RESILIENT DISTRIBUTED DATASETS

Resilient Distributed Datasets

We've talked a little bit about RDDs before

Recap

Resilient Distributed Datasets

RDDs are the main programming abstraction in Spark

Recap

Resilient Distributed Datasets

RDDs are in-memory objects and all data is processed using them

Recap

Resilient Distributed Datasets

in-memory

In-memory, yet resilient!

RDDs are designed to be fault-tolerant while dealing with vast amounts of data

Recap

Resilient Distributed Datasets

There are a lot of complexities involved in dealing with vast amounts of data

Recap

Resilient Distributed Datasets

Lot of complexities

1. Distributing data across a cluster
2. Fault tolerance (if in-memory data is lost)
3. Efficiently processing billions of rows of data

Recap

Resilient Distributed Datasets

1. Distributing data across a cluster
2. Fault tolerance
3. Efficiently processing billions of rows of data

Think of other environments
like Java or Excel

It would be the responsibility of the
user to deal with most of these issues

Recap

Resilient Distributed Datasets

1. Distributing data across a cluster
2. Fault tolerance
3. Efficiently processing billions of rows of data

With RDDs, you can interact and play with billions of rows of data

...without caring about any of the complexities

Recap

Resilient Distributed Datasets

1. Distributing data across a cluster
2. Fault tolerance
3. Efficiently processing billions of rows of data

Spark takes care of all the complexities under the hood

The user is completely abstracted!

Resilient Distributed Datasets

RDDs have 3 defining characteristics

partitions

read-only

lineage

Resilient Distributed Datasets

partitions

read-only

lineage

Let's understand each of these

RDDs partitions

RDDs represent data in-memory

1	Swetha	30	Bangalore
2	Vitthal	35	New Delhi
3	Navdeep	25	Mumbai
4	Janani	35	New Delhi
5	Navdeep	25	Mumbai
6	Janani	35	New Delhi

RDDs partitions

When you have large amounts of data

1. It can take very long to process them on a single machine

2. The data may be too large to even be stored and processed on a single machine

1	Swetha	30	Bangalore
2	Vitthal	35	New Delhi
3	Navdeep	25	Mumbai
4	Janani	35	New Delhi
5	Navdeep	25	Mumbai
6	Janani	35	New Delhi

RDDs partitions

Instead you can divide the data into partitions

and distribute them to multiple machines

1	Swetha	30	Bangalore
2	Vitthal	35	New Delhi
3	Navdeep	25	Mumbai
4	Janani	35	New Delhi
5	Navdeep	25	Mumbai
6	Janani	35	New Delhi

RDDs partitions

These partitions are distributed to multiple machines/nodes

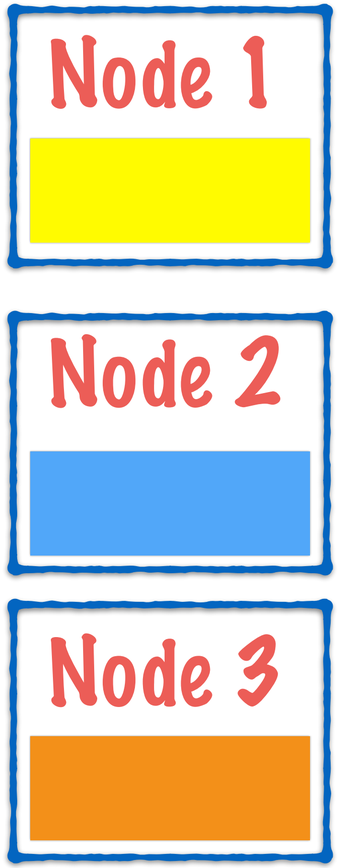
1	Swetha	30	Bangalore
2	Vitthal	35	New Delhi

3	Navdeep	25	Mumbai
4	Janani	35	New Delhi

5	Navdeep	25	Mumbai
6	Janani	35	New Delhi

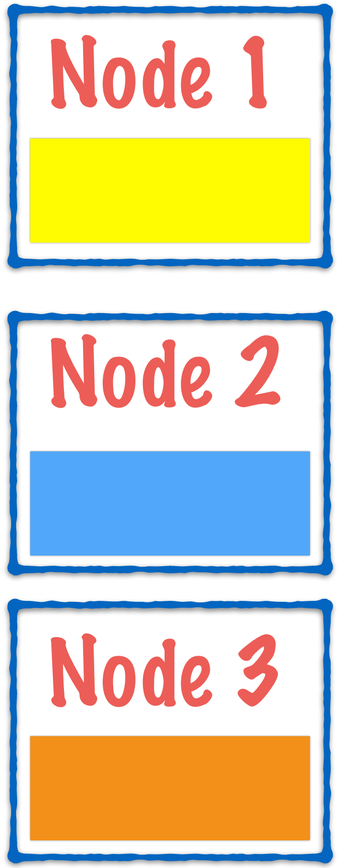
RDDs partitions

These partitions are distributed to multiple machines/nodes



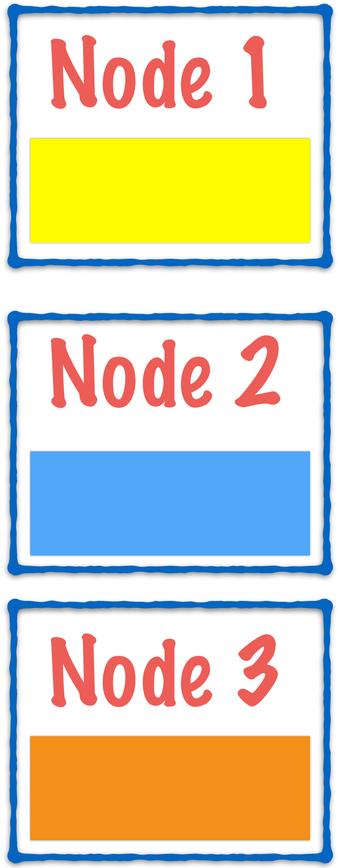
RDDs partitions

The nodes can work in parallel, making the processing much faster than if it were on a single machine



