Presentation Abstract

Centaurus: A Just-in-time Parallel-parser Generator for Ad Hoc Data Processing

Shigeyuki Sato[1,a)] Hiroka Ihara[2] Kenjiro Taura[1,b)]

Presented: July 26, 2019

It is important to handle data in text formats such as XML, JSON, and CSV because these data very often appear in the context of data exchange. Only parts of these data are typically used afterwards so that it is not worth ingesting the whole of them into databases. It is therefore desired to match and extract the concerned part in a lightweight ad hoc manner. Classically used for such a purpose are linewise regular expression tools such as grep, sed, and awk. These are, however, not powerful enough for text formats commonly used for data exchange because they cannot recognize nested structures in general. To support a lightweight ad hoc data processing, we present Centaurus, a just-in-time parallel-parser generator library. By generating native scannerless LL(*) parsers dynamically, our library enables us to process input data in parallel merely by calling Python functions with LL(*) grammars and Python actions. This presentation gives the design and implementation of Centaurus and reports its experimental performance on data filtering.

This is the abstract of an unrefereed presentation, and it should not preclude subsequent publication.

[1] Graduate School of Information Science and Technology, The University of Tokyo, Bunkyo, Tokyo 113–8656, Japan

[2] Center for Technology Innovation, Hitachi Ltd., Yokohama, Kanagawa 244–0817, Japan

[a)] sato.shigeyuki@mi.u-tokyo.ac.jp

[b)] tau@eidos.ic.i.u-tokyo.ac.jp