I have used 4 optimization methods

First I’ve used caching to cache movie, star and genre table , I cached their names and ids into

3 hash maps, in order to avoid duplication records.

Then during the insertion process, I disabled auto-commit, to reduce overhead in DBMS. And I have 5 batches of insert query, such as insert movie, insert star, insert genre, link genre and link star, each batch has a distinct prepared statement.

I think this optimization is acceptable for this kind of task, but the tradeoff for performance is memory. As I chosen the DOM parser and caching technic(abusing Hash maps which is a very memory efficient data structure), the memory usage is large, so this method might not be suit for large dataset.