

# ABSTRACT

This paper describes a workflow manager developed and deployed at Yahoo called *Nova*, which pushes continually-arriving data through graphs of Pig programs executing on Hadoop clusters. Nova is like data stream managers in its support for stateful incremental processing, but unlike them in that it deals with data in large batches using disk-based processing. Batched incremental processing is a good fit for a large fraction of Yahoo's data processing use-cases, which deal with continually-arriving data and benefit from incremental algorithms, but do not require ultra-low-latency processing.