Big Data

Hands on PySpark

How do we process Big Data?

Main issues

- Where do we store the data?
- How do we process it?

Big Data greatly exceeds the size of the typical drives

- Even if a big drive existed, it would be too slow (at least for now)

Year: 1990
Size: 1.3 GB
Speed: 4,4 MB/s

5 minutes

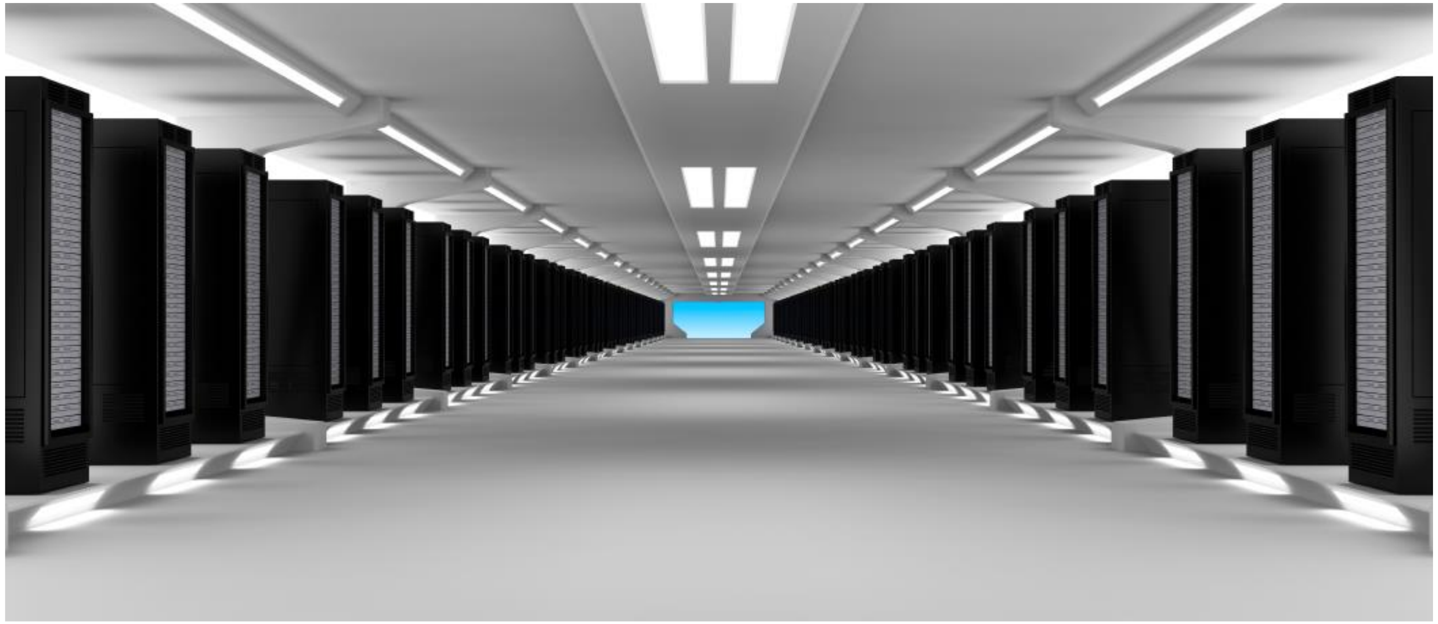
Year: 2014
Size: 1 TB
Speed: 100 MB/s

3 hours

Year: 2015
Size: 1 TB
Speed: 600 MB/s

30 minutes

The answer: cluster computing



100 hard disks? 2 mins to read 1TB

Commodity hardware

You are not tied to expensive, proprietary offerings from a single vendor

You can choose standardized, commonly available hardware from a large range of vendors to build your cluster

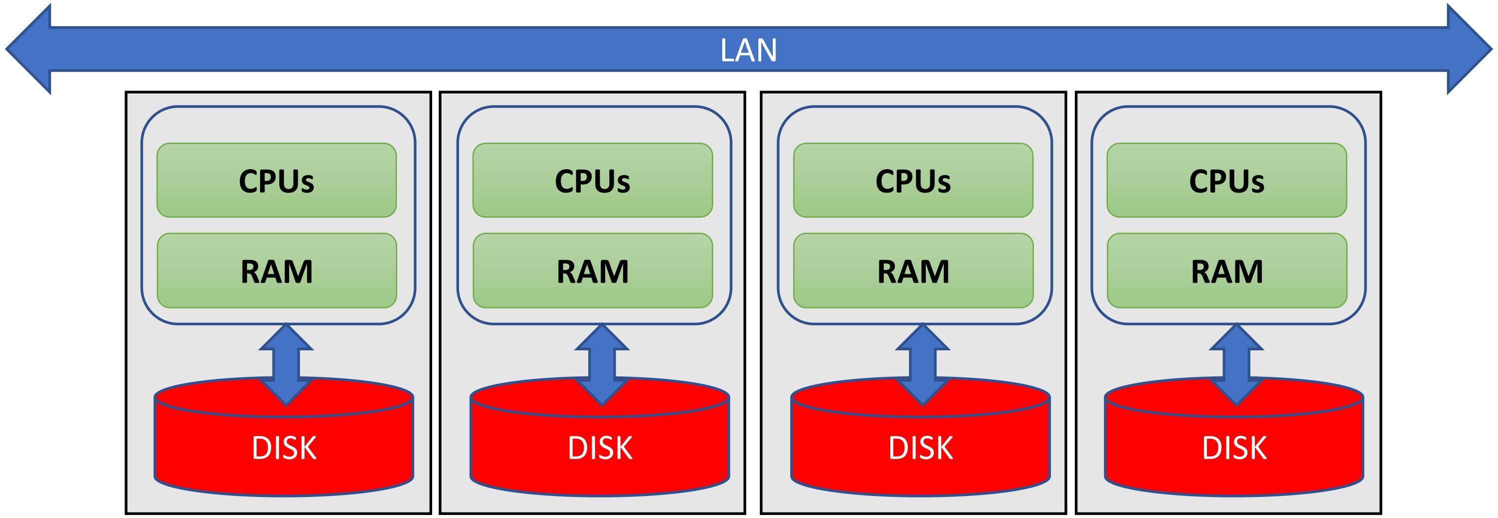
Commodity ≠ Low-end!

