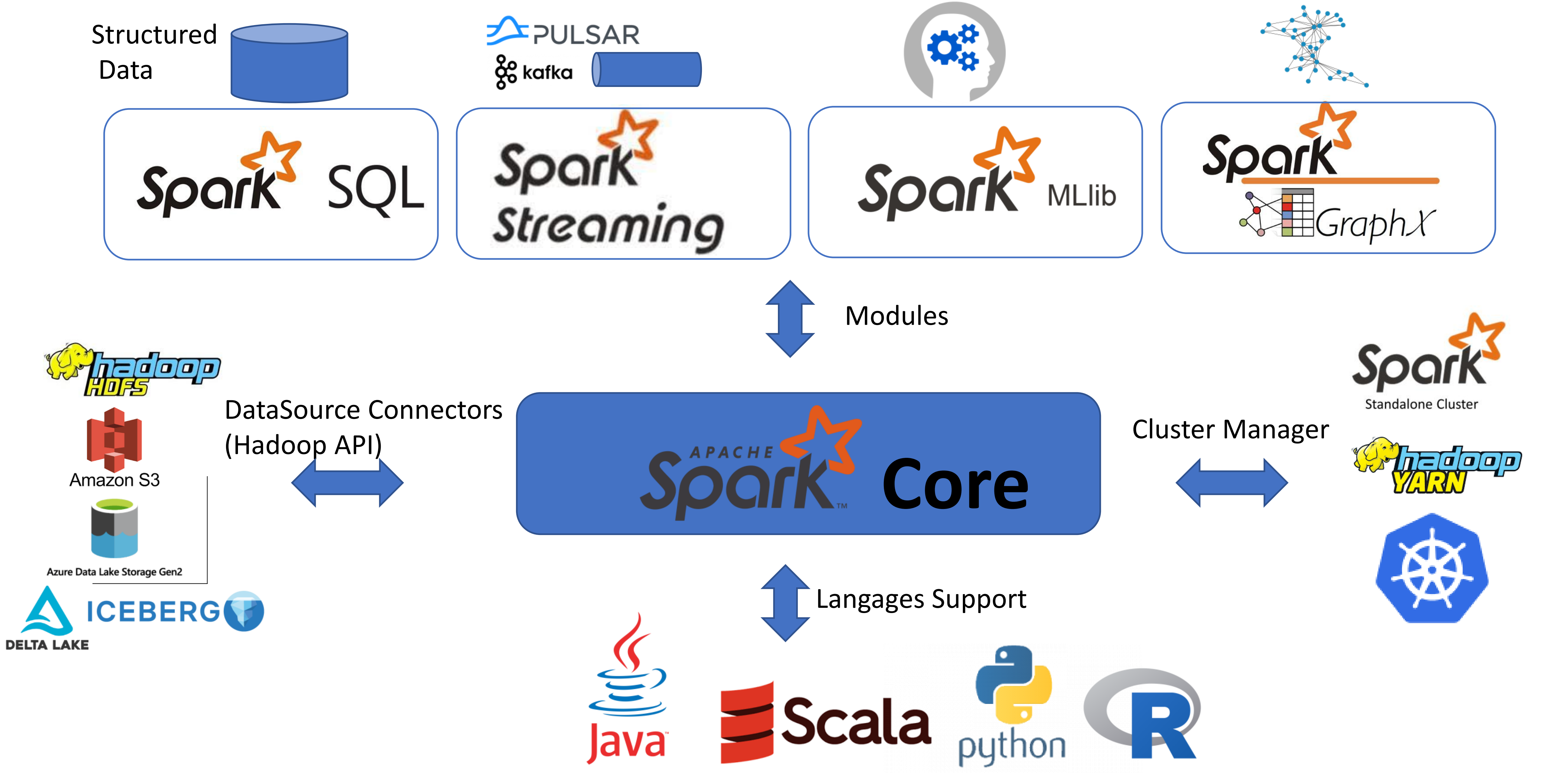
public class Dataset {

... hundred methods ...

public Y **method123** (X x) {
  ... internals sparkContext...
}

Scala **SparkContext**

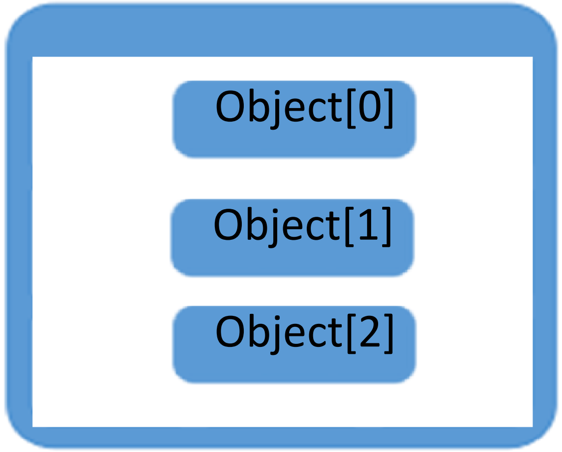
Spark-Core + …


Structured Data
PULSAR
kafka
Spark SQL
Spark Streaming
Spark MLlib
Spark GraphX
Modules
hadoop HDFS
Amazon S3
Azure Data Lake Storage Gen2
DELTA LAKE
ICEBERG
DataSource Connectors
(Hadoop API)
APACHE Spark Core
Cluster Manager
Spark Standalone Cluster
hadoop YARN
Langages Support
Java
Scala
python
R

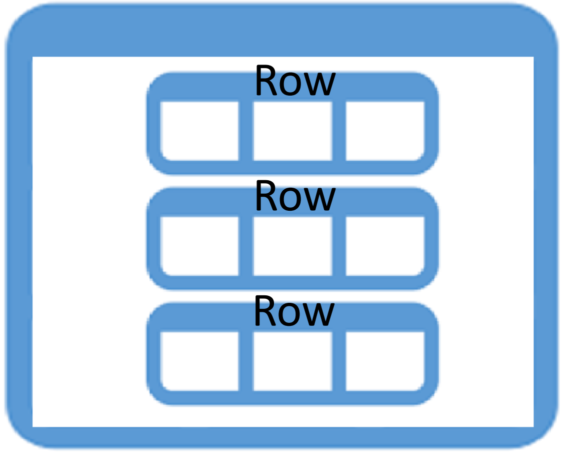
What are Dataset/RDD ?
How they work ?

DataSet ~ Set of <Data>

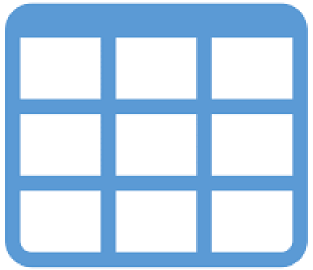
Dataset<UserDefinedClass>



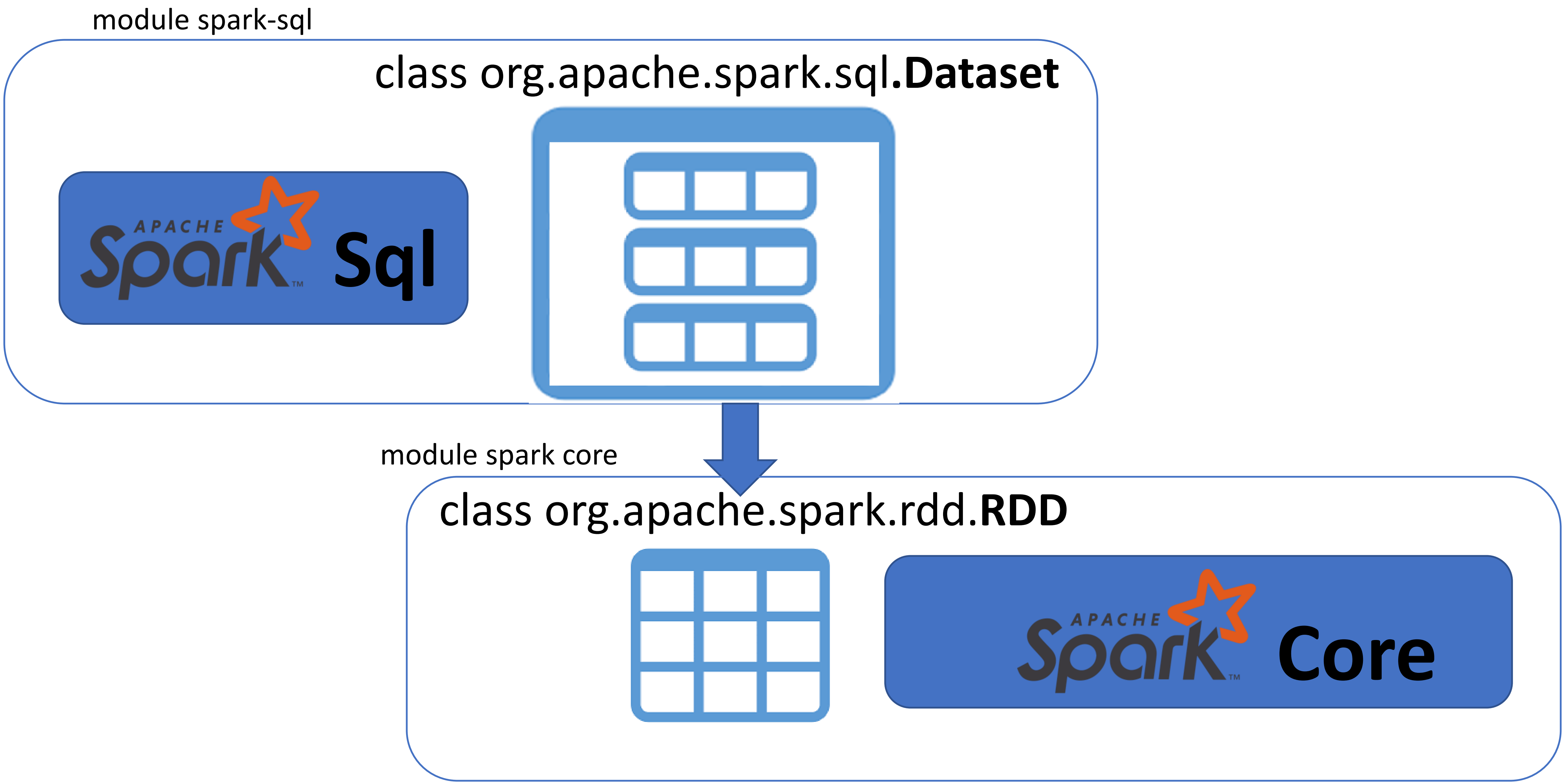
Dataset<Row> = « DataFrame »



Internal RDD



DataSet = sql view for RDD



RDD = ?

Resilient	Resist to failure
Distributed	Horyzontal scale
Dataset	

RDD source doc

Search or jump to... Pull requests Issues Marketplace Explore

apache / spark Public

Edit Pins Watch 2.1k Fork 26.3k Star 34.2k

Code Pull requests 236 Actions Projects Security Insights

master spark / core / src / main / scala / org / apache / spark / rdd / RDD.scala

Go to file

```scala
/**
 * A Resilient Distributed Dataset (RDD), the basic abstraction in Spark. Represents an immutable,
 * partitioned collection of elements that can be operated on in parallel. This class contains the
 * basic operations available on all RDDs, such as `map`, `filter`, and `persist`. In addition,
 * [[org.apache.spark.rdd.PairRDDFunctions]] contains operations available only on RDDs of key-value
 * pairs, such as `groupByKey` and `join`;
 * [[org.apache.spark.rdd.DoubleRDDFunctions]] contains operations available only on RDDs of
 * Doubles; and
 * [[org.apache.spark.rdd.SequenceFileRDDFunctions]] contains operations available on RDDs that
 * can be saved as SequenceFiles.
 * All operations are automatically available on any RDD of the right type (e.g. RDD[(Int, Int)])
 * through implicit.
 *
 * Internally, each RDD is characterized by five main properties:
 *
 *  - A list of partitions
 *  - A function for computing each split
 *  - A list of dependencies on other RDDs
 *  - Optionally, a Partitioner for key-value RDDs (e.g. to say that the RDD is hash-partitioned)
 *  - Optionally, a list of preferred locations to compute each split on (e.g. block locations for
 *    an HDFS file)
 *
 * All of the scheduling and execution in Spark is done based on these methods, allowing each RDD
 * to implement its own way of computing itself. Indeed, users can implement custom RDDs (e.g. for
 * reading data from a new storage system) by overriding these functions. Please refer to the
 * <a href="http://people.csail.mit.edu/matei/papers/2012/nsdi_spark.pdf">Spark paper</a>
 * for more details on RDD internals.
 */
abstract class RDD[T: ClassTag](
    @transient private var _sc: SparkContext,
    @transient private var deps: Seq[Dependency[_]]
  ) extends Serializable with Logging {
```

RDD Doc (1/3)

A **Resilient Distributed Dataset** (RDD),
the **basic abstraction** in Spark.

Represents
an **immutable**,
partitioned collection of elements
that can be operated on **in parallel**.

This class contains the basic operations available on all RDDs,
such as \`map\`, \`filter\`, and \`persist\`.
In addition, PairRDDFunctions (..) of key-value pairs,
(..contains) \`groupByKey\` and \`join\` (..)

