CS 378 – Big Data Programming

Lecture 26
Review

• Assignment 11 – Custom Input Format
  – Generate random messages
    • From a set of ten words
    • Random selection with replacement
    • Specific distribution for message length
    • Mean 50, standard deviation 10
    • Message lengths between 1 and 99
  – Word count statistics on the random messages
  – Statistics on message length
Assignment 11

• Some important points

• In the RecordReader
  – Try to limit small object allocation by reusing objects

• In your mapper, limit small object allocation

• Use a combiner to limit data transfer
Assignment 12

• Utilize multiple patterns/techniques
  – Filtering, inverted index
  – Reduce-side join
  – Summarization
  – Job chaining
Assignment 12 - Task

• Collect data on the price ranges of vehicles of interest to users
• “Interest” -> user clicked on the vehicle to view the details (VDP impression)

• Answer questions like:
  – For users searching for a vehicle around $15K
  – How broad is the range of prices they consider
Data

• What data do we have for this task?
  – User sessions contain VDPs
  – For a user, we have VDPs
  – Each VDP has an ID

• What else do we need?
  – Data that associates a price with the ID

• We’ll join using the common data: ID
Step 1

• Identify IDs (vehicles) viewed by the user
  – Filtering pattern

• We need an ID as the key (for join)

• The value will be a userID
  – This is essentially the inverted index pattern

• Output: ID and a userId
Step 2

• Join the two data sources (keyed by ID)
  – Data created in step 1
  – The ID/price file
    • Data format: ID,price

• Reduce-side join pattern

• Output: userId and price
Step 3

• Aggregate the price data for each user
  – Each user viewed one or more vehicles (IDs)

• Compute the statistics
  – Number, min, max, mean, median, standard deviation, skewness, kurtosis
  – Summarization pattern
Recommendations

• Use the small data sets to test your code
• Implement the solution as multiple steps
  – Independent jobs initially
  – Then combine using job chaining
• Use `DoubleArrayWritable` for final output
• Use `DescriptiveStatistics` for stats