- Cheap components with high failure rate can be a false economy

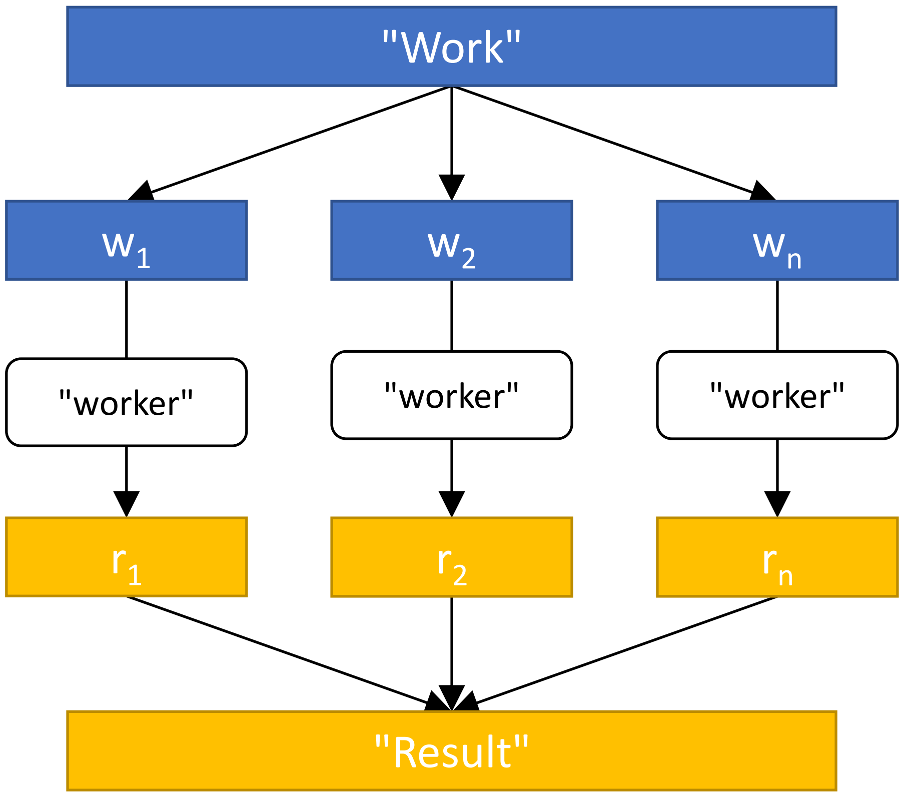
Cluster Computing Architecture

A computer cluster is a group of linked computers (nodes), working together closely so that in many respects they form a single computer

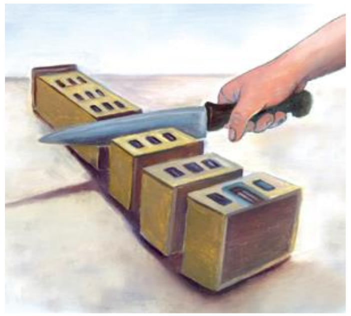
- Typically connected to each other through fast LAN
- Every node is a system on its own, capable of independent operations
  - Unlimited scalability, no vendor lock-in
- Number of nodes in the cluster >> Number of CPUs in a node



Distributed computing: an old idea



Divide



Conquer

What is the solution?

Hide system-level details from the developers

- No race conditions, lock contention, etc.
- No need to become hardcore techies

Separate the *what* from the *how*

- Developer specifies the computation that needs to be performed
- Execution framework (“runtime”) handles the actual execution

The datacenter IS the computer!

Spark

It is a fast and general-purpose execution engine

- In-memory data storage for very fast iterative queries
- Easy interactive data analysis
- Combines different processing models (machine learning, SQL, streaming, graph computation)
- Provides (not only) a MapReduce-like engine...
- ... but it's up to 100x faster than Hadoop MapReduce

Compatible with Hadoop's storage APIs

- Can run on top of a Hadoop cluster
- Can read/write to any database and any Hadoop-supported system, including HDFS, HBase, Parquet, etc.

What does Spark offer?

In-memory data caching

- HDD is scanned once, then data is written to/read from RAM

Lazy computations

- The job is optimized before its execution

Efficient pipelining

- Writing to HDD is avoided as much as possible

Spark pillars

Two main abstractions of Spark

RDD – Resilient Distributed Dataset

- An RDD is a collection of data items
- It is split into partitions
- It is stored in memory on the worker nodes of the cluster

DAG – Direct Acyclic Graph

- A DAG is a sequence of computations performed on data
- Each node is an RDD
- Each edge is a transformation of one RDD into another

RDD

RDDs are immutable distributed collection of objects

- **Resilient**: automatically rebuild on failure
- **Distributed**: the objects belonging to a given collection are split into *partitions* and spread across the nodes
  - RDDs can contain any type of Python, Java, or Scala objects
  - Distribution allows for scalability and locality-aware scheduling
  - Partitioning allows to control parallel processing

Fundamental characteristics (mostly from *pure functional programming*)

- **Immutable**: once created, it can’t be modified
- **Lazily evaluated**: optimization before execution
- **Cacheable**: can persist in memory, spill to disk if necessary
- **Type inference**: data types are not declared but inferred (≠ dynamic typing)

RDD operations

RDDs offer two types of operations: *transformations* and *actions*

Transformations construct a new RDD from a previous one

- E.g.: map, flatMap, reduceByKey, filtering, etc.
- https://spark.apache.org/docs/latest/programming-guide.html#transformations

Actions compute a result that is either returned to the driver program or saved to an external storage system (e.g., HDFS)

- E.g.: saveAsTextFile, count, collect, etc.
- https://spark.apache.org/docs/latest/programming-guide.html#actions

RDD operations

RDDs are **lazily evaluated**, i.e., they are computed when they are used in an action

- Until no action is fired, the data to be processed is not even accessed

Example (in Python)

```
sc = new SparkContext
rddLines = sc.textFile("myFile.txt")
rddLines2 = rddLines.filter (lambda line: "some text" in line)
rddLines2.first()
```

Transformations

Action

There is no need to compute and store everything

- In the example, Spark simply scans the file until it finds the first matching line

DAG

Based on the user application and on the lineage graphs,
Spark computes a logical execution plan in the form of a DAG

- Which is later transformed into a physical execution plan

The DAG (Directed Acyclic Graph) is a sequence of computations performed on data

- Nodes are **RDDs**
- Edges are operations on RDDs
- The graph is Directed: transformations from a partition A to a partition B
- The graph is Acyclic: transformations cannot return an old partition

Application decomposition

Application

- Single instance of SparkContext that stores data processing logic and schedules series of jobs, sequentially or in parallel

Job

- Complete set of transformations on RDD that finishes with action or data saving, triggered by the driver application

Stage

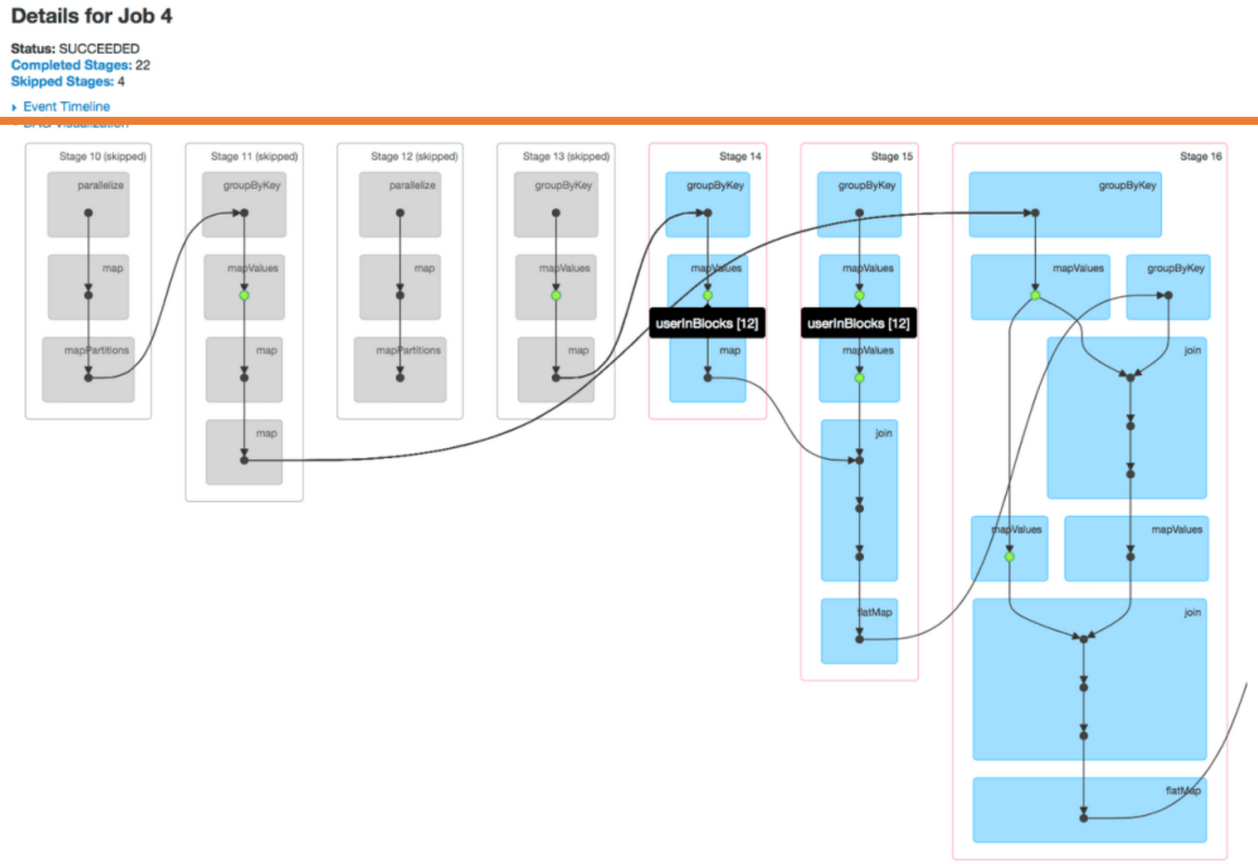
- Set of transformations that can be pipelined and executed by a single independent worker

