THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

ST446 – Distributed Computing for Big Data

Seminar 3

Dr. Marcos Ennes Barreto

Assistant Professorial Lecturer

4 February 2021

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Dr. Marcos Ennes Barreto

Web: https://www.lse.ac.uk/Statistics/People/Dr-Marcos-Barreto

Contact: m.e.barreto@lse.ac.uk

Room: COL 5.04

Office hours (book through LSE Student Hub):

✓ Tuesday – 09:00 – 10:00 / 14:00 pm – 15:00 pm

slack

ST446 Q&A

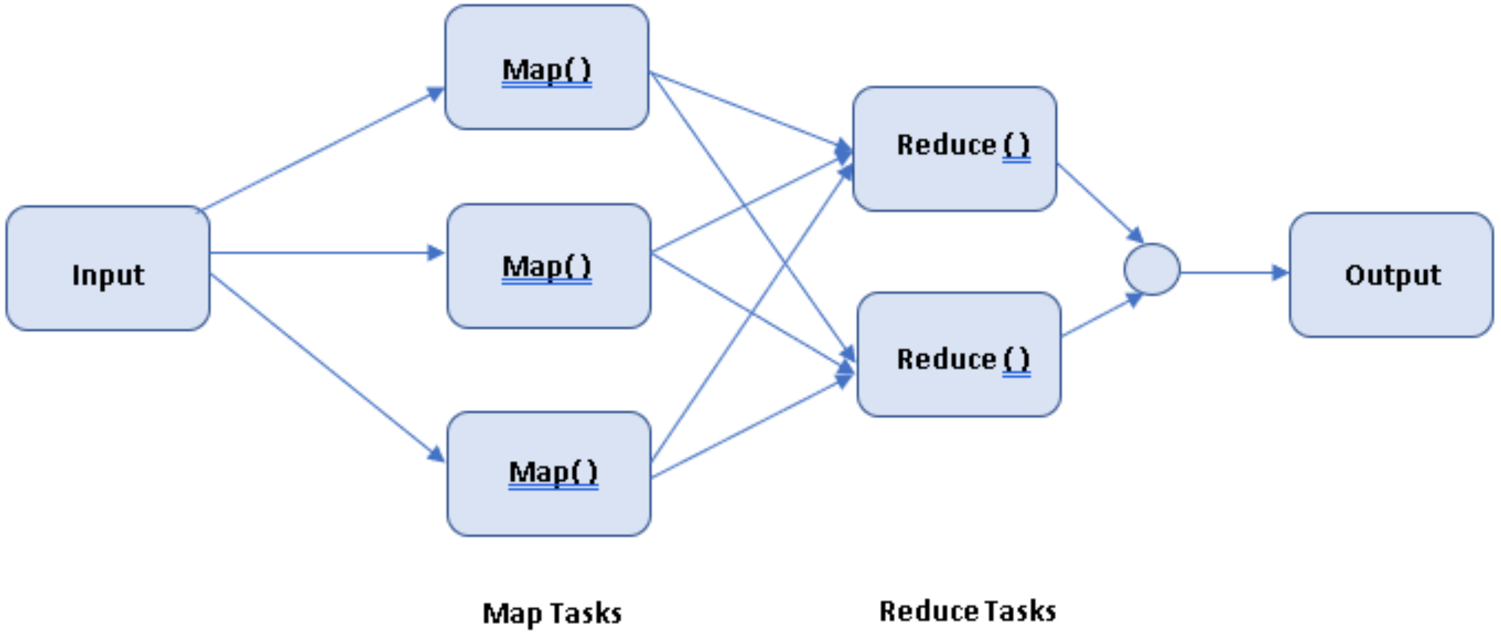
st446-lt2021-qa.slack.com

4 February 2021

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

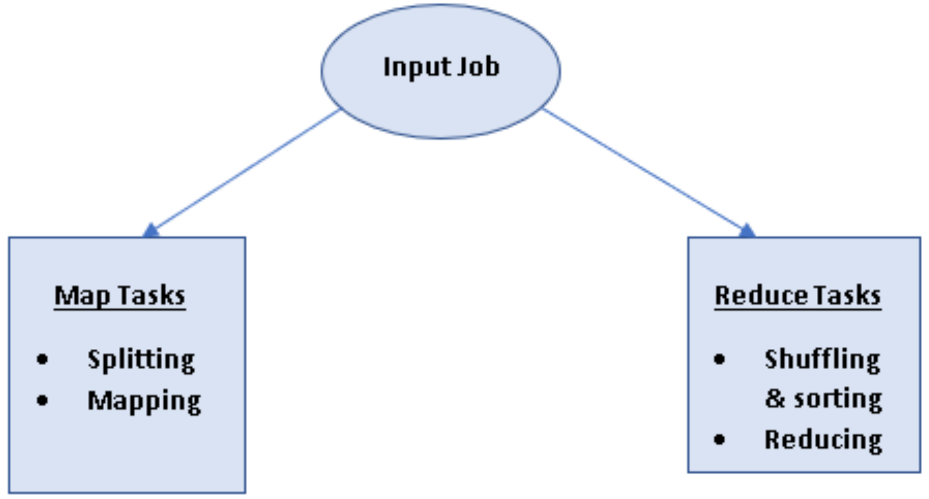
MapReduce

A **programming framework** that allows us to perform **distributed** and **parallel** processing on large data sets in a distributed environment.



https://www.datasciencecentral.com/profiles/blogs/hadoop-for-beginners-part-2

How MapReduce works?



Map Tasks

- **Splitting:** Input data is divided into fixed-size chunks called input splits.
- **Mapping:** In this phase each input split is passed to a mapping function which divides the split into *List (Key, Value)*.

Reduce Tasks

- **Shuffling and Sorting:** Reduce tasks are the combination of shuffle/sort and reduce. This phase consumes output of the Mapping phase. Its main task is to club together the relevant record in sorting manner from the output of mapping phase. The output is in the form of *Key, List (Value)*.

- **Reducing:** In this phase, output from shuffling and sorting are aggregated and returns single *(Key, Value)* output value. This final output value is then written in the output file in HDFS.

https://www.datasciencecentral.com/profiles/blogs/hadoop-for-beginners-part-2

Working example

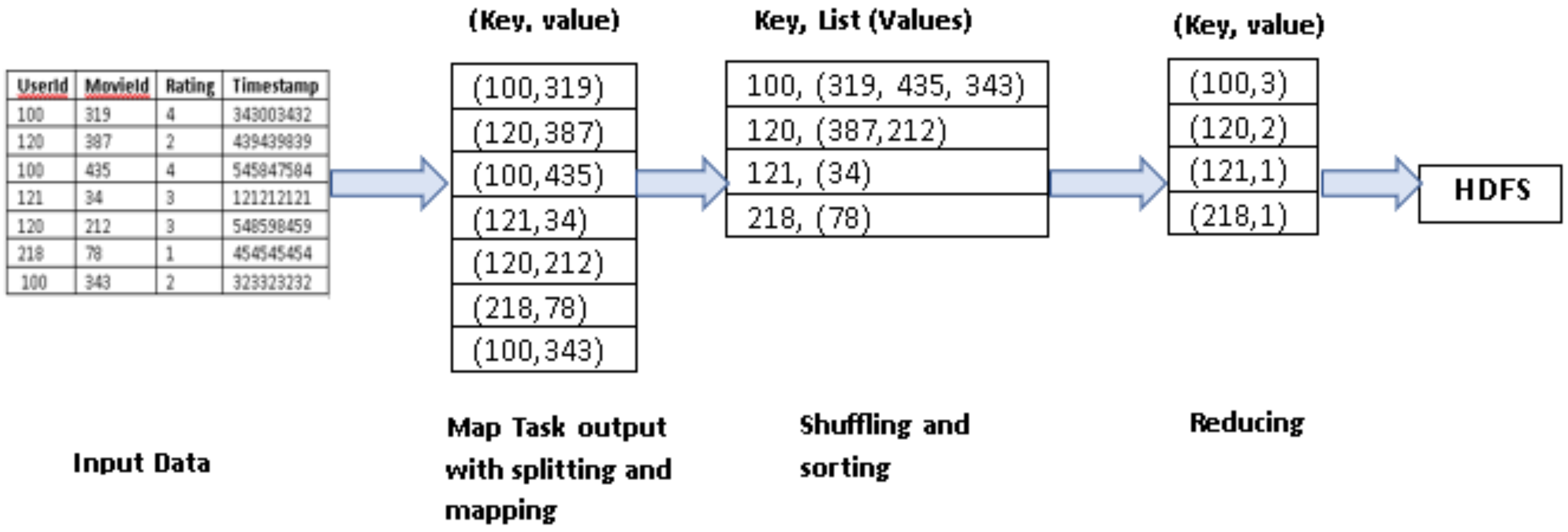
- **Task** – How many movies did each user rate in the Movie data set?
- **Sample Dataset (Input File)**-

