* Jaccard similarity is the similarity of sets by looking at the relative size of their intersection, measured by the ratio of the size of the intersection of the two sets to the size of their union.
* A variation of shingling produces a bag, rather than a set, each shingle would appear in the result as many times as it appears in the document. But most of the time we use set other than bag.
* When picking k, if we pick it too small, we would expect most sequences of k characters to appear in most documents. k should be picked large enough that the probability of any given shingle appearing in any given document is low.
* Since it needs large space to store sets of shringles, we want to replace large sets by much smaller representations called “signatures.” The important property we need for signatures is that we can compare the signatures of two sets and estimate the Jaccard similarity of the underlying sets from the signatures alone.
* The signatures we desire to construct for sets are composed of the results of a large number of calculations, say several hundred, each of which is a “minhash”.
* The remarkable connection between Jaccard similarity and Minhash is that, the probability that the minhash function for a random permutation of rows produces the same value for two sets equals the Jaccard similarity of those sets.
* There are two ways that queries get asked about streams, standing queries are permanently executing, and produce outputs at appropriate times, while ad-hoc queries are a question asked once about the current state of a stream or streams
* Examples of Stream Sources includes sensor data, image data, internet and web traffic.
* The purpose of the Bloom filter is to allow through all stream elements whose keys are in S, while rejecting most of the stream elements whose keys are not in S.