Task

- Basic unit of scheduling: executes the stage on a single data partition

Application decomposition

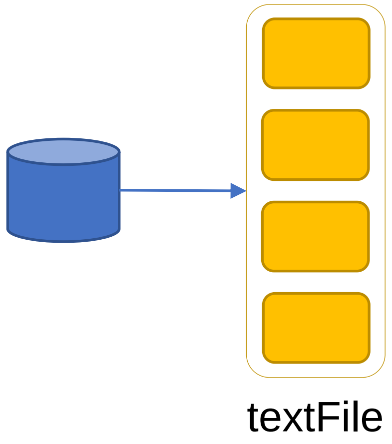
Job



DAG example

Word count in Scala

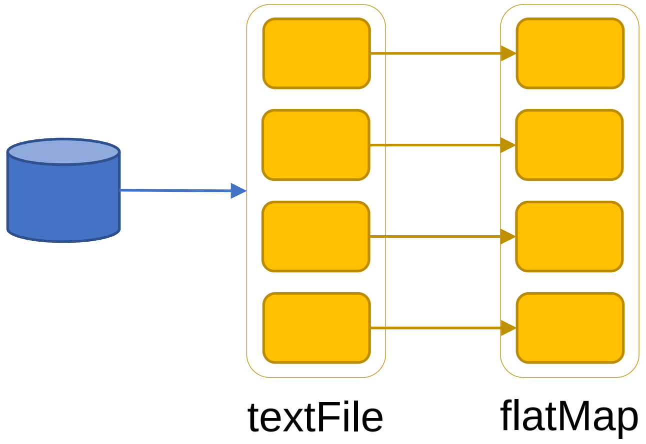
textFile = sc.textFile("hdfs://...")



DAG example

Word count in Scala

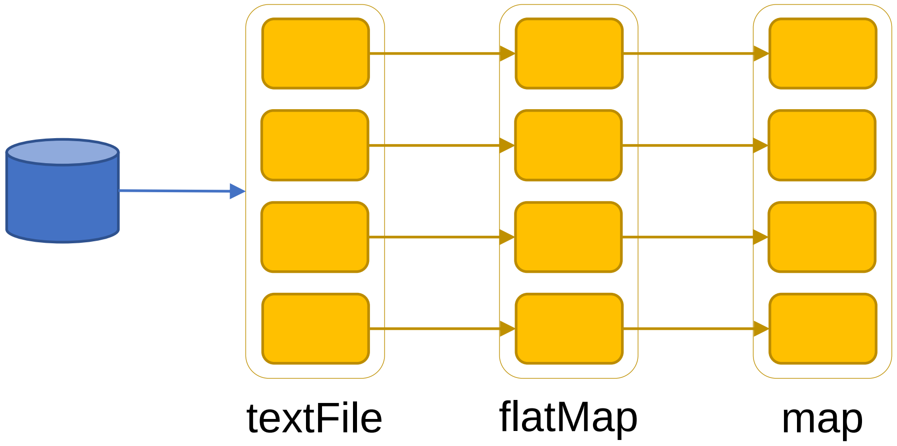
```
textFile = sc.textFile("hdfs://...")
counts = textFile
    .flatMap(line => line.split(" "))
```



DAG example

Word count in Scala

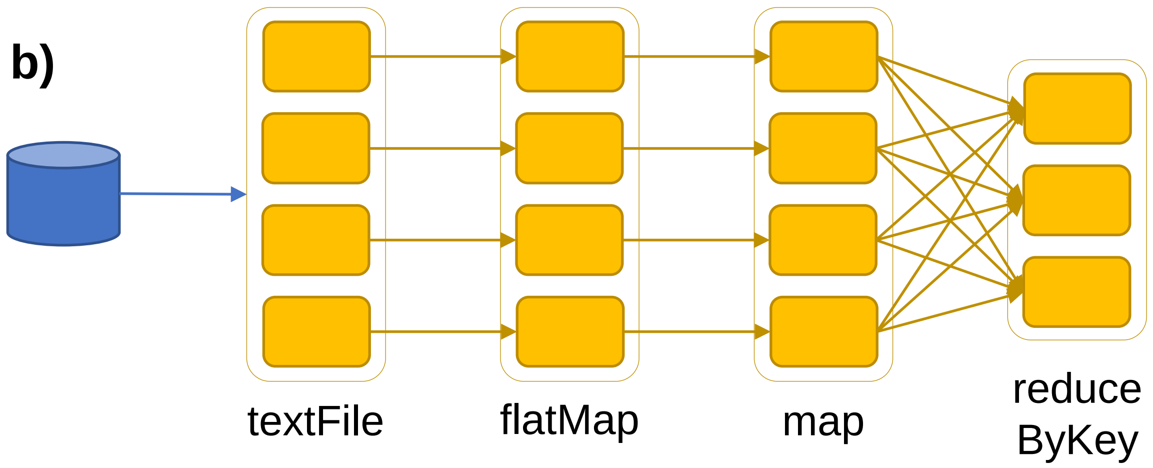
```
textFile = sc.textFile("hdfs://...")
counts = textFile
    .flatMap(line => line.split(" "))
    .map(lambda word: (word, 1))
```



