1. Welcome to Lesson 4
2. Overview of Patterns
   1. Filtering patterns
      1. Sampling
      2. top-N list
   2. Summarization patterns
      1. Counting
      2. min/max
      3. Statistics
      4. Index
   3. Structural patterns
      1. Combining data sets
3. Filtering Patterns
   1. Don’t change actual details of data. It just filters
   2. Simple filter (returns and throw away records)
   3. Bloom filter
   4. Sampling: make smaller data set from bigger data set
   5. Random sampling
   6. Top 10 of something
4. Filtering Exercise
   1. Filter short posts one sentence or less
5. Top 10
   1. RDMS
      1. Sort data
      2. Pick top N records
   2. MapReduce
      1. Each mapper generate top N list
      2. Reducer finds global top N

1. Summarization Patterns
   1. Inverted index
   2. Numerical summarizations
      1. Count
      2. min/max
      3. first/last
      4. Mean, median
2. Inverted Index
3. Numerical Summarizations
   1. Word/record count
      1. Key: thing you want
      2. Value: 1
   2. Mean, median, SD
4. Finding mean
   1. Is there any correlation between the day of the week and how much people spend on items?
   2. *Mapper*: `Monday`: 5.20
   3. *Reducer*: All math
5. Combiners
   1. To calculate the mean (what happened in IX):
      1. Mappers go through data output `Monday`: 6.35
      2. For each day, your reducer kept a sum and a count
      3. Divide sum by count
   2. What if before reducer comes into play, use *combiners*?
      1. Combines the input records
6. Structural Patterns
   1. Used to “migrate” RDBMS to hadoop
   2. Reformatting data.
   3. In order to use this pattern,
      1. Data sources linked by foreign keys
      2. Data must be structured and row based
7. Combine Datasets
8. Conclusion