RDD Doc (2/3)

Internally, each RDD is characterized by :

- A **list of partitions**

- A **function for computing** each split

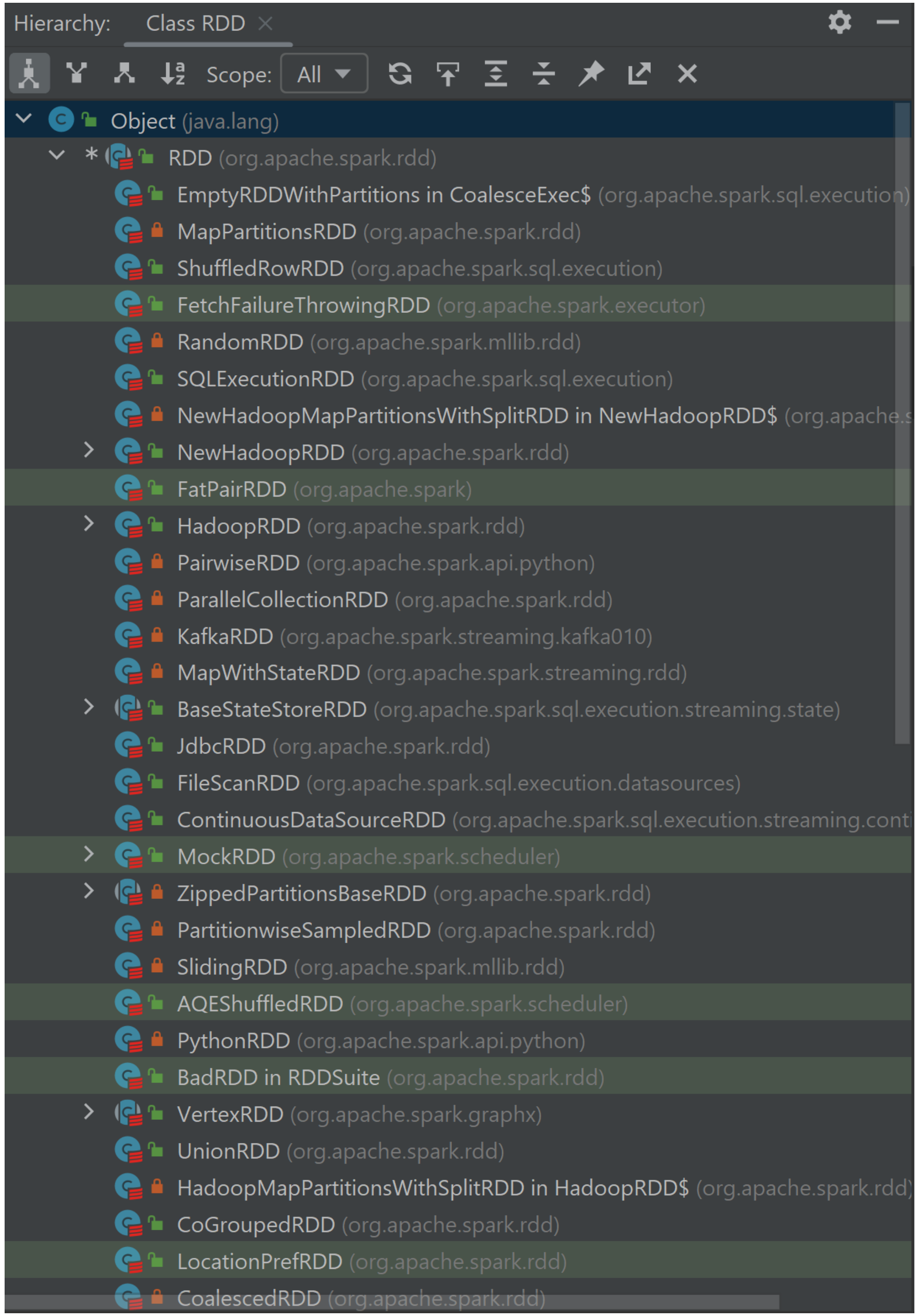
- A list of **dependencies on other RDDs**

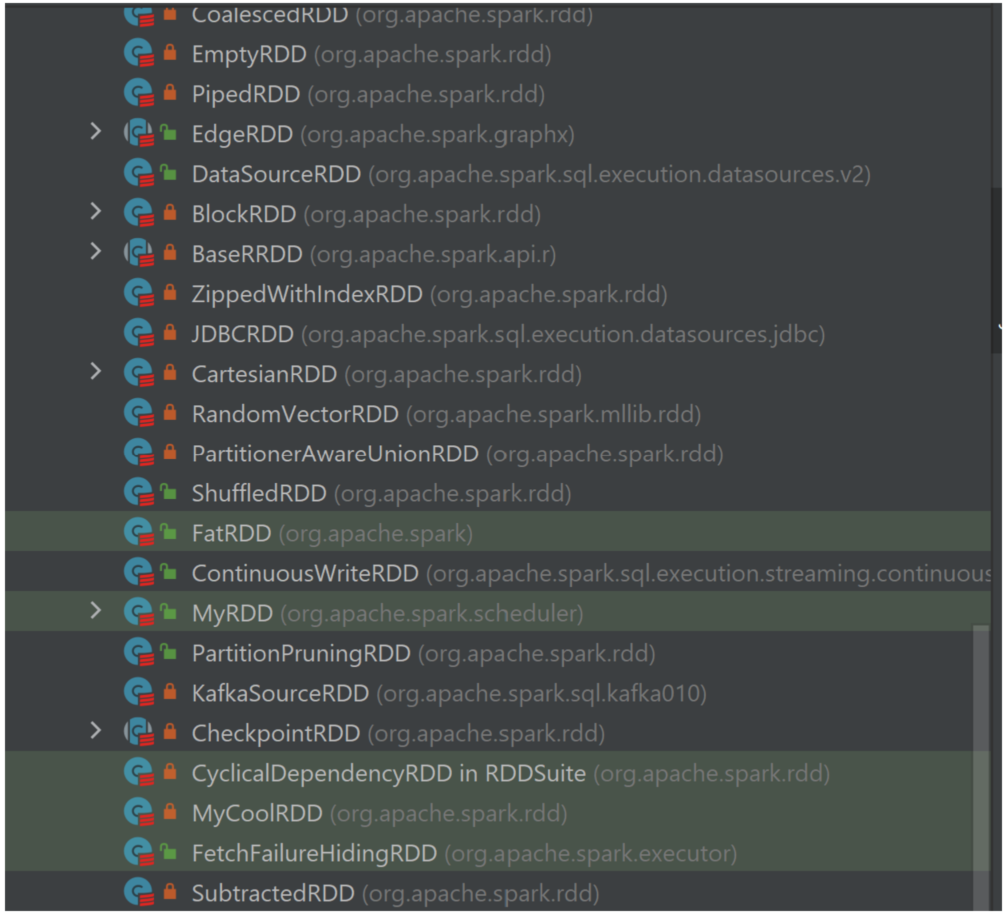
- Optionally, a **Partitioner**
- Optionally, a list of **preferred locations**

RDD Abstract methods

```scala
// =======================================================================
// Methods that should be implemented by subclasses of RDD
// =======================================================================

/**
 * :: DeveloperApi ::
 * Implemented by subclasses to compute a given partition.
 */
@DeveloperApi
def compute(split: Partition, context: TaskContext): Iterator[T]

/**
 * Implemented by subclasses to return the set of partitions in this RDD. This method will only
 * be called once, so it is safe to implement a time-consuming computation in it.
 *
 * The partitions in this array must satisfy the following property:
 *   `rdd.partitions.zipWithIndex.forall { case (partition, index) => partition.index == index }`
 */
protected def getPartitions: Array[Partition]

/**
 * Implemented by subclasses to return how this RDD depends on parent RDDs. This method will only
 * be called once, so it is safe to implement a time-consuming computation in it.
 */
protected def getDependencies: Seq[Dependency[_]] = deps

/**
 * Optionally overridden by subclasses to specify placement preferences.
 */
protected def getPreferredLocations(split: Partition): Seq[String] = Nil

/** Optionally overridden by subclasses to specify how they are partitioned. */
@transient val partitioner: Option[Partitioner] = None
```

Abstract RDD class
=> (many) concrete sub-classes





1 algorithm / transformation => 1 RDD Sub-class

Example: rdd.map(func) or rdd.flatMap(func)

```scala
/**
 * Return a new RDD by applying a function to all elements of this RDD.
 */
def map[U: ClassTag](f: T => U): RDD[U] = withScope {
  val cleanF = sc.clean(f)
  new MapPartitionsRDD[U, T]( prev = this, (_, _, iter) => iter.map(cleanF))
}

/**
 *  Return a new RDD by first applying a function to all elements of this
 *  RDD, and then flattening the results.
 */
def flatMap[U: ClassTag](f: T => TraversableOnce[U]): RDD[U] = withScope {
  val cleanF = sc.clean(f)
  new MapPartitionsRDD[U, T]( prev = this, (_, _, iter) => iter.flatMap(cleanF))
}
```

```scala
/**
 * An RDD that applies the provided function to every partition of the parent RDD.
 *
 * @param prev the parent RDD.
 * @param f The function used to map a tuple of (TaskContext, partition index, input iterator) to
 *          an output iterator.
 * @param preservesPartitioning Whether the input function preserves the partitioner, which should
 *                              be `false` unless `prev` is a pair RDD and the input function
 *                              doesn't modify the keys.
 * @param isFromBarrier Indicates whether this RDD is transformed from an RDDBarrier, a stage
 *                      containing at least one RDDBarrier shall be turned into a barrier stage.
 * @param isOrderSensitive whether or not the function is order-sensitive. If it's order
 *                         sensitive, it may return totally different result when the input order
 *                         is changed. Mostly stateful functions are order-sensitive.
 */
private[spark] class MapPartitionsRDD[U: ClassTag, T: ClassTag](
    var prev: RDD[T],
    f: (TaskContext, Int, Iterator[T]) => Iterator[U],  // (TaskContext, partition index, iterator)
    preservesPartitioning: Boolean = false,
    isFromBarrier: Boolean = false,
    isOrderSensitive: Boolean = false)
  extends RDD[U](prev) {
```

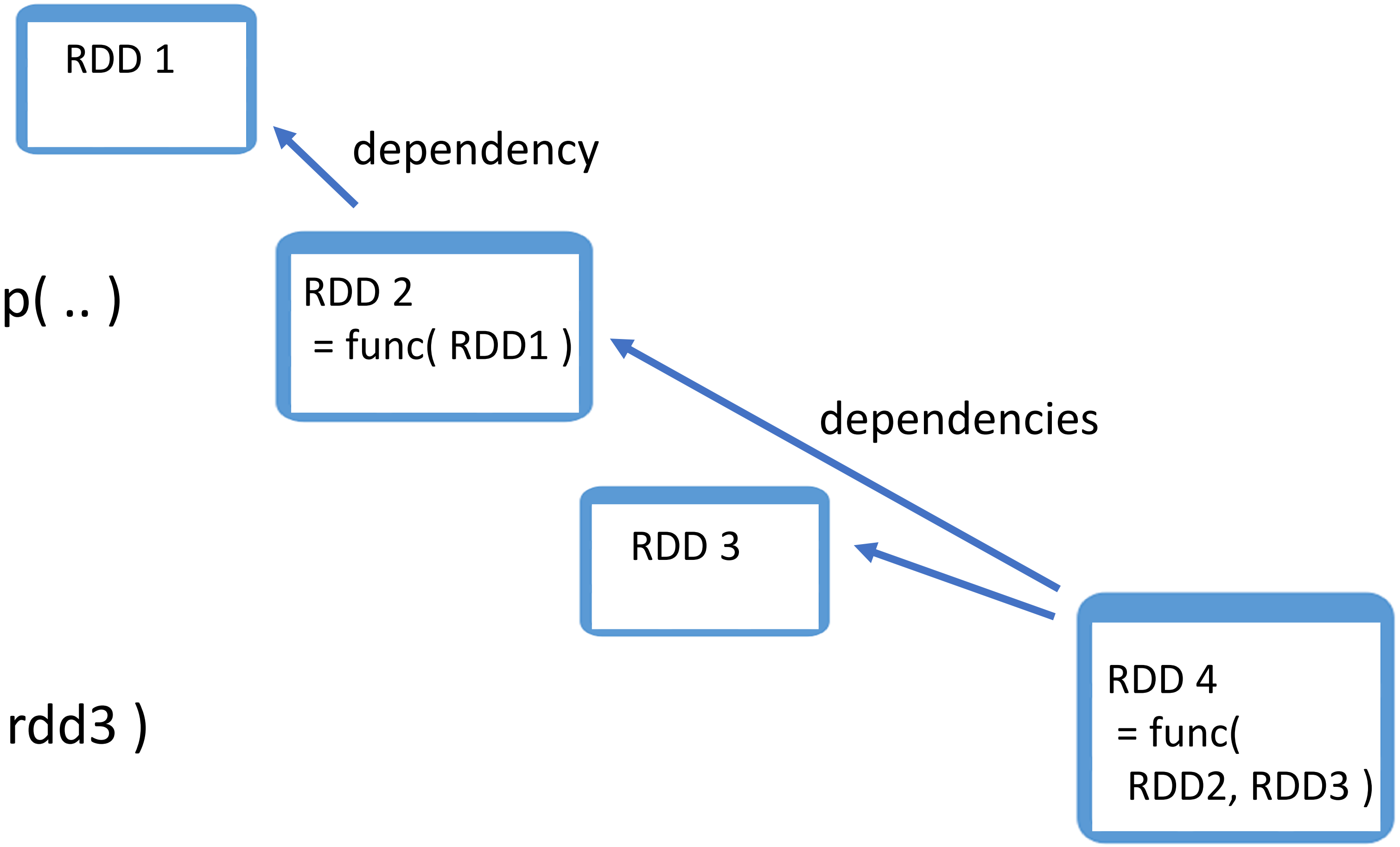
Call transform function => Create new RDD (linked)