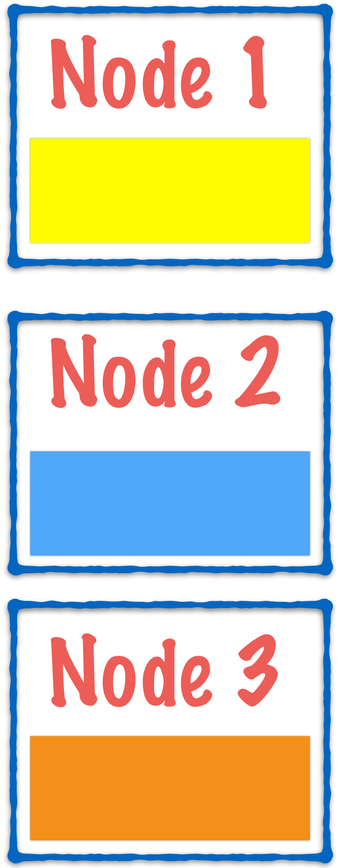
RDDs partitions

Each partition, is kept in-memory on a node in the cluster



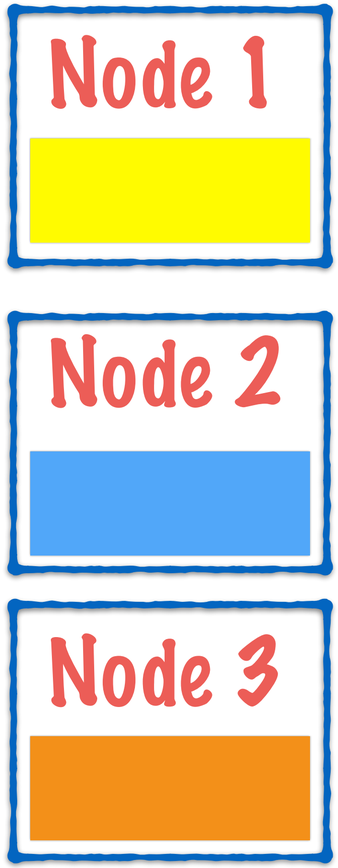
RDDs partitions

How is the data partitioned and distributed?



RDDs partitions

Usually the data comes from a distributed filesystem like HDFS where the data is already partitioned!