UserId	**MovieId**	**Rating**	**Timestamp**
100	319	4	343003432
120	387	2	439439839
100	435	4	545847584
121	34	3	121212121
120	212	3	548598459
218	78	1	454545454
100	343	2	323323232

https://www.datasciencecentral.com/profiles/blogs/hadoop-for-beginners-part-2

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Working example

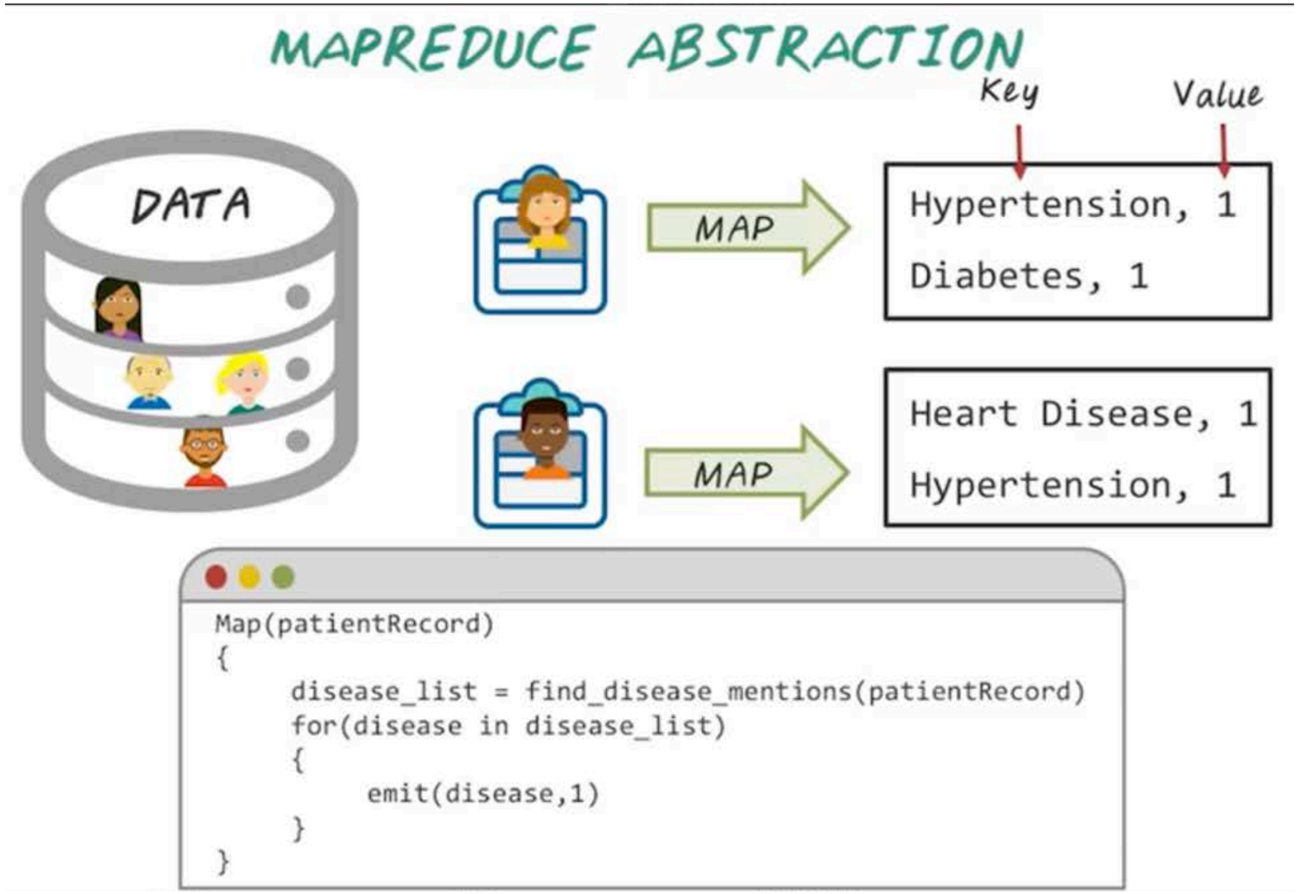


UserId	MovieId	Rating	Timestamp
100	319	4	343003432
120	387	2	439439839
100	435	4	545847584
121	34	3	121212121
120	212	3	548598459
218	78	1	454545454
100	343	2	323323232

https://www.datasciencecentral.com/profiles/blogs/hadoop-for-beginners-part-2

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

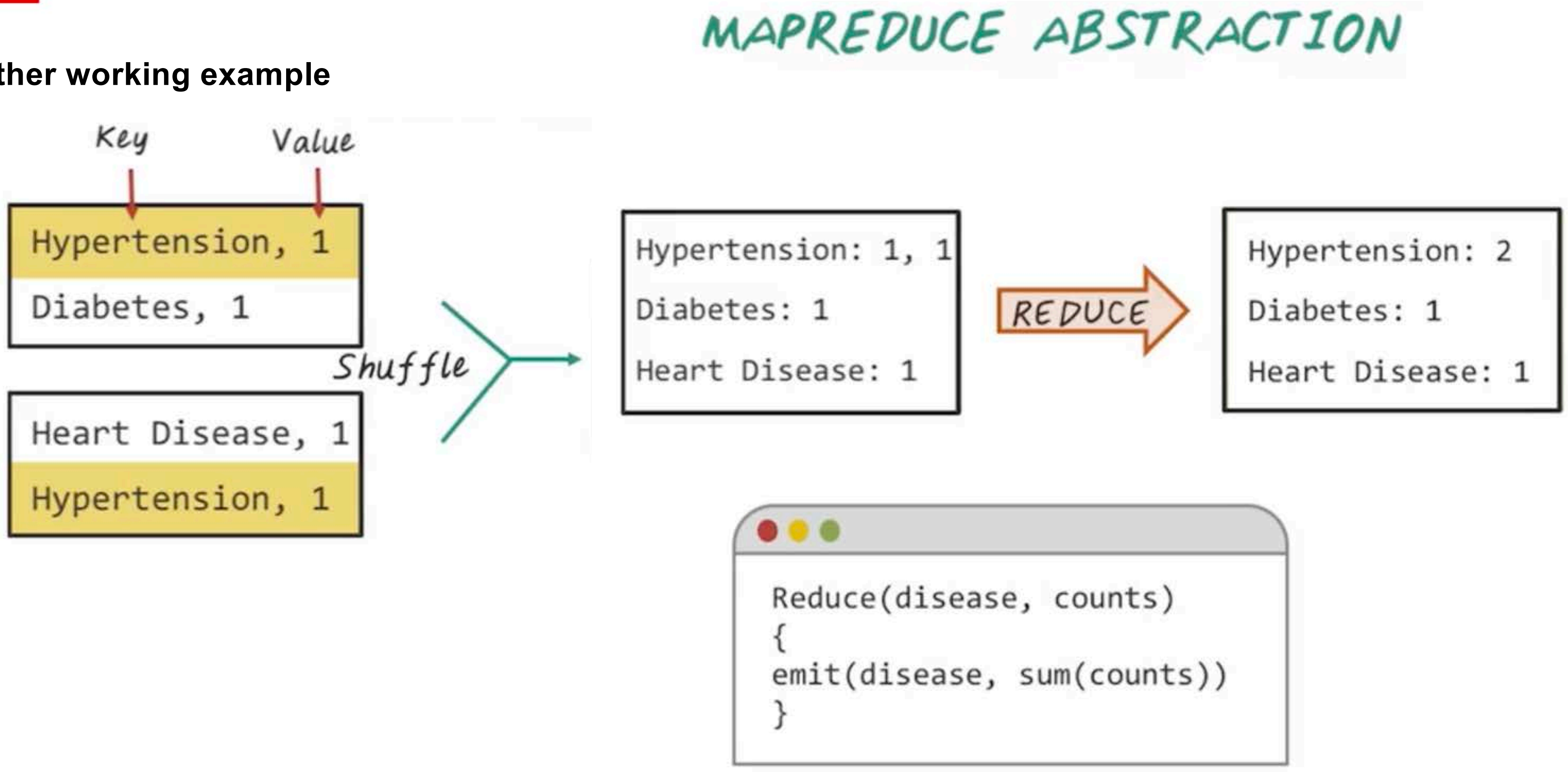
Another working example



```
Map(patientRecord)
{
      disease_list = find_disease_mentions(patientRecord)
      for(disease in disease_list)
      {
            emit(disease,1)
      }
}
```

