Big data Management Analytics and Management

Please apply Hadoop map-reduce to derive some statistics from **Yelp Dataset.**
The dataset files are as follows and columns are separate using '**::**'
business.csv.
review.csv.
user.csv.

Dataset Description.
The dataset comprises of **three** csv files, namely user.csv, business.csv and review.csv.

Business.csv file contain basic information about local businesses.
Business.csv file contains the following columns "business_id"::"full_address"::"categories"

'business_id': (a unique identifier for the business)
'full_address': (localized address),
'categories': [(localized category names)]

review.csv file contains the star rating given by a user to a business. Use user_id to associate this review with others by the same user. Use business_id to associate this review with others of the same business.

review.csv file contains the following columns "review_id"::"user_id"::"business_id"::"stars"
 'review_id': (a unique identifier for the review)
 'user_id': (the identifier of the reviewed business),
 'business_id': (the identifier of the authoring user),
 'stars': (star rating, integer 1-5),the rating given by the user to a business

user.csv file contains aggregate information about a single user across all of Yelp
user.csv file contains the following columns "user_id"::"name"::"url"
user_id': (unique user identifier),
'name': (first name, last initial, like 'Matt J.'), this column has been made anonymous to preserve privacy
'url': url of the user on yelp

Hadoop Map-Reduce programs in Java to find the following information.

NB: :: is Column separator in the files.

Q1. List the unique categories of business located in "Palo Alto"

Sample output:

Automotive
Auto Detailing

…….

Q2

Find the top ten rated businesses using the average ratings. Top rated business will come first. Recall that 4ᵗʰ column in review.csv file represents the rating.

Please answer the question by calculating the average ratings given to each business using the review.csv file.

Sample output:
business id   rating
xdf12344444444   4.9

Q3:
List the business_id , full address and categories of the Top 10 businesses using the average ratings.

This will require you to use **review.csv** and **business.csv files.**

Please use reduce side join and job chaining technique to answer this problem.

Sample output:

business id	full address	categories	avg rating
xdf12344444444,	CA 91711	List['Local Services', 'Carpet Cleaning']	5.0

Q4:
List the 'user id' and 'rating' of users that reviewed businesses located in Stanford
Required files are 'business' and 'review'.

Please use In Memory Join technique to answer this problem.
Hint: Please load all data in business.csv file into the distributed cache.

Sample output

User id	Rating
0WaCdhr3aXb0G0niwTMGTg	4.0