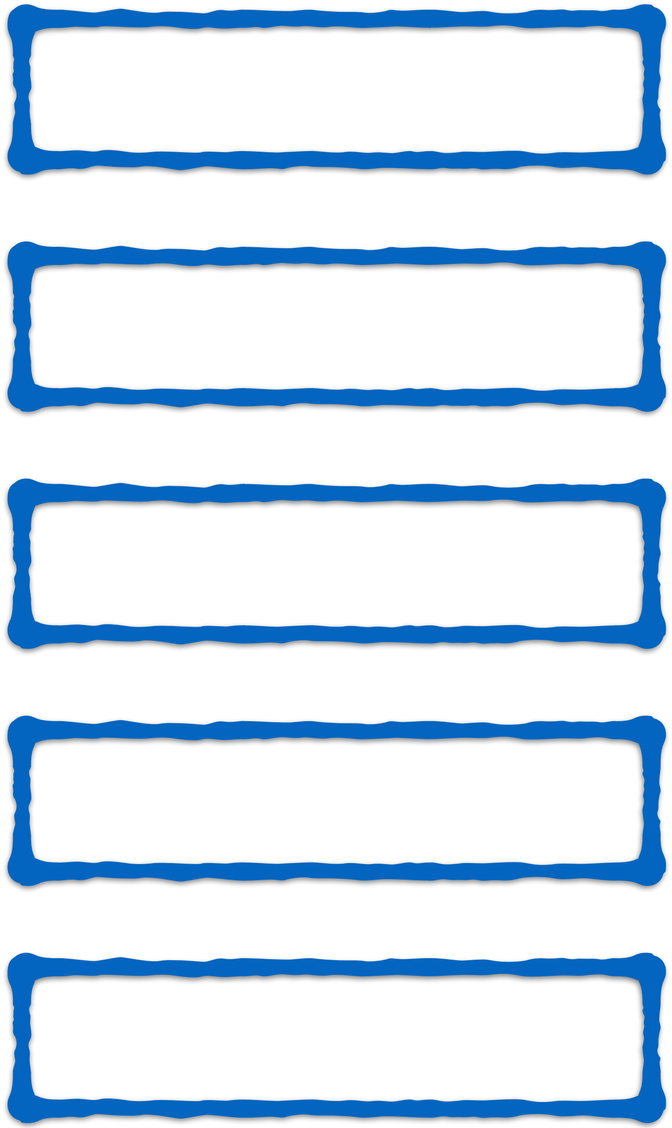


HDFS

The Hadoop Distributed File System

Hadoop uses this to store data across multiple disks

HDFS

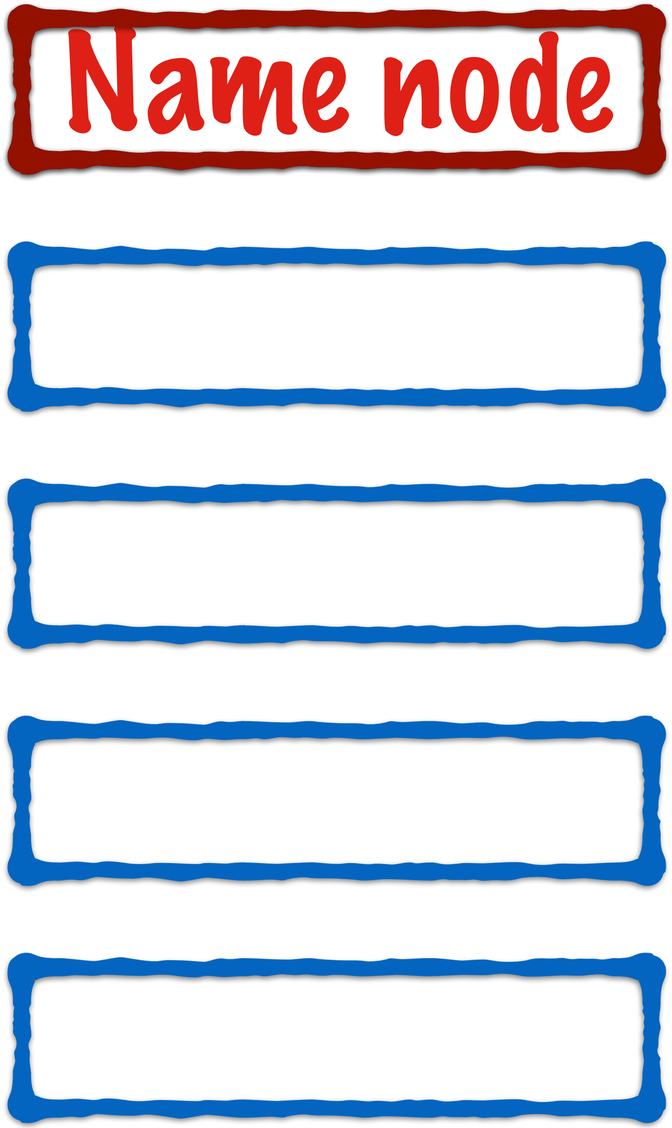


Hadoop is normally deployed on a group of machines

Cluster

Each machine in the cluster is a node

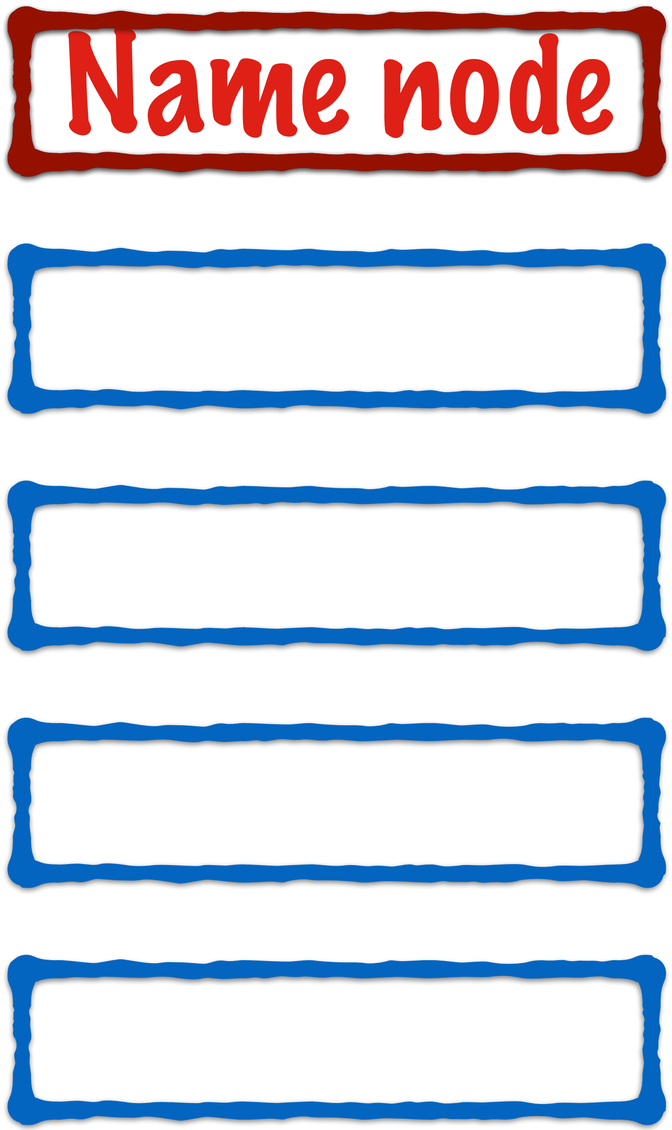
HDFS



One of the nodes acts as the **master node**

This node **manages** the overall file system

HDFS

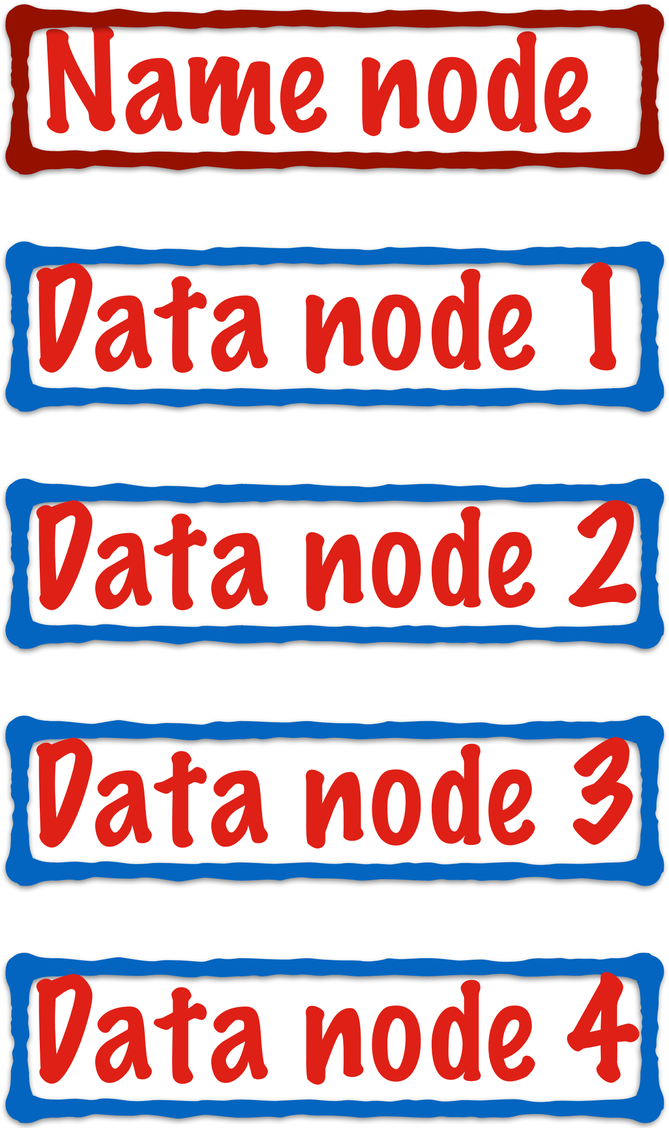


The name node stores

1. The directory structure

2. Metadata for all the files

HDFS



Other nodes are called data nodes

The data is physically stored on these nodes

HDFS

Here is a large text file

next up previous contents index
Next: Dynamic indexing Up: Index construction Previous: Single-pass in-memory indexing Contents Index

Distributed indexing

Collections are often so large that we cannot perform index construction efficiently on a single machine. This is particularly true of the World Wide Web for which we need large computer clusters [*]to construct any reasonably sized web index. Web search engines, therefore, use distributed indexing algorithms for index construction. The result of the construction process is a distributed index that is partitioned across several machines - either according to term or according to document. In this section, we describe distributed indexing for a
term-partitioned index . Most large search engines prefer a document-partitioned index (which can be easily generated from a term-partitioned index). We discuss this topic further in Section 20.3 (page [*]).