DAG example

Word count in Scala

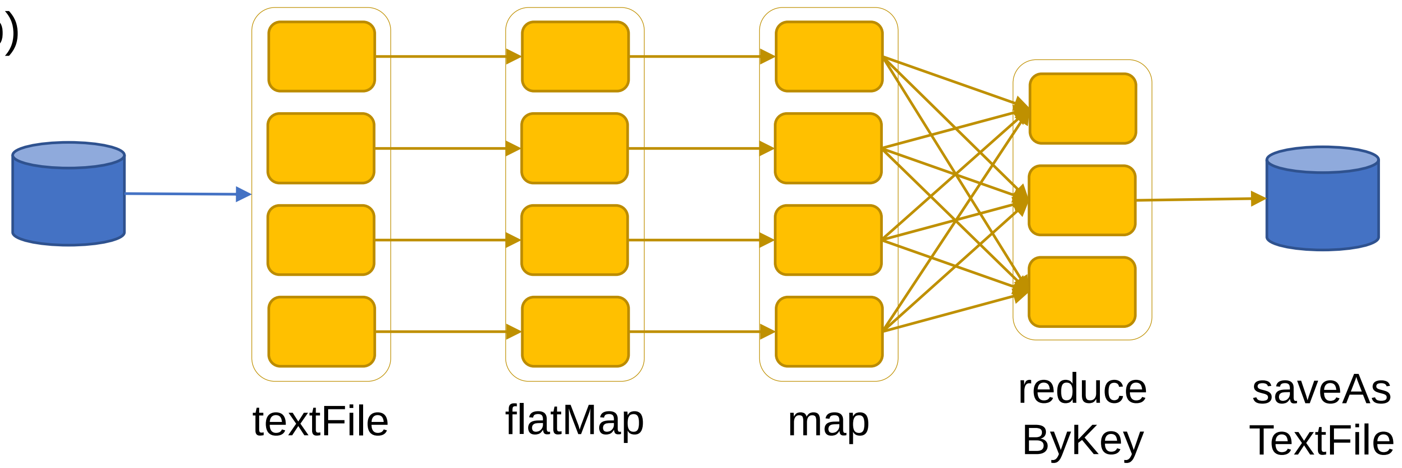
```
textFile = sc.textFile("hdfs://...")
counts = textFile
    .flatMap(line => line.split(" "))
    .map(lambda word: (word, 1))
    .reduceByKey(lambda a, b: a + b)
```



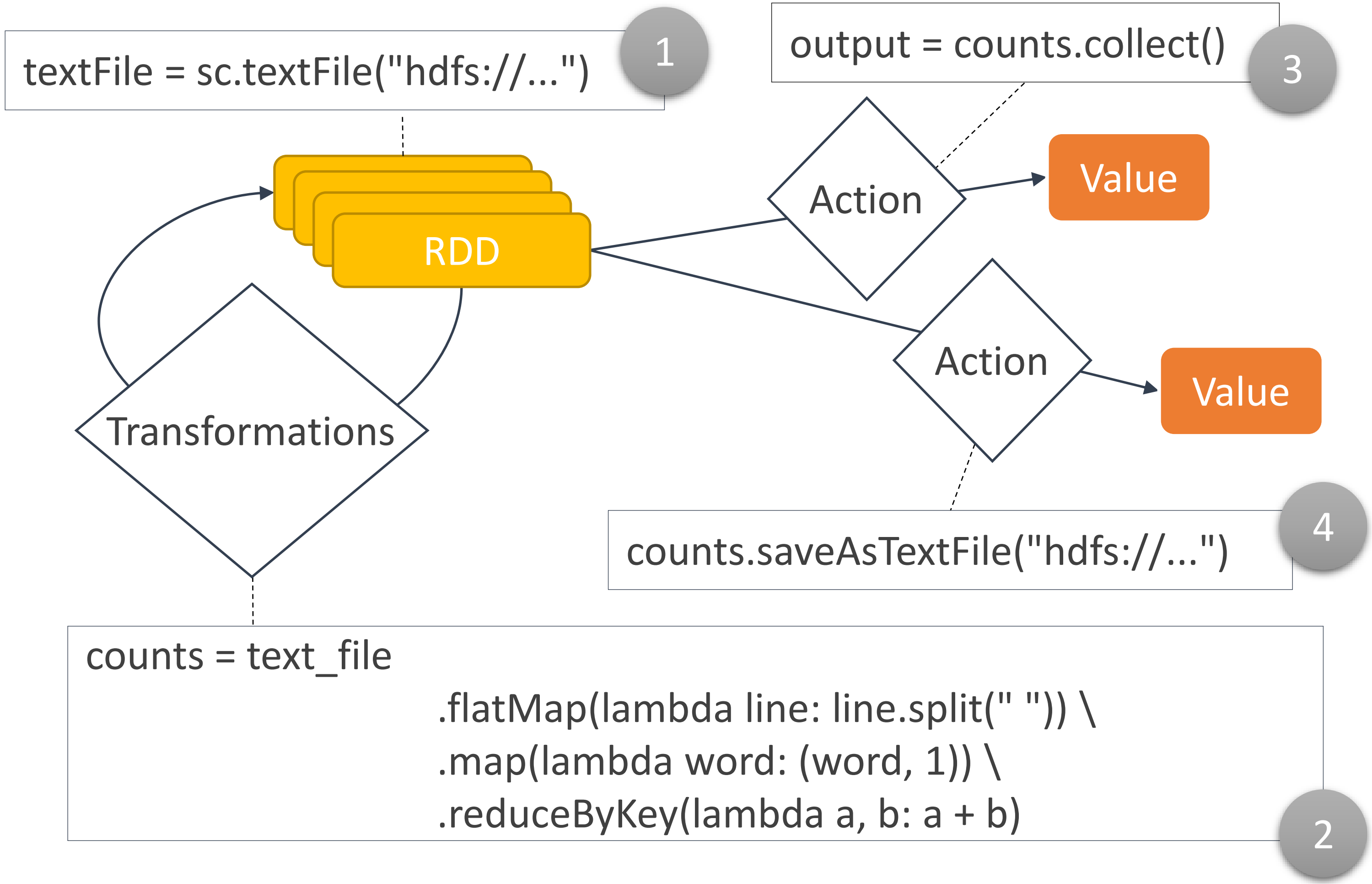
DAG example

Word count in Scala

```
textFile = sc.textFile("hdfs://...")
counts = textFile
    .flatMap(line => line.split(" "))
    .map(lambda word: (word, 1))
    .reduceByKey(lambda a, b: a + b)
counts.saveAsTextFile("hdfs://...")
```



Conceptual representation



DataFrame and DataSet

RDDs are immutable distributed collection of objects

DataFrames and DataSets are immutable distributed collection of records organized into named columns (i.e., a table)

- **Simply put, RDDs with a schema attached**
- Support both relational and procedural processing (e.g., SQL, Scala)
- Support complex data types (struct, array, etc.) and user defined types
- Cached using columnar storage