https://www.udacity.com/course/big-data-analytics-in-healthcare--ud758

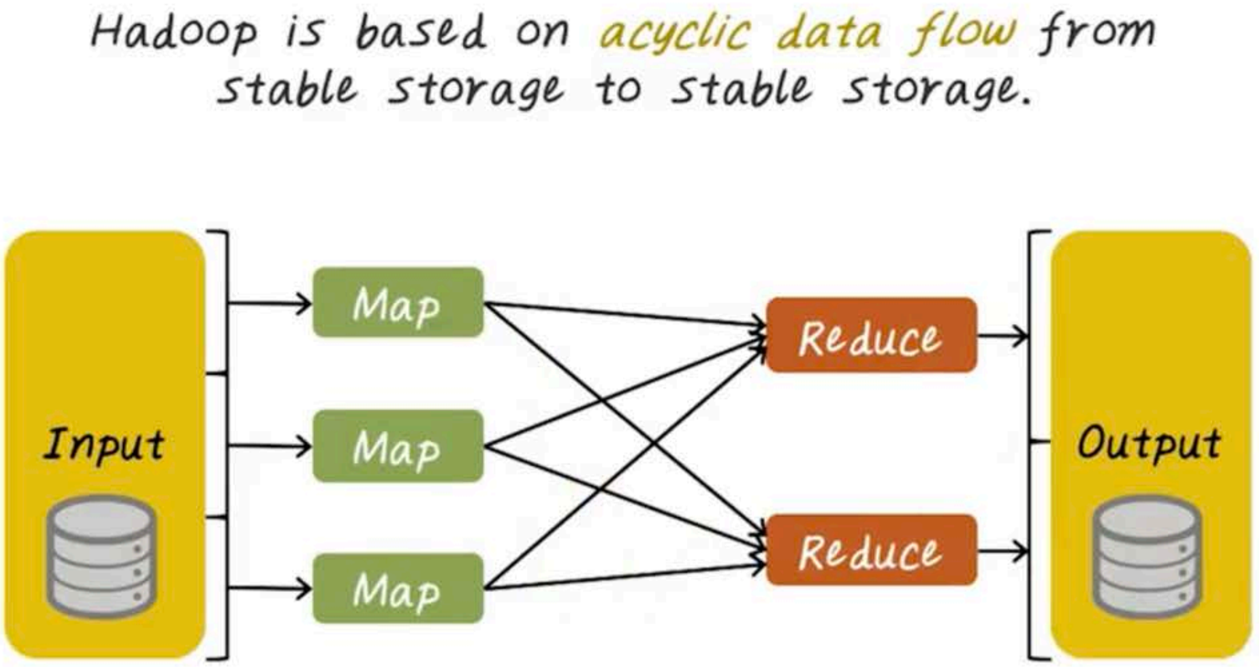
Another working example

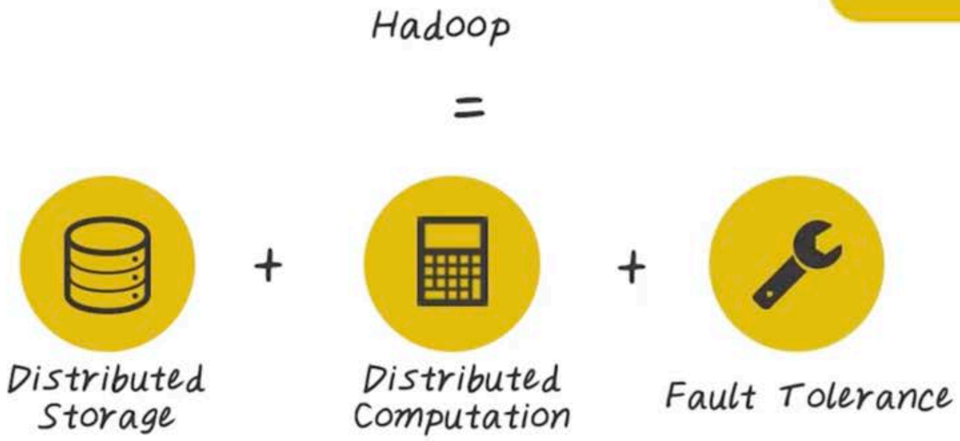


```
Reduce(disease, counts)
{
emit(disease, sum(counts))
}
```

https://www.udacity.com/course/big-data-analytics-in-healthcare--ud758

When to use Hadoop?