RDD rdd1 = … read

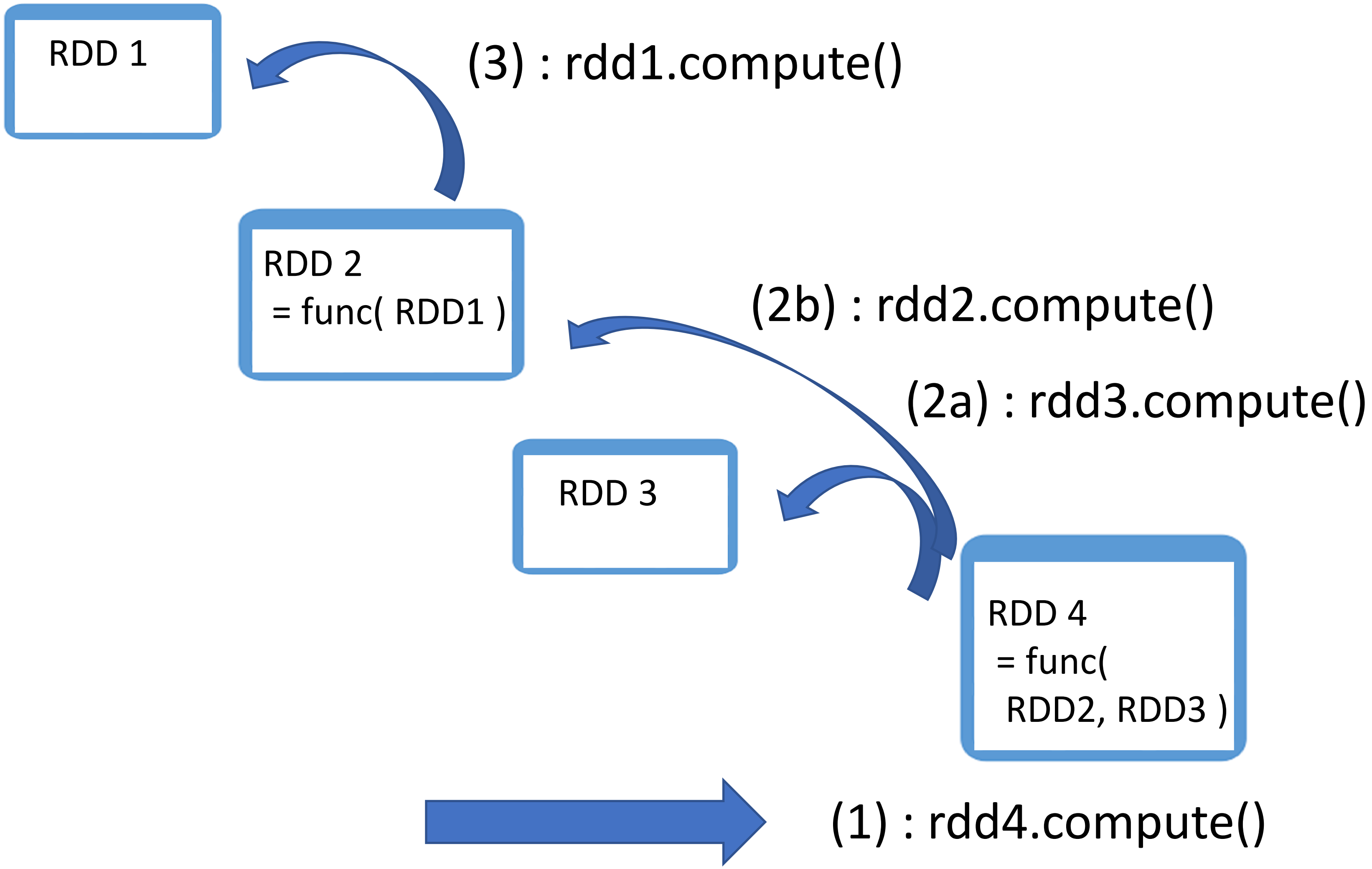
RDD rdd2   = rdd1 .map(..)

RDD rdd3 = .. read

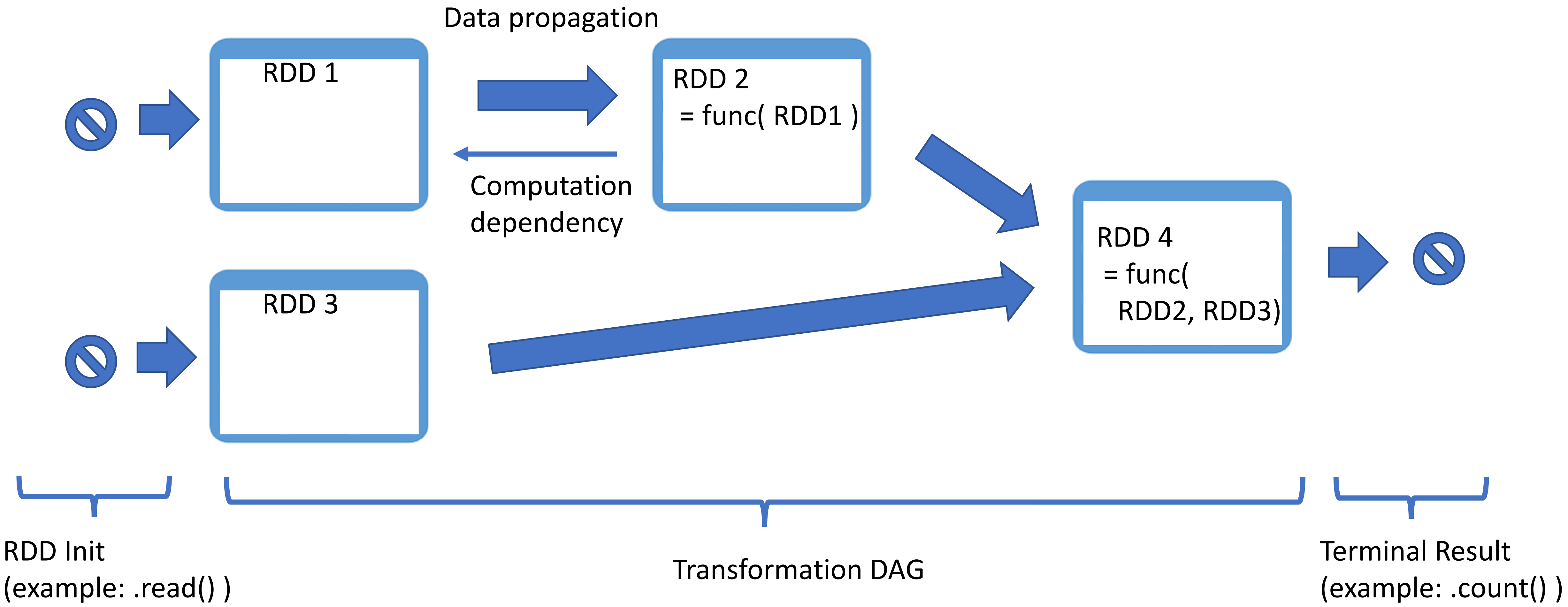
RDD rdd4 = rdd2 .join(rdd3)



Call compute() => … dependency.compute()



RDD Dependencies : DAG (Directed Acyclic Graph)



3 equivalent formalisms:
SSA create Api, Expression Algebra, DAG

SSA = Single State Assignments

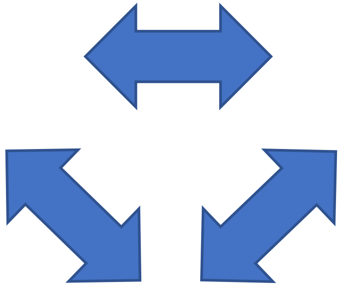
RDD API

```
RDD rdd1 = ... read
RDD rdd2   =  rdd1 .map( .. )
RDD rdd3 = .. read
RDD rdd4 = rdd2 .join( rdd3 )
```

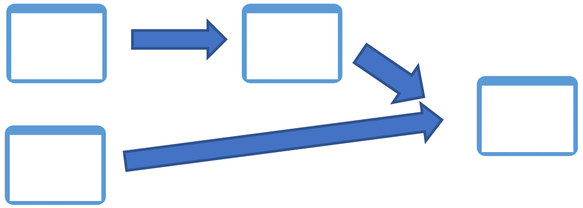
Expression Algebra, Sql

```
SELECT  map(t1) FROM Table1 t1
JOIN Table2 t2 on ..
```

```
new JoinRDD(
   new MapRDD( readRDD(table1) ),
   readRDD(table2)
   )
```



DAG



RDD Doc (3/3)

All (..) in Spark is done based on these methods,
allowing each RDD **to implement** its own **way of computing** itself.

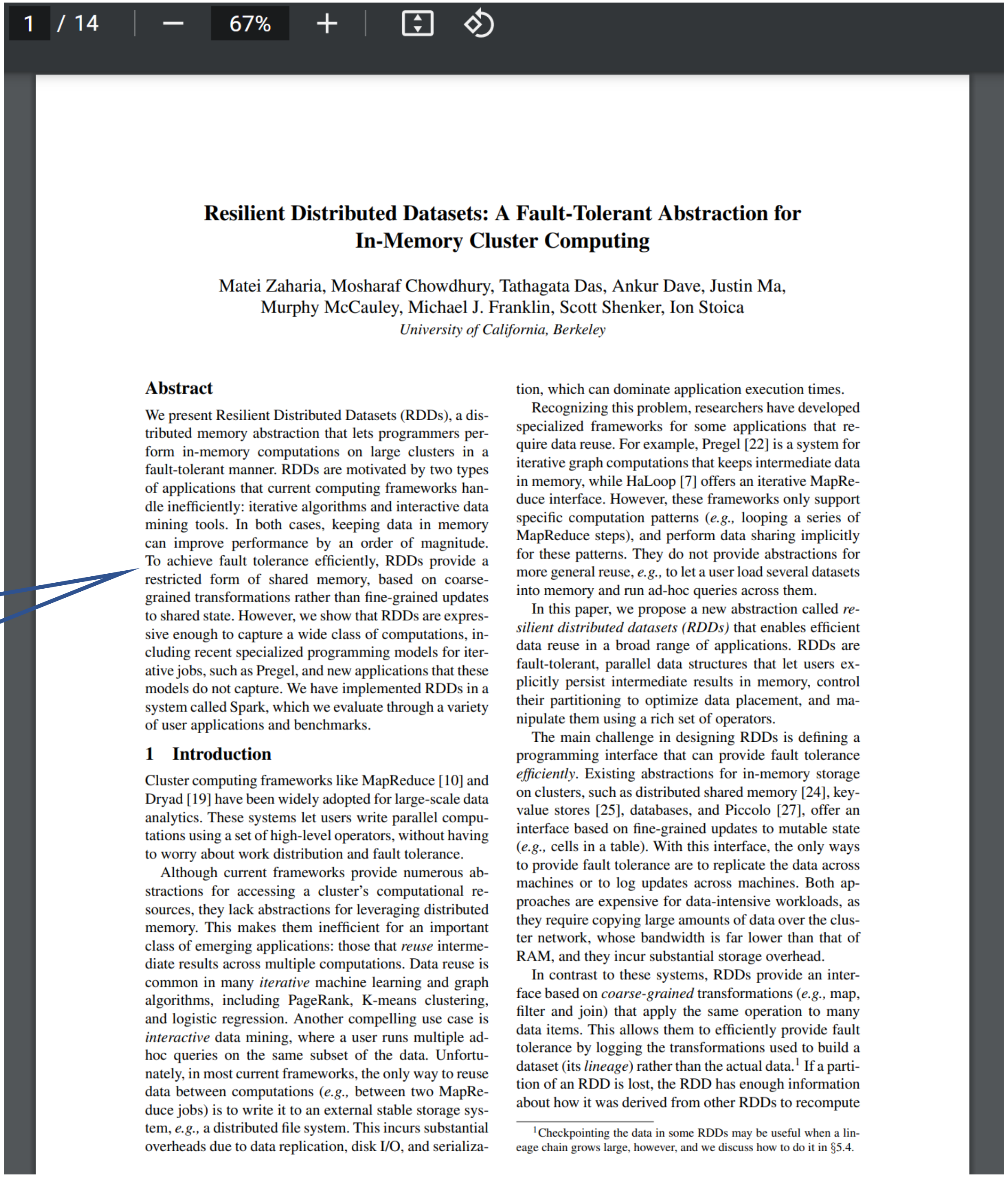
Indeed, users **can implement** custom RDDs
(e.g. for reading data from a new storage system)
by **overriding** these functions.

Please refer to the
Spark paper
for more details on RDD internals.

RDD Paper

A Fault-Tolerant Abstraction For In-Memory Cluster computing

To achieve fault tolerance efficiently,
RDDs provide a restricted form of shared memory,
based on coarse-grained transformations



1 / 14 67%

Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing

Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma,
Murphy McCauley, Michael J. Franklin, Scott Shenker, Ion Stoica
University of California, Berkeley

Abstract

We present Resilient Distributed Datasets (RDDs), a distributed memory abstraction that lets programmers perform in-memory computations on large clusters in a fault-tolerant manner. RDDs are motivated by two types of applications that current computing frameworks handle inefficiently: iterative algorithms and interactive data mining tools. In both cases, keeping data in memory can improve performance by an order of magnitude. To achieve fault tolerance efficiently, RDDs provide a restricted form of shared memory, based on coarse-grained transformations rather than fine-grained updates to shared state. However, we show that RDDs are expressive enough to capture a wide class of computations, including recent specialized programming models for iterative jobs, such as Pregel, and new applications that these models do not capture. We have implemented RDDs in a system called Spark, which we evaluate through a variety of user applications and benchmarks.

1 Introduction

Cluster computing frameworks like MapReduce [10] and Dryad [19] have been widely adopted for large-scale data analytics. These systems let users write parallel computations using a set of high-level operators, without having to worry about work distribution and fault tolerance.

Although current frameworks provide numerous abstractions for accessing a cluster's computational resources, they lack abstractions for leveraging distributed memory. This makes them inefficient for an important class of emerging applications: those that *reuse* intermediate results across multiple computations. Data reuse is common in many *iterative* machine learning and graph algorithms, including PageRank, K-means clustering, and logistic regression. Another compelling use case is *interactive* data mining, where a user runs multiple ad-hoc queries on the same subset of the data. Unfortunately, in most current frameworks, the only way to reuse data between computations (*e.g.,* between two MapReduce jobs) is to write it to an external stable storage system, *e.g.,* a distributed file system. This incurs substantial overheads due to data replication, disk I/O, and serialization, which can dominate application execution times.

Recognizing this problem, researchers have developed specialized frameworks for some applications that require data reuse. For example, Pregel [22] is a system for iterative graph computations that keeps intermediate data in memory, while HaLoop [7] offers an iterative MapReduce interface. However, these frameworks only support specific computation patterns (*e.g.,* looping a series of MapReduce steps), and perform data sharing implicitly for these patterns. They do not provide abstractions for more general reuse, *e.g.,* to let a user load several datasets into memory and run ad-hoc queries across them.