Can be built from many different sources

- DBMSs, files, other tools (e.g., Hive), RDDs

Type conformity is checked

- At *compile time* for DataSets; at *runtime* for DataFrames

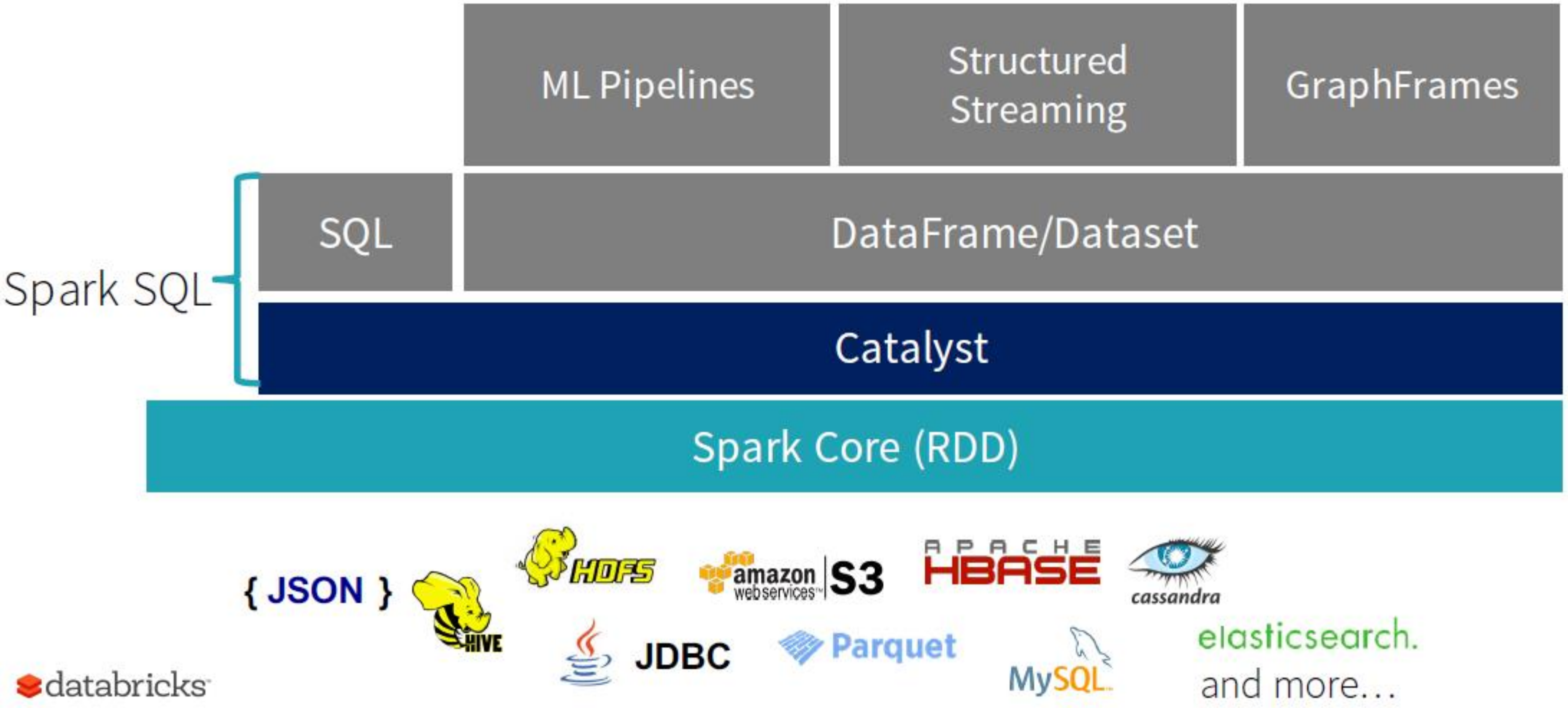
DataFrame and DataSet

Still lazily evaluated…

…but supports under-the-hood optimizations and code generation

- **Catalyst optimizer creates optimized execution plans**
  - IO optimizations such as skipping blocks in parquet files
  - Logic push-down of selection predicates
- JVM code generation for all supported languages
  - Even non-native JVM languages; e.g., Python

Spark structured



Why structure?

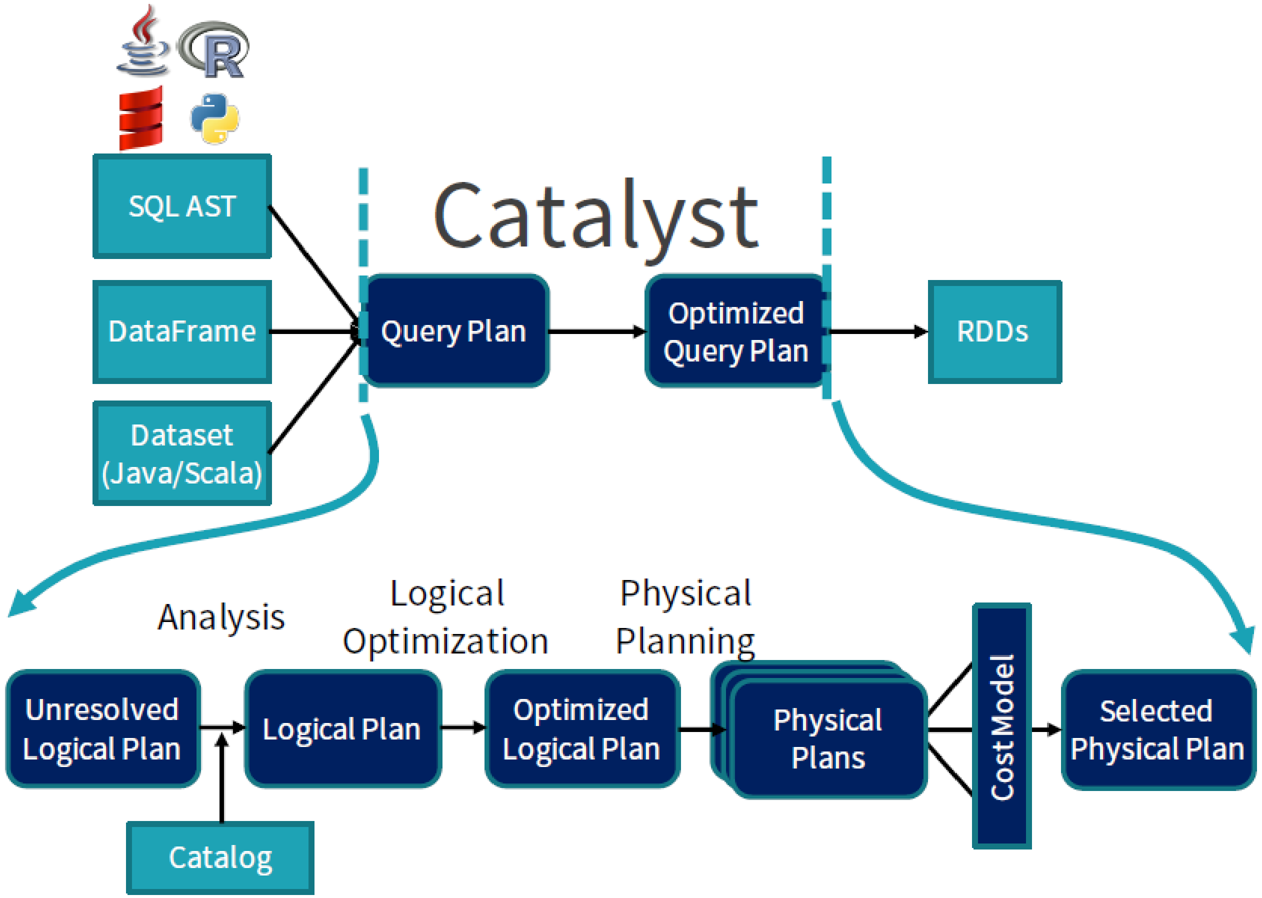
Cons

- **Structure imposes some limits**
  - RDDs enable any computation through user defined functions