The distributed index construction method we describe in this section is an application of MapReduce , a general architecture for distributed computing. MapReduce is designed for large computer clusters. The point of a cluster is to solve large computing problems on cheap commodity machines or nodes that are built from standard parts (processor, memory, disk) as opposed to on a supercomputer with specialized hardware. Although hundreds or thousands of machines are available in such clusters, individual machines can fail at any time. One requirement for robust distributed indexing is, therefore, that we divide the work up into chunks that we can easily assign and - in case of failure - reassign. A master node directs the process of assigning and reassigning tasks to individual worker nodes.

The map and reduce phases of MapReduce split up the computing job into chunks that standard machines can process in a short time. The various steps of MapReduce are shown in Figure 4.5 and an example on a collection consisting of two documents is shown in Figure 4.6 . First, the input data, in our case a collection of web pages, are split into n splits where the size of the split is chosen to ensure that the work can be distributed evenly (chunks should not be too large) and efficiently (the total number of chunks we need to manage should not be too large); 16 or 64 MB are good sizes in distributed indexing. Splits are not preassigned to machines, but are instead assigned by the master node on an ongoing basis: As a machine finishes processing one split, it is assigned the next one. If a machine dies or becomes a laggard due to hardware problems, the split it is working on is simply reassigned to another machine.

Figure 4.5: An example of distributed indexing with MapReduce. Adapted from Dean and Ghemawat (2004).
\includegraphics[width=11.5cm]{art/mapreduce2.eps}
In general, MapReduce breaks a large computing problem into smaller parts by recasting it in terms of manipulation of key-value pairs . For indexing, a key-value pair has the form (termID,docID). In distributed indexing, the mapping from terms to termIDs is also distributed and therefore more complex than in single-machine indexing. A simple solution is to maintain a (perhaps precomputed) mapping for frequent terms that is copied to all nodes and to use terms directly (instead of termIDs) for infrequent terms. We do not address this problem here and assume that all nodes share a consistent term $\rightarrow$ termID mapping.

The map phase of MapReduce consists of mapping splits of the input data to key-value pairs. This is the same parsing task we also encountered in BSBI and SPIMI, and we therefore call the machines that execute the map phase parsers . Each parser writes its output to local intermediate files, the segment files (shown as \fbox{a-f\medstrut} \fbox{g-p\medstrut} \fbox{q-z\medstrut} in Figure 4.5).

For the reduce phase , we want all values for a given key to be stored close together, so that they can be read and processed quickly. This is achieved by partitioning the keys into j term partitions and having the parsers write key-value pairs for each term partition into a separate segment file. In Figure 4.5 , the term partitions are according to first letter: a-f, g-p, q-z, and $j=3$. (We chose these key ranges for ease of exposition. In general, key ranges need not correspond to contiguous terms or termIDs.) The term partitions are defined by the person who operates the indexing system (Exercise 4.6). The parsers then write corresponding segment files, one for each term partition. Each term partition thus corresponds to r segments files, where r is the number of parsers. For instance, Figure 4.5 shows three a-f segment files of the a-f partition, corresponding to the three parsers shown in the figure.

Collecting all values (here: docIDs) for a given key (here: termID) into one list is the task of the inverters in the reduce phase. The

