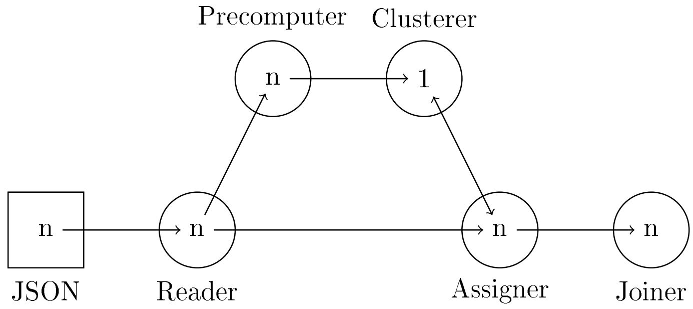
Chapter 4

Approach

4.1 Apache Storm Topology



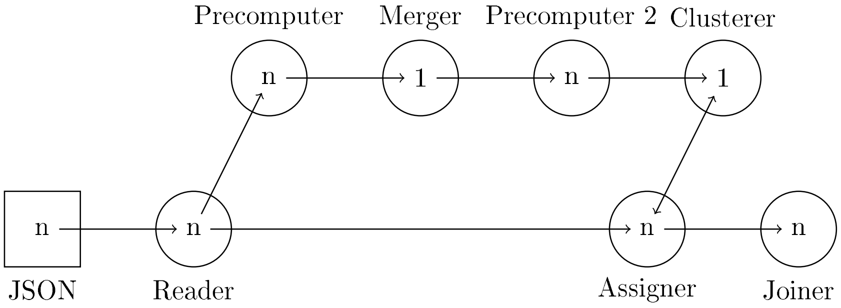
Precomputer: Computes the attribute-value CoOccurances over a small set of documents and does precomputation (e.g splitting common attributes). This process can be parallelized. !potentially problematic!

Clusterer: merges the tables obtained from the precomputer and performs the clustering and partitioning

Assigner: dispatches documents to the joiner. Informs the clusterer on the quality of the partitions

Joiner: performs the actual join of the documents

4.2 Alternative Apache Storm Topology



Precomputer: Computes the attribute-value CoOccurances over a small set of documents. This process can be parallelized.

Merger: Merges the tables from the precomputer and splits common attributes

Precomputer 2: Recomputes the attribute-value CoOccurances over a small set of documents. This process can be parallelized.

Clusterer: merges the tables obtained from the Precomputer 2 and performs the clustering and partitioning

Assigner: dispatches documents to the joiner. Informs the clusterer on the quality of the partitions

Joiner: performs the actual join of the documents