Pros

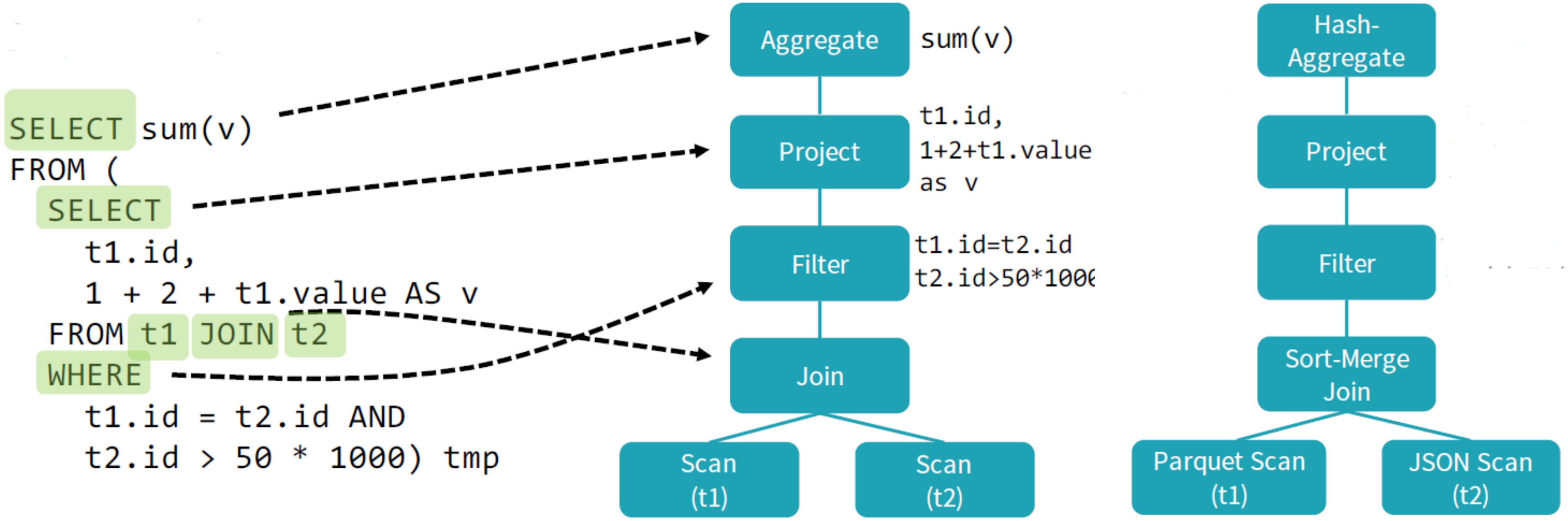
- The most common computations are supported
- Language simplicity
- **Opens the room to optimizations**
  - Hard to optimize a user defined function

Catalyst



SQL AST
DataFrame
Dataset
(Java/Scala)
Catalyst
Query Plan
Optimized
Query Plan
RDDs
Analysis
Logical
Optimization
Physical
Planning
Unresolved
Logical Plan
Logical Plan
Optimized
Logical Plan
Physical
Plans
Cost Model
Selected
Physical Plan
Catalog

Logical and Physical Plan



Logical Plan

Describes **what** computation must be done

Physical Plan

Describes **what** computation must be done and **how** to conduct it (i.e., which algorithms are used)

Logical optimization

Based on rules

- **A rule is a function** that can be applied on a portion of the logical plan

Implemented as Scala functions

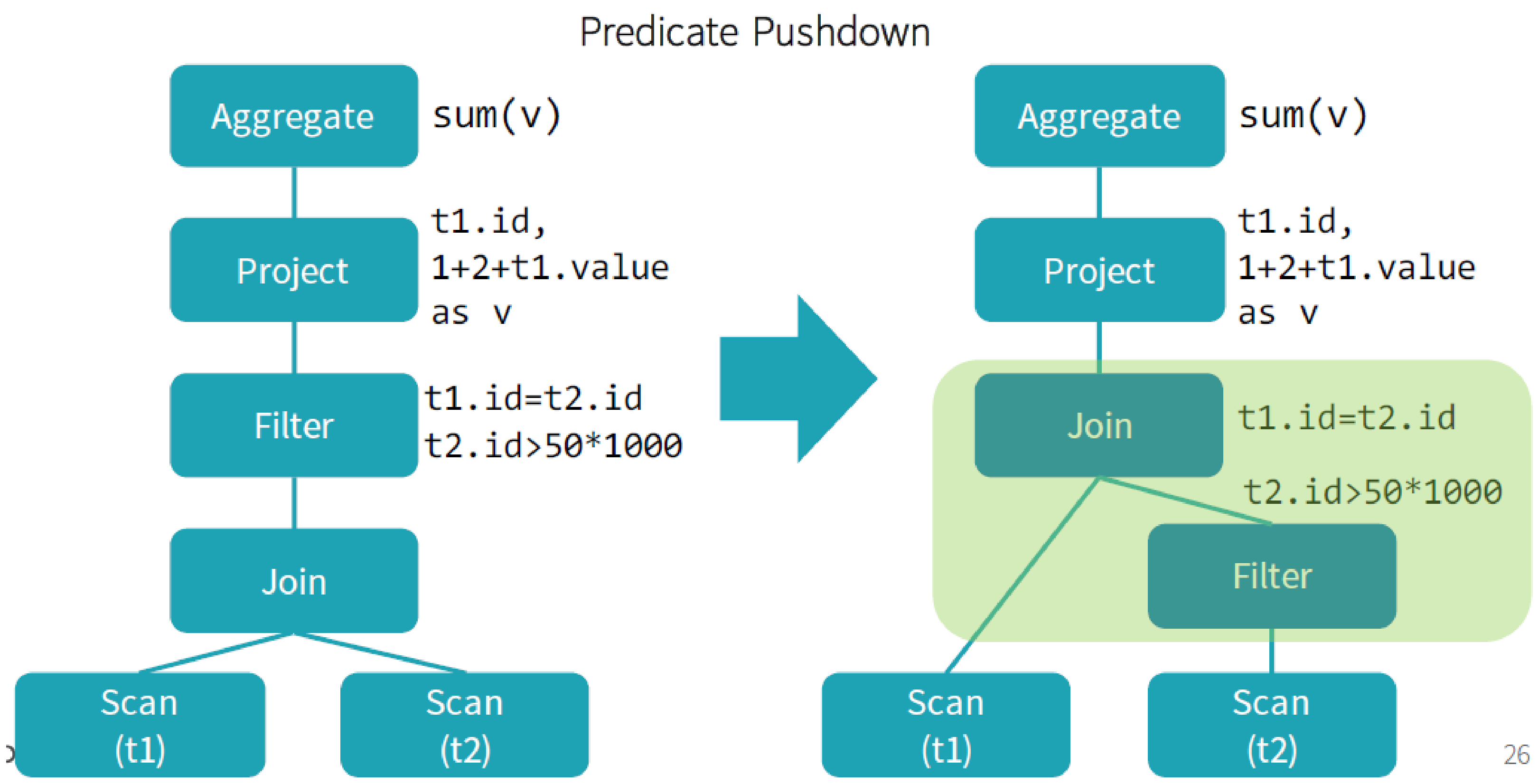
```
val expression: Expression = ...
expression.transform {
  case Add(Literal(x, IntegerType), Literal(y, IntegerType)) =>
    Literal(x + y)
}
```

