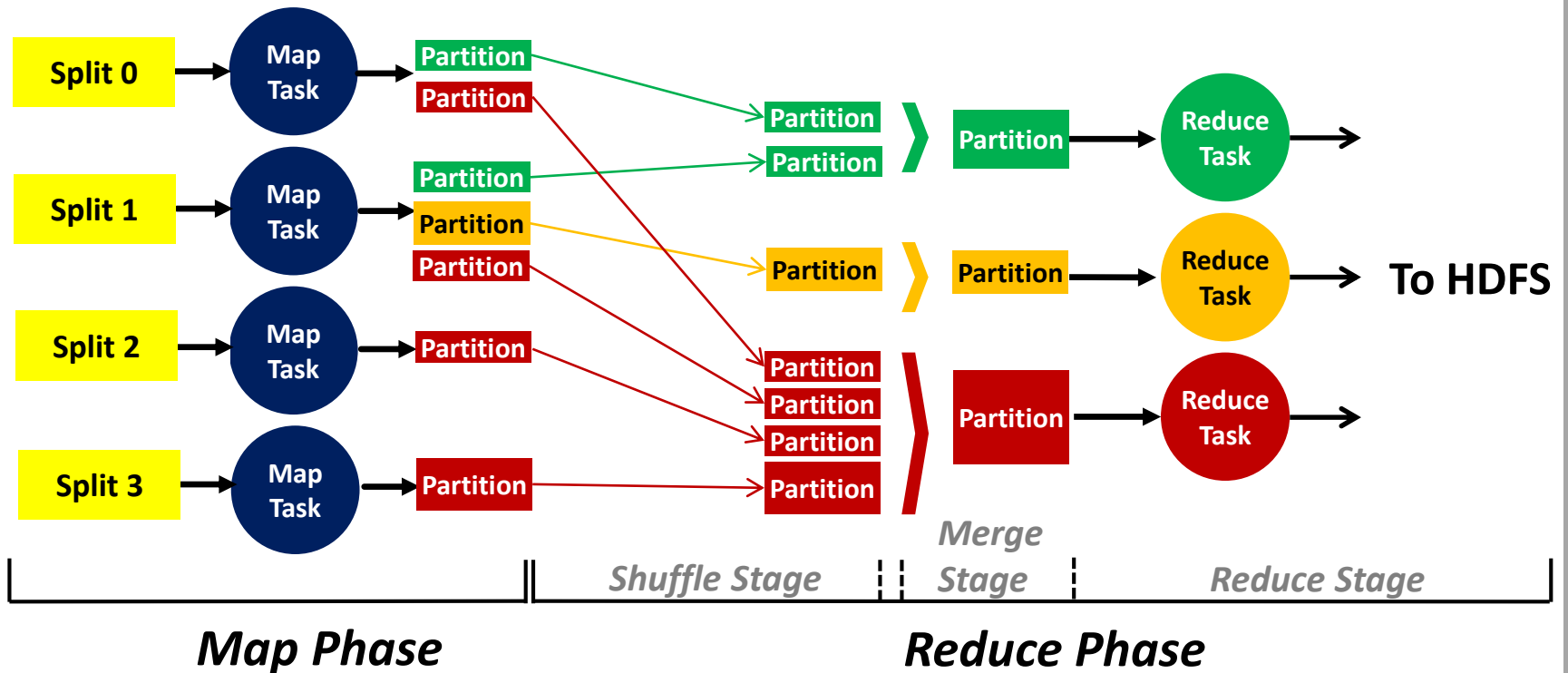# 15-440
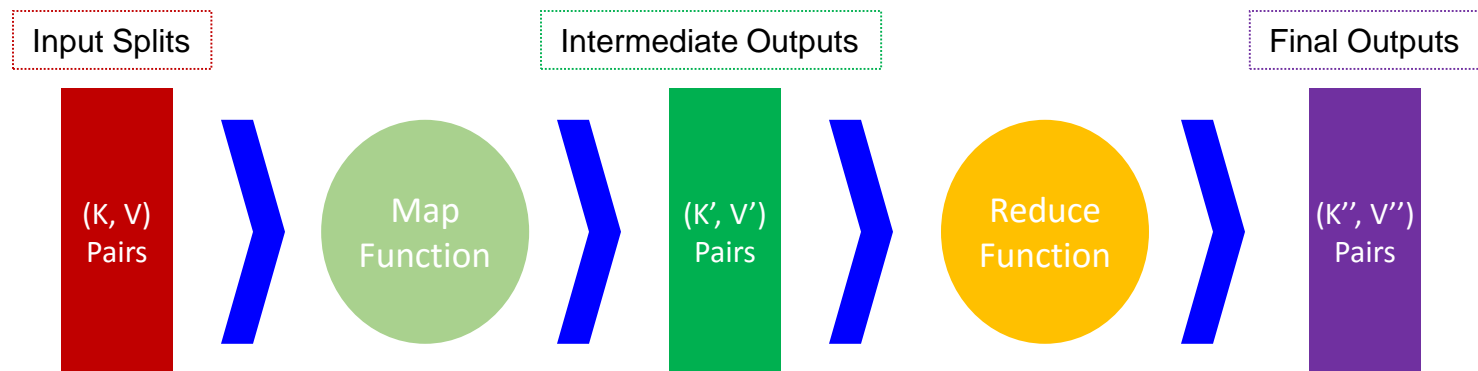# Distributed Systems
Recitation 12

Tamim Jabban

Project 4

- Apply MapReduce to cluster analysis, using the **K-Means** algorithm
- Due date: December 3rd
  - Please start early!

MapReduce: A Systems View

Data Structure: Keys and Values

- In a MapReduce program, the programmer has to specify <u>two functions</u>: the **Map function** and the **Reduce function** that implement the *Mapper* and the *Reducer*, respectively

- In MapReduce, data elements are always structured as <u>key-value (i.e., (K, V)) pairs</u>

- The Map and Reduce functions receive and *emit* (K, V) pairs

Input Splits		Intermediate Outputs		Final Outputs
(K, V) Pairs	Map Function	(K', V') Pairs	Reduce Function	(K'', V'') Pairs

MapReduce: An *Application* View

A *Map* Function

A Chunk of File

Tamim is delivering a recitation to the 15-440 class

Key1	Value1
0	Tamim is
20	delivering a
38	recitation to
60	the 15-440 class

Parse & Count

Key2	Value2
Tamim	1
is	1
delivering	1
a	1
recitation	1
to	1
the	1
15-440	1
class	1

A *Map* Function

A Chunk of File

The course name of 15-440 is Distributed Systems

Key1	Value1
0	*The course*
17	*name of 15-440*
40	*is Distributed*
58	*Systems*

Parse & Count

Key2	Value2
The	1
course	1
name	1
of	1
15-440	1
is	1
Distributed	1
Systems	1

A *Reduce* Function

Iterate & Sum

Key2	Value2
Tamim	1
is	2
delivering	1
a	1
recitation	1
to	1
the	2
15-440	2
class	1
course	1
name	1
of	1
Distributed	1
Systems	1