**Towards Autonomous Infrastructure Learning:   
Data Observability and SLA-Aware Alerting Frameworks for Analytics Pipelines**

Andrew Espira

Saint Peter’s University

# Abstract

Modern analytics pipelines face growing challenges in ensuring reliability, performance, and alignment with business Service Level Agreements (SLAs). Traditional monitoring tools often fail to provide context-aware insights that connect technical metrics with business outcomes. This proposal outlines a research effort to design and implement an intelligent data observability and SLA-driven alerting framework. The framework leverages techniques from distributed systems monitoring, anomaly detection, and data pipeline optimization. By prototyping the framework using BigQuery and Looker Studio as an example platform, we aim to demonstrate how infrastructure-level signals can be mapped to business-relevant metrics, enabling more proactive and context-rich operational decision-making. Expected contributions include a generalizable architecture, empirical evaluation of observability techniques, and exploration of adaptive alerting strategies.

# Introduction

Data observability has emerged as a critical challenge in modern analytics ecosystems. With organizations increasingly relying on data pipelines to drive decision-making, failures in these pipelines can lead to operational disruptions, missed SLAs, and downstream business impact (Polyzotis et al., 2019). Existing monitoring systems often capture raw metrics such as latency or throughput, but they lack mechanisms to contextualize these measurements against business goals. This gap motivates the need for frameworks that connect infrastructure monitoring with SLA-aware alerting.

# Problem Statement

The reliability of data pipelines depends on timely detection of anomalies and efficient response mechanisms. However, current observability practices are fragmented across platforms, with a lack of unified approaches to mapping technical metrics to business outcomes. For instance, a query slowdown may only be flagged when it violates a threshold, yet the broader context—such as its impact on executive dashboards—remains hidden. The problem can be summarized as the absence of intelligent observability frameworks that integrate distributed systems telemetry with SLA-driven alerting.

# Objectives

This research proposes the following objectives:  
1. To design a generalizable architecture for data observability that bridges infrastructure-level monitoring with business SLA requirements.  
2. To evaluate anomaly detection and alerting techniques in the context of data pipelines.  
3. To demonstrate a prototype implementation on cloud-native platforms, with BigQuery and Looker Studio as examples.  
4. To assess the effectiveness of context-rich alerts in improving operational response times and business alignment.

# Literature Review

Several streams of research intersect with this work. Gu et al. (2019) explore large-scale monitoring architectures, highlighting the challenges of high-dimensional metrics. Polyzotis et al. (2019) introduce the concept of data quality monitoring in pipelines, laying foundations for data observability. Bodik et al. (2010) discuss anomaly detection in distributed systems, offering statistical approaches relevant to SLA-aware monitoring. Recent work by Breck et al. (2020) also emphasizes practical challenges in deploying machine learning monitoring, further reinforcing the need for adaptive frameworks.

# Proposed Methodology

The proposed research will follow a design and evaluation methodology. First, we will derive requirements for SLA-aware data observability through a survey of existing monitoring tools and case studies. Second, we will design a layered architecture that incorporates data collection, metric aggregation, anomaly detection, and SLA mapping. Third, we will prototype the framework using BigQuery and Looker Studio metadata as a representative use case. Finally, we will evaluate the framework against synthetic and real workloads, measuring detection accuracy, alert precision, and response impact.

# Expected Contributions

This research is expected to contribute:  
1. A conceptual framework for SLA-driven data observability.  
2. A prototype system demonstrating the feasibility of mapping infrastructure metrics to business KPIs.  
3. Empirical insights into the trade-offs of anomaly detection and alerting strategies in analytics pipelines.  
4. Recommendations for integrating intelligent observability into enterprise data architectures.

# References

Bodik, P., Goldszmidt, M., Fox, A., & Woodard, D. B. (2010). Fingerprinting the datacenter: Automated classification of performance crises. Proceedings of EuroSys, 111–124.  
  
Breck, E., Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2020). Data validation for machine learning. Proceedings of MLSys.  
  
Gu, J., et al. (2019). Eagle: Efficient monitoring at scale for distributed systems. Proceedings of NSDI, 469–484.  
  
Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2019). Data management challenges in production machine learning. Proceedings of SIGMOD, 1723–1726.