Several types of rules

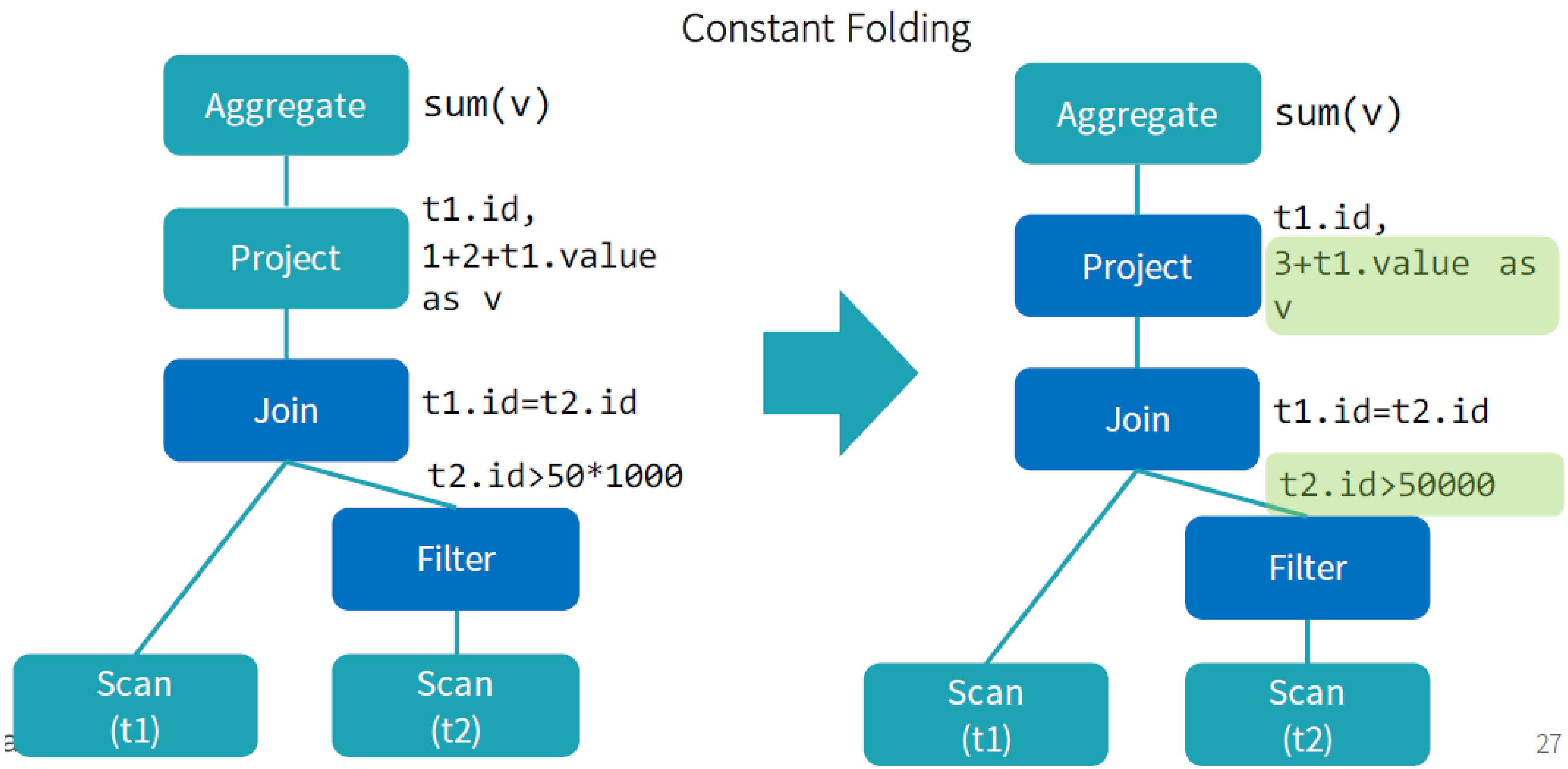
- **Constant folding**: resolve constant expressions at compile time
- **Predicate pushdown**: push selection predicates close to the sources
- **Column pruning**: project only the required column
- **Join reordering**: change the order of join operations