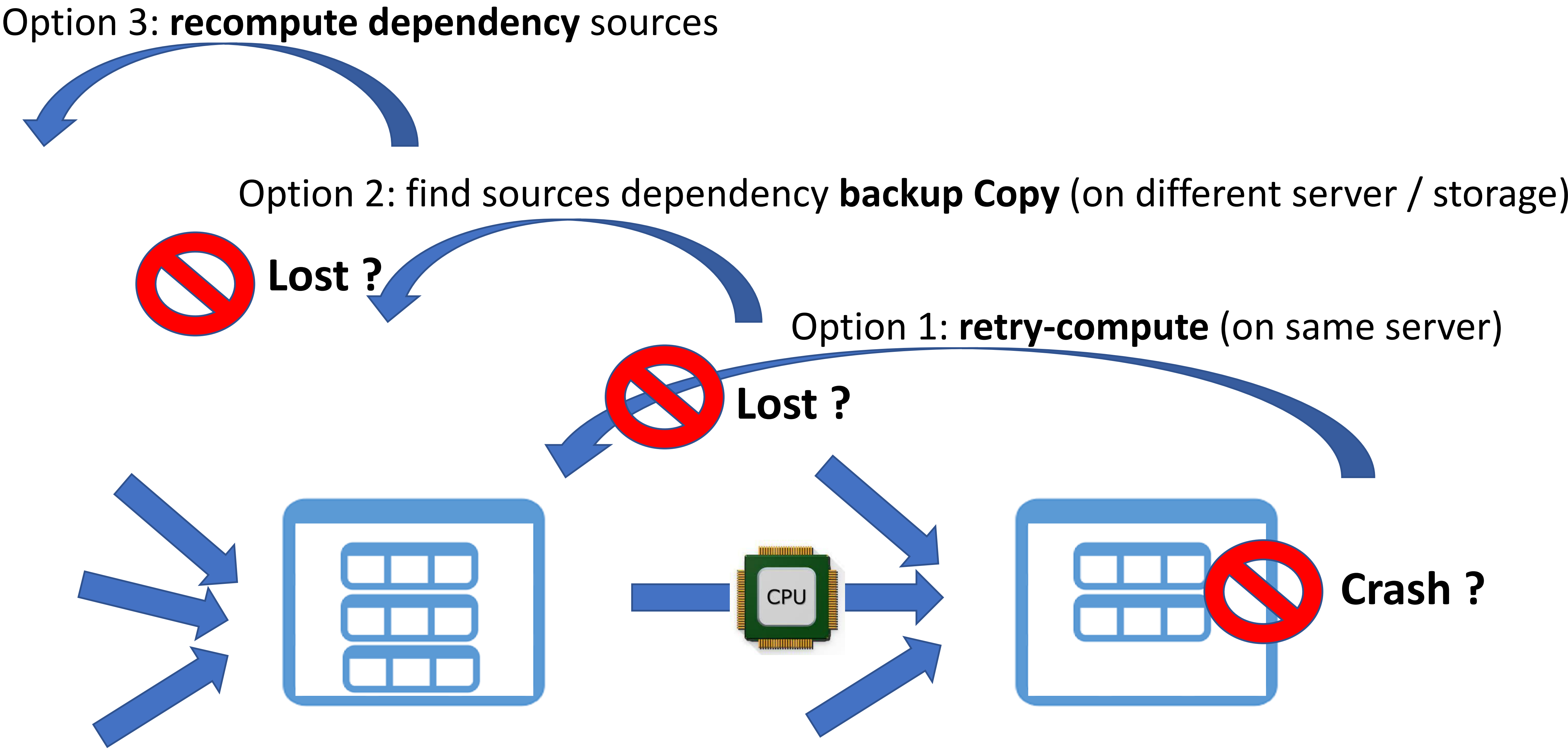
In this paper, we propose a new abstraction called *resilient distributed datasets (RDDs)* that enables efficient data reuse in a broad range of applications. RDDs are fault-tolerant, parallel data structures that let users explicitly persist intermediate results in memory, control their partitioning to optimize data placement, and manipulate them using a rich set of operators.

The main challenge in designing RDDs is defining a programming interface that can provide fault tolerance *efficiently*. Existing abstractions for in-memory storage on clusters, such as distributed shared memory [24], key-value stores [25], databases, and Piccolo [27], offer an interface based on fine-grained updates to mutable state (*e.g.,* cells in a table). With this interface, the only ways to provide fault tolerance are to replicate the data across machines or to log updates across machines. Both approaches are expensive for data-intensive workloads, as they require copying large amounts of data over the cluster network, whose bandwidth is far lower than that of RAM, and they incur substantial storage overhead.

In contrast to these systems, RDDs provide an interface based on *coarse-grained* transformations (*e.g.,* map, filter and join) that apply the same operation to many data items. This allows them to efficiently provide fault tolerance by logging the transformations used to build a dataset (its *lineage*) rather than the actual data.[1] If a partition of an RDD is lost, the RDD has enough information about how it was derived from other RDDs to recompute

[1] Checkpointing the data in some RDDs may be useful when a lineage chain grows large, however, and we discuss how to do it in §5.4.

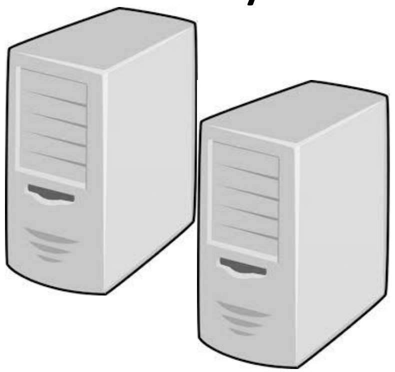
Fault Tolerant - Computation



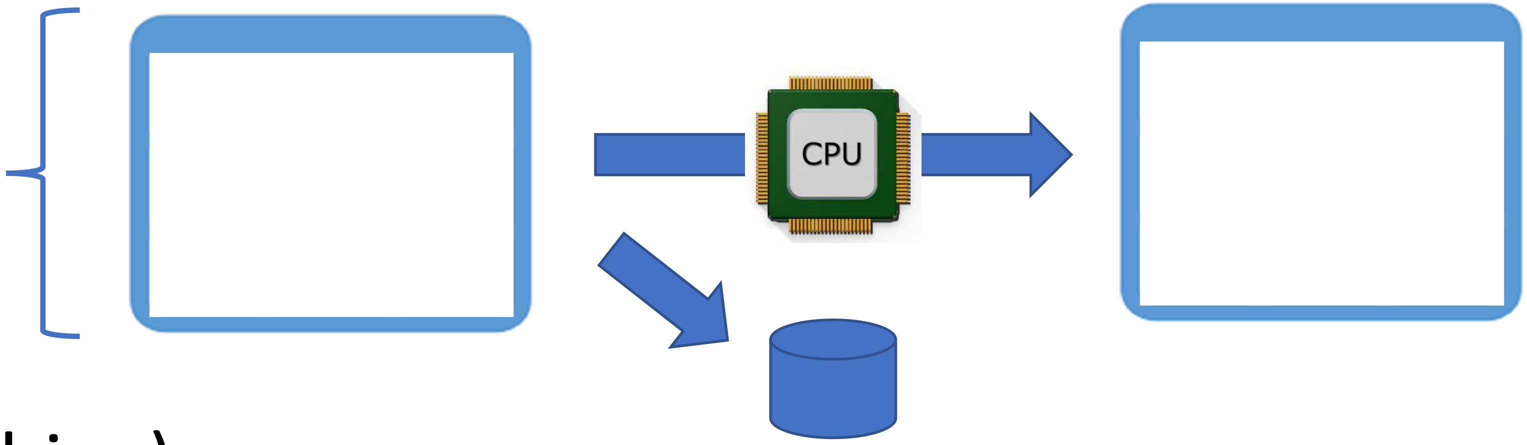
CoarseGrain … Scheduler/Executer

Spark-driver
Implements Fault Tolerance+Distribution
… internally called « CoarseGrainScheduler »

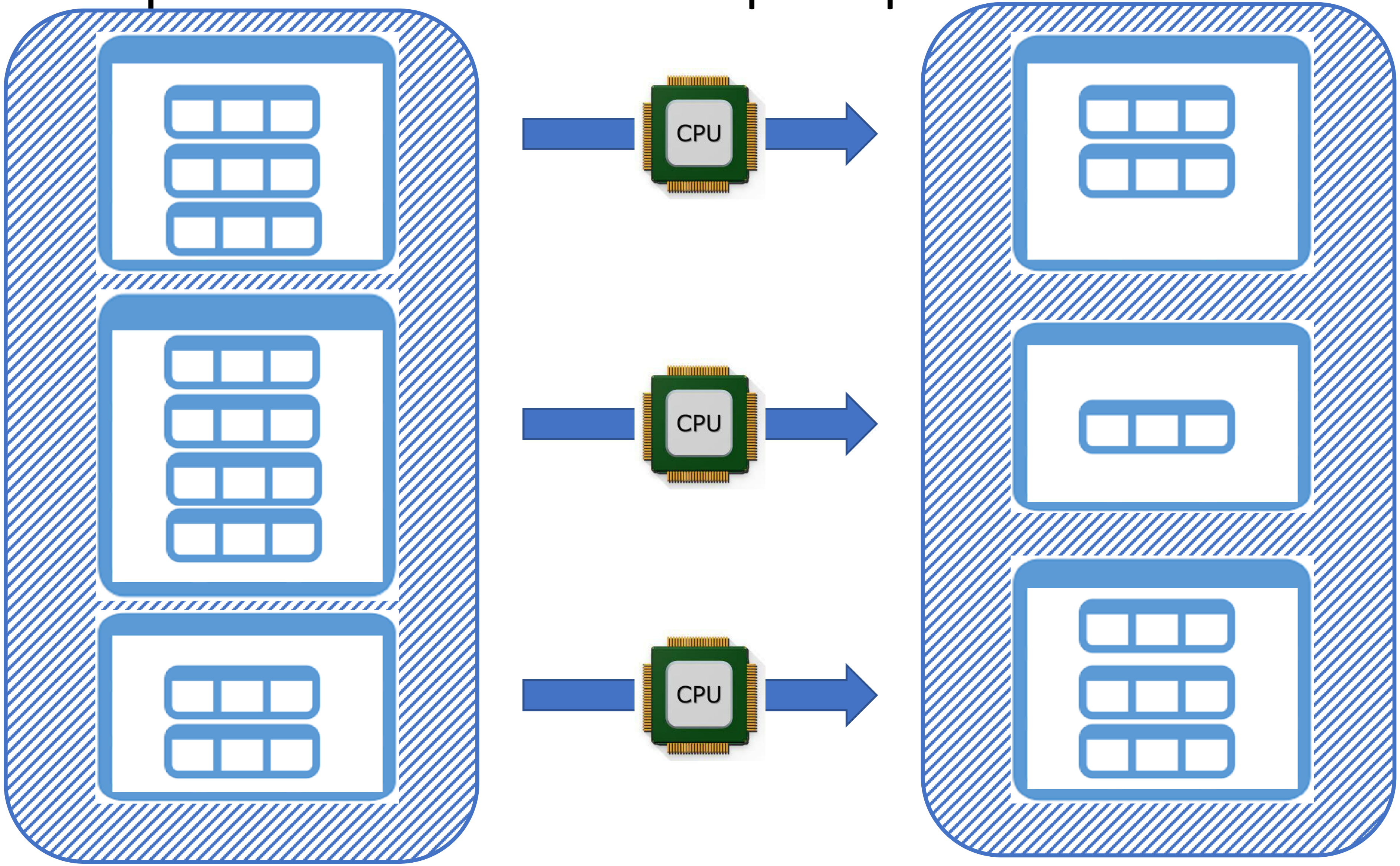
Spark-executor
Implements task main loop
… internally called « CoarseGrainExecutor »



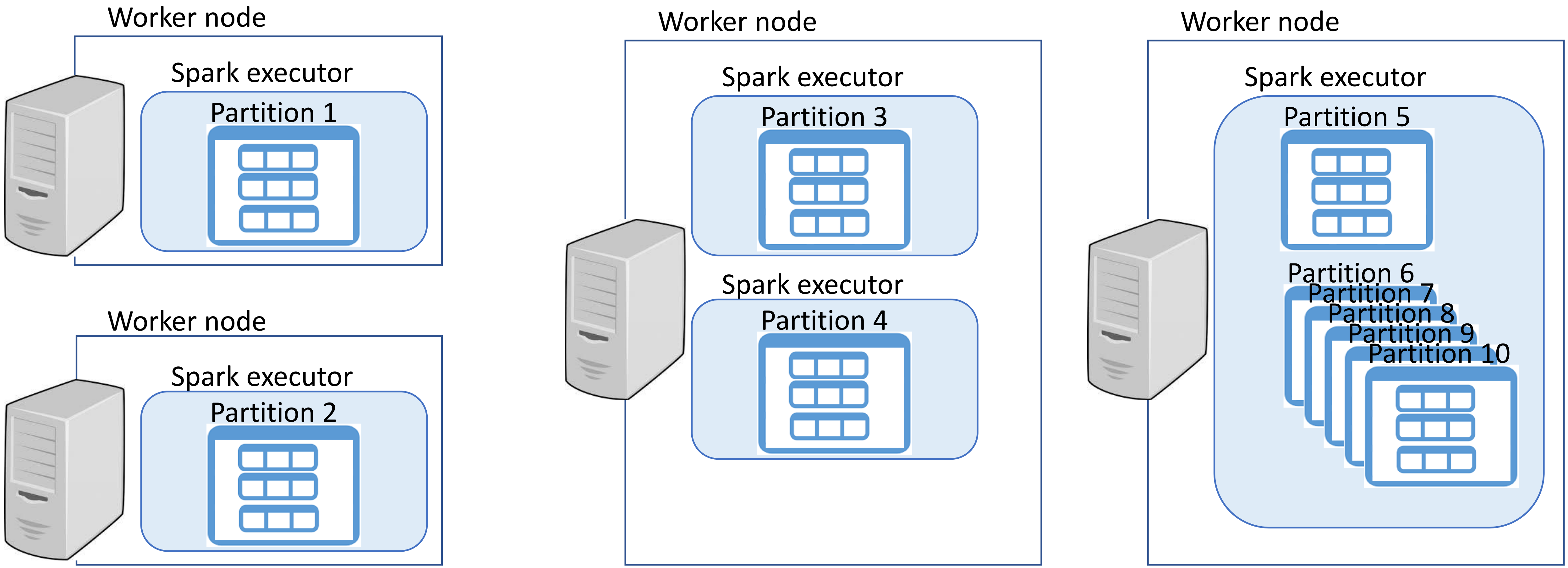
CoarseGrain
partition
= unit of caching/
recomputation
(all elements or nothing)