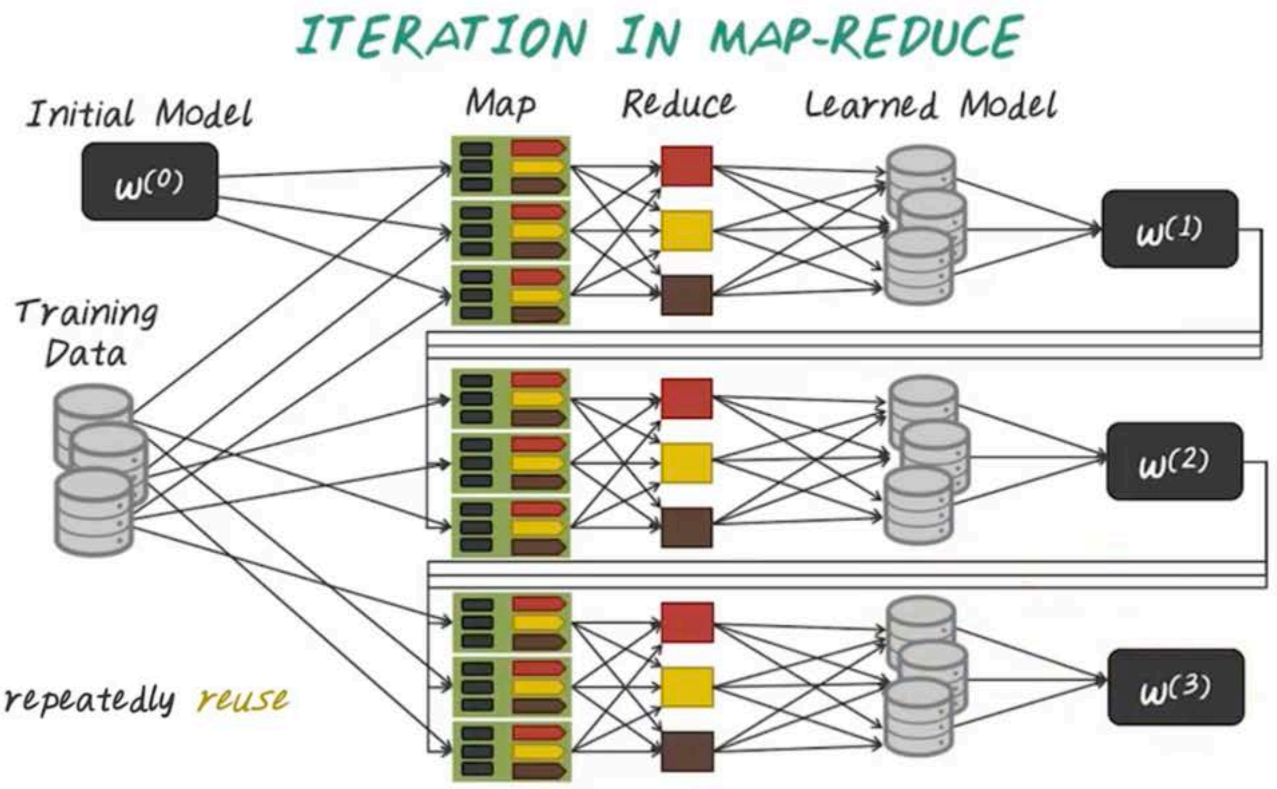

Hadoop is based on acyclic data flow from stable storage to stable storage.
Input
Map
Map
Map
Reduce
Reduce
Output



https://www.udacity.com/course/big-data-analytics-in-healthcare--ud758

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

When not to use Hadoop?



Hadoop is inefficient for applications that repeatedly reuse a working set of data:

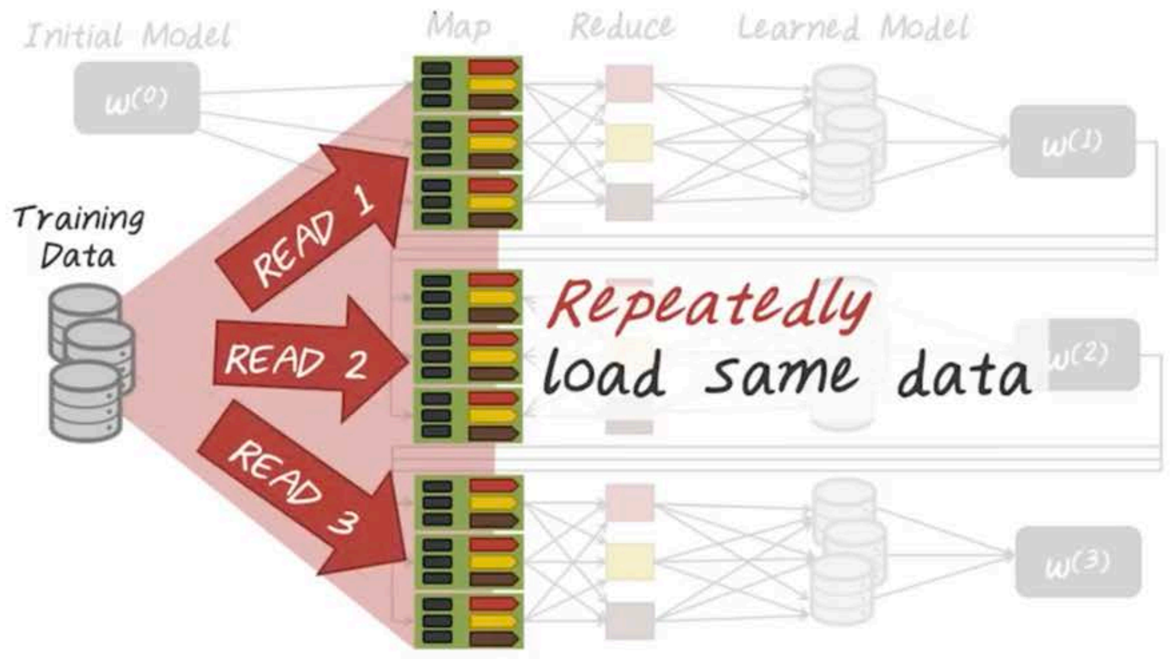
Iterative Algorithms

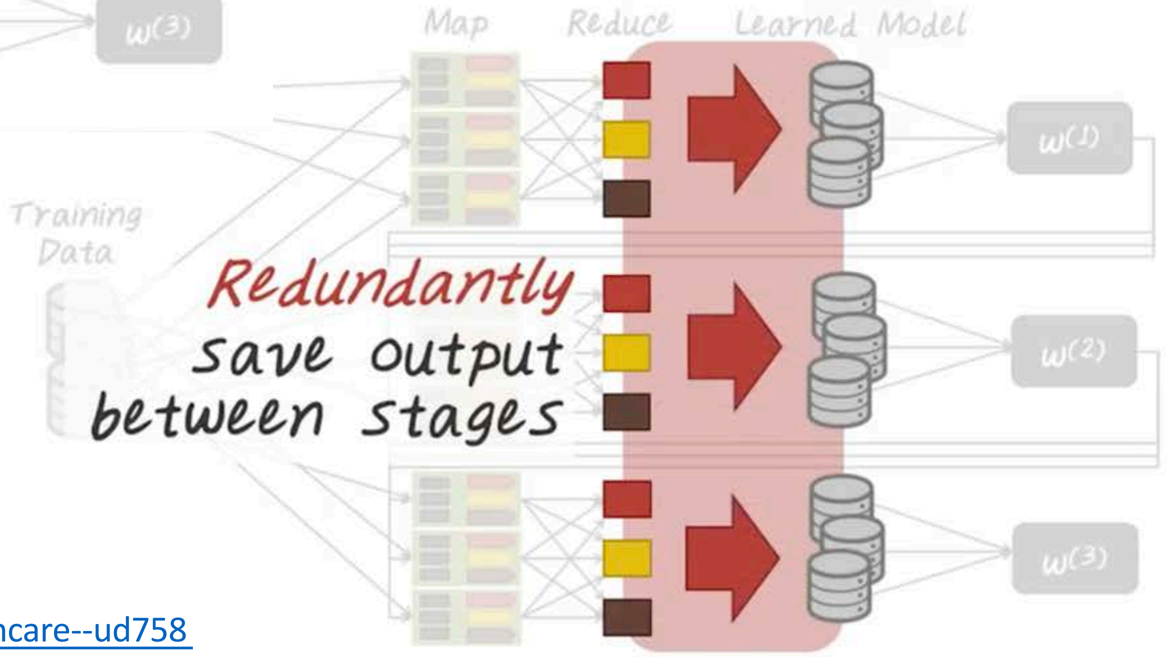
- Machine learning
- Graph analysis

Interactive Data Mining Tools

- R
- Python

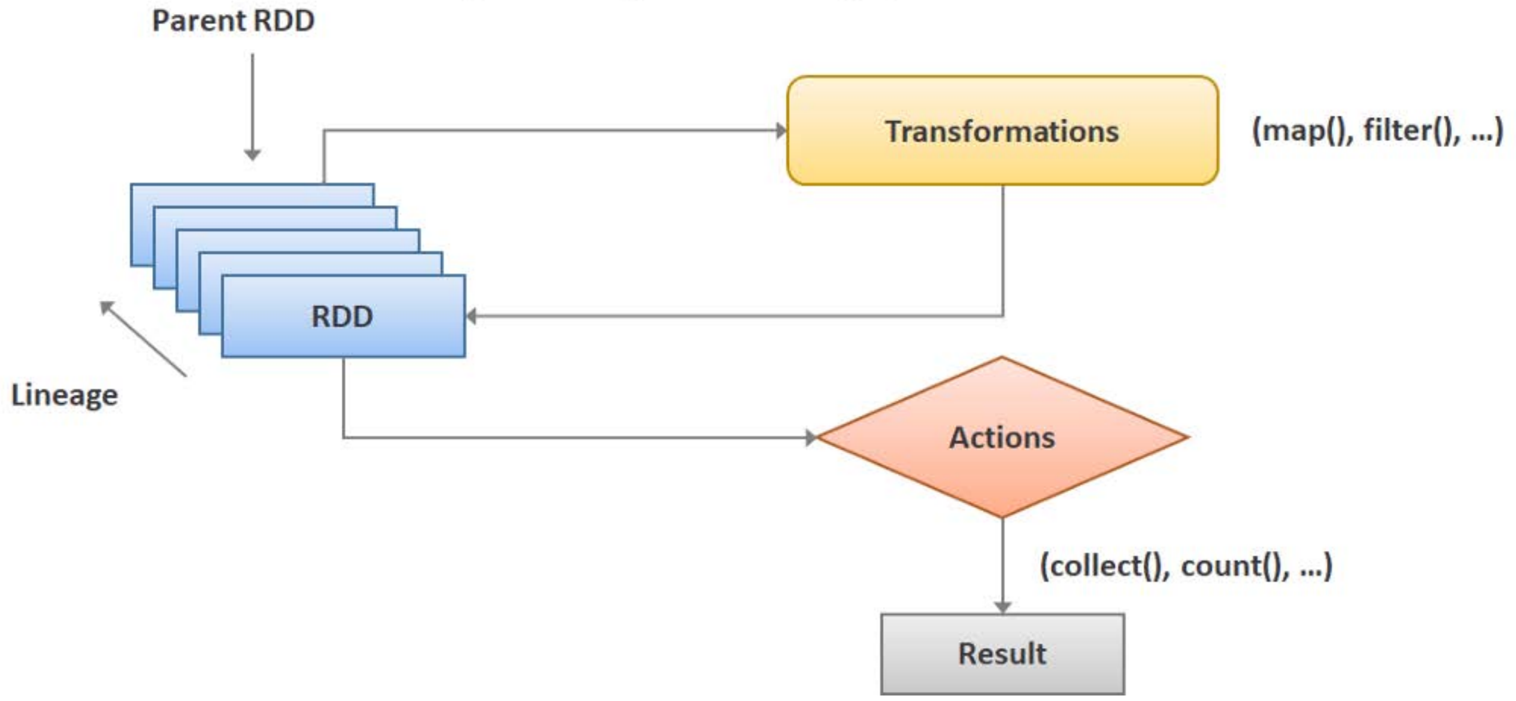
https://www.udacity.com/course/big-data-analytics-in-healthcare--ud758



