# The Role of Performance in Streaming Analytics Projects: Expert Interviews on Current Challenges and Future Research Directions

Johannes Rank, Andreas Hein, Helmut Krcmar
Technical University of Munich
85748 Garching, Germany
{johannes.rank, andreas.hein, helmut.krcmar}@tum.de

## Abstract

Stream processing systems (SPS) are becoming more frequent due to current trends such as Industry 4.0 or the Internet of Things. These systems' performance is particularly important, as their timely processing is a crucial capability. At the same time, these systems are often combined with novel machine learning approaches (steaming analytics) that have high-performance demands. This combination poses potential challenges for performance management. In this paper, we have conducted expert interviews in the industry to identify performance challenges in streaming analytics implementations and to derive future research directions to address them. Our analysis shows that while the experts had different opinions on the role of performance in project management, they agreed on five common challenges.

## 1 Introduction

Stream Processing Systems (SPS) are a backbone technology to ensure timely processing, especially in areas such as the Internet of Things (IoT) or Industry 4.0. While these systems can achieve impressive performance results, they often run distributed and require careful performance settings e.g., for parallelization and state management. At the same time, streaming analytics, especially the integration of machine learning (ML) concepts, is becoming increasingly important. This combination of distributed and highly configurable systems, coupled with the complexity of ML approaches and low latency requirements, can pose significant challenges to performance management. For this reason, we wanted to explore how industry experts manage streaming analytics implementation projects. Our focus is to identify the current role of performance in project management, what challenges the industry is currently facing, and what future research directions could solve these. To this end, we conducted a semi-structured interview with industry experts in streaming analytics projects and compared the results to the state of the research.

Table 1: Interview Partners

| Expert | Business Sector | Job Profile | Experience |
|---|---|---|---|
| A | IT-Consulting | Architect | >5 years |
| B | IT-Consulting | Project Manager | >10 years |
| C | Manufacturing (Industry Automation) | Project Manager | >10 years |
| D | Manufacturing (Industry Automation) | Developer | >5 years |
| E | Manufacturing (Automobile) | Project Manager | >5 years |
| F | Banking | Developer | >10 years |
| G | IT-Consulting | Architect | >5 years |
| H | Telecommunication | Developer | >10 years |

## 2 Related Work

Performance research in SPS has typically focused on individual areas of performance management, such as evaluating the streaming application [5], benchmarking the streaming engine [7] or predicting upscaling scenarios [2]. However, to the best of our knowledge, there are no studies on the performance challenges of streaming analytics implementation projects. Streitz et. al (2018) conducted semi-structured expert interviews on performance improvement barriers in the context of SAP Enterprise Resource Planning systems [4]. The methodology of our interviews are similar in its design and approach.

## 3 Methodology

We performed a semi-structured interview with eight industry experts. Therefore, we pre-defined an interview guide composed of thirteen questions to ensure comparability among the candidates. However, we allowed spontaneous additions of questions to explore topics further as long as they fit the discussion. As our target group for the interviews, we identified Software Developers, Solution Architects, and Project Managers in the area of streaming analytics projects. We considered five years of working experience in the SPS field to classify a person as an expert. As shown in Table 1 these come from four industries all located in Germany. All interviews were conducted in 2020. Due to the COVID-19 pandemic, only four took place in person, while the others were conducted online. On average, an interview lasted 30 minutes, which we recorded and then transcribed. The interviews were conducted in German. The statements we include in this paper were translated by the author.

## 4 Findings

The expert's opinions differed regarding whether performance receives special consideration in the context of streaming analytics projects. Some experts believe that performance is a general requirement of any system. Suggesting that performance is not treated differently from other implementation projects. "*Performance is, of course, one of the design criteria (..), it is one of the requirements of all projects.*" Others stated that it deserves special attention e.g., that it is important to ensure performance already at an early stage. "*(..)we are talking about a data volume of 200m data records within a short time. Of course, we make sure from the beginning that our applications are designed in a performance-optimized way*". One also added that the systems with which the SPS are are integrated require special attention. "*The performance issues we are currently working on are related with the systems with which the SPS is integrated*" **(C2)**.