DataSet : Collection of Objects, parallelize 1 CPU per partition



Distribution: Partition < Executor < Node



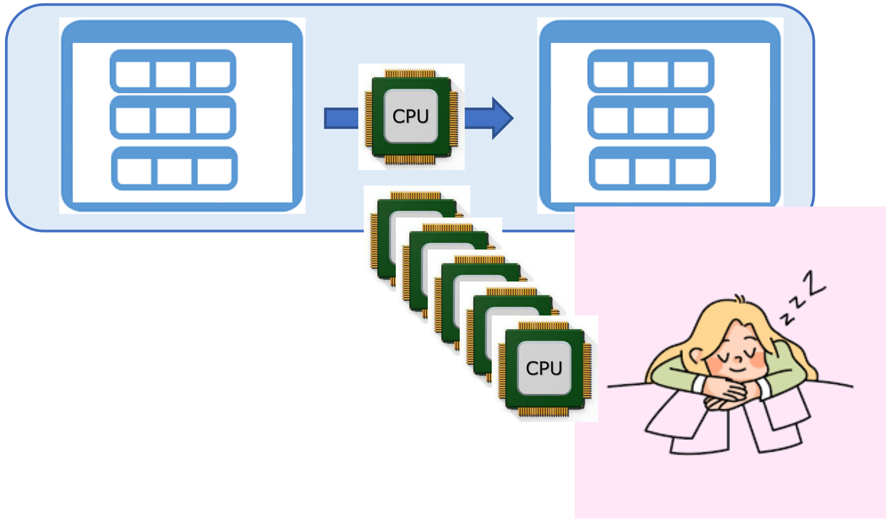
several worker nodes

Several spark-executor processes per nodes

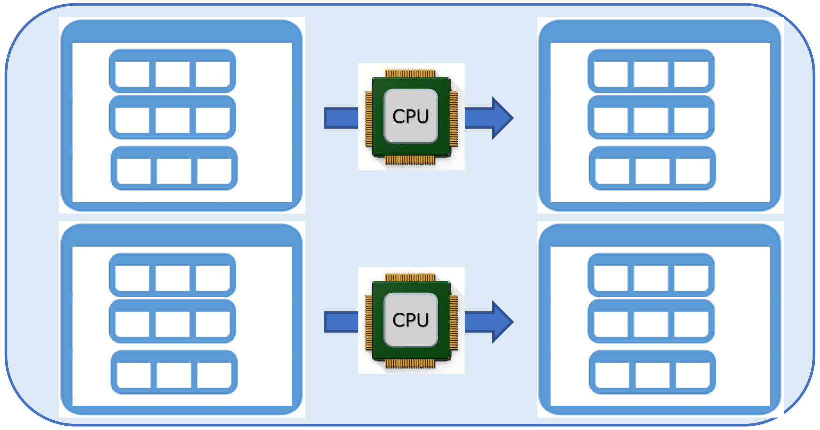
Several partitions Per spark-executor

Optimize parralelism: Adapt partitions to number of Cores

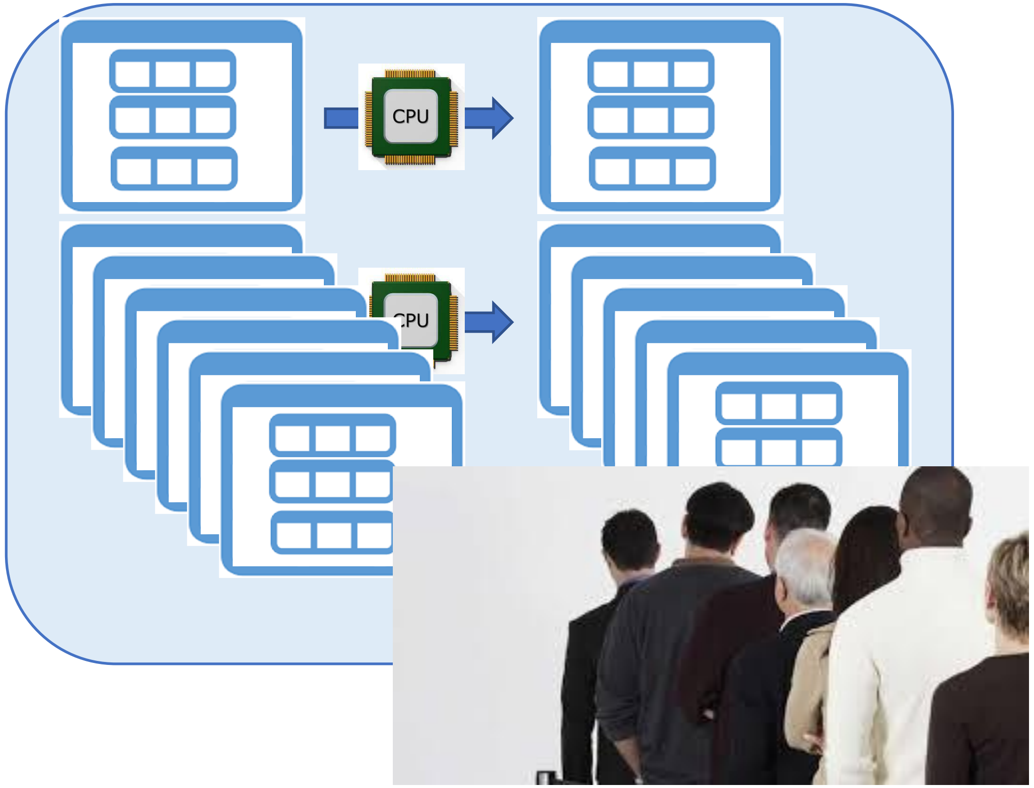
Spark-executor: 1 partition << N cores



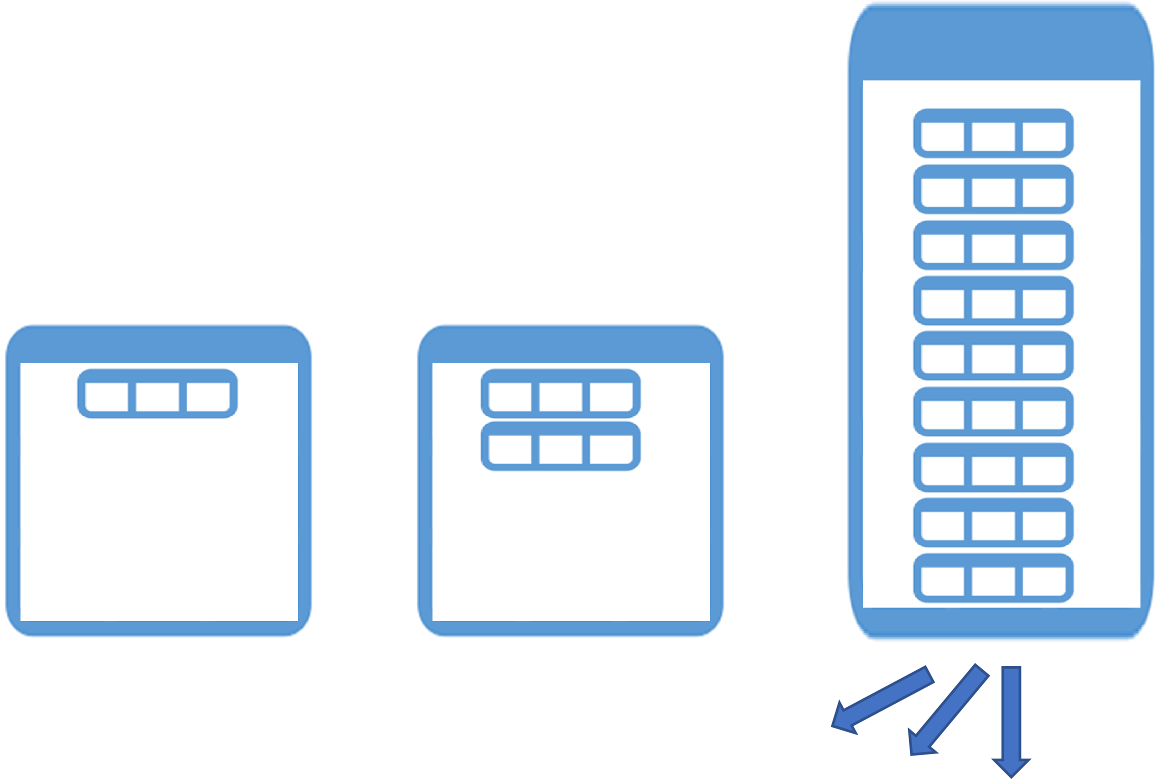
Spark executor: N partitions ~ N cores



Spark executor: N partitions >> few cores

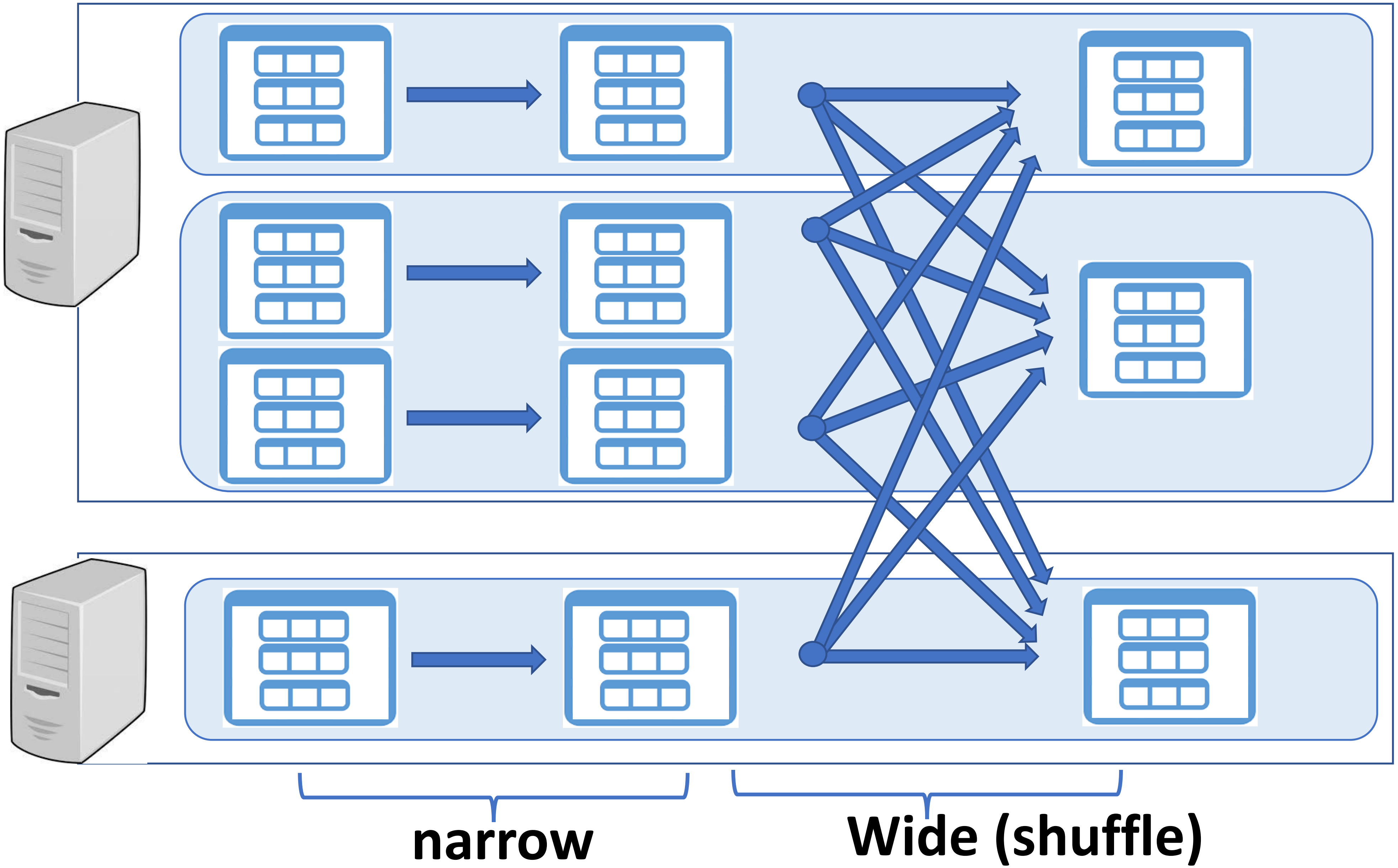


Skewed Data … need Repartitioned equally



Target: move each row[i] to node[j] j = « rowId modulo N »

« Narrow » / « Wide » Transformations



RDD … OK
so What is a Dataset ?

