Using the distributed cache in MapReduce to find lines that contain matching keywords.

The distributed cache in MapReduce is almost always required for any complex assignment involving dependent libraries and code. One very common operation is passing cache files for use in each map/reduce task JVM.

This recipe will use the MapReduce API and the distributed cache to associate abbreviated data to detail file which is present in the Cache.

Say we have data related to state-wise population of India as follows,

up 199654321

ma 112328654

bi 103876487

wb 91765349

If we need to come out with detailed names for these abbreviated data we can simply pass the detail data to cache which is as follows.

up Uttar\_Pradesh

ma Maharashtra

bi Bihar

wb WestBengal

Then read this file in Mapper or Reducer program where we can write logic to associate this detail information.