Abstract

Data stream processing systems (DSPSs) compute real-time queries over constantly changing streams of data. A stream is a possibly infinite sequence of tuples, or timestamped data items. In contrast to traditional databases, in which queries are issued over stored data, results in DSPSs are generated continuously as new data enters the system.

The constant increase in data volume renders the provisioning of a DSPS difficult and costly. It may require computing resources that may not be available or too costly to purchase. Even if a user opts for a cloud deployment, thus renting all resources on demand, perfect processing may still not be possible due to the financial cost. For this reason, *overload* should be considered a common operating condition for such DSPS and not an exception.

Overload can be considered a type of failure because the system is not able to fully carry out the required computation. A system operating under constant overload is thus subject to continuous failure. In this situation, the system needs to discard some of its input data, an operation called *load shedding*.

Many streaming applications are able to produce valuable results even after some failure has occurred during the processing. Examples of such applications are meso-scale weather prediction, tornadoes and hurricanes forecasting, and real-time social media monitoring. An approximated result may still be useful to the user, as long as it is delivered with a low latency and it contains some information about its quality. In many cases, an imperfect result is better than no result at all.

We propose a new stream processing model under overload. The system constantly estimates the impact of overload on the computation and reports to the user the achieved quality-of-service. We introduce a quality metric called *Source Information Content (SIC)*. This can be used by the user as an indicator for the achieved quality-of-service and by the system to implement intelligent shedding policies and to better allocate the system resources among users. When an overloaded system performs *load-shedding*, the choice of how much and what to discard is crucial for the correct functioning of the system. The SIC quality metric helps the system make more informed decisions when to shed data. It allows the implementation of a *fair-shedding* policy, giving an equal quality-of-service to all users, without penalising certain kinds of queries.

We develop these ideas as part of a research prototype called DISSP, the Dependable Internet-Scale Stream Processing engine. With it we explore the issues related to overload management and fair resource allocation. We show that augmenting streams with the SIC metric allows the system to make better load-shedding decisions, leading to more accurate results for many queries. It also allows the user to reason about the amount of processing resources that are needed to run a given query, striking a balance between the quality of the delivered results and the cost of operating the system.