Let's see how this file is stored in HDFS

HDFS

Block 1

Block 2

Block 3

Block 4

Block 5

Block 6

Block 7

Block 8

First the file is broken into blocks of size 128 MB

HDFS

Block 1

Block 2

Block 3

Block 4

Block 5

Block 6

Block 7

Block 8

First the file is broken into

blocks of size 128 MB

This size is chosen to minimize the time to seek to the block on the disk

HDFS

Block 1

Block 2

Block 3

Block 4

Block 5

Block 6

Block 7

Block 8

These blocks are then stored across the data nodes

HDFS

Data node 1

Block 1 | Block 2

Data node 3

Block 5 | Block 6

Data node 2

Block 3 | Block 4

Data node 4

Block 7 | Block 8

Name node

The name node stores metadata

HDFS

Block locations for each file are stored in the name node

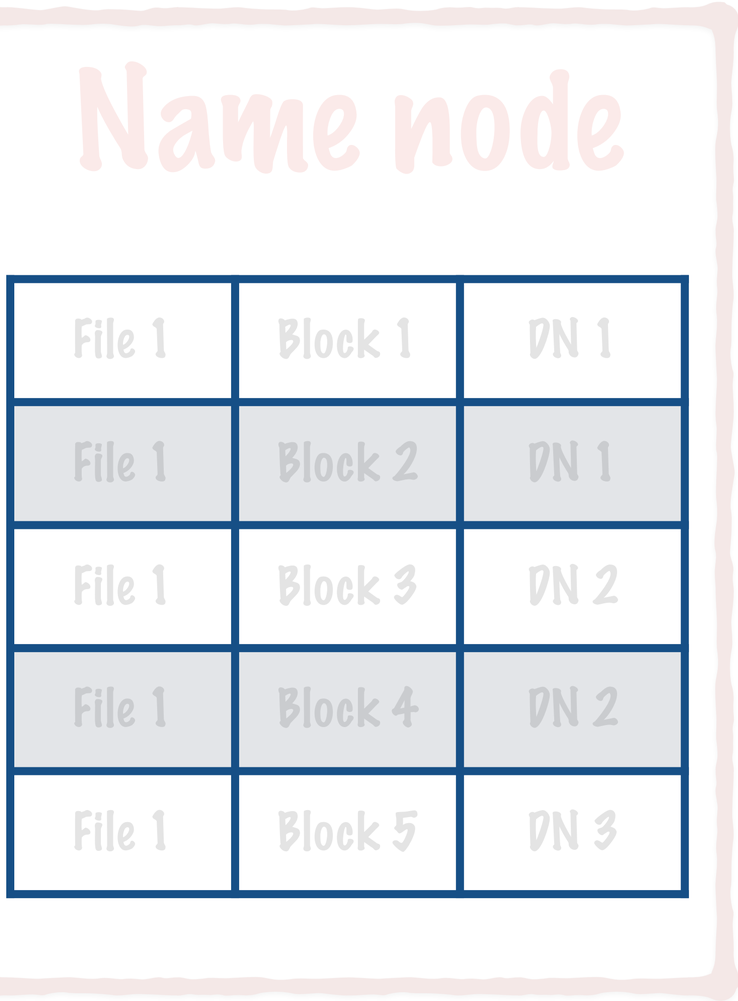
Name node

File 1	Block 1	DN 1
File 1	Block 2	DN 1
File 1	Block 3	DN 2
File 1	Block 4	DN 2
File 1	Block 5	DN 3

HDFS

A file is read using

1. The **metadata** in name node
2. The **blocks** in the data nodes



HDFS

Data node 1

Block 1 | Block 2

Data node 3

Block 5 | Block 6

Data node 2

Data node 4

Name node

File 1	Block 1	DN 1
File 1	Block 2	DN 1
File 1	Block 3	DN 2
File 1	Block 4	DN 2
File 1	Block 5	DN 3

What if one of the blocks gets corrupted?

HDFS

Data node 1

Block 1 | Block 2

Data node 3

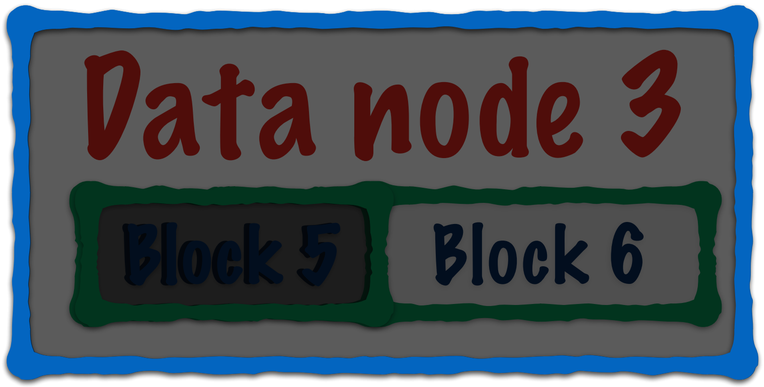
Block 5 | Block 6

Name node

File 1	Block 1	DN 1
File 1	Block 2	DN 1
File 1	Block 3	DN 2
File 1	Block 4	DN 2
File 1	Block 5	DN 3

Or one of the data nodes crashes?

HDFS



This is one of the key challenges in distributed storage

Name node

File 1	Block 1	DN 1
File 1	Block 2	DN 1
File 1	Block 3	DN 2
File 1	Block 4	DN 2
File 1	Block 5	DN 3

HDFS

Data node 1

Block 1

Data node 3

Block 6

Data node 2

Block 3

Block 4

Data node 4

Block 7

Block 8

You can define a replication factor in HDFS

Name node

File 1	Block 1	DN 1
File 1	Block 2	DN 1
File 1	Block 3	DN 2
File 1	Block 4	DN 2
File 1	Block 5	DN 3

HDFS

Data node 1

Block 1 | Block 2

Data node 2

Block 3 | Block 4

Block 1 | Block 2

Data node 3

Block 5 | Block 6

Name node

Each block is replicated, and the replicas are stored in different data nodes

HDFS

Data node 1

Block 1 | Block 2

Data node 3

Block 5 | Block 6

Name node

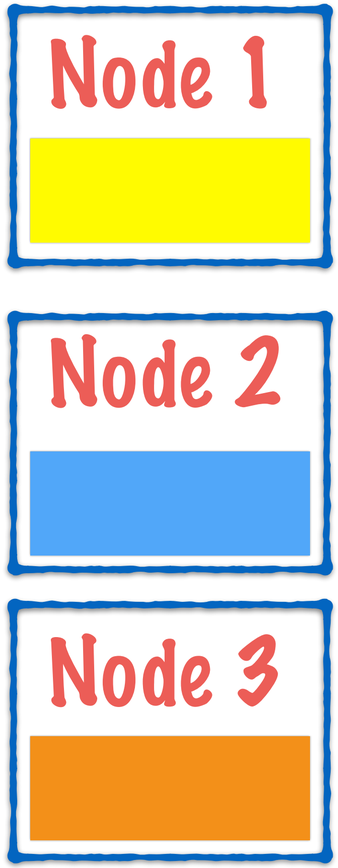
File 1	Block 1	Master	DN 1
File 1	Block 1	Replica	DN 2
..	..	..	..
..	..	..	..
..	..	..	..

The replica locations are also stored in the name node

RDDs partitions

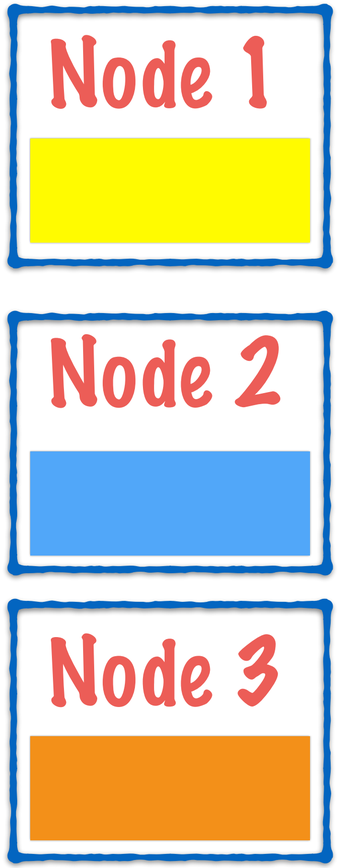
If Spark is reading data from HDFS
the data is already partitioned!

