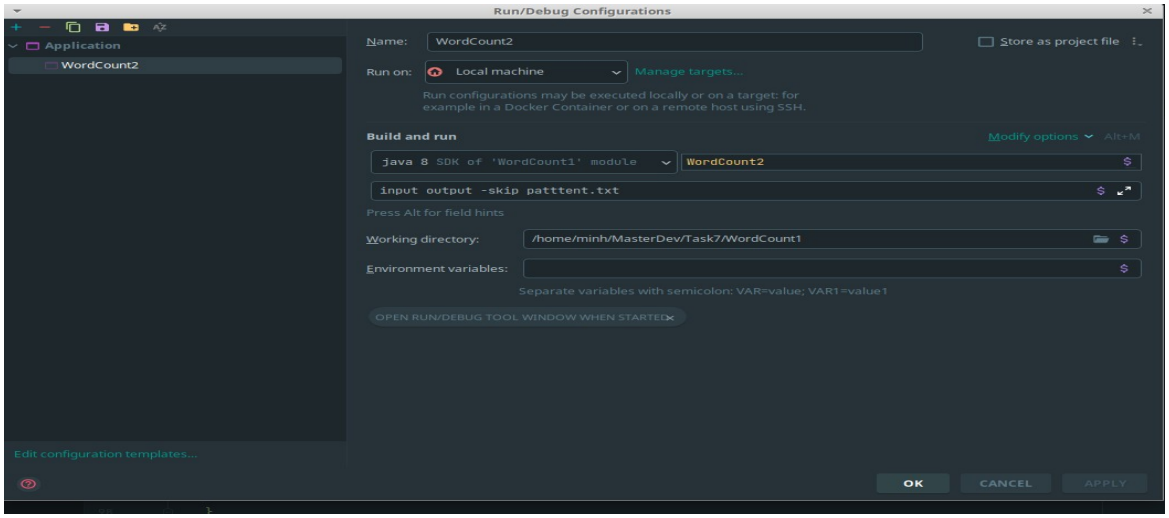
Bài 1: WordCount

- Em sử dụng ví dụ WordCount2 trên trang chủ apache
  Apache Hadoop 3.3.3 – MapReduce Tutorial

- **Lưu ý: Để chạy được trên Intellij cũng như trên HDFS các folder ouput cần phải chưa tồn tại nên khi chạy code phải xoá các folder output đã có hoặc sửa lại tên folder output**

- **Trên server trong folder /home/hadoop/minhnx12/bai1**
  - Hai câu lệnh chạy như sau:
    - yarn jar WordCount.jar WordCount /minhnx12/WordCount/Input/text /minhnx12/WordCount/output -skip /minhnx12/WordCount/patterns.txt
    - -skip để bỏ qua các kí tự trong file patterns.txt
    - yarn jar WordCount.jar WordCount -Dwordcount.case.sensitive=false /minhnx12/WordCount/Input/text /minhnx12/WordCount/output -skip /minhnx12/WordCount/patterns.txt
    - -Dwordcount.case.sensitive=false không phân biệt chữ hoa và chữ thường
  - Output nằm ở phần em gạch chân /part-r-0000 trên HDFS.
  - Em đã -get về thư mục **/home/hadoop/minhnx12/bai1/** 2 file output của 2 câu lệnh trên

- Cách chạy trên Intellij cấu hình các argument như ảnh dưới:
  - -Dwordcount.case.sensitive=false input output -skip patterns.txt

