University of
South Australia

What is k-means clustering?
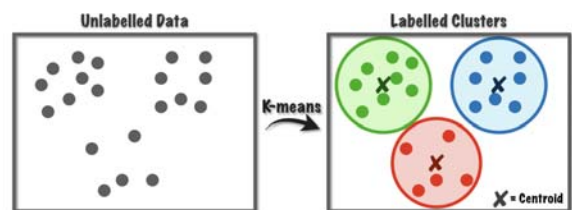

K-means clustering

https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c

Goal:
Clustering the data into K groups

Method:
- Select K points as cluster centres
- Assign observations to their closest cluster centre using some distance function
- Calculate the centroid of all observations in each cluster
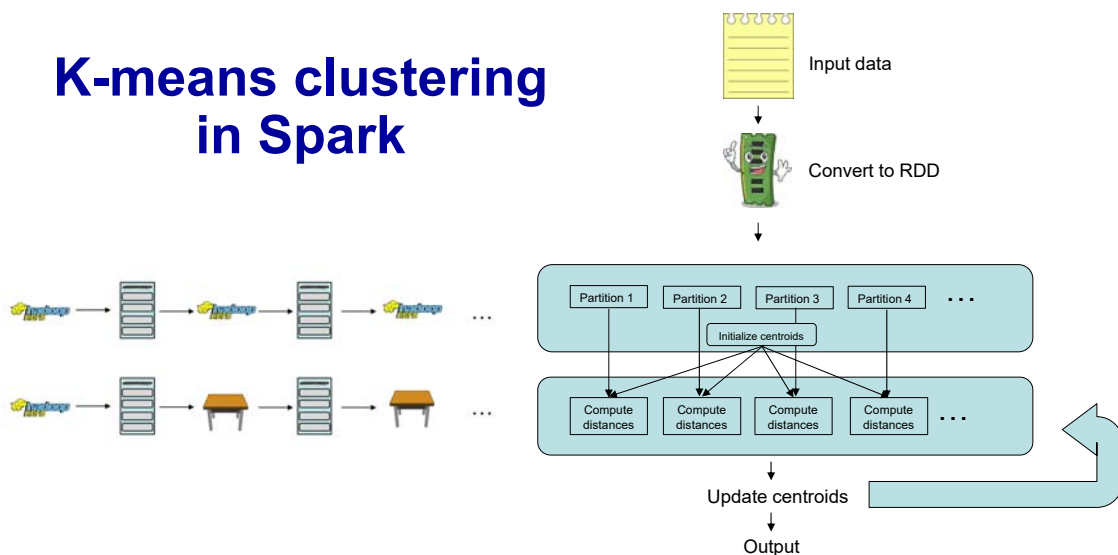- Repeat previous steps until stopping conditions are met

University of
South Australia

Question time

Q What parameters do we need to set?

Q In which part of the algorithm would MLlib leverage Spark's architecture to speed up K-means clustering?

University of
South Australia

3

K-means clustering in Spark

Input data

Convert to RDD

Partition 1 | Partition 2 | Partition 3 | Partition 4 · · ·
Initialize centroids

Compute distances | Compute distances | Compute distances | Compute distances · · ·

Update centroids

Output

University of
South Australia

4

University of
South Australia

5