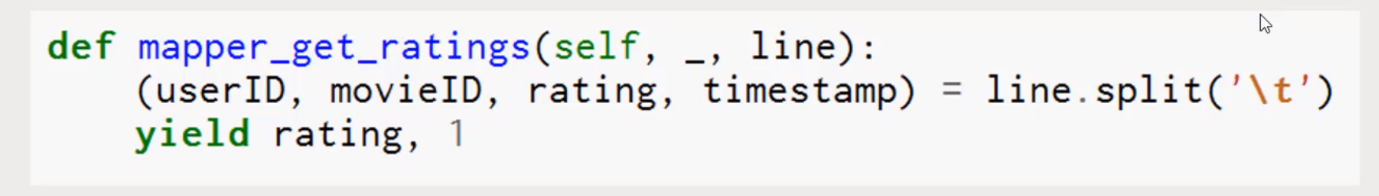
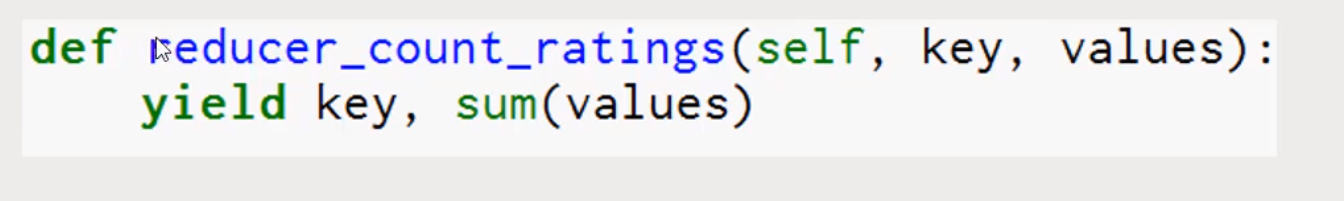
Keys are Ratings, Values are no.1s ,

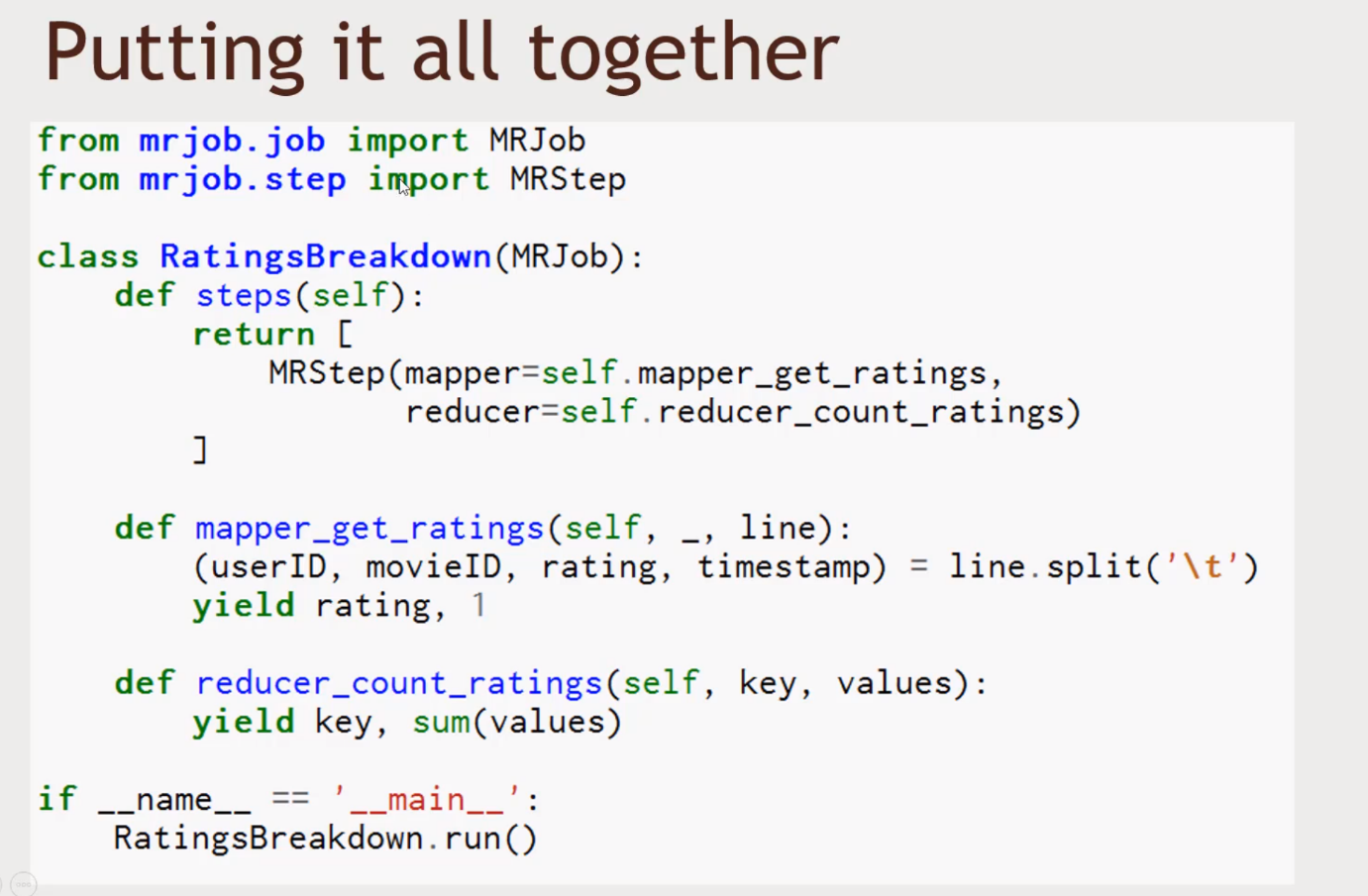
Reducer will add them all up.

