

2) 3 reducers
 # invocation of reduce?
 4: Luca, Carmen, Danilo, Pablo
 (C)

Dataset

Customers.txt
 CID, Name, Surname, City, Country

TVSeries.txt
 SID, Title, Genre → can be only one

Episodes.txt
 SID, SeasonNum, EpisodeNum, Title, AirDate

Customers Watched.txt
 CID, StartTime, SID, SeasonNum, EpNum
 → can watch multiple times the same episode.

MapReduce

Longest life-spanning comedy TV series.
 lifespan = difference between first and last air date

Store SID

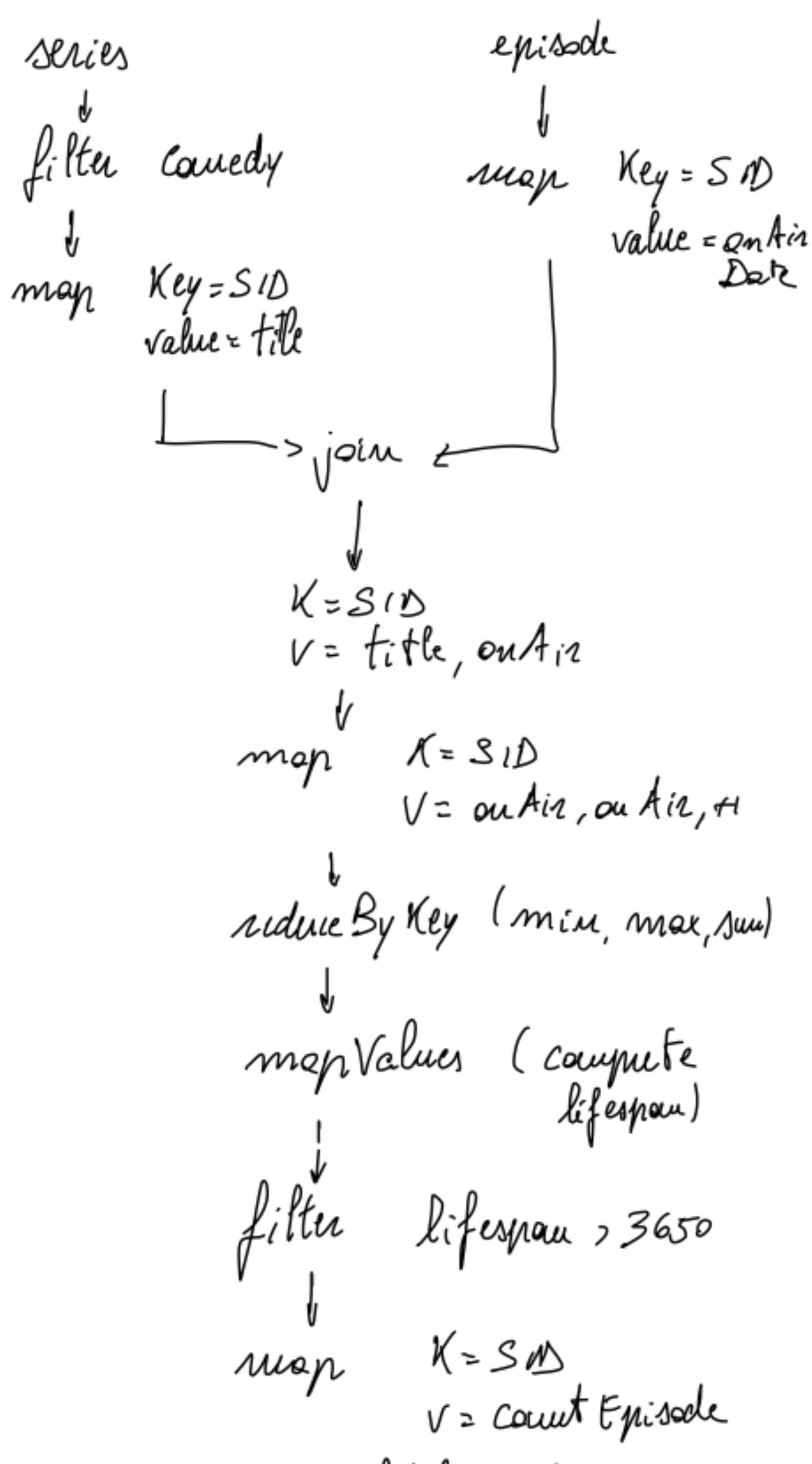
2 jobs (for efficiency)

job #1: # reducers > 1
 mapper: emit (SID, OnAirDate)
 reducer: compute min and max
 compute lifespan
 if lifespan > this.maxLifespan
 OR (lifespan == this.maxLifespan AND SID < this.SIDMax)
 found new max
 in cleanup
 emit (SIDMax, lifespanMax)

job #2: # reducers = 1
 mapper: emit (SID, lifespan)
 reducer: compute max
 emit (SIDMax, NullWait)

Spark

1) tot # of episodes for each comedy TV series with lifespan > 3650 days.



2) # of TV series completely watched by each customer.

CID, # all-watched series