Applied recursively and iteratively until the plan reaches a *fixed point*

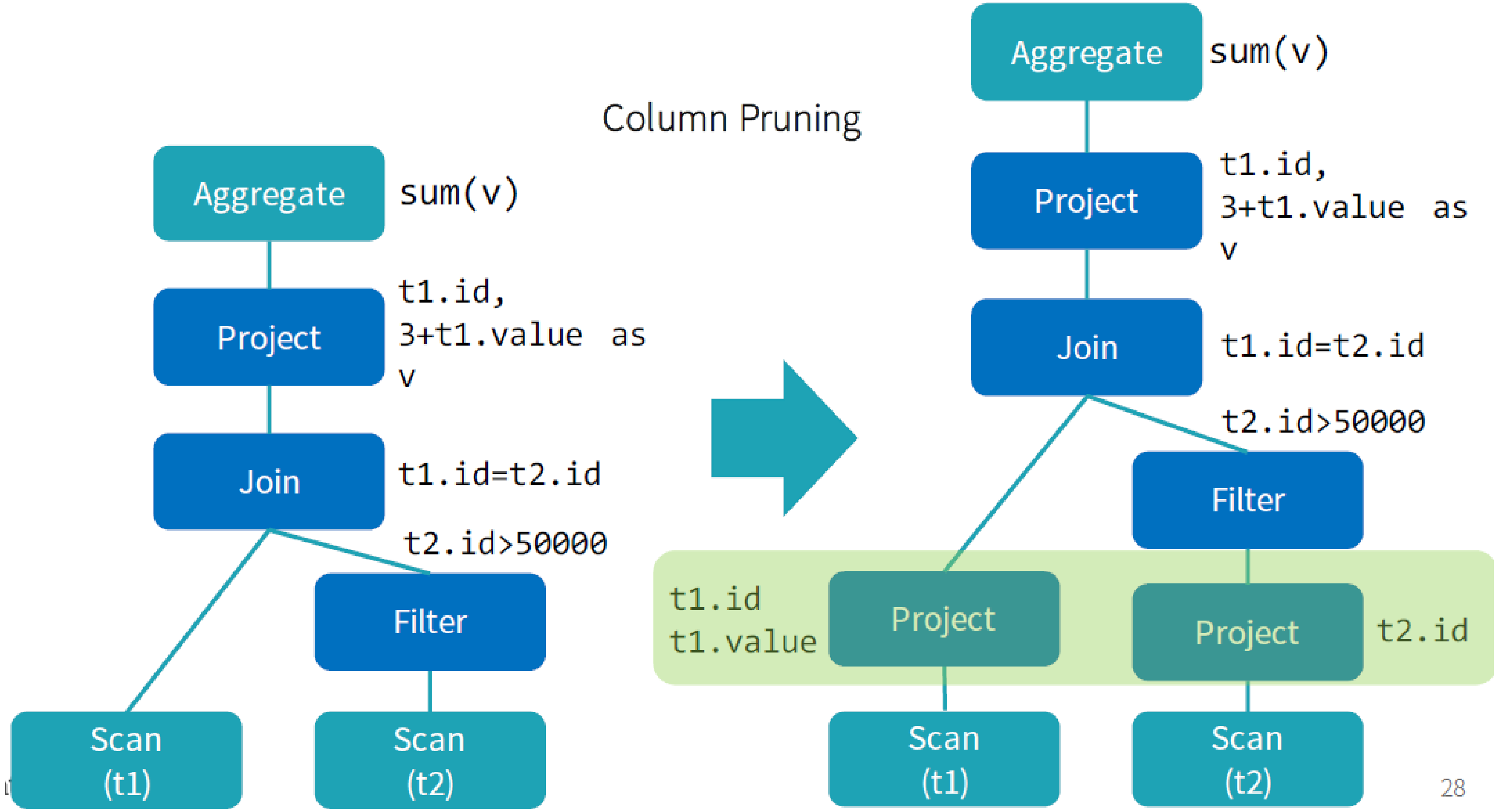
Logical optimization



Logical optimization



Logical optimization

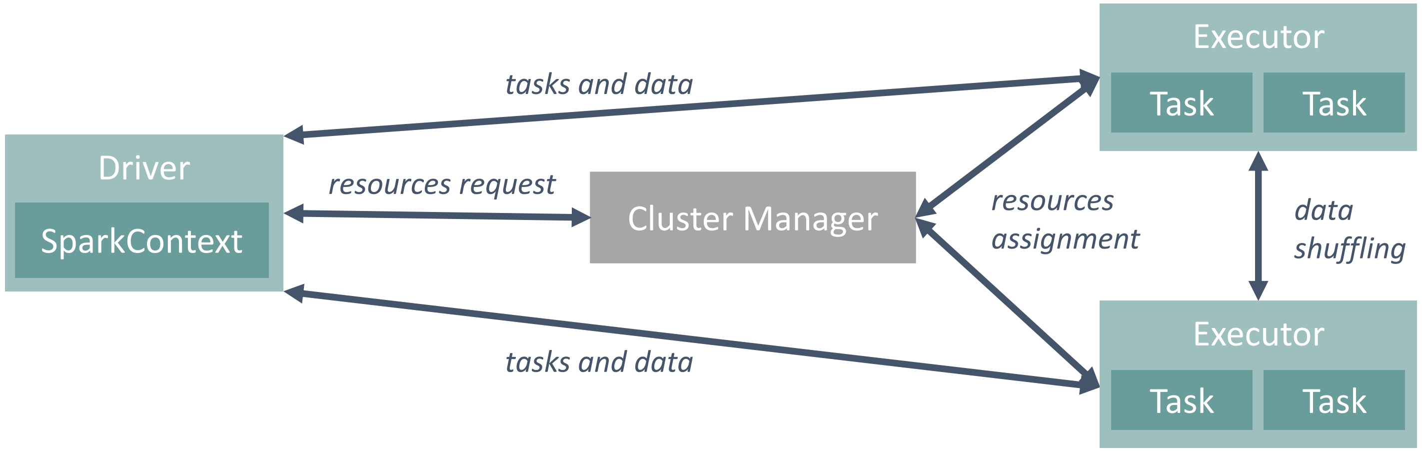


Spark architecture

Spark uses a *master/slave architecture* with one central coordinator (*driver*) and many distributed workers (*executors*)

- The driver and each executor are independent Java processes
- Together they form a Spark *application*

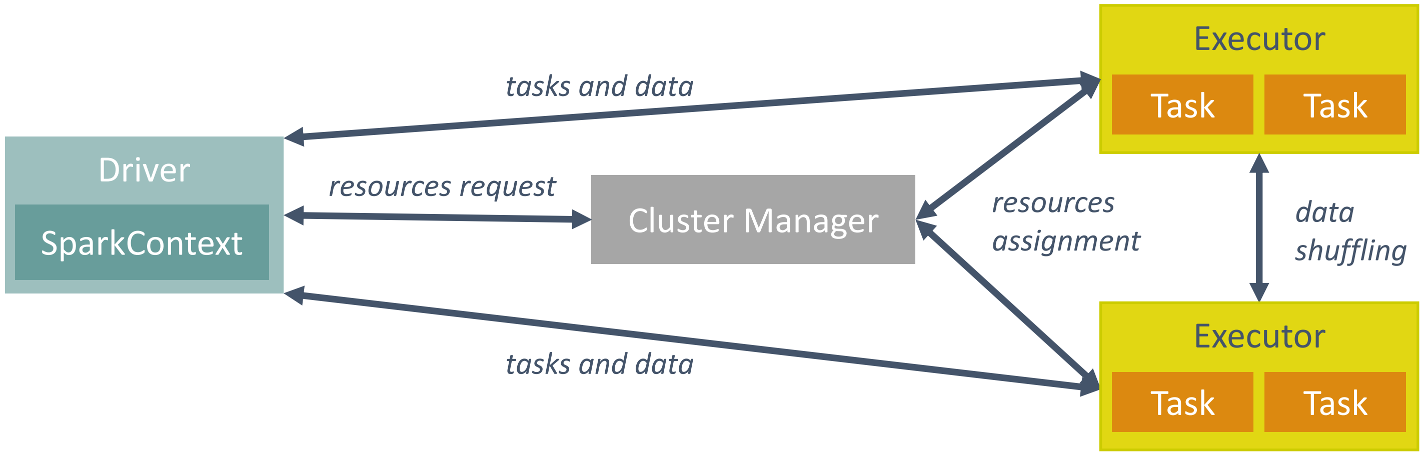
The architecture is independent of the cluster manager that Spark runs on



Spark architecture

Executor: a process responsible for executing the received tasks

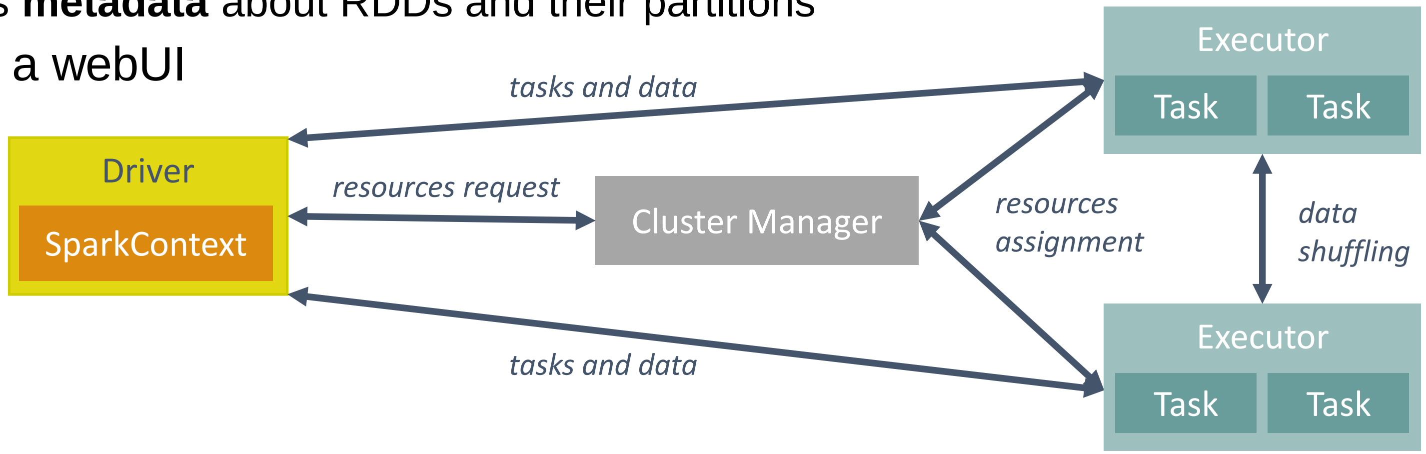
- Each spark application can have (and usually has) multiple executors, and each worker node can host many executors
- Typically runs for the entire duration of the application
- Stores (caches) RDD data in JVM heap
- Tasks are the smallest unit of work and are carried out by executors



Spark architecture

Driver Program (a.k.a. *Spark Driver*, or simply *Driver)*

- Each spark application can only have one driver (entry point of Spark Shell)
- Converts user program into tasks
  - Creates the **SparkContext**, i.e., the object that handles communications
  - Computes the logical **DAG** of operations and converts it into a physical **execution plan**
- Schedules tasks on executors
  - Has a **complete view** of the available executors and schedules tasks on them
  - Stores **metadata** about RDDs and their partitions
- Launches a webUI



Spark

Suggested reading and resources