Dataset source doc

Search or jump to... Pull requests Issues Marketplace Explore

apache / spark Public

Edit Pins | Watch 2.1k | Fork 26.2k | Star 34.2k

Code | Pull requests 238 | Actions | Projects | Security | Insights

master | spark / sql / core / src / main / scala / org / apache / spark / sql / Dataset.scala | Go to file | ...

```
/**
 * A Dataset is a strongly typed collection of domain-specific objects that can be transformed
 * in parallel using functional or relational operations. Each Dataset also has an untyped view
 * called a `DataFrame`, which is a Dataset of [[Row]].
 *
 * Operations available on Datasets are divided into transformations and actions. Transformations
 * are the ones that produce new Datasets, and actions are the ones that trigger computation and
 * return results. Example transformations include map, filter, select, and aggregate (`groupBy`).
 * Example actions count, show, or writing data out to file systems.
 *
 * Datasets are "lazy", i.e. computations are only triggered when an action is invoked. Internally,
 * a Dataset represents a logical plan that describes the computation required to produce the data.
 * When an action is invoked, Spark's query optimizer optimizes the logical plan and generates a
 * physical plan for efficient execution in a parallel and distributed manner. To explore the
 * logical plan as well as optimized physical plan, use the `explain` function.
 *
 * To efficiently support domain-specific objects, an [[Encoder]] is required. The encoder maps
 * the domain specific type `T` to Spark's internal type system. For example, given a class `Person`
 * with two fields, `name` (string) and `age` (int), an encoder is used to tell Spark to generate
 * code at runtime to serialize the `Person` object into a binary structure. This binary structure
 * often has much lower memory footprint as well as are optimized for efficiency in data processing
 * (e.g. in a columnar format). To understand the internal binary representation for data, use the
 * `schema` function.
```

Dataset Doc (1/4)

A strongly typed **collection of domain-specific objects**

that can be **transformed in parallel**

using **functional or relational operations**

Associated untyped view: DataFrame = Dataset<Row>

Dataset Doc (2/4)

Operations on Datasets:

Transformations = produce new Datasets

Actions = trigger computation and return results.

Example transformations : map, filter, select, groupBy ...
Example actions : count, show, write ...

Datasets are "lazy",
i.e. computations are only triggered **when an action is invoked**.

Transformation… « produce » new Dataset
NO UPDATE method
Dataset are computable / « **immutable** »

dataset. <noSetter> ();

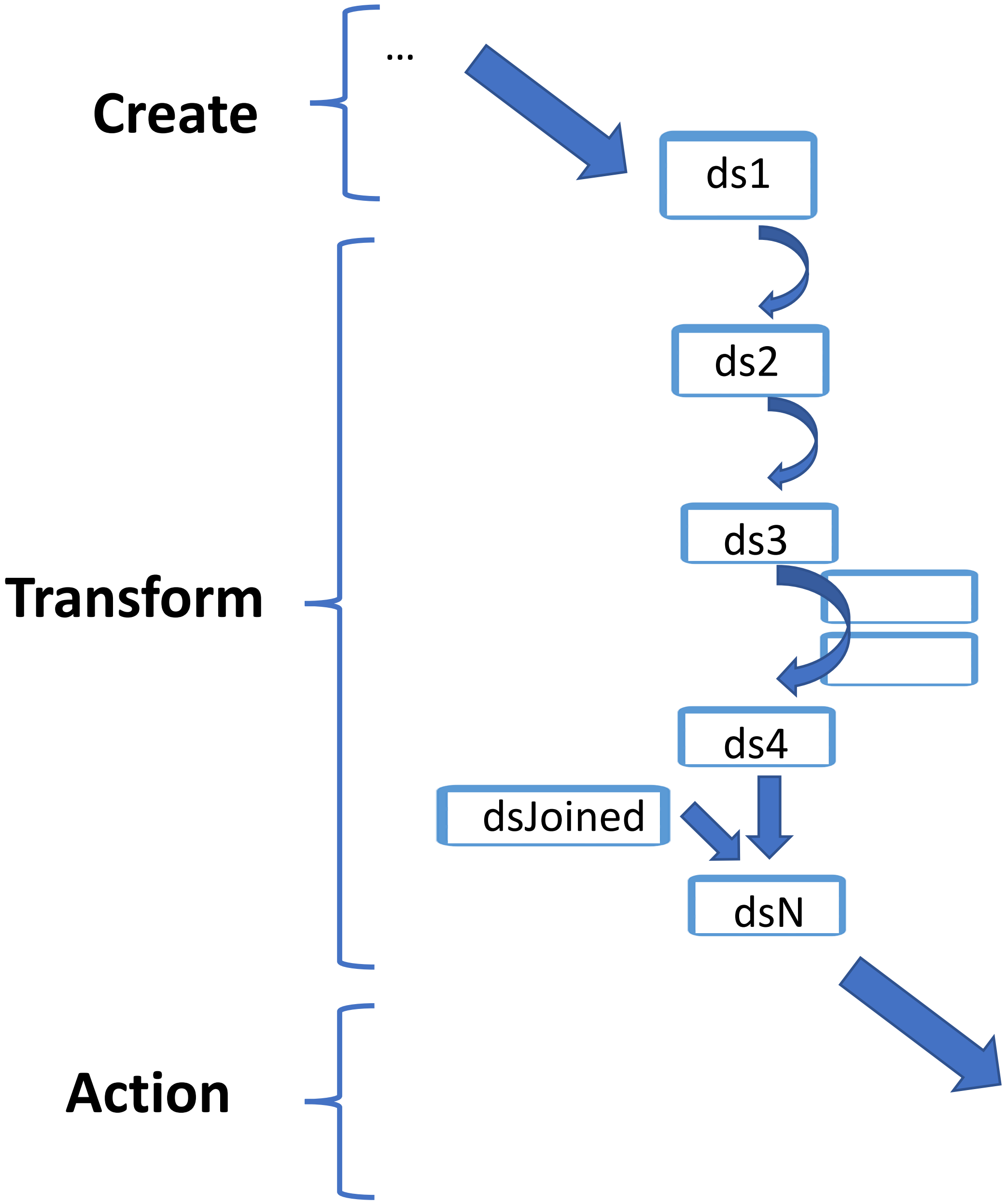
Transform

Dataset newDataset = dataset . <createNewWithTransform> (…)
Example: chain method
dataset = ds .filter(..) .filter(..) .map(..) .groupBy(..),

Action

<Result> = dataset. <action> ();
example: long result = dataset. count ();

Dataset operations…

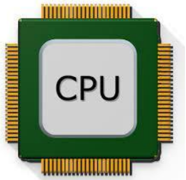



```
Dataset<T1>   ds1 =  spark.read ….

Dataset<T2>    ds2 =  ds1.map(   x -> f(x)  );
```

```
Dataset<Row>  ds3 =  ds2.toDF();

Dataset<Row>  ds4 =  ds3 .filter(  « col >= value »)
                         .filter( x -> g(x) )
                         .map( y -> h(y) );
```

```
…
Dataset<Row>  dsN =   ds4  .    join( dsJoined)
```

ds . **show**(); // => TRIGGER COMPUTE !!



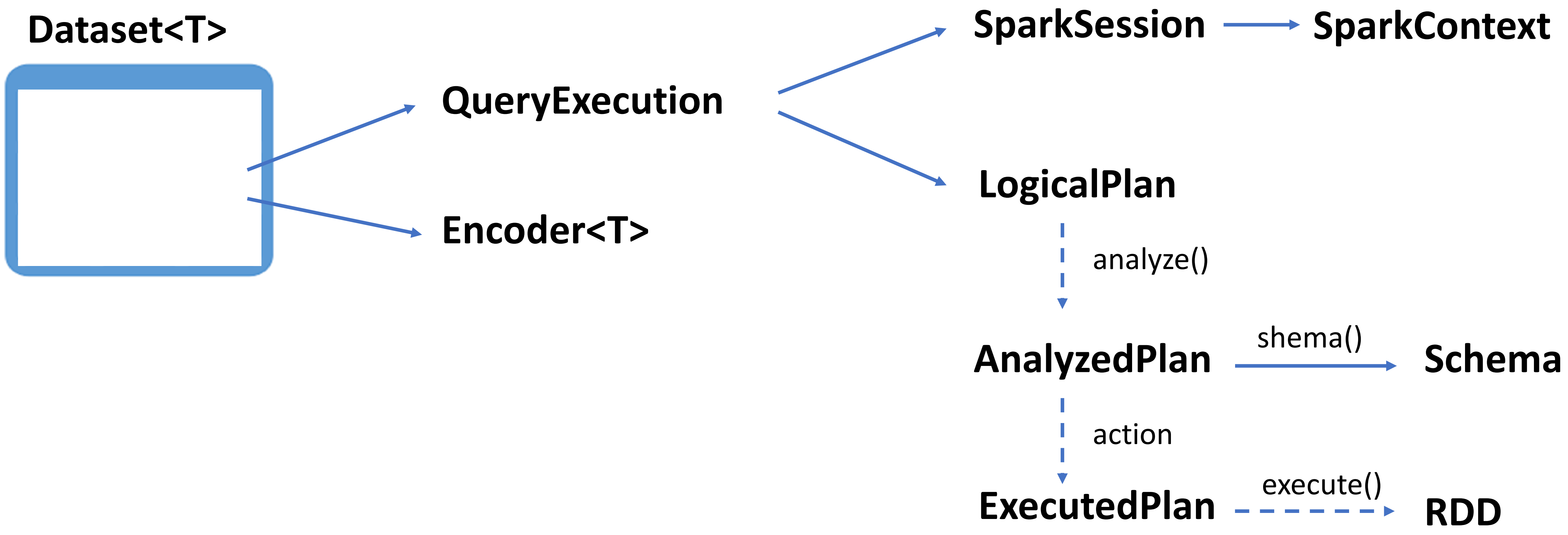
Dataset Doc (3/4)

Internally, a Dataset represents a logical plan
that **describes** the computation required **to produce the data**.

When an action is invoked,
Spark's query optimizer optimizes the **logical plan**
and generates a **physical plan**
for efficient execution in a parallel and distributed manner.

To explore.. Use ‘explain plan’

Internally…


Dataset<T>
QueryExecution
Encoder<T>
SparkSession
SparkContext
LogicalPlan
analyze()
AnalyzedPlan
shema()
Schema
action
ExecutedPlan
execute()
RDD

Dataset Doc 4/4

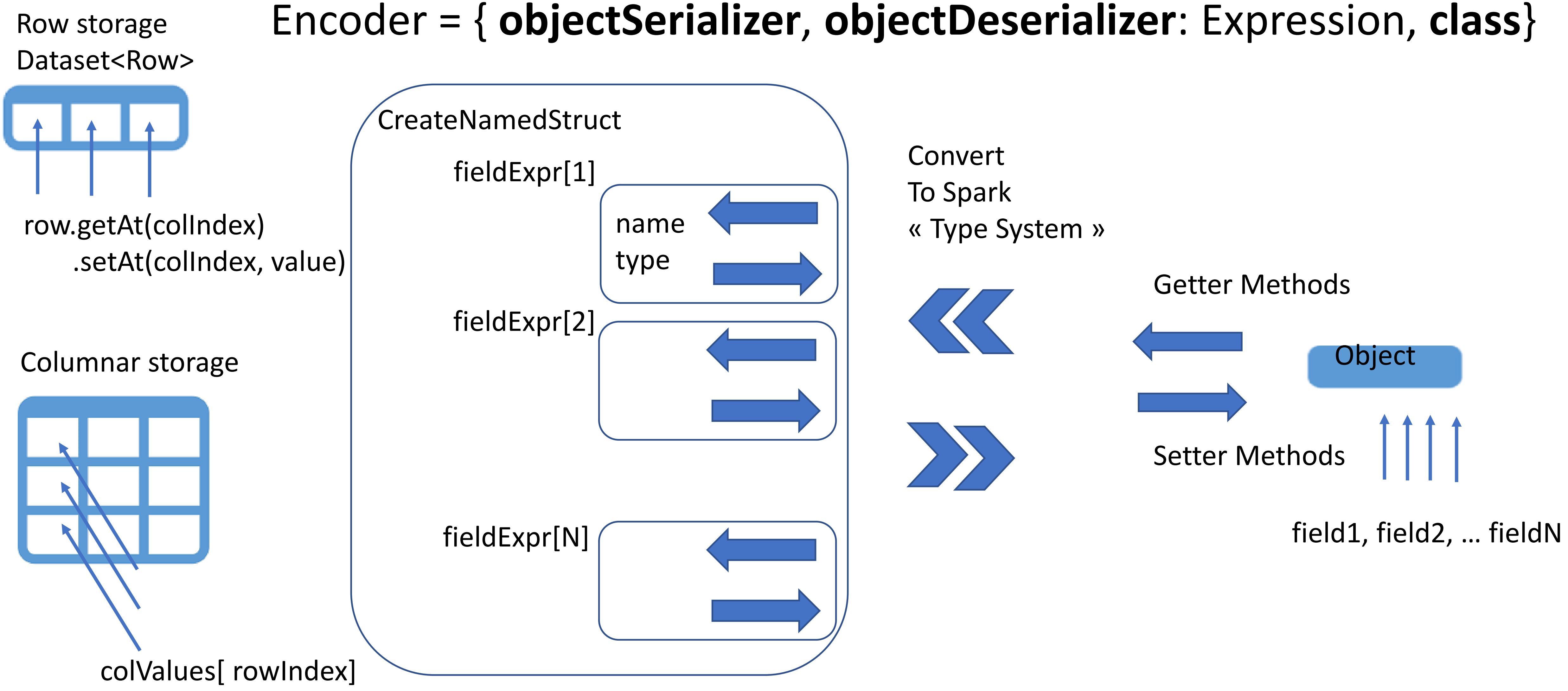
To efficiently support domain-specific objects, an ‘Encoder’ is required.

The encoder maps the domain specific type `T` to Spark's internal type system.
(…) to generate code at runtime
to serialize object into a much efficient binary structure
(..) with lower memory footprint
(..) optimized for processing (e.g. in a columnar format).

To explore representation of data, use `schema` function.

