Big Data Analytics 2013

Invited talk

Online Data Processing with S4 and Omid

Flavio Paiva Junqueira, *Microsoft Research*

A number of online applications available on the Web rely upon processing streams os events. Such streams include user generated content in social networking applications, clickthrough events of online services, and even documents crawled to populate a search index. One natural way of processing such streams is to map events to keys and have the same processing element manipulating events with the same key, similar to offline platforms based on MapReduce.

S4 is a platform for building stream processing applications that follows this model. It enables applications to use resources in a flexible manner by allocating clusters of S4 nodes and to partition the load of a stream across nodes. Each application can produce one or more output streams and other applications can subscribe to receive the events of such streams The ability of having output streams enables applications to dynamically compose.

A platform like S4, however, does not provide directly any facility for the processing elements to access shared state. In the case such an access is necessary, approaches that process events against a data store directly are more suitable. Given the concurrent nature of accesses, it is desirable to provide guarantees such as atomicity and isolation of transactions when modifying the data store. We have consequently designed and implemented Omid to be a transaction manager for NoSQL data stores. Omid is a lock-free, centralized implementation of transactions that enables low latency. Solutions based on Omid add to the solution space of event processing by offering the ability of processing events against shared state.