The file blocks in HDFS
are the partitions



RDDs partitions

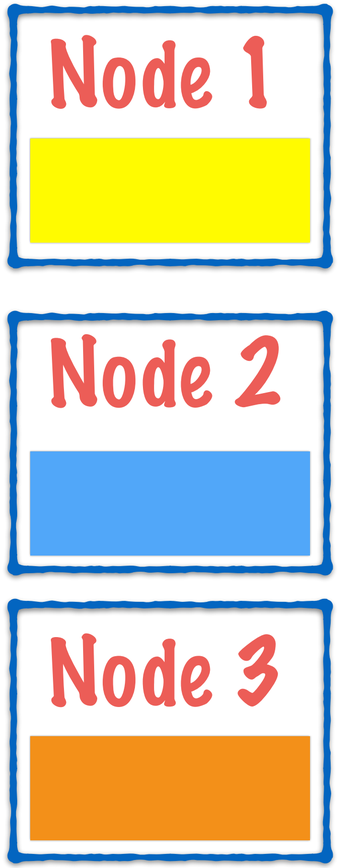
Each block is loaded into memory in the same node where it is stored



RDDs partitions

If the same node doesn't have enough free memory, then another node is chosen

The choice is made such that network transfer time for the data is minimized



Resilient Distributed Datasets

partitions

read-only

lineage

RDDs read-only

RDDs are immutable

RDDs read-only

Once an RDD is created, you can only do 2 things

Read data from it

Action

Transform it to another RDD

Transformation

Resilient Distributed Datasets

partitions

read-only

lineage

Recap

RDDs lineage

All operations on RDDs are either Transformations or Actions

Recap

RDDs lineage

When created, an RDD just holds metadata

1. A transformation
2. It's parent RDD

Recap

RDDs lineage

AirlineFiltered RDD

filter

airlines

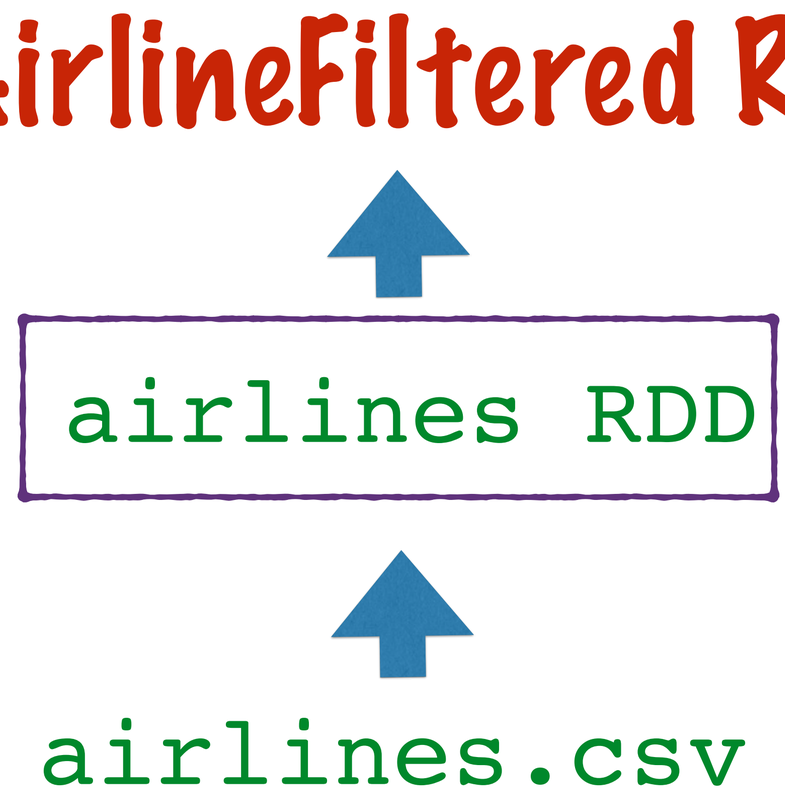
This way, every RDD knows where it came from

ie. RDDs know their lineage

Recap

RDDs lineage

This lineage can be traced back to the source (in this case, the original file that held the data)



RDDs lineage

lineage

is a really powerful concept!

lineage is the thing that

1. makes RDDs Resilient

2. makes lazy evaluation possible

Resilient

One of the key challenges
in distributed computing

How do you recover data if
one of the nodes crashes?

Resilient

In Hadoop for instance, this is managed by

A. Replication in HDFS

B. Writing data to disk
after every operation

Resilient

So, how do RDDs:

being in-memory

provide fault tolerance?

lineage

Resilient

lineage

Since RDDs know their lineage,
they can always be reconstructed
from the source

RDDs lineage

lineage is the thing that

1. makes RDDs Resilient

2. makes Lazy evaluation possible

Lazy evaluation

The other advantage of lineage is that it allows RDDs to be lazily evaluated

Lazy evaluation

When created, RDDs are not materialized

ie. they only consist of metadata

Lazy evaluation

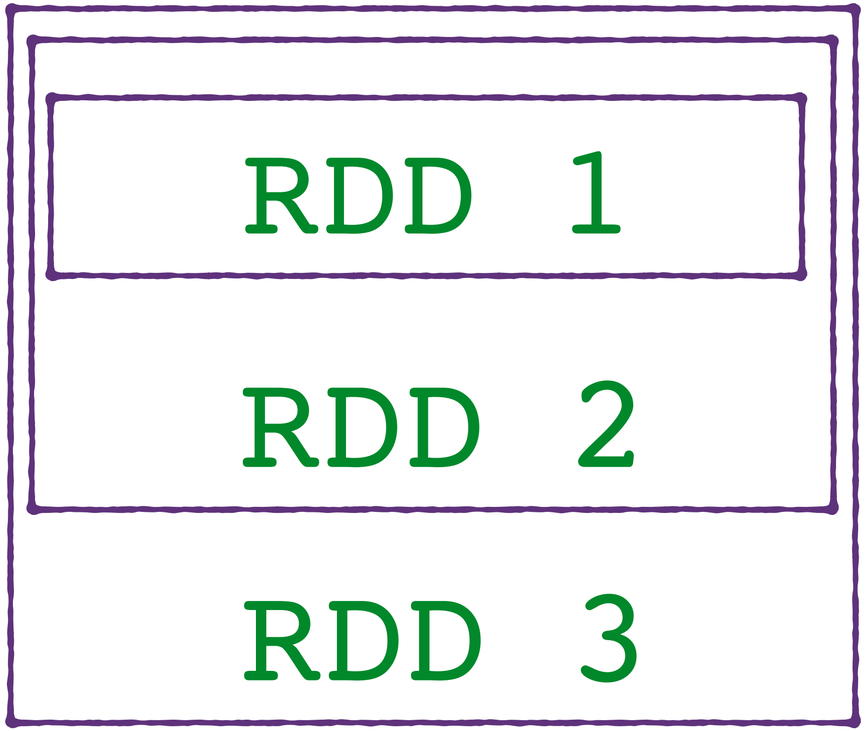
RDDs are only materialized when user requests a result using Actions