Encoder<T>

Encoder = { **objectSerializer**, **objectDeserializer**: Expression, **class**}



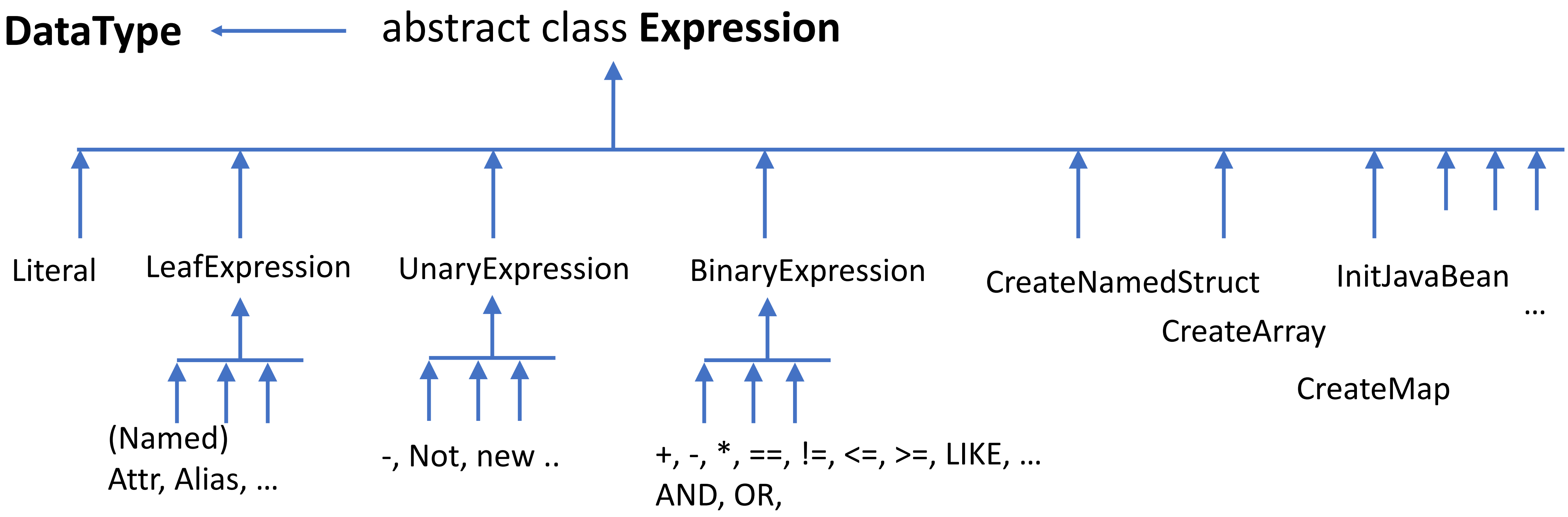
Annexe

Spark « Type System »
DataTypes ...

Annexe

Encoder .. Internal Expression with DataType

Expression abstract class AST (Abstract Syntaxic Tree)
for Sql / CodeGenerator / Java Getter-Setter



Expression ...

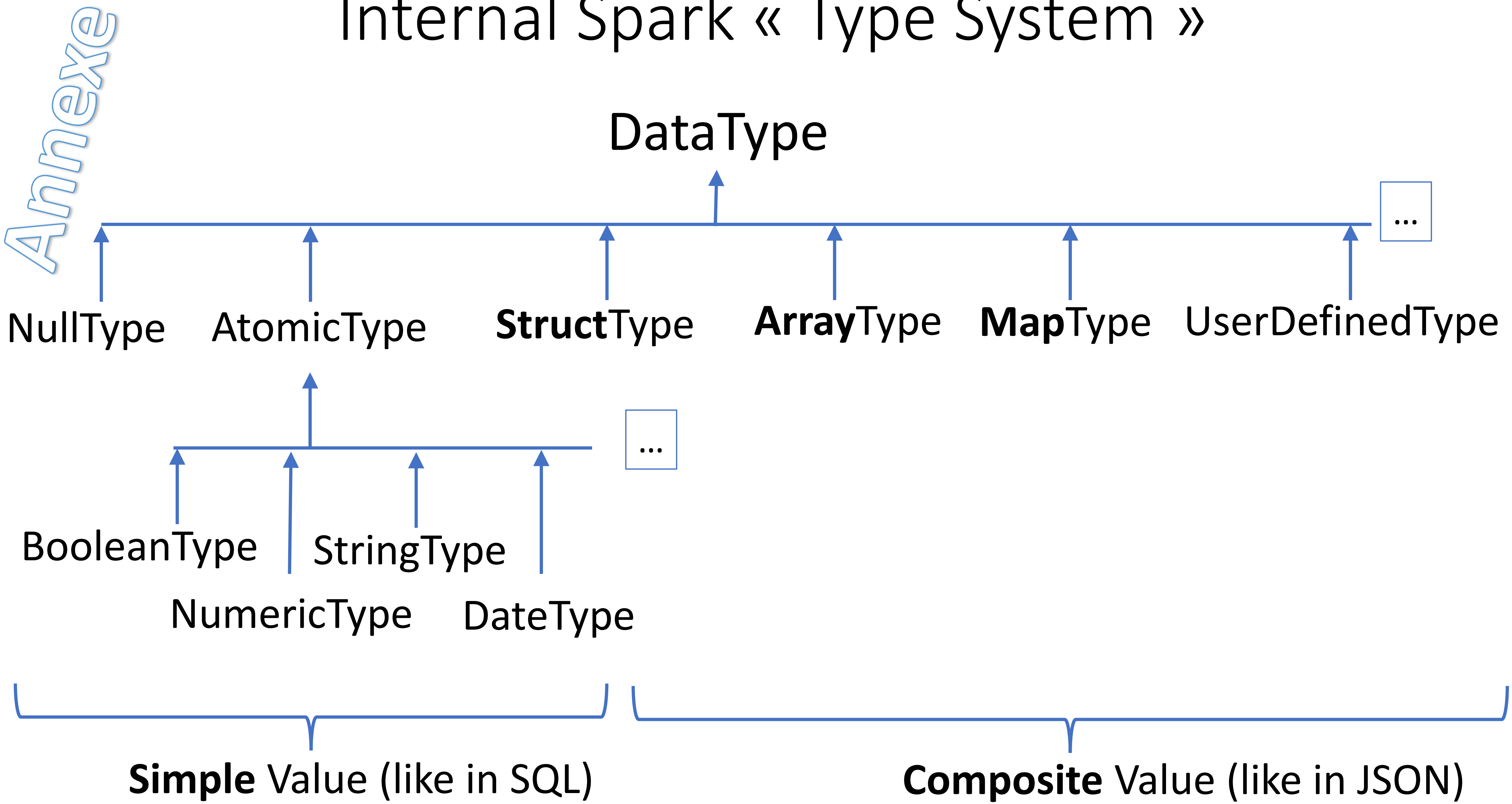
Complex … INTERNAL …
(not need to understand, to use Spark)

Looks like a « **SQL Compiler** »

… support **NOT ONLY Sql literal values**,
but also **NamedStruct, Array, Map, Java Objects** !!!

Used internally to **compile + generate code**

Internal Spark « Type System »



Annexe

Spark DataType

- AbstractDataType (org.apache.spark.sql.types)
  - * DataType (org.apache.spark.sql.types)
    - ObjectType (org.apache.spark.sql.types)
    - UserDefinedType (org.apache.spark.sql.types)
    - CalendarIntervalType (org.apache.spark.sql.types)
    - MapType (org.apache.spark.sql.types)
    - NullType (org.apache.spark.sql.types)
    - AtomicType (org.apache.spark.sql.types)
      - BinaryType (org.apache.spark.sql.types)
      - BooleanType (org.apache.spark.sql.types)
      - NumericType (org.apache.spark.sql.types)
      - DatetimeType (org.apache.spark.sql.types)
      - AnsiIntervalType (org.apache.spark.sql.types)
      - VarcharType (org.apache.spark.sql.types)
      - CharType (org.apache.spark.sql.types)
      - StringType (org.apache.spark.sql.types)
    - StructType (org.apache.spark.sql.types)
    - ArrayType (org.apache.spark.sql.types)

Hive - Spark SQL supports Struct, List, Map …

Example:

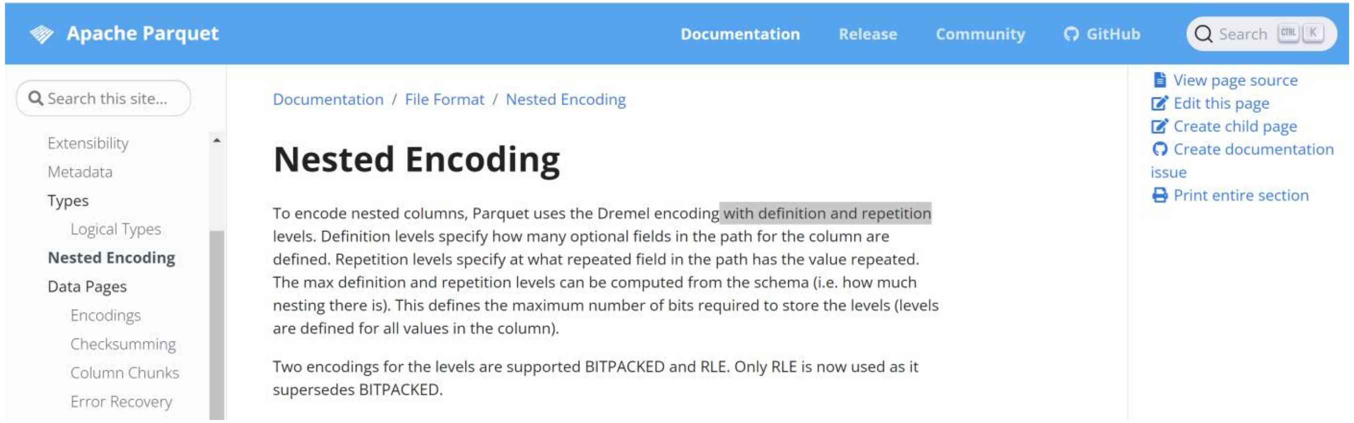
CREATE EXTERNAL TABLE `student` (
firstName string, lastName string,

practicedSports **array< named_struct<** name: string, numberYear: int **> >**,

diploma **map<**string, **named_struct<**mention: string, obtentionDate: Date **> >**

)

Nested fields in File Format: Parquet / Orc / Json



Apache Parquet   Documentation   Release   Community   GitHub   Search

Search this site...

- Extensibility
- Metadata
- Types
  - Logical Types
- Nested Encoding
- Data Pages
  - Encodings
  - Checksumming
  - Column Chunks
  - Error Recovery

Documentation / File Format / Nested Encoding

Nested Encoding

To encode nested columns, Parquet uses the Dremel encoding with definition and repetition levels. Definition levels specify how many optional fields in the path for the column are defined. Repetition levels specify at what repeated field in the path has the value repeated. The max definition and repetition levels can be computed from the schema (i.e. how much nesting there is). This defines the maximum number of bits required to store the levels (levels are defined for all values in the column).

Two encodings for the levels are supported BITPACKED and RLE. Only RLE is now used as it supersedes BITPACKED.

- View page source
- Edit this page
- Create child page
- Create documentation issue
- Print entire section

Parquet DataType ~ Spark DataType

Nested Encoding with « definiton » + « repetition »

DataType

JSON DataType ~ Spark DataType

(map with string only)

Nested Fields in Spark SQL UDF

```
SELECT ename, dept_list
FROM employee
```

```
+-----------+--------------+
| ename  | dept_list  |
+-----------+--------------+
| Tom      | [20]           |
| Jerry    | [10,20]      |
| Riley    | [20,30,40] |
+-----------+--------------+
```

```
SELECT ename,
    exists(dept_list, x -> x = 10) as found10
FROM employee
```

```
+-----------+--------------+
| ename  | found10   |
+-----------+--------------+
| Tom      | false         |
| Jerry    | true           |
| Riley    | false          |
+-----------+--------------+
```

SQL Grammar Extension: « lateral view »

```
SELECT ename, dept_list
FROM employee
```

```
+-----------+--------------+
| ename     | dept_list    |
+-----------+--------------+
| Tom       | [20]         |
| Jerry     | [10,20]      |
| Riley     | [20,30,40]   |
+-----------+--------------+
```

```
SELECT ename, dept_id
FROM employee
LATERAL VIEW explode(dept_list) depts AS dept_id;
```

```
+---------+----------+
|ename    |dept_id   |
+---------+----------+
| Tom     | 20       |
| Jerry   | 10       |
| Jerry   | 20       |
| Riley   | 20       |
| Riley   | 30       |
| Riley   | 40       |
+---------+----------+
```

Annexe

More SQL: collect_list(row) -> List

```
SELECT ename, dept_list
FROM employee
```

```
+-----------+--------------+
| ename     | dept_list    |
+-----------+--------------+
| Tom       | [20]         |
| Jerry     | [10,20]      |
| Riley     | [20,30,40]   |
+-----------+--------------+
```

```
SELECT ename, collect_list(dept_id + 1) as ls
FROM ( SELECT employee
    LATERAL VIEW explode(dept_list) depts AS dept_id )
GROUP BY ename
```

```
+-----------+--------------+
| ename     | ls           |
+-----------+--------------+
| Tom       | [21]         |
| Jerry     | [11,21]      |
| Riley     | [21,31,41]   |
+-----------+--------------+
```

UDF / UDAF   (User Defined Aggregate Function)

Example UDF :   f(x, y) { return x + y }

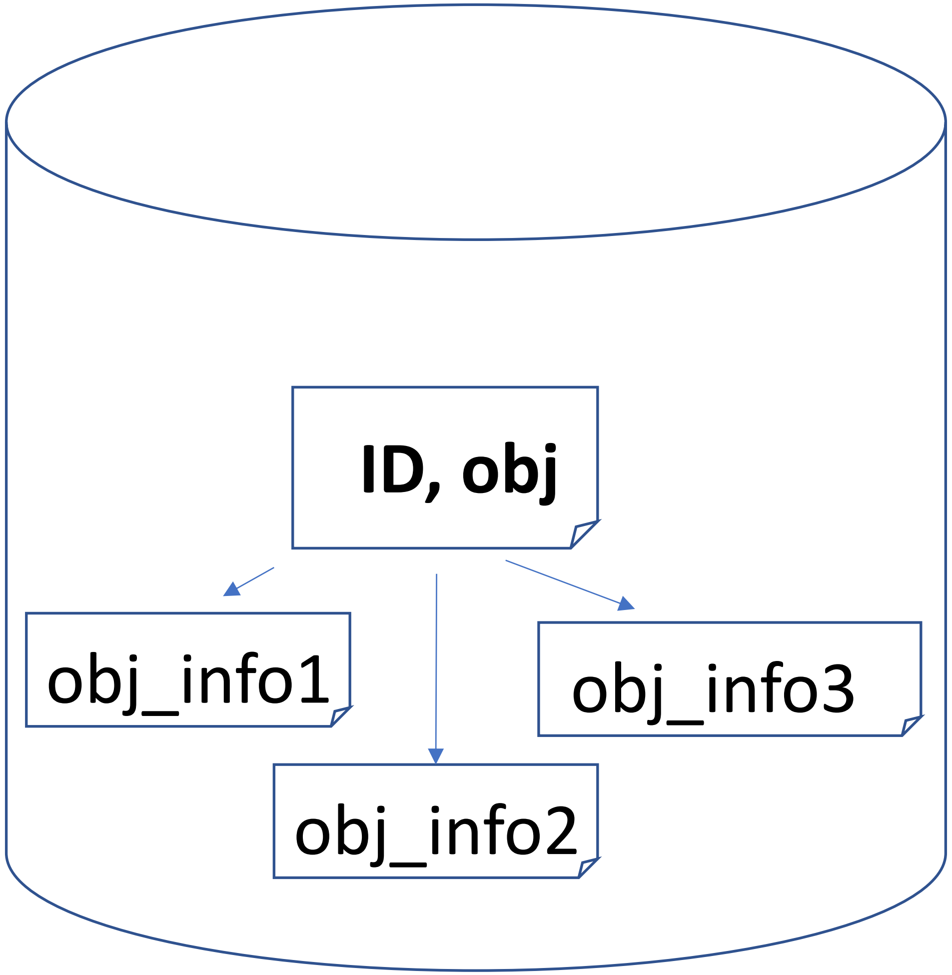
Function like in Math : idempotent, side-effect less, ..