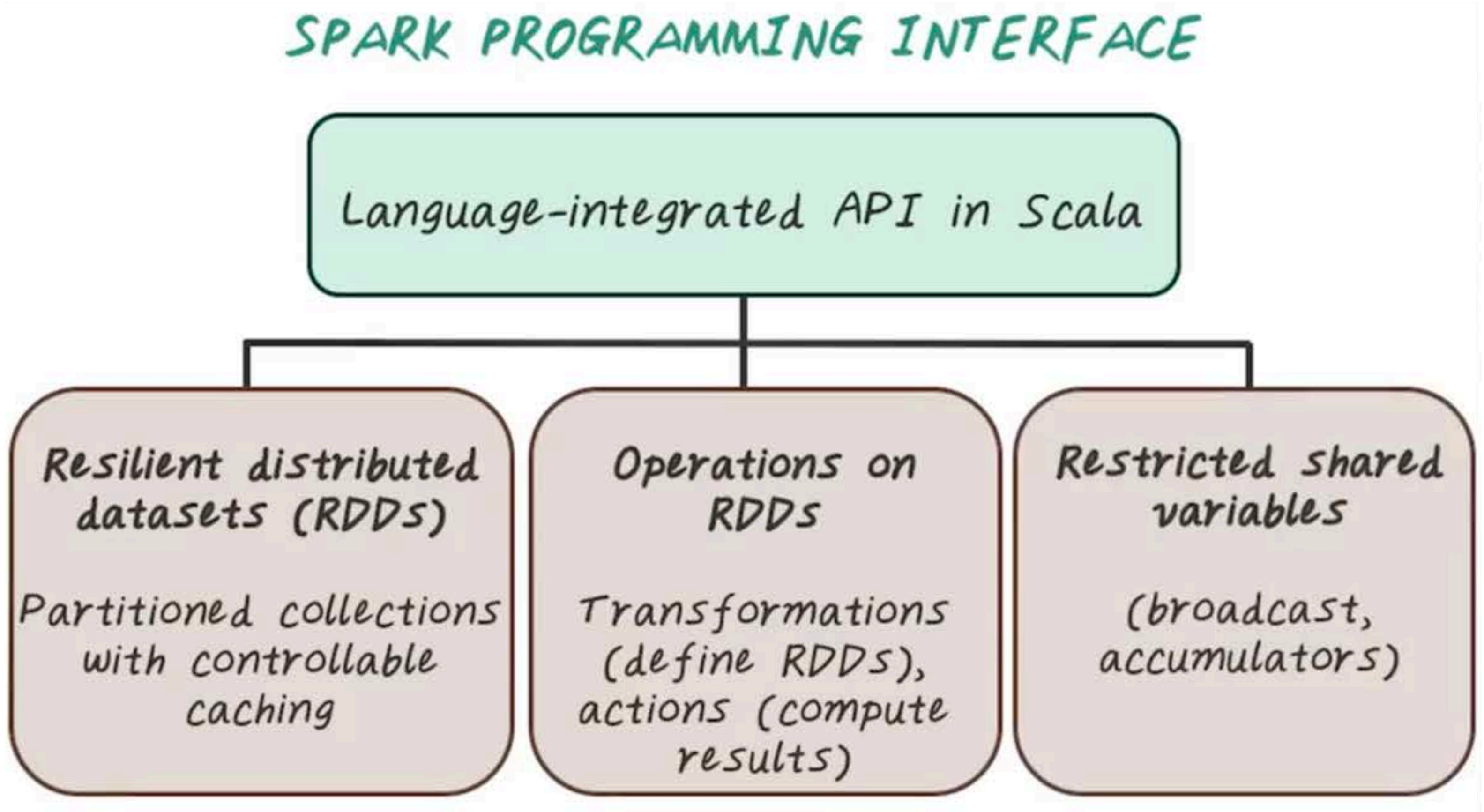


https://www.udacity.com/course/big-data-analytics-in-healthcare--ud758

Spark RDD (Unstructured) Operations



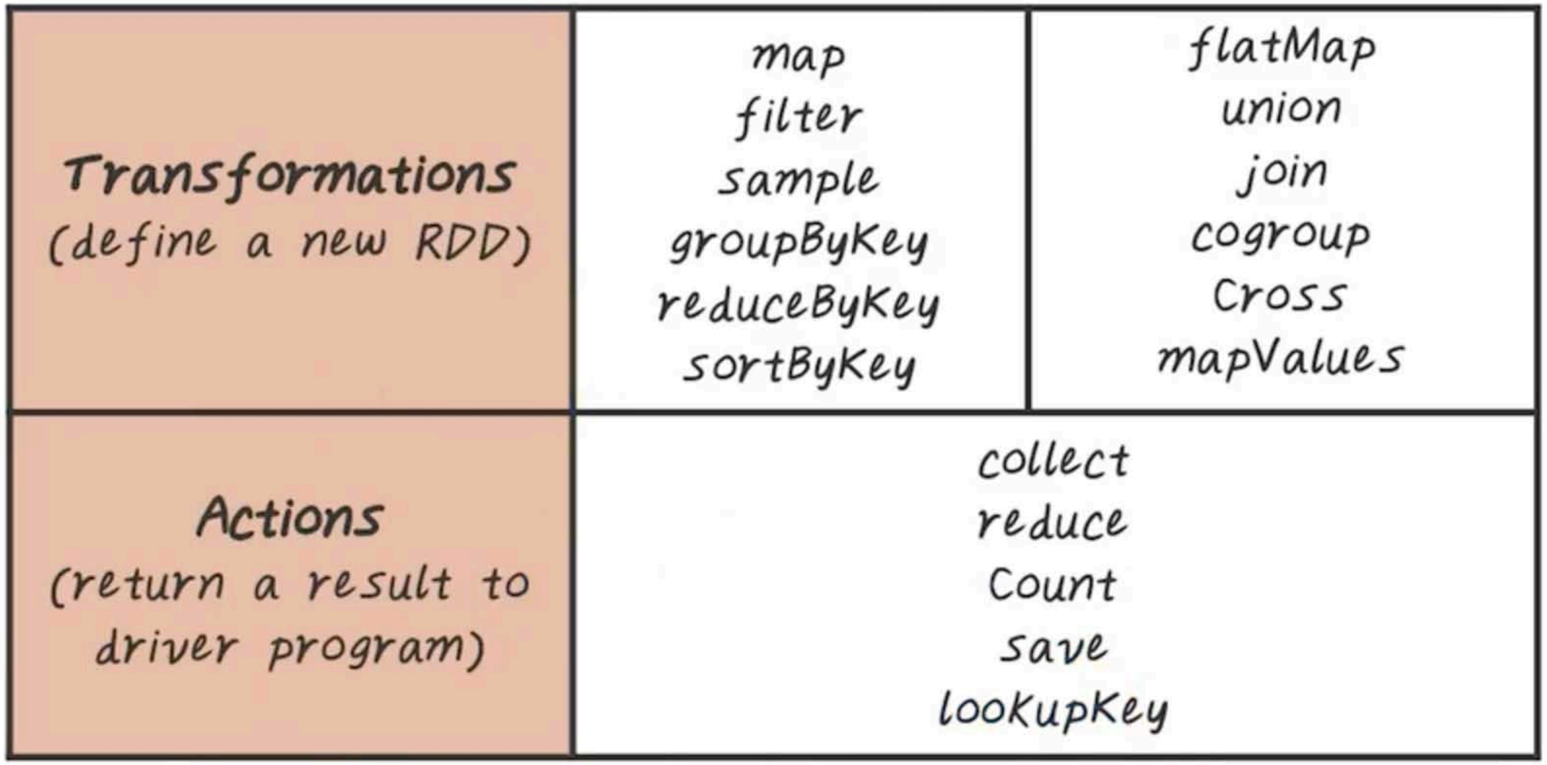
https://medium.com/analytics-vidhya/spark-rdd-low-level-api-basics-using-pyspark-a9a322b58f6


SPARK PROGRAMMING INTERFACE
Language-integrated API in Scala
Resilient distributed datasets (RDDs)
Partitioned collections with controllable caching
Operations on RDDs
Transformations (define RDDs), actions (compute results)
Restricted shared variables
(broadcast, accumulators)

https://www.udacity.com/course/big-data-analytics-in-healthcare--ud758

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

SPARK OPERATIONS



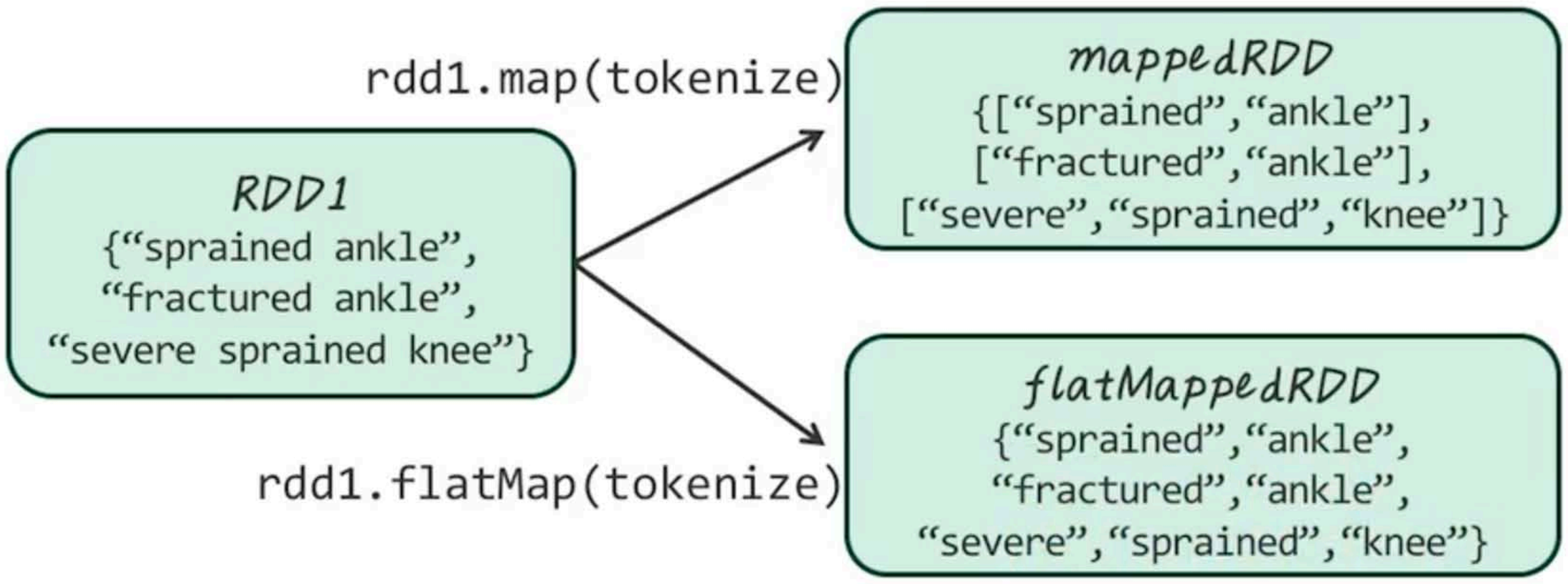
Transformations (define a new RDD)	map filter sample groupByKey reduceByKey sortByKey	flatMap union join cogroup Cross mapValues
Actions (return a result to driver program)	collect reduce Count save lookupKey	