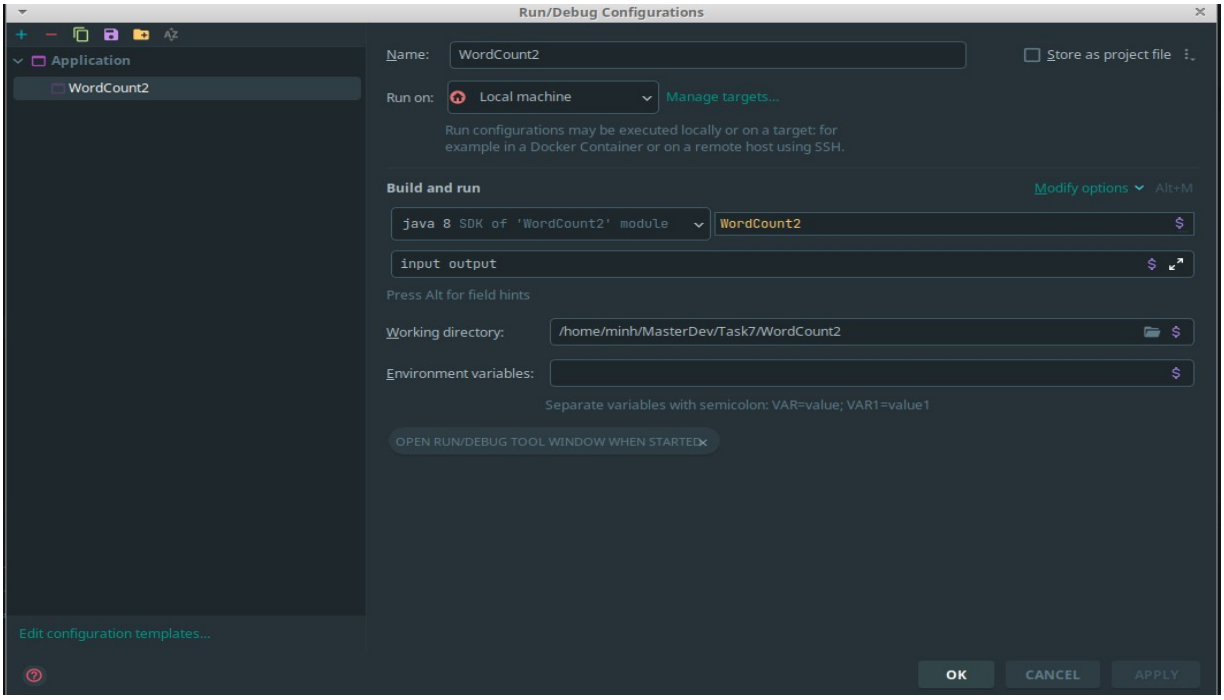


Bài 2: NumCount

- Em sử dụng ví dụ WordCount ở link trên và sửa code phần reduce.

- **Trên Server trong folder /home/hadoop/minhnx12/bai2**
    - Câu lệnh chạy như sau:
        - yarn jar WordCount2.jar WordCount2 /minhnx12/WordCountNumber/Input/count_distinct.csv /minhnx12/WordCountNumber/output
    - Output nằm ở phần em gạch chân /part-r-0000 trên HDFS
    - Em đã -get về thư mục **/home/hadoop/minhxn12/bai2/** file output của câu lệnh trên

- Cách chạy trên Intellij cấu hình các argument như ảnh dưới:
    - input output



Bài 3: JoinTable

**** reduce_join***

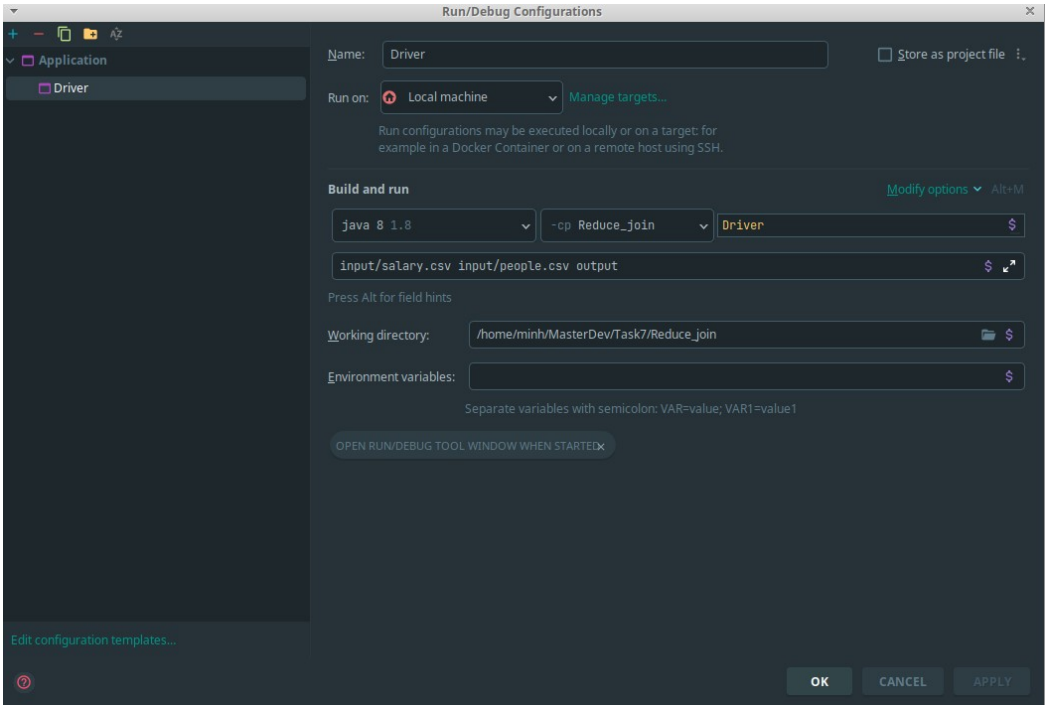
- Sử dụng 2 mapper cho 2 file csv để output ra <JobKey, JoinGenericWritable>
- Tạo 2 biến final SALARY_RECORD = 0 và PEOPLE_RECORD = 1 để lấy ra giá trị salary trước sau đó append vào các bản ghi people

- **Trên Server trong folder /home/hadoop/minhnx12/bai3**

    - Câu lệnh chạy như sau:
        - yarn jar reduce_join.jar Driver /minhnx12/map_reduce_join/input/salary.csv /minhnx12/map_reduce_join/input/people.csv /minhnx12/map_reduce_join/output

    - Output nằm ở phần em gạch chân /part-r-0000 trên HDFS
    - Em đã -get về thư mục **/home/hadoop/minhnx12/bai3/** file output của câu lệnh trên

- Cách chạy trên Intellij cấu hình các argument như ảnh dưới:
    - input/salary.csv input/people.csv output



* ***reduce_join***

- ***Trên Server trong folder /home/hadoop/minhnx12/bai4***

Câu lệnh chạy như sau:

yarn jar map_join.jar Driver /minhnx12/bai3/salary.csv /minhnx12/map_reduce_join/input/people.csv /minhnx12/map_reduce_join/output

Output nằm ở phần em gạch chân /part-r-0000 trên HDFS

- *Em đã -get về thư mục **/home/hadoop/minhnx12/bai4/** file output của câu lệnh trên*

Cách chạy trên Intellij cấu hình các argument như ảnh dưới:

- *input/salary.csv input/people.csv output*