! = UDAF : Aggregate / Accumulator

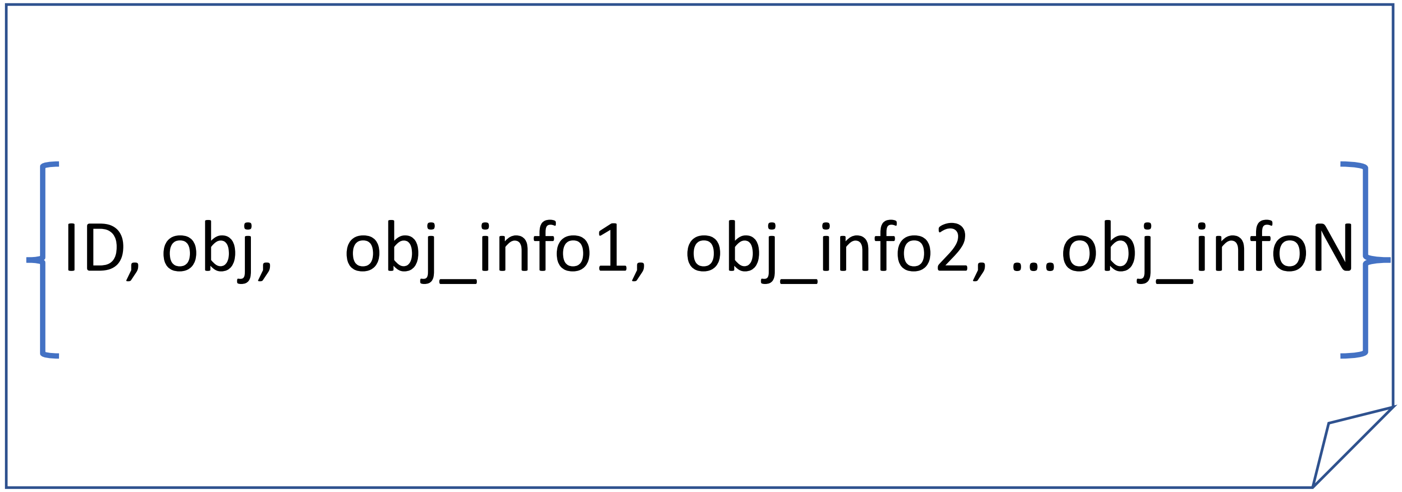
like in « SELECT count(..), sum(..), average(..) FROM .. GROUP BY .. »

Object instance, Class with 3 methods:
init()
add(value)
Result getResult()

List, Map, Struct … denormalize data, avoid Joins



Normalized relationnal database



Efficient DE-normalized analytics system

Enough for Today !

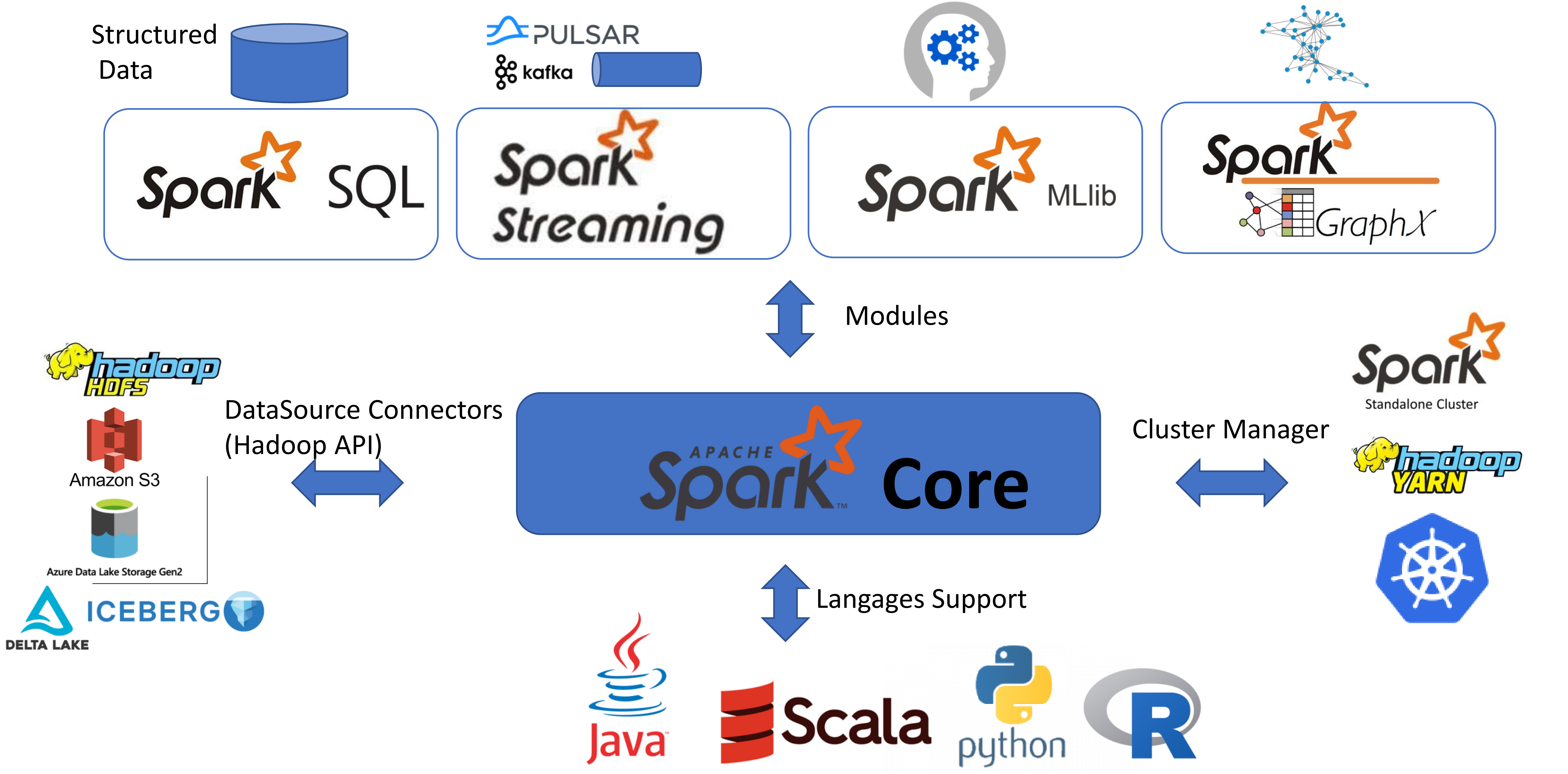
… but more to come

Questions are Welcome

Take Away

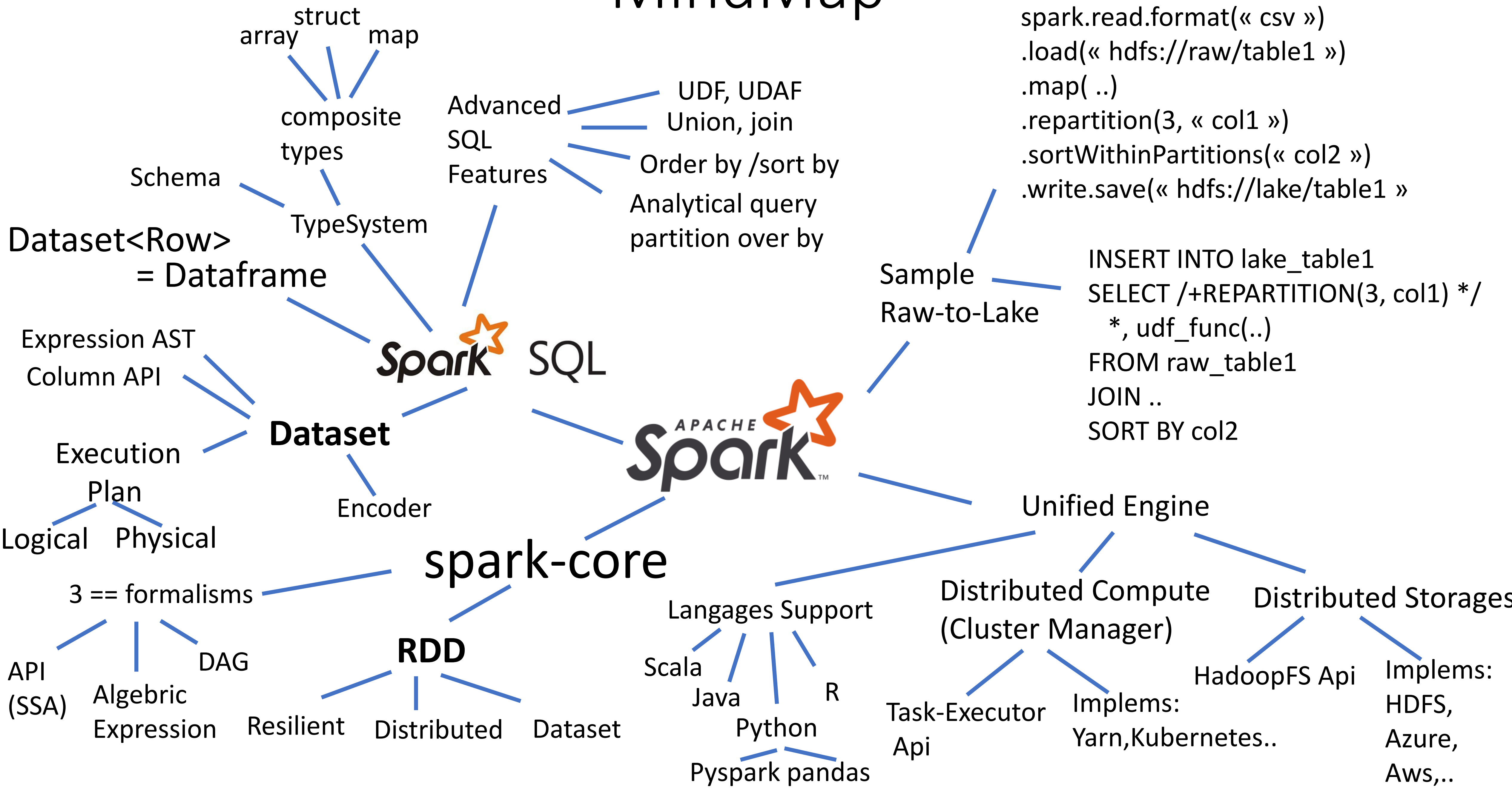
What Did you learn ?

Spark-Core + …



Structured
Data
PULSAR
kafka
Spark SQL
Spark Streaming
Spark MLlib
Spark GraphX
Modules
hadoop HDFS
Amazon S3
Azure Data Lake Storage Gen2
DELTA LAKE
ICEBERG
DataSource Connectors
(Hadoop API)
APACHE Spark Core
Cluster Manager
Spark Standalone Cluster
hadoop YARN
Langages Support
Java
Scala
python
R

MindMap


struct
array
map
composite types
Schema
TypeSystem
Advanced SQL Features
UDF, UDAF
Union, join
Order by /sort by
Analytical query partition over by
Dataset<Row> = Dataframe
Spark SQL
Expression AST
Column API
Dataset
Execution Plan
Logical
Physical
Encoder
APACHE Spark
spark.read.format(« csv »)
.load(« hdfs://raw/table1 »)
.map(..)
.repartition(3, « col1 »)
.sortWithinPartitions(« col2 »)
.write.save(« hdfs://lake/table1 »
Sample Raw-to-Lake
INSERT INTO lake_table1
SELECT /+REPARTITION(3, col1) */
*, udf_func(..)
FROM raw_table1
JOIN ..
SORT BY col2
spark-core
3 == formalisms
API (SSA)
Algebric Expression
DAG
RDD
Resilient
Distributed
Dataset
Unified Engine
Langages Support
Scala
Java
Python
R
Pyspark pandas
Distributed Compute (Cluster Manager)
Task-Executor Api
Implems: Yarn,Kubernetes..
Distributed Storages
HadoopFS Api
Implems: HDFS, Azure, Aws,..

Next..

Lesson 1 : introduction to BigData / Hadoop cluster / Spark

Lesson 2 :   this document : spark-core / spark-sql

Lesson 3 : cluster management, tuning args, deploy-mode

Lesson 4 : Advanced Spark: UI, parquet PPD, java api, spark-streaming, ..