We received mixed opinions on whether performance KPIs or SLAs are defined as part of the streaming analytics project. Three experts stated that this is not the case. "*There are usually no hard limits that are defined in the project (..)*" **(C1)**. The other five experts agreed that such definitions are important. "*Yes, such goals are firmly defined at the beginning of a project*". Regarding performance metrics, most experts named resource utilization and cost efficiency as major performance criteria next to latency, especially in cloud deployments. "*Performance is often a cost issue. The customer wants reasonable response times, but the instance should not be too expensive or over dimensioned*". "*Scaling a streaming system in the cloud is not difficult. But it's not cheap (..) customers sometimes ask whether the system can also run on a smaller instance*" **(C2)**. All eight experts agreed that the responsibility for performance never lies with any one person but that each developer is responsible for the proper performance of his component. "*It is important for us that every developer is responsible for performance so that the topic of performance is already considered in the conception (..)*". However, regarding how the expert would rate the know-how of the project members in performance engineering, we received mixed experiences. "*I worked with two data scientists who knew what they were doing and what could be done to improve performance. Before these specialists joined our project, it was much more uncertain.*" "*(..)especially young colleagues tend to choose the first working design and think about performance when it is too late*". We conclude that there is a potential gap of people responsible for the performance but do not have the experience to ensure it **(C1+C3)**. According to the experts, analytical SPS are complex regarding their performance management. "*These systems are mostly distributed, and the configuration is important e.g., parallelization (..). The search for the performance bottleneck can become very time-consuming*" **(C2)**. All experts stated that performance is tested during the development. Different approaches are used, but all have in common that they are based exclusively on measurements, and use instruments developed in-house, e.g., stress tests or profiling. "*Every release and every change is retested for performance. For this purpose, load tests and performance-oriented unit tests are performed*". Following up on this, we asked if the future workload can be estimated in advance and if differences between the development and production environment cause uncertainties. The experts agreed that workload estimations are quite reliable. "*I don't think you can define it to the fifth decimal place in advance, but you can at least estimate it roughly.*" However, most experts explained that the differences between development and production system cause uncertainties that complicate performance estimations. "*"We often have weaker hardware in development systems, but we also often test with smaller amounts of data (..). With the combination of weaker hardware and less data, it's difficult to make predictions*" **(C4)**. Regarding performance tools, only three experts were aware of performance benchmarks for SPS, and none used one. They were considered not suitable due to limited result transferability and missing advantages over stress tests. "*Yes, I know that such benchmarks are used in research (..) the question is to what extent the results are transferable and what advantage they would bring*". None of the experts uses simulation tools either. Some experts consider them too complex or that measuring is the better alternative. "*We have not yet used performance simulation tools, the complexity of setting up the models seems too great, we rather test the actual behavior with a load generator*". However, planning tools are used by four experts. "*Yes, we sometimes use planning tools to get a proposal for the instance size*" **(C5)**.

Finally, we asked how such problems are approached and what could help to improve performance management. They suggested that it should always be the first step to optimize the software before increasing hardware resources "*In the first stage, this is usually software optimization, hardware optimization comes later.*" Some also emphasized that they try to solve performance problems early during the design phase. "*We try to fix such problems already in the design phase and not during implementation, then it would be too late.*" The majority stated that the most considerable potential for improvement lies in raising developers' awareness and improving their know-how in performance engineering **(C3)**. "*Significant potential for improvement lies in raising developer awareness of performance and providing training in performance engineering.*" Several experts also mentioned the need for better measurement tools. "*Probably better tools that allow testing the software with a self-designable set of mock data*" **(C1+C5)**.

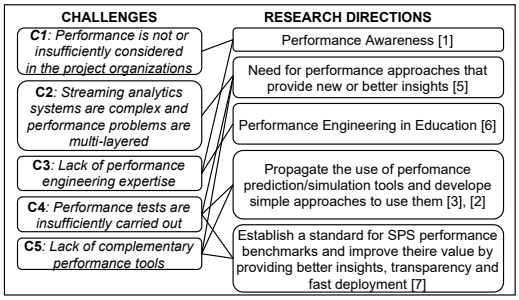


Figure 1: Challenges and Research Directions

## 5 Discussion

Based on the analysis of the interview results, we identified five key performance challenges. A summary of these challenges and potential research directions to address them are illustrated in Figure 1.

