Making Data Analysis More Efficient with Noir

Luca De Martini[1], Alessandro Margara[1] and Gianpaolo Cugola[1]

[1]*Politecnico di Milano*

Abstract

Modern distributed platforms for scalable data analysis offer a high-level programming model that results in simple and concise definitions of the processing tasks, abstracting away most of the concerns associated to concurrency and distribution, but at the cost of a large performance gap with custom programs that use low-level primitives to control distribution and resource usage.

To investigate if it is possible to reduce this performance gap without affecting ease of use we developed Noir, a novel platforms for data analysis that combines the simple programming model of state-of-the-art solutions with an efficient engine, exploiting implementation strategies mutuated from high-performance computing. Our results show that it is indeed possible to achieve a level of performance that is comparable to custom implementations without abandoning a simple programming model.

1. Motivations and Research Question

Most modern applications are data-intensive [1]: they continuously analyze large volumes of data to extract valuable knowledge for the environment in which they operate. Over the last decade, many platforms for distributed data processing have been proposed to effectively support data-intensive applications [2, 3]. These platforms offer a simple programming model that hides most of the complexity associated to parallel and distributed computations, thus making data analysis at scale more accessible. However, they cannot provide a level of performance that is comparable to custom programs designed from the specific problem at hand.

Our research investigates the design of a data analysis platform capable of reducing the performance gap with custom programs without negatively affecting simplicity and ease of use. Our research efforts resulted in Noir [4], a data processing platforms that retains the same programming model of alternative solutions while offering up to more than one order of magnitude higher throughput.

2. Noir: an Efficient Platform for Data Analysis

Noir is a data analysis platform designed to offer the same expressivity and simplicity of state-of-the-art solutions while providing a more efficient implementation. To do so, it mostly relies on the dataflow model, a simple programming model that defines computations on data in the form of directed graphs, where vertices represent operators and edges represent the flow of data from operator to operator [5]. Operators do not share state: the model promotes parallelism

BigHPC2023: Special Track on Big Data and High-Performance Computing

luca.demartini@polimi.it (L. De Martini); alessandro.margara@polimi.it (A. Margara); gianpaolo.cugola@polimi.it (G. Cugola)

CEUR Workshop Proceedings (CEUR-WS.org)

by deploying operators independently on the same or on different hosts and by instantiating multiple copies of each operator that process independent partitions of the input data in parallel. Noir supports the analysis of both static and dynamic (streaming) datasets, offers a rich library of operators for data transformation, and it enables disciplined forms of iterative computations.

At the same time, Noir provides a level of performance that is close to custom algorithms written using low-level programming promitives such as C/MPI. In fact, Noir achieves up to more than one order of magniture better throughput that competing platforms. These results derive from key design and implementation strategies that are inspired by HPC solutions. (i) While most alternative solutions rely on JVM-based platforms, Noir is written in Rust, a compiled programming language that offers high-level abstractions with vitually no run-time overhead. (ii) Noir adopts a lightweight approach to resource management, which delegates critical scheduling decisions to the operating system. (iii) Noir laverages the mechanisms embedded into TCP to implement flow control between operators.

Noir is available as an open-source project[1]. It has been used to compete to the DEBS 2022 Grand Challenge, winning the performance award for the solution with the highest throughput and lowest latency [6]. We are currently investigating strategies to dynamically scale the use of resources and to migrate part of the computation to different hosts, with the goal of adapting to changes in the workload and in the compute infrastructure. We are also exploring the integration of hardware accelerators such as GPUs within the same programming model.

In summary, Noir aims to integrate the efficiency of high-performance computing and the simple programming model of data analysis platform, allowing developers to easily implement and maintain their processing tasks without sacrificing absolute performance.

References

[1] A. Margara, G. Cugola, N. Felicioni, S. Cilloni, A model and survey of distributed data-intensive systems, ACM Comput. Surv. (2023).

[2] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker, I. Stoica, Apache spark: A unified engine for big data processing, Comm. ACM 59 (2016) 56–65.

[3] P. Carbone, A. Katsifodimos, S. Ewen, V. Markl, S. Haridi, K. Tzoumas, Apache flink™: Stream and batch processing in a single engine, IEEE Data Eng. Bulletin 38 (2015) 28–38.

[4] L. D. Martini, A. Margara, G. Cugola, M. Donadoni, E. Morassutto, The noir dataflow platform: Efficient data processing without complexity (2023). URL: https://arxiv.org/abs/2306.04421.

[5] T. Akidau, R. Bradshaw, C. Chambers, S. Chernyak, R. J. Fernández-Moctezuma, R. Lax, S. McVeety, D. Mills, F. Perry, E. Schmidt, S. Whittle, The dataflow model: A practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing, VLDB 8 (2015) 1792–1803.

[6] L. De Martini, A. Margara, G. Cugola, Analysis of market data with noir: Debs grand challenge, in: DEBS 2022, ACM, 2022, p. 139–144.

[1] https://github.com/deib-polimi/noir