https://www.udacity.com/course/big-data-analytics-in-healthcare--ud758

RDD TRANSFORMATIONS

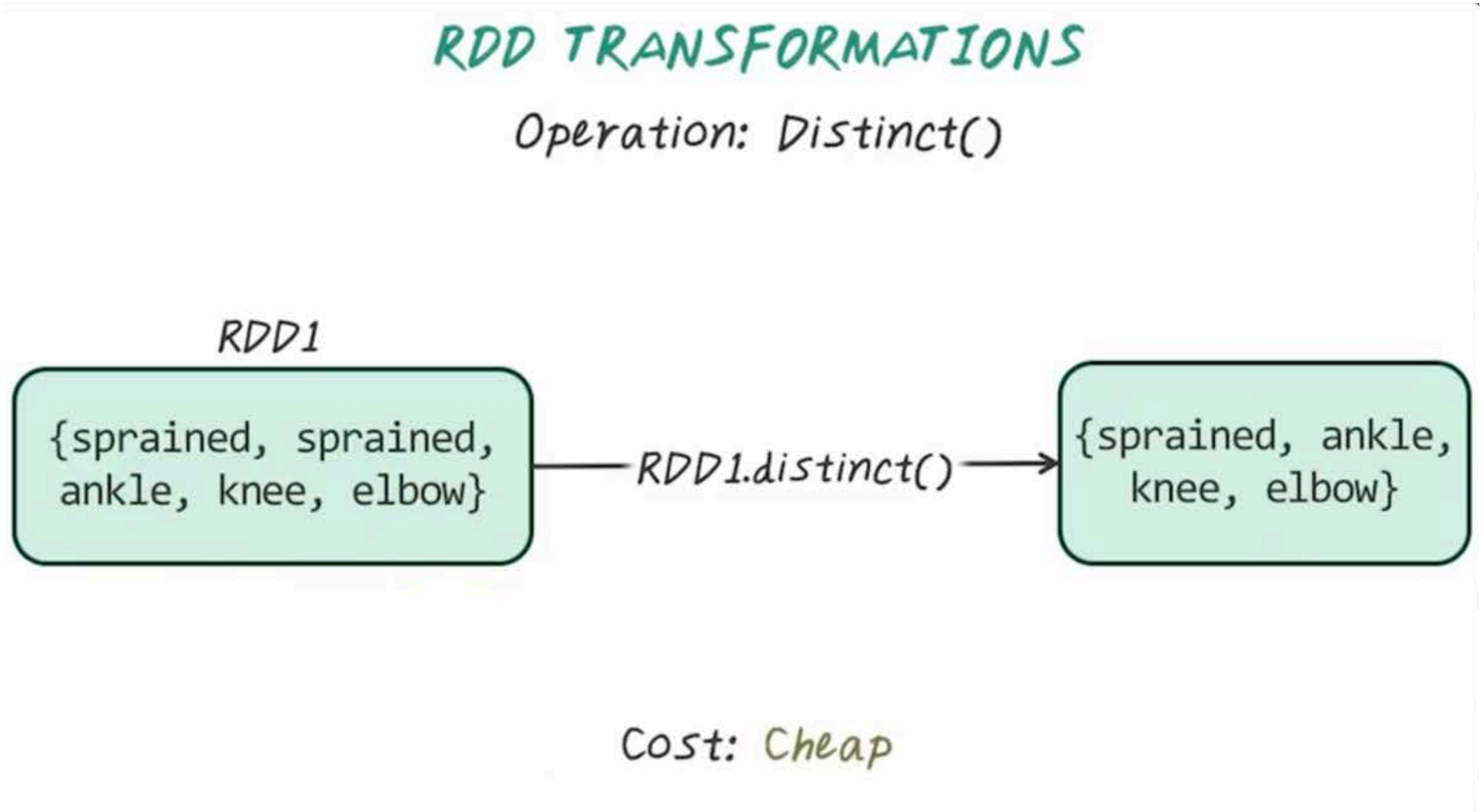
map() vs flatmap()

```
tokenize("sprained ankle")=List("sprained","ankle")
```



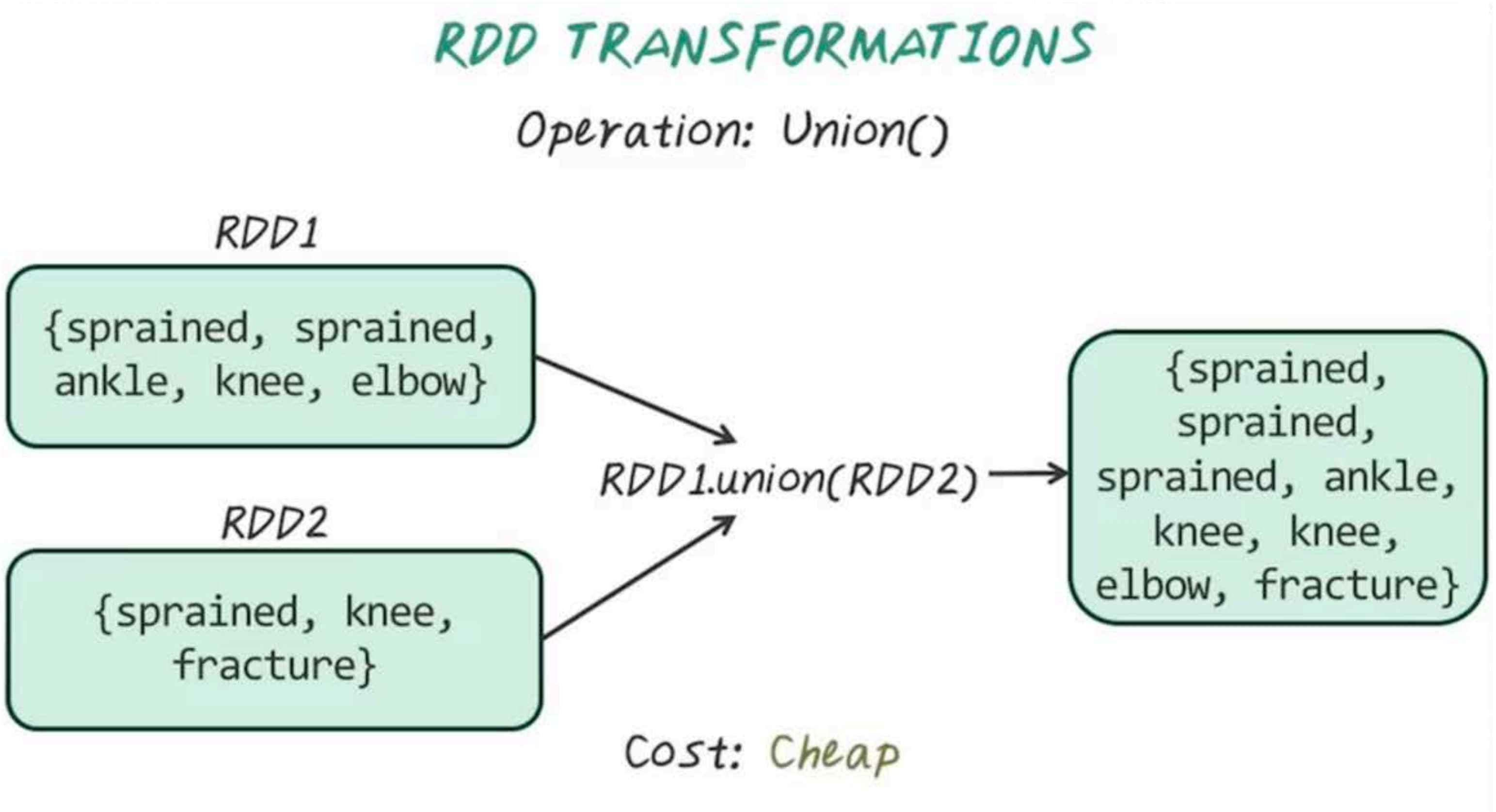
```
RDD1
{"sprained ankle",
"fractured ankle",
"severe sprained knee"}

rdd1.map(tokenize)

mappedRDD
{["sprained","ankle"],
["fractured","ankle"],
["severe","sprained","knee"]}

rdd1.flatMap(tokenize)

flatMappedRDD
{"sprained","ankle",
"fractured","ankle",
"severe","sprained","knee"}
```