self object, \_ unused in case of mapper, if we chaining mapping, then \_ will represent something else.

Split it on a tab character (tab delimited data) – Yield key value pair, yield key, value, yield rating and number 1



Each unique Key, values – list of all the values associated with the key. Iteratable values.



Connect to Putty

Root and change password into stevenpyae

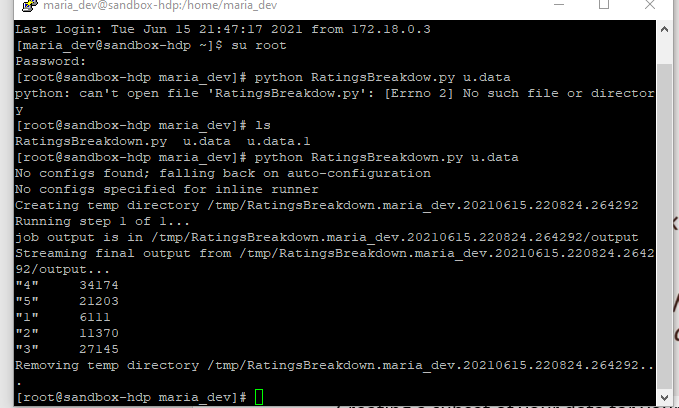
And follow the above screenshot

Running on your own.

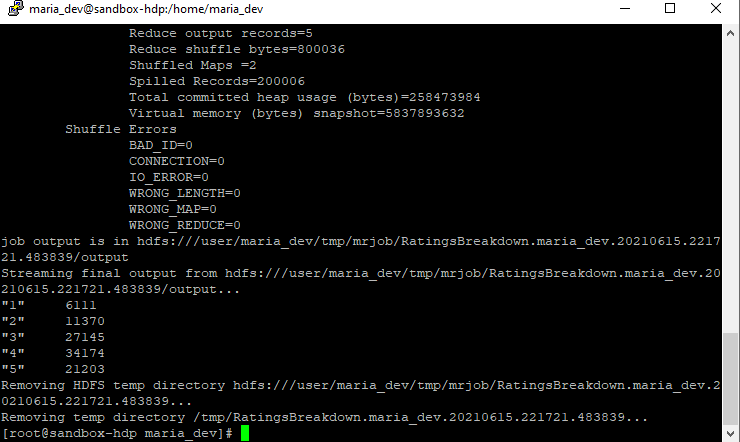
Creating a subset of your data for your testing on your own.

