CS4225: Massive Data Processing Techniques in Data Science

Tutorial 9: Stream Processing

1. The reservoir sampling algorithm maintains a sample *S* (of *s* elements) with the desired property: After *n* elements, the sample contains each element seen so far with probability *s/n*. Prove this property. [Hints: We prove this by induction.]

2. Consider a bloom filter with an m-bit bit vector, and k hash functions: h_1, h_2, …, h_k. What is the false positive probability for a membership operation?

3. We wish to use **Flajolet-Martin** counter algorithm to count the number of distinct elements in a stream. Suppose that there are ten possible elements, 1, 2, …, 10, that could appear in the stream, but only four of them have actually appeared. To make our estimation of the count of distinct elements, we hash each element to a 4-bit binary number. The element x is hashed to 3x + 7 (modulo 11). For example, element 8 hashes to 3*8+7 = 31, which is 9 modulo 11 (i.e., the remainder of 31/11 is 9). Thus, the 4-bit string for element 8 is 1001.

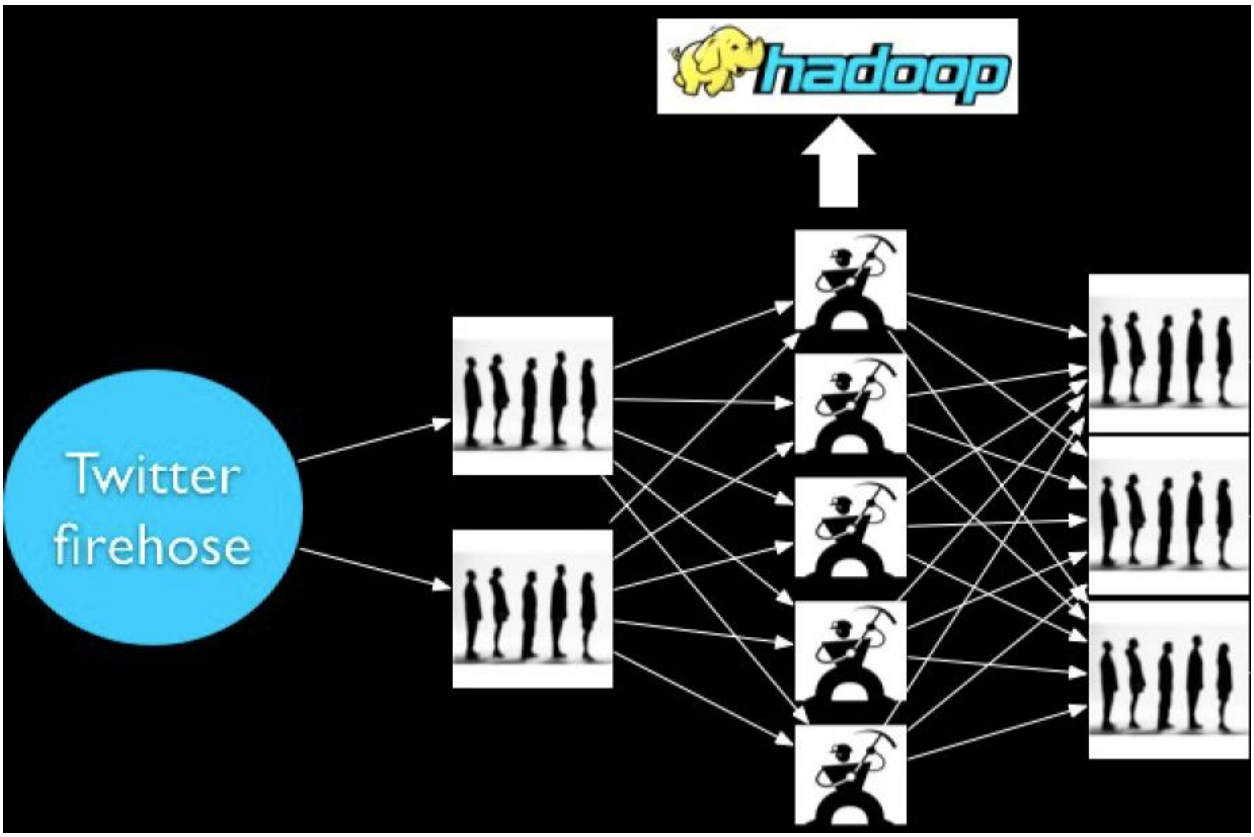
A set of four of the elements 1 through 10 could give an estimate that is exact (if the estimate is 4), or too high, or too low. You should figure out under what circumstances a set of four elements falls into each of those categories. Identify the set of four elements that gives the exactly correct estimate.

4. Consider Storm architecture and the job topologies in tweet processing in Twitter. Prior to Storm, the task of "schemify tweets and append to Hadoop" is implemented in Hadoop, as illustrated in the below picture. There are three major issues of implementing the task using Hadoop. Explain the reasons.

a) Scaling is painful

b) Poor fault-tolerance

c) Coding is tedious


hadoop
Twitter
firehose