https://www.udacity.com/course/big-data-analytics-in-healthcare--ud758

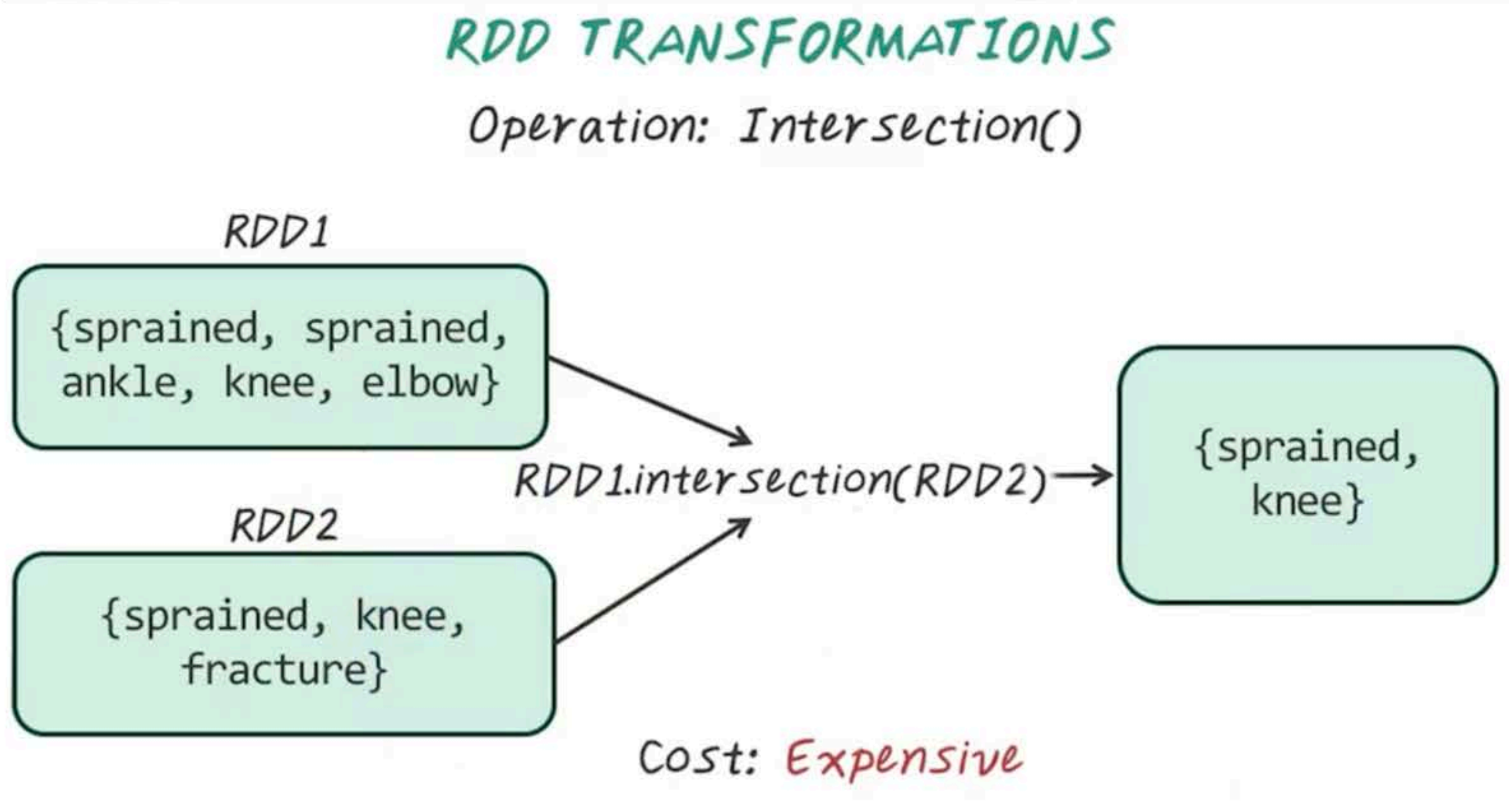

RDD TRANSFORMATIONS
Operation: Distinct()
RDD1
{sprained, sprained, ankle, knee, elbow}
RDD1.distinct()
{sprained, ankle, knee, elbow}
Cost: Cheap

https://www.udacity.com/course/big-data-analytics-in-healthcare--ud758

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE


RDD TRANSFORMATIONS
Operation: Union()
RDD1
{sprained, sprained, ankle, knee, elbow}
RDD2
{sprained, knee, fracture}
RDD1.union(RDD2)
{sprained, sprained, sprained, ankle, knee, knee, elbow, fracture}
Cost: Cheap

https://www.udacity.com/course/big-data-analytics-in-healthcare--ud758

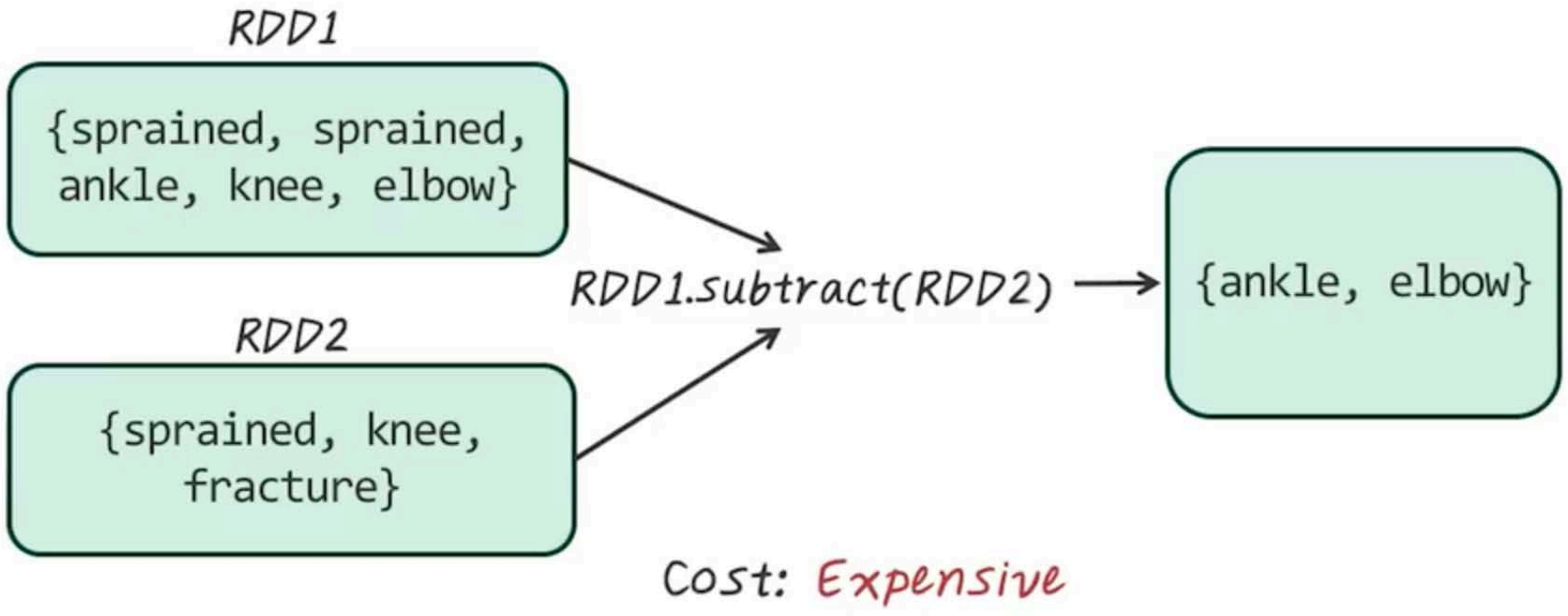

RDD TRANSFORMATIONS
Operation: Intersection()
RDD1
{sprained, sprained, ankle, knee, elbow}
RDD2
{sprained, knee, fracture}
RDD1.intersection(RDD2)
{sprained, knee}
Cost: Expensive

https://www.udacity.com/course/big-data-analytics-in-healthcare--ud758

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

RDD TRANSFORMATIONS

Operation: Subtract()



RDD1
{sprained, sprained, ankle, knee, elbow}
RDD2
{sprained, knee, fracture}
RDD1.subtract(RDD2)
{ankle, elbow}
Cost: Expensive

https://www.udacity.com/course/big-data-analytics-in-healthcare--ud758