Lazy evaluation

Say you read data from a file

Filter the header row

Split the row by commas

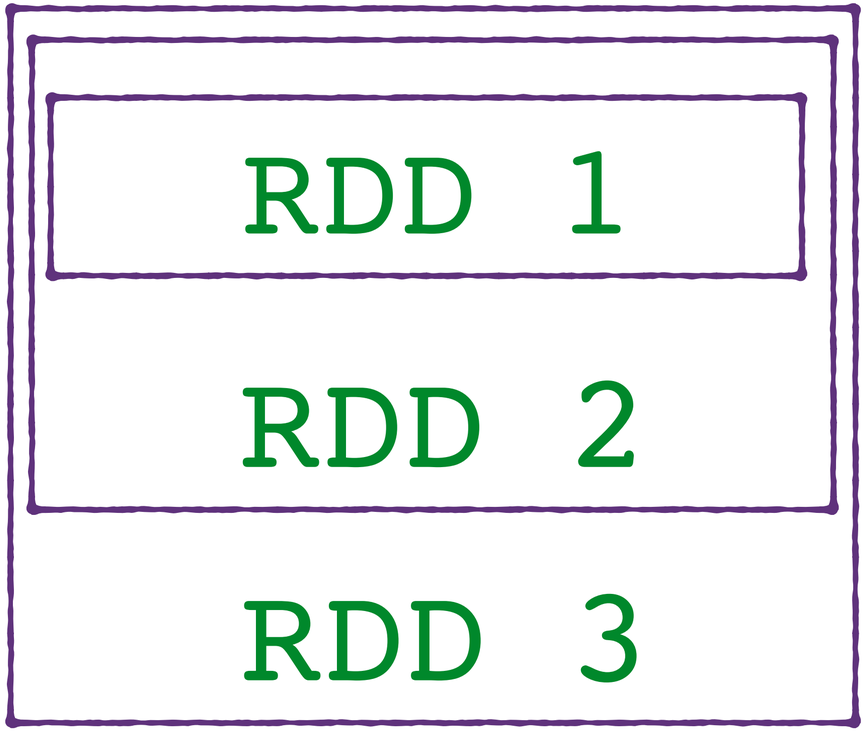


3 RDDs are created, none are materialized

Lazy evaluation

3 RDDs are created,
none are materialized

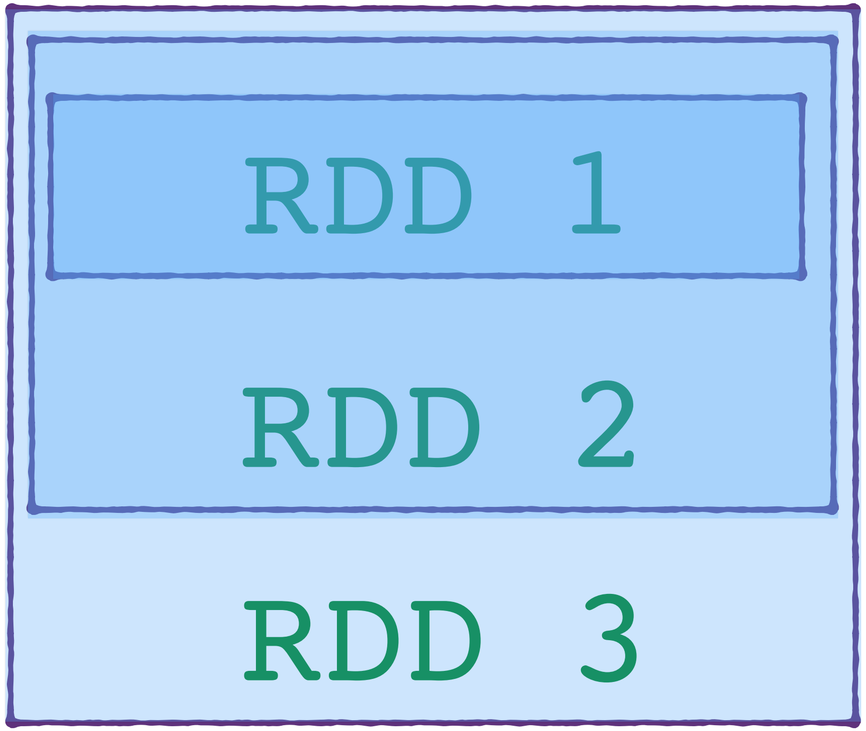
User asks for first
10 rows from RDD 3



Lazy evaluation

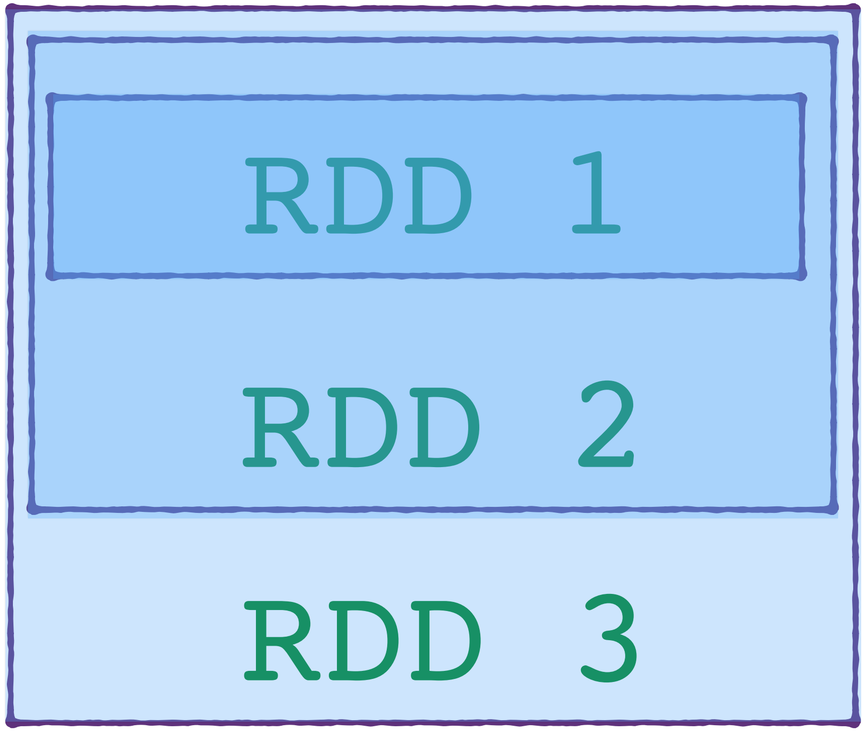
User asks for first 10 rows from RDD 3

Spark will now materialize all it's parent RDDs



Lazy evaluation

Lazy evaluation makes Spark very efficient



RDDs lineage

lineage is the thing that

1. makes RDDs Resilient

2. makes Lazy evaluation possible

Resilient Distributed Datasets

partitions

read-only

lineage

Things you can do with RDDs

Things you can do with RDDs

Create an RDD

Transformations

Actions

Persistence

Things you can do with RDDs

Create an RDD

Transformations

Actions

Persistence

Create an RDD

You can create an RDD in 2 ways

From a data source (usually on disk)

```
airlines=sc.textFile(airlinesPath)
```

From another RDD (using Transformations)

```
airlines.filter(lambda x:'Description' not in x)
```

Create an RDD

You can create an RDD in 2 ways

From a data source (usually on disk)

```
airlines=sc.textFile(airlinesPath)
```

The data source could also be a collection in your program

```
myList = [1,2,5,10]
myListRDD=sc.parallelize(myList)
```

Things you can do with RDDs

Create an RDD

Transformations

Actions

Persistence

Things you can do with RDDs

Create an RDD

Transformations

Actions

Persistence

Transformations

Transformations convert one RDD to another

```
airlines.filter(lambda x:'Description' not in x)
```

The most common transformations on an RDD are

filter

map

flatMap

Transformations

filter filters records from the RDD based on a condition

```
airlines.filter(lambda x:'Description' not in x)
```

map applies a function on each record

```
airlines.map(len)
```

flatMap takes each record and converts it to multiple records

Ex: from an RDD of lines of text to an RDD of words

Transformations

filter Takes a boolean function

map Any function that returns a single object

flatMap Any function that returns an iterator

Each of these take a **function** and **apply it on each** record in the RDD

Transformations

filter

map

flatMap

In addition to these, we have transformations that act on 2 RDDS

union

intersection

subtract

cartesian

Transformations

union union of 2 RDDs

intersection intersection of 2 RDDs

subtract Remove the contents of 1 RDD from the other

cartesian cartesian product : Creates an RDD of tuples

These act just like set operations

Transformations

Transformations are
lazily evaluated

ie. they are only executed
when an Action is performed

Things you can do with RDDs

Create an RDD

Transformations

Actions

Persistence

Things you can do with RDDs

Create an RDD

Transformations

Actions

Persistence

Actions compute a result from the RDD and return it to the user