Ratings, 1 (1 through 5)

Yield the key, 1 through 5 and find the count of values.

Running Internally

Running under Hadoop Environment

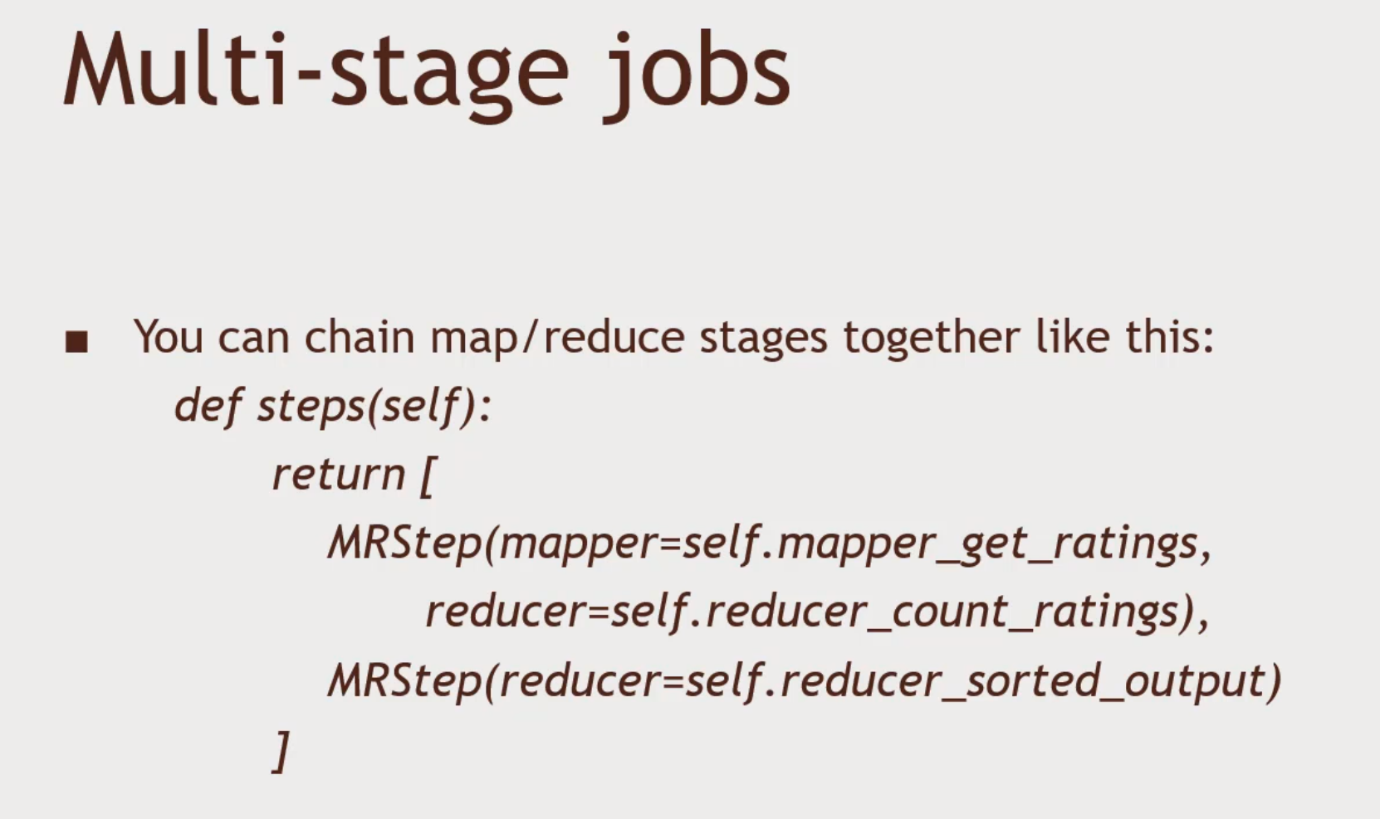
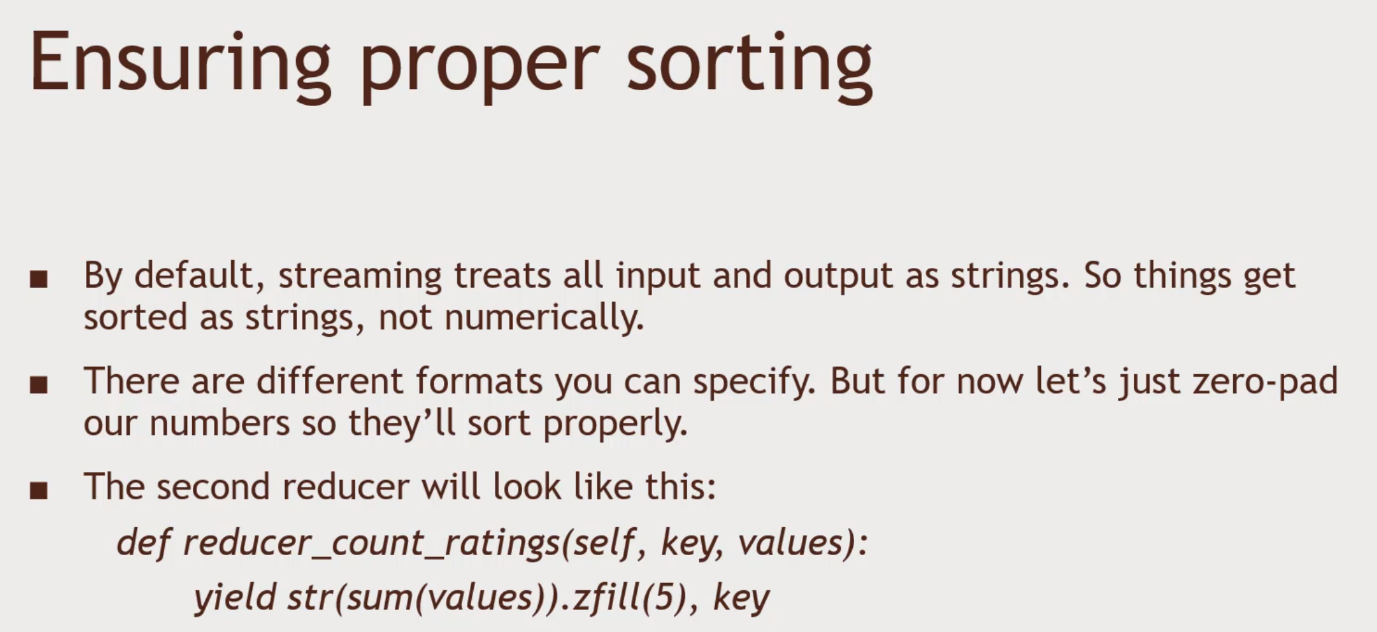
Count up ratings given for each movie

* All you need is to change one thing in the mapper – we don’t care about ratings now, we care about Movie ID’s
* Start with this and make sure you can do it
* You can use nano to edit the existing RatingsBreakdown.py script

Advanced

Sort the movies by their numbers of ratings

Strategy:

* Map to (movieID, 1) key/value pairs
* Reduce with output of (rating count, movieID)
* Send this to a second reducer so we end up with things sorted by rating count!