*C1: Performance is not or insufficiently considered in the project organization*: Performance goals should always be formulated as part of the project. Unknown expectations often cause the absence of such. From a research perspective, a key driver to cope with this challenge is the rise of performance awareness to emphasize proper performance design at the early stages of the software lifecycle [1].

*C2: Streaming analytics systems are complex, and performance problems are multi-layered*: Response time and cost efficiency are important requirements for industry implementations. At the same time, many factors influence the performance of an SPS, and it is not easy to find the right spot to tune the system. In terms of performance evaluation, research should not only focus on response time, but also on the efficient utilization of resources. At the same time, better tooling support is required to identify performance bottlenecks. There are already concepts for measuring task-level CPU demands [5].

*C3: Lack of performance engineering expertise*: Performance engineering is an important skill for any developer. Performance awareness and a greater focus on performance engineering could address this from an educational point of view. [6].

*C4: Performance tests are insufficiently carried out*: All experts used performance measurement approaches. However, there was uncertainty because the testing environment did not adequately reflect the production environment. A primary reason for this is that building a quality assurance system that reflects the sizing of the production system is cost-intensive. The use of performance simulation could be a good way to address this problem. Model-based prediction tools such as the Palladio Component Model [3] have only a low entry barrier and can achieve accurate performance predictions [2]. However, we found that experts mistakenly believed that simulation approaches are always complex. Research should therefore develop simple approaches and propagate the advantages of these methods more strongly.

*C5: Lack of complementary performance tools*: Simulation tools can be a valuable addition to the existing approaches. However, also performance benchmarks were not applied. Some experts did not know that such are available in the context of streaming. Others felt that the benchmark results are not transferable. Research should focus on establishing an industry-standard benchmark. For the benchmark to have an advantage over self-developed load tests, it should offer more result transparency, new insights and deployable with little effort. An early concept for a kit-based streaming benchmark was proposed in [7].

## 6 Conclusion

Streaming analytics projects face several challenges concerning performance management. In this work, we performed semi-structured interviews with industry experts to shed light on the topic. As a result, we identified five challenges and compared them with the current state of research. This approach allowed us to identify five promising research directions to address these challenges in the future. These are raising performance awareness, developing new performance approaches that facilitate performance management, increasing the focus on performance engineering in education, disseminating fast and simple prediction approaches and establishing an industry standard for SPS performance benchmarks.

## References

[1] P. Tůma. "Performance awareness: keynote abstract". In: *Proceedings of the 5th ACM/SPEC international conference on Performance engineering*. 2014, pp. 135–136.

[2] J. Kroß and H. Krcmar. "Modeling and simulating Apache Spark streaming applications". In: *Softwaretechnik-Trends* 36.4 (2016), pp. 1–3.

[3] R. H. Reussner et al. *Modeling and simulating software architectures: The Palladio approach*. MIT Press, 2016.

[4] A. Streitz et al. "Performance Improvement Barriers for SAP Enterprise Applications: An Analysis of Expert Interviews". In: *Proceedings of the 2018 ACM/SPEC International Conference on Performance Engineering*. 2018, pp. 223–228.

[5] J. Rank, A. Hein, and H. Krcmar. "A Dynamic Resource Demand Analysis Approach for Stream Processing Systems". In: *Softwaretechnik-Trends* 40.3 (2020), pp. 40–42.

[6] C. U. Smith. "Software Performance Engineering Education: What Topics Should be Covered?" In: *International Conference on Performance Engineering*. 2021, pp. 131–132.

[7] S. Chintapalli et al. "Benchmarking Streaming Computation Engines: Storm, Flink and Spark". In: *2016 IEEE Intern. Parallel and Distributed Processing Workshops*, pp. 1789–1792.