```
# View a few lines
airlines.take(10)
```

```
[u'Code,Description',
 u'"19031","Mackey International Inc.: MAC"',
 u'"19032","Munz Northern Airlines Inc.: XY"',
 u'"19033","Cochise Airlines Inc.: COC"',
 u'"19034","Golden Gate Airlines Inc.: GSA"',
 u'"19035","Aeromech Inc.: RZZ"',
 u'"19036","Golden West Airlines Co.: GLW"',
 u'"19037","Puerto Rico Intl Airlines: PRN"',
 u'"19038","Air America Inc.: STZ"',
 u'"19039","Swift Aire Lines Inc.: SWT"']
```

Actions

collect

first

take

reduce

aggregate

count

countByValue

are some commonly used actions

Actions

first

take

collect

count

countByValue

reduce

aggregate

Collect will return all the elements of the RDD

Actions

first

take

count

countByValue

reduce

aggregate

collect

We usually don't use as it may overwhelm the system..

.....unless we know that the entire dataset is small enough fit onto 1 machine

collect

countByValue

reduce

aggregate

Actions

first The first row

take A specified number of rows

count The total number of rows

These are useful to get a quick sense of the data

collect

first

take

count

reduce

aggregate

Actions

countByValue

This will count the number of times a value appears in the RDD

collect

first

take

count

reduce

aggregate

Actions

countByValue

It can be used to summarize a dataset/ build a histogram

collect

first

take

count

countByValue

aggregate

Actions

reduce

This is one of the most common actions

collect

first

take

count

countByValue

aggregate

Actions

reduce

It will combine all the values in the RDD in a specified manner

collect

first

take

count

countByValue

aggregate

Actions

reduce

We can use this for instance to sum up all the values in an RDD of numbers

collect

first

take

count

countByValue

aggregate

Actions

reduce

Reduce takes a function
that acts on 2 elements
.. and returns another
element of the same type

collect

first

take

count

countByValue

reduce

Actions

aggregate

Similar to reduce

reduce will only take functions that return the same type as the RDD elements

Ex. Return a number using an RDD of numbers

collect

first

take

count

countByValue

reduce

Actions

aggregate

aggregate allows you to return elements of a different type

Ex. take an RDD of numbers and return a tuple

collect

first

take

count

countByValue

reduce

Actions

aggregate

We'll see some examples of reduce and aggregate later

Things you can do with RDDs

Create an RDD

Transformations

Actions

Persistence

Things you can do with RDDs

Create an RDD

Transformations

Actions

Persistence

Persistence

Let's say there is an RDD
that you need to use a lot

It will be **materialized**
from scratch every time
you perform an action on it

Persistence

Instead, you could persist the RDD in memory

This will force the RDD to be materialized

Things you can do with RDDs

Create an RDD

Transformations

Actions

Persistence

Let's see all of this in action